Proceedings of ISSI 2007

11th International Conference of the International Society for Scientometrics and Informetrics

CSIC, Madrid, Spain
25-27 June 2007

VOLUME II

Edited by

» Daniel Torres-Salinas

- EC3 Research Group (Evaluación de la Ciencia y la Comunicación Científica)
  Universidad de Granada, Granada, Spain
- Centro de Investigación Médica Aplicada
  Universidad de Navarra, Pamplona, Spain

» Henk F. Moed

- Centre for Science & Technology Studies (CWTS)
  Leiden University, Leiden, the Netherlands

Published for the International Society for Informetrics and Scientometrics (ISSI) by the Centre for Scientific Information and Documentation (CINDOC) of the Spanish Research Council (CSIC), Madrid, Spain

All rights reserved. No part of this book may be reproduced in any form without the the written permission of the authors.

VOLUME II

V. Markusova, M. Jansz, I. Libkind and A.Varshavsky. *Trends in Russian Research Output in Post-soviet Era* 542
Elba Mauleón and María Bordons. *Women Involvement in Editorial Boards of Mathematics Journals* 552
Katherine W. McCain. *Analysing Influence Over Time: An Historiographic Mapping of the Research of Conrad Hal Waddington (1905-1975)* 558
Lokman Meho and Kiduk Yang. *Fusion Approach to Citation-Based Quality Assessment* 568
Félix Moya-Anegón, Zaida Chinchilla-Rodríguez, Benjamín Vargas-Quesada, Elena Corera-Álvarez, Antonio Muñoz-Molina, Francisco José Muñoz-Fernández and Rocío Gómez-Crisóstomo. *Scientific Output by Gender in Spain (Web of Science, 2004)* 582
Ahat A. Nabiullin. *Emergence of a New Discipline in the Earth Sciences: Bibliometric Analysis of Photogrammetry and Remote Sensing Literature* 594
Ed C.M. Noyons and Clara Calero-Medina. *Applying Bibliometric Mapping in a High Level Science Policy Context. Mapping the Research Areas of Three Dutch Universities of Technology* 599
José Luis Ortega and Isidro Aguillo. *Linear Analysis of Cybermetric Data: Quantifying The European University Web Space* 608
Xavier Polanco and Eric Sanjuan. *Hypergraph Modelling and Graph Clustering Process Applied to Co-word Analysis* 613
Anastassios Pouris and Anthipi Pouris. *The State of Science and Technology in Africa (2000-2004): A Scientometric Assessment* 619
Ismael Rafols and Martin Meyer. *Diversity Measures and Network Centralities as Indicators of Interdisciplinarity: Case Studies in BionanoScience* 631
Suzy Ramanana-Rahary, Michel Zitt and Ronald Rousseau. *Aggregation Properties of Relative Impact and other Classical Indicators: Convexity Issues and the Yule-Simpson Paradox* 643
I.K. Ravichandra Rao. *Distributions of Hirsch-Index and G-index: An Empirical Study* 655
Victor Rodriguez, Frizo Janssens, Koenraad Debackere and Bart De Moor. *On Material Transfer Agreements and Visibility of Researchers in Biotechnology* 659
Adelaida Román, Mª-Dolores Alcain and Elea Giménez. *Evaluation of Scientific Publications in the Humanities* 672
Ana Romero-de-Pablos and Joaquín M. Azagra-Caro. *Internationalisation of Patents by Public Research Organisations from a Historical and an Economic Perspective* 677
Ulf Sandström and Martin Hällsten. *Gender, Funding Diversity and Quality of Research* 685
Edgar Schiebel and Marianne Hörlesberger. *About the Identification of Technology Specific Keywords in Emerging Technologies: The Case of "Magnetoelectronic"* 691
Jesper Wiborg Schneider, Birger Larsen and Peter Ingwersen. *Comparative Study between First and All-Author Co-Citation Analysis Based on Citation Indexes Generated from XML Data* 696
R. D. Shelton, Patricia Foland, and Roman Gorelsky. *Do New SCI Journals Have a Different National Bias?* 708
Henry Small and Phineas Upham. *Citation Structure of an Emerging Research Area: Organic Thin Film Transistors* 718
Alastair Smith. *Issues in "Blogmetrics" - Case Studies Using BlogPulse to Observe Trends in Weblogs* 726
David Stuart and Mike Thelwall. *University-Industry-Government Relationships Manifested Through MSN Reciprocal Links* 731
Bart Thijs and Wolfgang Glänzel. *A Structural Analysis of Benchmarks on Different Bibliometrical Indicators for European Research Institutes Based on their Research Profile* 736
Robert J.W. Tijssen and Thed. N. van Leeuwen. *Research Cooperation within Europe: Bibliometric Views of Geographical Trends and Integration Processes* 740
Liwen Vaughan and Justin You. *Content Assisted Web Co-Link Analysis For Competitive Intelligence* 745
Peter Vinkler. *Introducing the Contemporary Contribution Index for Characterizing the Recent, Relevant Impact of Journals* 753
Martijn S. Visser, Clara M. Calero Medina and Henk F.Moed. *Beyond Rankings: the Role of Large Research Universities in the Global Scientific Communication System* 761

Liying Yang, Steven A. Morris and Elizabeth M. Barden. *Mapping Institutions and Their Weak Ties in a Research Specialty: A Case Study of Cystic Fibrosis Body Composition Research* 766
Yang Zhiping, Fang Shu, Cheng Yunwei, Wang Chun,Wen Yi, Hu Zhengyi and Zheng Yi. *Profiles of Technological Capabilities of the Chinese Academy of Sciences (CAS). A Comparison of Patenting Activities of the CAS with other National Level Institutions* 776
Fuyuki Yoshikane, Takayuki Nozawa, Susumu Shibui and Takafumi Suzuki. *An Analysis of the Connection between Researchers' Productivity and their Co-authors' Past Attributions, Including the Importance in Collaboration Networks* 783
Weiping Yue, Yuntao Pan, Zheng Ma and Xishan Wu. *A Bibliometric Analysis of siRNA Fundamental Research and Patent Activities* 792
Michel Zitt,, Elise Bassecoulard, Ghislaine Filliatreau and Suzy Ramanana-Rahary. *Revisiting Country and Institution Indicators from Citation Distributions: Profile Performance Measures* 797
Alesia Zuccala and Peter van den Besselaar. *Mapping Review Networks: Exploring Research Community Roles and Contributions* 803

POSTER PAPERS 815

Isidro F. Aguillo, José Luís Ortega and Begoña Granadino. *BRIC: Academic Web Contents in the Four Largest Emerging Economies* 816
Rafael Aleixandre-Benavent, Gregório González-Alcaide, Alberto Miguel-Dasit, Miguel Castellano-Gómez and Juan Carlos Valderrama Zurian. *Citation Analysis and Impact Indicators of the Spanish Medical Journals (2001-2005)* 818
Farzaneh Aminpour and Payam Kabiri. *Research Performance in Isfahan University of Medical Sciences: Evaluation of Scientific Productivity* 820
Farzaneh Aminpour and Payam Kabiri. *Webometric Study on Iranian Universities of Medical Sciences* 822
Emma Angus, Mike Thelwall and David Stuart. *University Groups in Flickr: Tagging for Purpose or Pleasure?* 824
Judit Bar-Ilan. *The H-Index of H-Index and of Other Informetric Topics* 826
Maite Barrios, Angel Borrego, Andreu Vilagines, Marta Somoza and Candela Ollé. *Bibliometric Study of Psychological Research on Tourism* 828
Ana Isabel Bonilla, Tony Hernández and Isabel Gómez. *Scientometric Analysis on a Sample of Scientists in Astronomy, Particles Physics and Multidisciplinary Physics in arXiv.org (1996-2006)* 830
Manfred Bonitz. *Spontaneity of Consciousness: V.V. Nalimov's Most Ingenious Book* 832
R.K. Buter and E.C.M. Noyons. *Searching for Converging Research using Citations between Subject Categories: First Results and Challenges* 834
Rosaria Rosanna Cammarano, Maurella Della Seta. *A Review of New Web Tools for Citation Analysis and their Impact on Research* 836
Carolina Cañibano, Javier Otamendi and Inés Andújar. *Exploiting the Potential of Researchers CV Databases for Policy Making: Evidence from the Ramón y Cajal Programme in Spain* 838
Mª Angeles Coslado and Juan Miguel Campanario. *Contribution to the Impact Factor of Academic Journals in the Field of Education and Educational Psychology of Citations to Articles Authored by Editorial Board Members* 840
Dang Yaru, Wang Liya, Gao Feng, Li Kan, Huang Yue and He Fenglan. *Changes and Development of Key Indicators in JCR on the Web* 842
Lixin Chen, Zeyuan Liu and Liming Liang. *The Relationship of Price Index and Median Citation Age* 844
Ms. Sandhya Diwakar and Dr. KK Singh. *Capacity Building in Specialties towards National Health Research through Research Funding* 846
Daniela De Filippo, Sanz-Casado, Elías and Isabel Gómez. *Impact of the Research Stays on the Scientific Output* 848
E. García-Carpintero, B. Granadino, N. Sastre and L.M. Plaza. *The Spanish Presence in Editorial Boards of International Scientific Journals: A Tool for the Promotion of Spanish Science* 850
Marianne Gauffriau, Peder Olesen Larsen, Isabelle Maye, Anne Roulin-Perriard, Markus von Ins. *40 Years Discussion on the Counting of Publications* 852
Juan Gorraiz, Christian Schlögl. *Comparison of two counting houses in the field of pharmacology and pharmacy: Web of Science versus Scopus* 854
B.M.Gupta, S.M.Dhawan and R. P. Gupta. *S&T Research in India: An Overview of its Research Output and Quality* 856

Yusef Hassan-Montero and Víctor Herrero-Solana. *Visualizing Library and Information Science from the Practitioner's Perspective*.....858
Frank Havemann, Michael Heinz, Marion Schmidt, and Jochen Gläser. *Measuring Diversity of Research in Bibliographic-Coupling Networks*.....860
Juan E. Iglesias and Carlos Pecharromán. *Comparing H-indices for Scientists in Different ISI Fields*.....862
Peter Ingwersen, Jesper W. Schneider, Morten Scharff and Birger Larsen. *A National Research Profile-Based Immediacy Index and Citation Ratio Indicator for Research Evaluation*.....864
Abdelali Kaaouachi. *Institutional Evaluation in Moroccan Higher Educational System: A Proposed Model*.....866
Payam Kabiri and Farzaneh Aminpour. *Iranian Medical Research Performance: Does it Comply with the National Needs?*.....868
Hildrun Kretschmer, and Theo Kretschmer. *Distribution of Co-Author Pairs Frequencies of the Journal of Biological Chemistry Explained as Social Gestalt*.....870
Mojca Kotar. *Author Cocitation Analysis of Polyamides Specialty 1974–1999 and the Modification of Multivariate Statistics*.....872
Paula Leite and Jacqueline Leta. *Productivity and Prestige among Brazilian Scientists*.....874
Jonathan M. Levitt and Michael Thelwall. *Two New Indicators Derived from the h-Index for Comparing Citation Impact: Hirsch Frequencies and the Normalised Hirsch Index*.....876
Grant Lewison. *Breast Cancer Research, Death Rates and Life Expectancy in Iceland: Are they Linked?*.....878
Grant Lewison. *Counting Citations: Fractionation of Addresses and "World-Scale", a New Scalar*.....880
Xia Lin. *New Visual Interfaces for Author Co-Citation Mapping*.....882
Szu-chia Scarlett Lo. *Patent Indicators of Genetic Engineering Research of EU Countries - Overseas Performance: Preliminary Study*.....884
Diana Lucio-Arias. *A Validation Study of HistCite™: Using the Discoveries of Fullerenes and Nanotubes*.....886
Renato Fabiano Matheus. *Information Science Network – ISN: Social Network Analysis of Scientific Production of LIS Field in Brazil*.....888
Katherine W. McCain. *The Relationship between Influence and Image: Two Views of the Oeuvre of Conrad Hal Waddington using Historiographic Mapping and Author Tri-Citation Image Analysis*.....890
Katherine W. McCain and Scot Silverstein. *Using Historiographic Mapping to Trace Persistent Highly Visible Research Themes in Medical Informatics*.....892
C.A Macías-Chapula, J.A. Mendoza-Guerrero, I.P Rodea-Castro and A. Gutiérrez-Carrasco. *Institutional Health Research Collaboration in Mexico. A Bibliometric Study*.....894
Raúl I. Méndez-Vásquez, Eduard Suñén-Pinyol, Ginés Sanz and Jordi Camí. *Characterization of research groups in the cardio-cerebrovascular field. Spain 1996-2004.*.....896
Alberto Miguel-Dasit, Gregorio González-Alcaide, Juan Carlos Valderrama-Zurián, Adolfo Alonso-Arroyo and Rafael Aleixandre-Benavent. *Analysis of the Interdisciplinar Collaboration in Spanish Scientific Production on Radiology and Diagnostic Imaging*.....898
Fernanda Morillo and Daniela de Filippo. *The Role of Madrid in the Spanish Regional Collaboration*.....900
Rogério Mugnaini, Elías Sanz-Casado and Carlos García-Zorita. *Ways of Adequacy for Evaluation of Brazilian Scientific Production: National Impact versus International Impact*.....902
Rogério Mugnaini, Rogério Meneghini and Abel Packer. *Citation Profiles in Brazilian Journals of the SciELO Database in Different Scientific Areas*.....904
Ülle Must. *History Research in a Globalized World: a Bibliometric Approach*.....906
Ryosuke L. Ohniwa, Aiko Hibino and Kunio Takeyasu. *Perspective Factor: Past, Present and Future of Life Sciences*.....908
José Felipe Ortega, Jesús M. González-Barahona and Gregorio Robles. *In-Depth Analysis of Wikipedia Community*.....910
Jose Luis Ortega, Isidro Aguillo, Viv Cothey, Andrea Scharnhorst. *Exploring Visually the European Academic Web Space*.....912
Ana Patricia Ortiz, William A. Calo, Carlos A. Suárez, Erik Suárez, Isabel Iribarren and Elías Sanz-Casado. *Basic Characteristics of Cancer-related Research in Puerto Rico: An Approach to Local and International Journals*.....914
Pang Jing'an, Wang Lian and Cao Yan. *Quantification Evaluation of Technical Innovation Capabilities of Chinese Large and Medium Industrial Enterprises*.....916

K. G. Pillai Sudhier. *Ranking Journals in Physics: An Informetric Study of the Citations in the Doctoral Theses of the Indian Institute of Science* 918
Marios Poulos, Nikolaos Korfiatis and George Bokos. *Towards the Construction of a Global Bibliometric Indicator* 920
Hampus Rabow. *A Bibliometric Analysis of Book Based Literature* 922
Cristina Ramo. *Research in Doñana National Park: Preliminary data* 924
Claude Robert, Concepción S. Wilson, Jean-François Gaudy, and Charles-Daniel Arreto. *The Evolution of the Sleep Scientific Literature over 30 Years: A Bibliometric Analysis* 926
Gwendolyn Rogge and Ronald Rousseau. *Is the European Web British or American?* 928
Nadine Rons and Arlette De Bruyn. *Quantitative CV-Based Indicators for Research Quality, Validated by Peer Review* 930
Hervé Rostaing, Nicolas Barts, Valérie Léveillé. *Scientific Portfolio Analysis of a Scientific Area by a Competitive Position Approach* 932
Samile Andréa de Souza Vanz and Renato Fabiano Matheus. *Analyzing Grey Literature from Postgraduate Programs in Social Communication in Brazil: Network of Influence and Citation Analysis* 934
Yuan Sun, Masamitsu Negishi and Loet Leydesdorff. *National and International Dimensions of the Triple Helix in Japan: University-Industry-Government and International Co-Authorship Relations* 936
Dimitar T. Tomov and AbdulKader A. Murad. *International Communication Patterns in an Emerging Interdisciplinary Field - Applications of Geographic Information Systems in Public Health* 938
Ming-yueh Tsay. *An Analysis and Comparison of Citation Data between Journals of Physics, Chemistry and Engineering* 940
Yuen-Hsien Tseng, Yu-I Lin, Chun-Hsien Kuo and Yi-Yang Lee. *Verification of Increasing Trend Detection* 942
Ali Uzun. *Recent Trends in Renewable Energy Research: A Bibliometric Perspective* 944
Thanh-Trung Van and Michel Beigbeder. *From Web Citation to Web Co-citation: Discovering Relatedness on the Web* 946
Sonia Vasconcelos, Martha Sorenson and Jacqueline Leta. *English Proficiency: A Potential Science Indicator?* 948
Jakob Voss. *Common Statistics of Heterogeneous Tagging Systems* 950
Chun-Chieh Wang , Mu-Hsuan Huang and Dar-Zen Chen. *Innovative Capacity Evaluation of Main Countries Based on Patent Analysis* 952
Dietmar Wolfram. *Search Characteristics in Different Types of Internet-Based IR Environments: Are They the Same?* 954
Yasuhiro Yamashita, Sen Ueno, Hiroyuki Tomizawa and Masayuki Kondo. *Influence of the International Migration of Researchers on National Publications in Three Fields of Engineering* 956
Li-li Yang, Dong-hong Fu, Xiao-ping Bai, Wei-gang Fang, Shu-yin Yao. *Citation Analysis of the Chinese Journal of Medical Science Management* 958
Lin Yang, Lili Yang and Chengzheng Zhao. *Assessment of Traditional Chinese Medicine for Substance-Related Disorders Based on Bibliometric Techniques* 960
Hairong Yu, Concepción S. Wilson, Mari Davis and Fletcher Cole. *Complex Data Modelling for Informetric Research* 962
Jian Zhang, Weizhong Zhu, Yunan Chen, Michael Vogeley and Chaomei Chen. *Analyzing the Impact of Sloan Digital Sky Survey on Astronomical Literature: A Multiple Perspective Approach* 964
Dangzhi Zhao. *Factor Rotation Methods in Author Co-citation Analysis: A Comparison* 966
M.A. Zulueta, G. García-Gómez, I. Doménech, M. Izquierdo and P. Moscoso. *Bibliometry Analysis of the Research on Women and Health* 968

AUTHOR INDEX 971

VOLUME I

KEYNOTES 19

Eugene Garfield. *From The Science of Science to Scientometrics. Visualizing the History of Science with HistCite Software* 21
Stevan Harnad. *Open Access Scientometrics and the UK Research Assessment Exercise* 27

FULL PAPERS AND RESEARCH-IN-PROGRESS PAPERS..... 35

Rafael Aleixandre-Benavent, Gregorio González-Alcaide, Alberto Miguel-Dasit, Carolina Navarro-Molina and Juan Carlos Valderrama-Zurián. *Full-Text Publications in Peer-Reviewed Journals Derived from Presentations at Two ISSI Conferences* 37

Éric Archambault and Vincent Larivière. *Origins of Measures of Journal Impact: Historical Contingencies and their Consequences on Current Use* 45

Judit Bar-Ilan and Bluma C. Peritz. *The Lifespan of "Informetrics" on the Web: an Eight Year Study (1998-2006)* 52

Franz Barjak and Simon Robinson. *International Collaboration, Mobility and Team Diversity in the Life Sciences: Impact on Research Performance* 63

Elise Bassecoulard, Alain Lelu, and Michel Zitt. *A Modular Sequence of Retrieval Procedures to Delineate a Scientific Field: from Vocabulary to Citations and Back* 74

Peter Van den Besselaar, Gaston Heimeriks, Koen Frenken. *Variety in Web Spheres between Research Fields: Content and Function* 85

Sujit Bhattacharya. *Impact of Indian Patents: Assessment through Citation Analysis* 95

Omwoyo Bosire Onyancha and Dennis N. Ocholla. *Is HIV/AIDS in Africa Distinct? What Can we Learn from an Analysis of the Literature?* 100

Kevin W. Boyack, Katy Börner and Richard Klavan. *Mapping the Structure and Evolution of Chemistry Research* 112

Kevin W. Boyack. *Using Detailed Maps of Science to Identify Potential Collaborations* 124

Robert Braam. *Everything about Genes: some Results on the Dynamics of Genomics Research* 136

Lutz Bornmann and Hans-Dieter Daniel. *Functional Use of Frequently and Infrequently Cited Articles in Citing Publications. A Content Analysis of Citations to Articles with Low and High Citation Counts* 149

Jenny-Ann Brodin Danell and Rickard Danell. *Spiritualised Medicine? A Bibliometric Study of Complementary and Alternative Medicine* 154

Quentin L. Burrell. *Hirsch's h-index and Egghe's g-index* 162

Linda Butler and Kumara Henadeera. *Is there a Role for Novel Citation Measures for the Social Sciences and Humanities in a National Research Assessment Exercise?* 170

Clara Calero Medina and Ed C.M. Noyons. *Combining Mapping and Citation Network Analysis for a Better Understanding of the Scientific Development: The case of the Absorptive Capacity Field.* 179

Mônica G. Campiteli, Pablo D. Batista and Alexandre S. Martinez. *A Research Productivity Index to Account for Different Scientific Disciplines* 184

Chaomei Chen, Il-Yeol Song and Weizhong Zhu. *Trends in Conceptual Modeling: Citation Analysis of the ER Conference Papers (1979-2005)* 189

Chen Li, Pan Yuntao, Ma Zheng, Su Cheng and Wu Yishan. *A Comparative Study between International and Domestic Interdisciplinary Journals and Specialty Journals: A Trial Analysis of Medical Journals, Philosophy Journals and Journals in Philosophy of Medicine* 201

Heting Chu and Thomas Krichel. *Downloads vs. Citations in Economics: Relationships, Contributing Factors and Beyond* 207

Mario Coccia. *Does Bureaucracy Affect Research Performance of Public Research Organizations?* 216

Rodrigo Costas and Maria Bordons. *A Classificatory Scheme for the Analysis of Bibliometric Profiles at the Micro Level* 226

Viv Cothey. *Applying Egghe's General Theory of the Evolution of Information Production Processes to the World Wide Web* 231

S.M. Dhawan and B.M. Gupta. *High Productivity Physics Institutions in India: A Study of their Performance in terms of Quantitative and Qualitative Indicators* 241

Leo Egghe. *Distributions of the h-index and the g-index* 245

Shu Fang, Xian Zhang and Guo-hua Xiao. *Research and Application of Patent Map Analysis* 254

Francisco Fernández-Izquierdo, Adelaida Román-Román, Cruz Rubio-Liniers, Francisco-Javier Moreno- Díaz-del-Campo, Carmen Martín-Moreno, Carlos García-Zorita, Maria Luisa Lascurain-Sánchez, Preiddy-Efraín García, Elisa Povedano and Elías Sanz-Casado. *Bibliometric Study of Early Modern History in Spain Based on Bibliographic References in National Scientific Journals and Conference Proceedings* 266

Enrico Forti, Chiara Franzoni and Maurizio Sobrero. *The Effect of Patenting on the Networks and Connections of Academic Scientists* 272

Chiara Franzoni, Christopher L. Simpkins, Baoli Li and Ashwin Ram. *Using Content Analysis to Investigate the Research Paths Chosen by Scientists Over Time* 285

Rainer Frietsch, Sybille Hinze and Pari Patel. *Using Patent Data for Monitoring the Globalisation of R&D* 295
Antonio García Romero, José Navarrete Cortés, Cristina Escudero , Juan Antonio Fernández López and Juan Antonio Chaichio Moreno. *Measuring the Contribution of Clinical Trials to Bibliometric Indicators: Citations and Journal Impact Factor®* 300
Monica Gaughan, Branco Ponomariov and Barry Bozeman. *Using Quasi-Experimental Design and the Curriculum Vitae to Evaluate Impacts of Earmarked Center Funding on Faculty Productivity, Collaboration, and Grant Activity* 305
Yves Gingras. *Mapping the Changing Centrality of Physicists (1900-1944)* 314
Wolfgang Glänzel,, Frizo Janssens and Bart Thijs. *A Comparative Analysis of Publication Activity and Citation Impact Based on the Core Literature in Bioinformatics* 321
Isabel Gómez, María Bordons, M.Teresa Fernández, Fernanda Morillo. *Structure and Research Performance of Spanish Universities* 335
Yuri Jack Gómez. *Revisiting the "Heroic" Age: From Externalism to Internalism in Serial History of Science.* 346
Iina Hellsten, Renaud Lambiotte, Andrea Scharnhorst and Marcel Ausloss. *Self-citation Networks as Traces of Scientific Careers'* 365
Jean-Pierre V. M. Hérubel. *Pre 1990 French Doctoral Dissertations in Philosophy: A Bibliometric Profile of a Canonical Discipline* 368
Kim Holmberg and Mike Thelwall. *Local Government Web Sites in Finland: A Geographic and Webometric Analysis* 378
Peter A. Hook. *Visualizing the Topic Space of the United States Supreme Court* 387
Stefan Hornbostel and Susan Böhmer. *Determinants for Young Researcher Careers in Germany. Comparative Evaluation of Postdoctoral Programmes* 397
Isabel Iribarren-Maestro, María Luisa Lascurain-Sánchez and Elías Sanz-Casado. *Are Multi-Authorship and Visibility Related? Study of Ten Research Areas at Carlos III University of Madrid* 401
Frizo Janssens, Wolfgang Glänzel and Bart De Moor. *A Hybrid Mapping of Information Science* 408
Evaristo Jiménez-Contreras, Rafael Bailón Moreno, Daniel Torres-Salinas, Rosario Ruiz Baños, Rafael Ruiz-Pérez, Mercedes Moneda Corrochano and Emilio Delgado López-Cózar. *Response Surface Methodology and its Application in Evaluating Scientific Activity* 421
Bihui Jin, Ronald Rousseau, Richard P. Suttmeier and Cong Cao. *The Role of Ethnic Ties in International Collaboration: The Overseas Chinese Phenomenon* 427
Richard Klavans and Kevin W. Boyack. *Is there a Convergent Structure of Science? A Comparison of Maps using the ISI and Scopus Databases* 437
Vincent Larivière, Éric Archambault and Yves Gingras. *Long-Term Patterns in the Aging of the Scientific Literature, 1900–2004* 449
Gavin LaRowe, Sumeet Ambre, John Burgoon, Weimao Ke and Katy Börner. *The Scholarly Database and Its Utility for Scientometrics Research* 457
Katarina Larsen. *Interdisciplinarity in Environmental Technology Applications – Examining Knowledge Interaction between Physics and Chemistry Research Teams* 463
Thed N. van Leeuwen and Robert J.W. Tijssen. *Strength and Weakness of National Science Systems. A Bibliometric Analysis through Cooperation Patterns* 469
Jacqueline Leta and Flávio Martins Teixeira. *Science in Brazil: Contribution of Male and Female Scientists* 480
Jonathan Levitt and Mike Thelwall. *Atypical Citation Patterns in the Twenty Most Highly Cited Documents in Library and Information Science* 485
Grant Lewison. *The References on UK Cancer Clinical Guidelines* 489
Loet Leydesdorff and Caroline Wagner. *Is the United States Losing Ground in Science? A Global Perspective on the World Science System in 2005* 499
Liming Liang. *Revealed Similarities between the Journals Nature and Science: Using a New Cluster of Rhythm Indicators* 508
Yuxian Liu and Ronald Rousseau. *Hirsch-Type Indices and Library Management: The Case of Tongji University Library* 514
Ma. Elena Luna-Morales, Francisco Collazo-Reyes and Jane M. Russell. *A Quantitative Hstoriography of Mexican Integration into the International Standards of Scientific Research.* 523

Trends in Russian Research Output in Post-soviet Era

V. Markusova[*], M. Jansz[**], I. Libkind[***] and A.Varshavsky[****]

[*]*science@viniti.ru*
VINITI (Russia)

[**]*m.jansz@stw.nl*
Technology Foundation STW (the Netherlands)

[***]*libkind@rfbr.ru, libkind_ilya@hotmail.com*
Russian Foundation for Basic Research (Russia), MISA (Russia)

[****]*varshav@cemi.rssi.ru*
Central Economics &Mathematics Institute (Russia)

Abstract
Recently, the Russian government has ordered evaluation and reform of the basic research system. As a consequence, the number of research staff at the Russian Academy of Sciences will be reduced by 20% by 2007. The basis for research evaluation and institute budgeting will be bibliometric indicators. In view of these changes we look at the Russian publication output and argue that

(1) publication output and citedness have to be considered in relation to the level of expenditure on R&D;
(2) bibliometric indicators depend strongly on the database used (ISI's databases are biased) and their interpretation can be confusing; better coverage of Russian publications or a Russian Science Citation Index are needed. Also, research results are communicated in more ways than paper publications.
(3) policy makers have misused ISI statistics to demonstrate "a low level" of Russian R&D.

Our paper is a part of a project designed to trace R&D development in a transition economy and knowledge transfer from basic research to innovation. Results of our project shed light on science policy and the social issues due to the indiscriminate introduction of quantitative indicators.

Keywords
research output; Russia, R&D expenditure; regions; science policy; bibliometric indicators; impact factor

Introduction

Russian science policy and R&D were the subject of many papers, described by C.Wilson and V.Markusova (2004). Since 2000 a decline in Russian output was observed in ISI Thomson Scientific data. Given the international importance of Russian R&D in many areas of science, this decline warrants a special investigation. Recently, the Minister of High Education and Science (MHES) of the Russian Federation, Dr.A.Fursenko, expressed "a deep concern" (www.sciencerf.ru) about this output decline. His comments have received attention at the highest political level. The government ordered evaluation and reform of the basic research system. As a result of this reform, research staff of the Russian Academy of Sciences (RAS), a main research body in Russia, will be cut by 20% in 2007. These measures will have serious social consequences and provoked a heated discussion in the Russian scientific community (www.poisknews.ru).

Russia has now, as for example Norway (Siversten G., 2006), implemented a bibliometric model for research evaluation and institute budgeting. Starting on January 1, 2007, the research staff of the RAS will be allowed to get an additional monthly income from the government based on ISI data. The selected indicators of research activity are : number of publications and citations in the last two years, as well as and journal impact factor. The MHES developed a methodology to give a researcher different scores depending on the type of publication (journal article, book, chapter in a book, conference proceedings and etc). Grants or awards are not included as indicators of research activity. According to this methodology, a paper published in a JCR journal yields six times the score of a paper presented at an international conference .

Although without doubt bibliometric indicators can be used as tools for monitoring and evaluation of research performance, they should be used carefully.
We argue that

(1) publication output and citedness have to be considered in relation to the level of expenditure on R&D. A method of calculation was developed by the Central Economics & Mathematics Institute (CEMI) of the RAS (Makarov V., 2005), which allows demonstrating the real place of Russian R&D among developed countries;
(2) bibliometric indicators depend strongly on the database used and their interpretation can be confusing . For a proper evaluation of science in Russia it is necessary to compile a national database and to use ISI statistics as an additional tool;
In addition, research results are communicated in more ways than paper publications only.
(3) policy makers have misused ISI statistics to demonstrate "a low level" of Russian R&D on the world science stage. This misuse had a strong impact on mass media, public attitude regarding the R&D level in Russia and resulted in negative consequences for Russian science.

Our paper is a part of a project, the main goals of which are to trace the R&D development in a transition economy and the knowledge transfer from basic research to innovation. Results of our project shed light on science policy and the social issues due to the indiscriminate introduction of quantitative indicators.

Methodology

As data sources for this study we used various information resources produced by ISI - the Web of Science (WOS), SCI on CD-ROM, the Essential Science Indicators (ESI), Journal Citation Reports and the National Science Indicators Deluxe (NSI) - and as a national database the Russian Foundation for Basic Research (RFBR) analytical system (Libkind I.& others, 2006). RFBR grantees represent about 60 % of all researchers in Russia. Some statistics were collected from the Directory of Statistics of the MHES, the OECD and the State Committee of Statistics of the Russian Federation (RF). We have to emphasize that the calculation of R&D personnel used by the MHES was methodologically inaccurate. The number of researchers counted in «full-time equivalents» (FTE) was higher by a factor 1.2 than the number of researchers by « heads». For all other countries the number of researchers counted in FTE is lower than in heads. For example, in transition economies the number of researchers in FTE is generally lower by a factor 1.5 than counted in heads. We used the method of calculation developed by the Central Economics & Mathematics Institute (CEMI) of the RAS. It allows us to demonstrate the real place of Russian R&D among developed countries. To get a reliable comparison of RAS researchers with their counterparts from the Max Plank Society (Germany) and the CNRS (France) an average FTE-value for countries with a transition economy was used.

Various variables were used for analysis: investment in R&D by region, annual output by region, share of journal papers, share of distribution by field of sciences, distribution by source, impact factor and share of highly cited papers by the ESI.

The Anglophone bias in ISI sources is well known. In average only about one hundred Russian journals (or 5 % of Russian peer reviewed journals) were covered by the WOS annually since 2000.

We identified and calculated the following indicators:

- Annual Russian expenditure on R&D and scientific publication output by region using the SCI on CD-ROM for 1994-2003;
- Number of papers and citations per researcher (from Web of Science) for OECD and some other countries, as well as for the RAS and the CNRS (France) for the time period 1996-2005 using an average indicator of FTE for transition economies;
- Number of papers and citations in 1996-2005 per $1mln of R&D by PPP (Power Purchase Parity) in 2004;
- Russian activity profile compared to that of other G-8 countries and China;
- RFBR grantees' output distribution by source of publication, activity profile, co-authorships.

There is only one national database in Russia which collects bibliographic records of the more active part of the research community, namely of grantees of the RFBR. We used the RFBR database to trace trends in research productivity.

Results and Discussion

Before the disintegration of the Former Soviet Union its annual share of scientific publication output was about seven percents. Reliable statistics on Russian output (RO) became available in ISI products since 1993. In our analyses we also had to cope with changes in journal coverage by ISI. As an example, since the middle of the 1990s ISI has made a shift from processing Russian journals in Cyrillic to using their English versions. As a result of this policy some Russian journals with a reasonably steady impact factor have disappeared from JCR for one or two years but reappear in the following years. According to NSI-Deluxe statistics, 1993-1997 saw the highest share of RO (3.80% of the world scientific publication production). The highest absolute value of RO was in 1996-2000 – 130,369 papers. This declined to 120,367 papers in 2001-2005 (see Fig.1). In 2001-2005 the Russian share dropped to 2.99 % of world output.

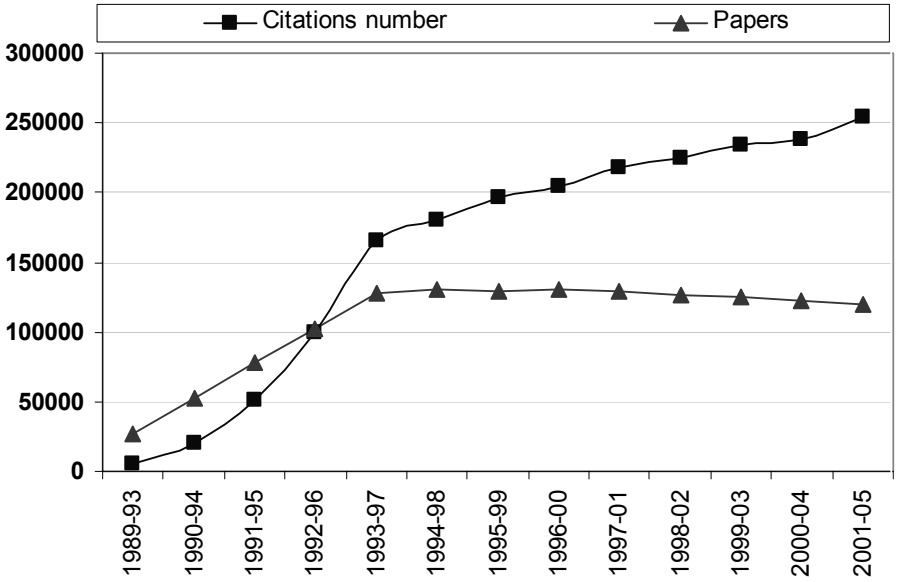


Figure 1. Russian Output and Citeness, NSI Deluxe

Analysis of output distribution by field of science using NSI-Deluxe revealed that there were nineteen disciplines whose share of papers in the field were higher than the average share of Russian output in the world for five-year overlapping periods. All of these disciplines, except “sociology and social sciences” (share 4.25%), belonged to “hard sciences” (e.g. aerospace engineering; chemistry, engineering, earth sciences).

A significant growth was observed only in Earth Sciences whose share of papers increased from 4.1% to 7.2%. This growth could be attributed partly to ISI's processing policy: in 2000 SCI processed eleven journals on Earth Sciences compared to fourteen in 2005. Relative citedness () in five-year overlapping periods had shown a significant growth from 32.62% in 1993-1997 to 42.98% in 2001-2005. In comparison, French citeness increased from 51.12% in 1981-85 to 63.66% in 2001-2005.
In 2001-2005, the number of citations to Russian papers in the same period had grown by 7.15% . Despite the limited output in clinical medicine its citedness was above average in 2001-2005. Among the “hard sciences” chemistry and earth sciences had a lower citedness relative to the field.

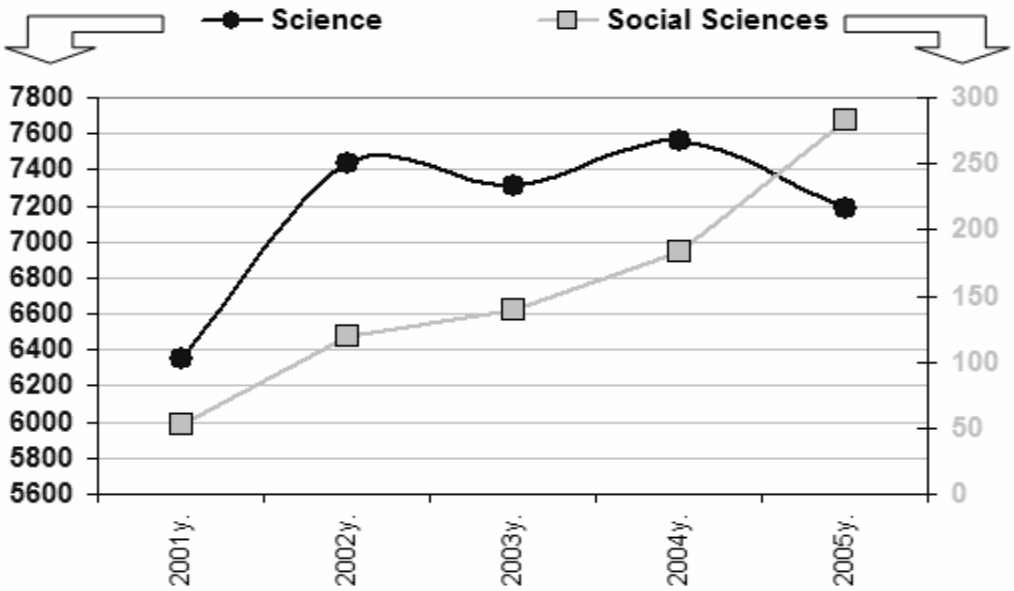


Figure 2. Russian output in ISI Proceedings

Conference proceedings play an important role in scientific communication. Contrary to paper output, a growth in RO in the ISI-covered Proceedings was observed (see Figure 2**)**. These changes were initiated by the end of the Cold War and more recently partly driven by a steady growth in collaboration of Russian researchers with their Western counterparts (Science & Engineering Indicators, 2006). We used the Science & Technology Indicators, 2004, to compare the USSR/Russia discipline profile with that of the USA. These data are presented inTable 1.

Table 1. Discipline profile USSR/Russia and USA, publication shares in 1988 and 2001 (S&E Indicators – 2004).

Disciplines	**USSR / Russia**		**USA**	
	1988	**2001**	**1988**	**2001**
Clinical medicine	14,3	3,2	31,9	31.7
Biomedicine	17,7	7,5	15,5	16,9
Biology	2,6	4,0	7,2	6,2
Chemistry	27,1	27,1	7,4	7,1
Physics	27,6	*35,6*	10,1	8,7
Earth Sciences	4,1	*8,1*	4,5	5,6
Engineering	4,1	*8,9*	6,7	6,9
Mathematics	0,9	*3,4*	2,2	1,8
Psychology	0,6	*1,3*	4,9	4,7
Social sciences	0,6	0,6	4,0	3,9
Health	0,0	0,1	1,6	2,5
Other fields of sciences	0,4	0,2	4,8	3,9

These data show that there were no major changes in USA output by discipline. In Russia, a significant growth in the share of RO can be observed in physics, engineering, mathematics and in psychology. In contrast, a drop in RO share was observed in biomedicine and in clinical medicine. Dr. Eugene Garfield (1979) emphasized that Soviet science was focused on hard science. Our data demonstrate that there are no changes in field of science development. This is in line with President V. Putin's list of priority research programs (decree N576, 2002), which stated that additional resources have to be given primarily not to life or environmental sciences but to “hard science”. It looks as if the

Russian government still lives in the first part of the XXth century, when physics played a leading role in science. But we live in the XXIst century and biosciences and biomedicine have taken over the leading role in contemporary society.
To get a reliable assessment of Russian R&D performance comparing it with the situation in other countries, we used a method developed by CEMI. The results of our calculations are displayed in Tables 2 and 3.

Table 2. Number of papers per researcher, based on ESI for 19996-2006

Country	Number of papers per researcher	Rank (32 countries and OECD)	Number of citations per researcher	Rank (32 countries and OECD)
China	0,41	33	1,43	33
Russia: - number of researchers in FTE used by MHES - number of persons involved - number of researchers in FTE (our estimation on the basis of the average FTE indicator used for transition economies	 0,57 0,68 1,05	32	 2,00 2,38 3,66	32
Japan	1,11	30	**7,64**	between 30 and 31
the Russian Academy of Sciences, RAS (researchers in terms of persons involved)	**2,00**	between 24 and 25	8,72	29
USA	2,17	24	**11,75**	between 24 and 25
OECD, total	2,41	23	**13,93**	between 23 and 24
Germany	2,65	20	25,55	13
France	2,71	18	25,92	12
RAS (our estimation based on the average "FTE" used for transition economies)	**3,08**	between 12 and 13	26,44	11
Canada	3,28	9	28,02	10
CNRS	**3,85**	between 5 and 6	34,70	8
Italy	4,86	3	43,97	4
Great Britain	4,87	4	54,63	3

These data show that Russian researchers, despite all economical turmoil, continue their devotion to science.

We compared the government expenditure on R&D by region in 1994-2003 with RO in the same period. These data are presented in Figure 3.

Table 3. Number of papers and citations based on ESI for 1996-2006 per $1mln of R&D by PPP in 2004

Country	Number of papers per $1mln of R&D by PPP	Rank (32 countries, RAS and CRNS	Number of citations per $1mln of R&D by PPP	Rank (32 countries, RAS and CRNS
China	3,7	37	12,9	37
Japan	6,6	36	52,2	34
USA	8,8	33	58,1	33
OECD, total	11,9	32	113,1	27
Germany	11,9	31	119,3	26
France	12,7	29	121,6	25
Russia	16,6	22	125,8	23
CNRS	16,9	21	177,0	15
Canada	19,1	16	184,9	13
Italy	19,5	15	202,2	11
Great Britain	22,8	12	255,5	5
RAS	70,7	1	269,5	4

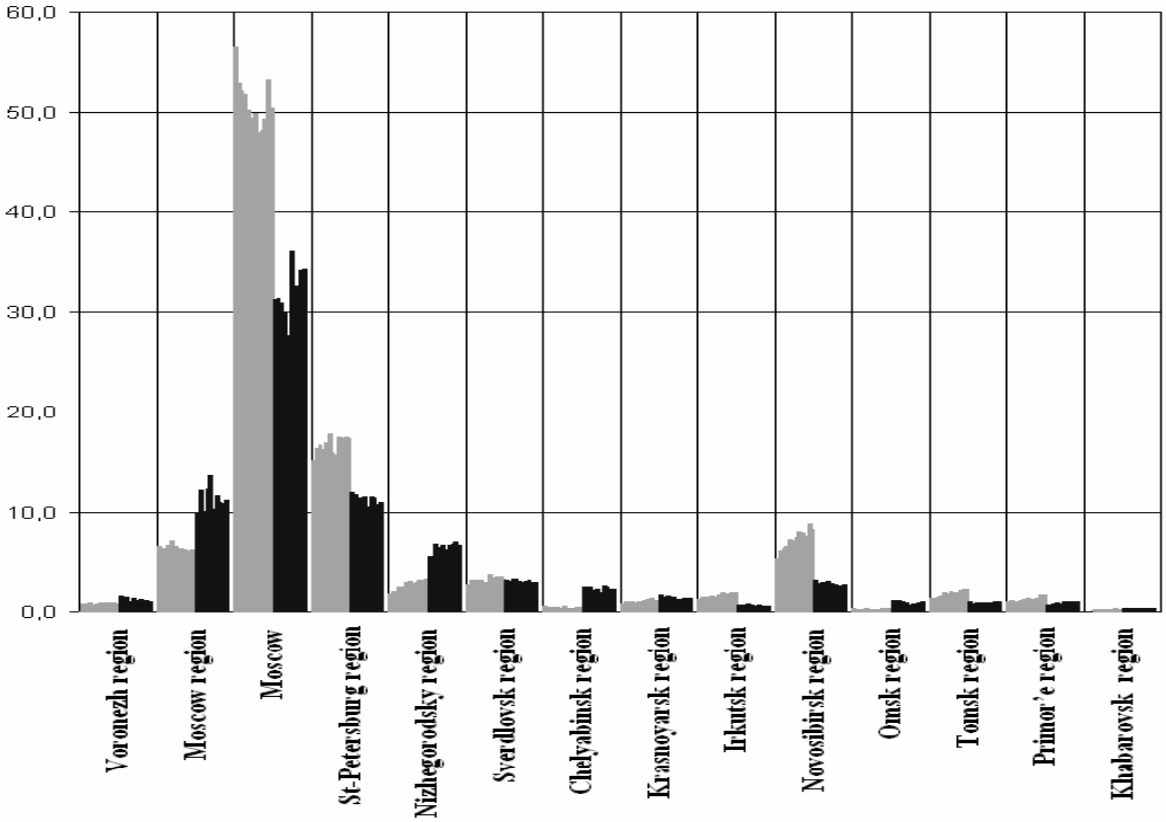


Figure 3. Expenditure and Output in Russian Regions (light: output 1994-2003; dark: expenditure 1994-2003)

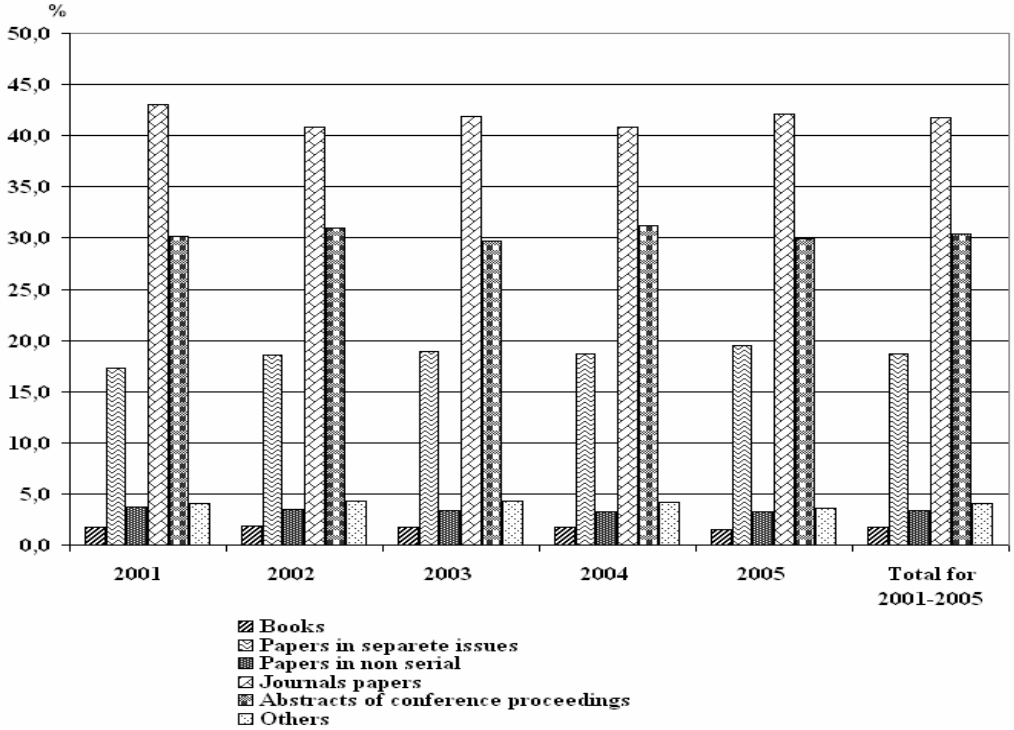
We observe that Moscow and St-Petersburg, where R&D was concentrated since the 1920's, still attract the main investments. Expenditure only grew significantly in Moscow city – its share increased by 3%. This growth could be attributed to various circumstances, such as the fact that the Moscow city government runs a few special programs to support R&D in the city and the presence of strong

research instituteswhich attract resources from various Federal Programs distributed by MHES. After the economic crisis in August 1998, the expenditure on R&D in any analyzed region except Moscow city was reduced. Later, the expenditure in those regions partly recovered. A different picture appeared when we studied the output share by regions. Moscow's share has diminished from 56% in 1994 to 49.2% in 2003. On average during 1994-2003 there were a few regions whose output share was significantly higher than the share of expenditures. This was the case for: Moscow city, Moscow region, St-Petersburg region, Irkutsk region, Tomsk region, Novosibirsk regions and Primorie. The regions whose share of output was significantly less than the investment were: Omsk region, Khabarovsk region, Chelyabinsk region and Nizhegorodsky region.

Tomsk is a well-known city in Siberia with a few strong research institutes under the auspices of the RAS and the State University of Tomsk. Tomsk region's share of investment (0.1%) was approximately the same as that of Omsk region. However Omsk's output share was only about one tenth of Tomsk's share. We assume that this difference could be attributed to the well-developed defense industry in Omsk region. The same phenomenon was observed in Chelyabinsk region. Our observations are in line with the opinion expressed by R.Shelton (2006) that non-defense R&D investments are more effective in producing public papers. Performing the search of output by region we included in the analysis all cities belonging to this region. It was interesting to observe a growing output from Sarov which belongs to Nizhegorodsky region. Sarov was, and still is, a closed city for foreigners. The city is famous for the Federal Nuclear Research Institute, in which the first Soviet nuclear and H-bombs were developed. Sarov's output increased from zero papers in 1994 to 98 papers in 2003, mainly due to a strong cooperation with the USA and European countries.

To find out what part of the Russian research output was covered by ISI databases and to trace trends in publishing activity we investigated Russian grantees' output, see Table 4 and Figure 4. These statistics were extracted from the annually updated RFBR database. Institutional addresses and source of publications have undergone a process of careful cleaning in this database. Total output for 2001-2005 consists of more than 296,100 publication, 10% of which did not have a full bibliographic description and were excluded from the analysis. The distribution of the output by source and year is presented in Figure 4.

Figure 4. Share of the RFBR Grantees Output Distribution by Source, 2001-2005



As we can see, there is no decline in grantees' output over time and no decline in journal papers. The number of journal papers (123,600) by grantees was approximately the same as the total Russian output for the same time period according to NSI data (122,300 papers); practically all active Russian researchers receive support from the RFBR. The share of journal's papers varies by field of science: from 35 % in Earth Sciences up to 49.9% in Physics. According to an analysis by C.Wilson (2004) and by our data, research in physics is more productive (highest share of publications – 24 %) and enjoys priority spending by government. By contrast, the life sciences, which were relatively neglected in the USSR, continue to be a "Cinderella" in research. Finally, among seven priority projects financed by the MHES, the research program "Life Systems" received $43.3 million (or 16.4% of the financial support) in 2005 and $45.4 mln (15.6%) in 2006. Hopefully, these additional investments will stimulate a growth in research productivity in the life sciences.
For the last five years, the share of conference proceedings varied according to field of science**.** the highest belonging to Chemistry. In the future we will investigate what part of RFBR papers in conference proceedings is covered by ISI Proceedings.

Most papers are now written in co-authorship. According to ISI data the number of authors per paper increased from 1.9 in 1960 to 4.3 in 2003. This is the first time that data on co-authorship practices have been collected among Russian researchers, particularly among RFBR grantees. The pattern of research team co-authorship by field of science in 2001-2005 is displayed in Fig. 5.

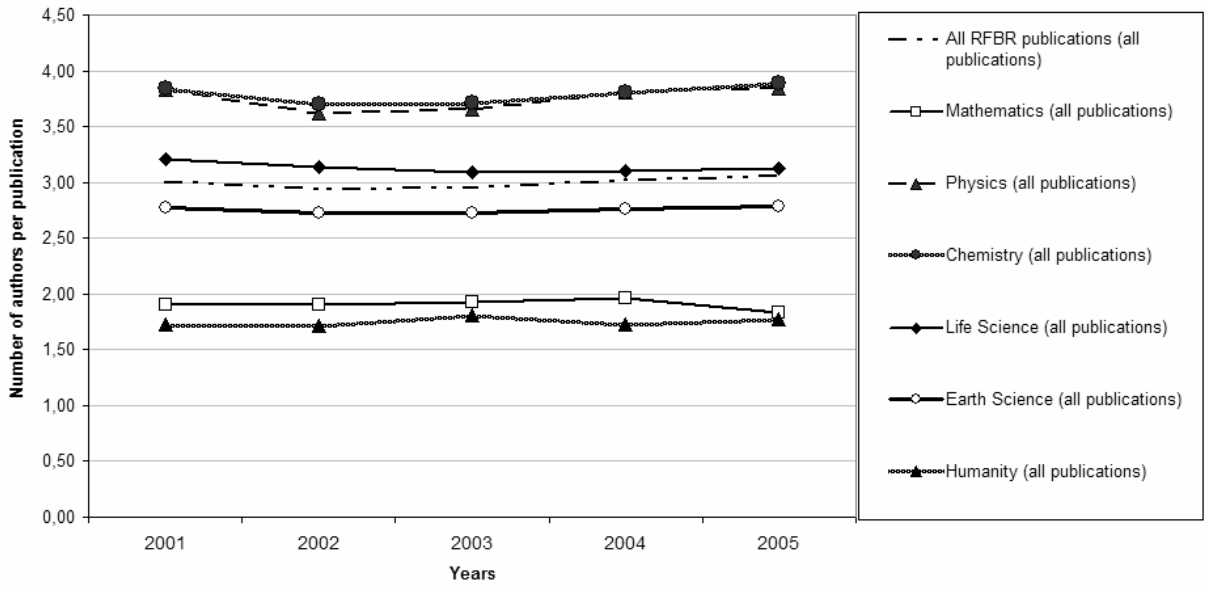


Figure 5. Co-authorship per publication of RFBR grantees, 2001-2005

Two fields of science, Mathematics and the Humanities, show significantly less co-authorships than other fields of sciences; in Physics and Chemistry the number of co-authorships is higher than average.

Our findings show that the average number of co-authors is significantly less than in SCI data.The number of co-authorships appear to be similar to an average number of co-authors in SCI data for 2003, only in internationally co-authored papers. Our findings have an important implication related to the individual performance evaluation. According to the MHES methodology, each co-author gets only a fraction of the score given to a specific article depending on journal origin – national or foreign. Many researchers argue that each co-author has to get at least half of the score. We will continue to monitor the practice of co-authorship.

An important indicator of performance activity is where articles are published. One of the goals of a researcher is a successful knowledge transfer by publishing in top journals indexed by WOS. We need

to emphasize that journal impact factor (IF) plays an important role in the evaluation of individual researchers in the new rules introduced by the MHES.

To evaluate what part of Russian research reaches a broad audience through journals as a main communication channel we investigated grantees' research output from 1997 to 2001. A list of journals, in which grantees published their research results was compiled and compared with the JCR Master List.

In general, Russian journals as well as Chinese journals have a significantly lower IF than Anglophones. This phenomenon was investigated by E. Garfield (1990) and by S. Ren and R. Rousseau (2004). In 2001 an average Russian journal's IF was 0.35; since then it has steadily grown to 0.43 in 2005.

Table 5. RFBR grantees outpt by JCR. 1997-2001

Year	Output	Share of Journal papers	Number of journal titles	Share of Journals covered by JCR	Mean weighted impact factors	Expected citations	Share of papers from total output covered by JCR
1997	56035	42.44	1916	54.23	1.58	11569.79	18.87
1998	38009	40.21	1609	57.92	1.46	8318.73	18.65
1999	35198	37.90	1583	58.05	1.67	7783.4	16.35
2000	43761	38.85	1901	56.97	1.63	10031.2	17.41
2001	53351	43.02	1878	55.56	1.67	12678.2	23.65

Russian papers were scattered among two thousand journals annually. Approximately 55% of those journals were processed by JCR. In general only 20% of the total grantees output will be accessible to the world science community. These findings confirmed that a large fraction of Russian external collaborative output is not available to the world community. This means that the Russian potential for contribution to knowledge transfer is significantly larger than could be judged by ISI statistics. These findings confirm the urgent need for a national database, which could be an important tool for bibliometric evaluation of research performance.

Statistics on each journal's IF were collected and a mean weighted impact factor (MWIF) for each year was calculated. MWIF and expected citations were calculated.

This MWIF is the ratio of the sum of the number of articles in each journal multiplied by the impact factor of this journal (taken for all journals) divided by the sum of the RFBR articles published in these journals in a particular year. As we can see from data presented in **Table 4**, Russian grantees shown a preference to publish their papers in journals with high IF. Expected citations (EC) were determined as the sum of number of articles published in a journal *i* multiplied by its IF . The expected impact for these scientists' publications is growing.

Conclusions

The key issues we address in this paper with regard to policy are:

- a strong necessity to build a national science citation database;
- the necessity to use various data sources for bilbiometric R&D evaluation: comprehensive national database and international (ISI), and to interpret the indicators with great care;

- the need to carefully revise the MHES methodology: how the contribution of Russian R&D should be evaluated, what method should be used and which indicators have to be implemented.

In following years we will see how the present method of research evaluation influences publication activity and how it impacts research quality.

References

Russian Federation Policy in R&D development till 2010 and in the future (2002). *Approved by President of RF and the Council of Ministry of RF,* Decree N-576.

Garfield E., Welljams-Doroff A.(1990). The Language Use in International Research: A Citation Analysis. *Essays of an Information Scientist,*. 13, 282-295.

Libkind I., Libkind A., Markusova V., Minin V., Jansz M.(2006). Index of RFBR as Scientific Information Search and Analysis Tool. *The 9th International Conference on Science and Technology Indicators*. Leuven, Belgium, 2006, September 7-9., 229-233.

Makarov V.L., Varshavsky A.E. (2004). Innovation management in Russia. Moscow, Nauka, 879 p.

Ren S., Rousseau R. (2004). The role of China's English-Language scientific journals in scientific communication. *Learned Publishing*, 17(2), 99-104.

Sivertsen G. (2006). A bibliometric model for performance based budgeting of research institutions.

Wilson C., Markusova V. (2004). Changes in the scientific output of Russia from 1980 to 2000, as reflected in the Science Citation Index, in relation to national politico-economic changes. *Scientometrics,* 59 (3), 345-389.

Women Involvement in Editorial Boards of Mathematics Journals[1]

Elba Mauleón and María Bordons

elba@cindoc.csic.es, mbordons@cindoc.csic.es
Centro de Información y Documentación Científica (CINDOC), Spanish Research Council (CSIC),
Joaquín Costa 22, 28002 Madrid (Spain)

Abstract
Members of editorial boards of journals play an essential role in science, since they contribute to guarantee the quality and relevance of publications. This study analyses whether women are well represented in editorial boards as a sign of her participation in responsibility positions in science. Women's representation on the editorial board of eight international high quality mathematics journals is studied and compared to women's presence as authors of papers in the same journals. On average only 4% of the members of the editorial boards analysed were women, while 10% of the authors with publications were female scientists. Inter-gender differences in impact of publications were not found, but less collaboration was observed in female publications. This study shows that women are under-represented as authors in Mathematics journals but her presence is even lower in leadership positions such as members of journal editorial boards.

Keywords
editorial boards; women and science; peer review; bibliometric indicators by gender

Introduction

Data supplied by the main statistics included in EU Reports on women and science have made clear an unequal distribution of men and women according to scientific disciplines. Women presence rate tends to be greater in areas such as Health Science, Social Science and Humanities, whereas it is very poor in Engineering, Computer Science or Mathematics. On the other hand, the shortage of female representation in high responsibility positions such as in the management boards of the entities sponsoring research, in the evaluation committees or in the editorial boards of scientific journals has been also pointed out in the literature (She Figures, 2006).

The editorial boards of scientific journals play a crucial role in science, since they contribute to guarantee the quality and relevance of the published research. The selection of referees is carried out by journal editors who usually give primacy to scientists' experience in the field and to their knowledge in a given area or discipline, so as to ensure sufficient authority to assess a piece of work. Since publications-based evaluations are essential for the promotion of scientists and for obtaining funds for their investigations, it is important to know whether women are well represented in the editorial positions that contribute to decisions about publication.

Women's involvement in editorial boards of scientific journals has been analysed by different authors in previous studies. Important differences among scientific fields in terms of women´s participation in editorial boards and in publication rates has been described (Stark, Spielmann and Ohnersorgen, 1994; Evans, Hsieh and Robinson, 2005).

Objectives

The aim of this work is to determine and evaluate the percentage of women on the editorial boards of eight mathematics journals as compared to women presence in the speciality. As a reference for the latter, the percentage of women as authors in the eight analysed journals was used. The following questions are addressed: Are women represented in the editorial boards as frequently as they are authors of papers in the corresponding journals? Are there inter-gender differences in collaboration habits and in impact of publications?

[1] This research was supported by a FPU scholarship and by the Research Project (I+D+I 87/02) supported by the Spanish Instituto de la Mujer.

Methodology

Source selection
The study was limited to a sample of high quality international journals specialised in Mathematics. The international database Science Citation Index (SCI), produced by Thomson Scientific, was used as a source of data. The Science Direct database, from the European Elsevier-Science distributor, was used to have access to full texts of documents and obtain the full name of the authors, not only the initials as recorded in the SCI. Mathscinet, the most important database in Mathematics, produced by the American Mathematical Society, was used to verify the interest of the selected journals for the area of Mathematics.

The Science Citation Index database classifies journals in categories. Mathematics had 181 publications in 2004. So as to select the sample of journals to study, the journals were ordered from higher to lower impact factor in JCR 2004. A total of eight journals that were listed among the 25% with the highest impact factor (first quartile) within the area and included at the same time in the three databases Science Citation Index, Science Direct and Mathscinet were selected.

Data processing
A database was created including bibliographic references of documents published in 2004 in the eight source journals. Only scientific articles were studied.

Main indicators
The following aspects were studied for each journal:

1. Editorial boards:
    a. International composition of the board: measured through the percentage of members from the publication country of the journal and from other countries.
    b. Men and women's presence in the committee (% men and %women).
    c. Men and women's roles in the editorial boards.
2. Documents' characteristics:
    a. Men and women's presence as authors in documents (percentage of documents signed only by women, only by men and by both).
    b. Impact by gender: percentage of non-cited documents and number of citations per document were calculated for different sets of papers (documents signed only by men, only by women and by both).
    c. Collaboration habits: average number of authors, centres and countries per document were obtained for men and women.

Authors and Editorial Board members' gender
The use of the variable of gender was a challenge, as far as authors and editorial board members' gender is not recorded in bibliographic databases or in journals. Different complementary approaches were followed to determine the gender of both groups of scientists.

Firstly, the gender of 22% of the authors was deduced directly from their names. However, most of the selected journals only showed the initial of the first name, which made the first approach impossible. In these cases, Internet was used, and the gender of 23% of the authors was identified. Finally, the gender of 51% of the scientists was identified by e-mail, since we wrote them informing about our research and data about gender was required (high response rate). Only the gender of 4% of authors remained unknown.

Results

Editorial board characteristics
The list of journals studied is shown in table 1, including the impact factor in 2004 (JCR 2004) and the rate of board members that come from the journal's country (domestic members) or from other

countries (foreign members). Considering all the journals, only 4% of the committee members are women. This rate varies from 0% (4 journals) to 11% (1 journal) (Table 1).

Table 1.International character of journals depending on the editorial committee member's country of origin

Journals	No. Members (%)			
	IF	Foreing	Domestic	Total
Computational Geometry-Theory and Applications	0.742	34 (94%)	2 (6%)	36
Journal of Combinatorial Theory, Series B	0.618	15 (71%)	6 (29%)	21
Topology	0.727	8 (53%)	7 (47%)	15
Journal of Functional Analysis	0.962	6 (46%)	7 (54%)	13
Advances in Mathematics	1.067	15 (30%)	35 (70%)	50
Journal of Differential Equations	0.877	4 (21%)	15 (79%)	19
Annales Scientifiques de l'École Normale Supérieure	1.186	2(20%)	8 (80%)	10
Journal de Mathématiques Pures et Appliqués	0.926	2 (20%)	8 (80%)	10
Total		86 (49%)	88 (51%)	174

Significant differences (Chi^2 =53.99; p=0.000)

Table 2. Presence of women in journals' editorial committees

Journal	Male	Female	Total
Topology	16 (89%)	2 (11%)	18
Advances in Mathematics	47 (94%)	3 (6%)	50
Computational Geometry-Theory And Applications	34 (94%)	2 (6%)	36
Journal of Combinatorial Theory, Series B	20 (95%)	1 (5%)	21
Journal of Differential Equations	19 (100%)	0 (0%)	19
Journal of Functional Analysis	14 (100%)	0 (0%)	14
Annales Scientifiques de l'École Normale Supérieure	10 (100%)	0 (0%)	10
Journal de Mathématiques Pures et Appliqués	10 (100%)	0 (0%)	10
Total	170 (96%)	8 (4%)	178

The percentage of women tend to increase when the size of the committee and its international character rise. However, conclusive data can not be obtained due to the small number of journals studied. Men and women's presence in the editorial committees proves that the rate of women decreases in the positions that reflect a higher degree of seniority. Therefore, no women were found in the categories of honorary member or editor (Table 3).

Table 3. Men and women's distribution by editorial role

Category	Female	Male	Total
Editorial Board	8 (6%)	138 (94%)	146
Editors	0	18 (100%)	18
Founding Editors	0	5 (100%)	5
Honorary Editor	0	9 (100%)	9
Total	8 (4%)	170 (96%)	178

Characteristics of the documents

Presence of men and women as authors

A total of 764 research articles were published in the studied journals in 2004. The study was limited to 733 papers (96% of the total) in which the gender of all the authors was identified. The percentage of authors identified ranged from 94% to 100%, depending on the journal. Table 4 shows the number of documents per journal and the presence of men and women as authors of papers. It is clear that documents signed only by men predominate in all the journals.

Table 4. Distribution of documents by gender of signing authors

Journal	All	Only men		Only women		Men and Women	
	Tot. Doc	No. Doc	%	No. Doc	%	No. Doc	%
Advances in Mathematics	131	115	87.79	5	3.82	11	8.4
Ann.Scientifiques l'École Normale Supérieure	27	22	81.48	3	11.11	2	7.41
Comp.Geometry-Theory and Appl.	44	29	65.91	1	2.27	14	31.82
J. Mathématiques Pures et Appliqués	43	30	69.77	3	6.98	10	23.26
J. Combinatorial Theory, Series B	60	55	91.67	0	0	5	8.33
J. Differential Equations	187	145	77.54	6	3.21	36	19.25
J. Functional Analysis	184	153	83.15	9	4.89	22	11.96
Topology	57	49	85.96	5	8.77	3	5.26
Total	733	598	81.58	32	4.37	103	14.05

Considering publications in the eight journals analysed, a total of 1358 researchers and 1448 authorships were found. Concerning gender, 10% of the researchers who published in these journals were women, increasing this percentage to 15% in the *Journal de Mathématiques Pures et Apliques.*

Impact by gender

Approximately 40% of documents had not received any citation one year and a half after their publication. This rate varies from 34% to 52%, according to journals.

A total of 886 citations were distributed among 445 cited documents, so on average each document received 1.95 citations. Documents signed only by women tend to receive a slightly lower number of citations (1.42±0.77) than those which were signed only by men (1.94±1.45) or those that were signed by both men and women (2.10±1.35), but the differences were not statistically significant.

Cooperation and internationalisation

Around 40% of documents were published by 2 authors and 35% by a single author. An average of 1.98 authors and 1.7 centres per document were found (Table 5). These values are coherent with those described in literature concerning the area of Maths. Women presented a lower trend toward collaboration than men as shown by the lower number of authors, centres and countries found in "only women" documents.

Conclusions

Methodological aspects

The lack of standardisation of authors' names in journals and bibliographic databases raises serious problems for the development of studies at the micro level, since we need to identify properly every author –which can appear under different variant names- and his/her production. Concerning the composition of editorial boards of journals, the lack of standardisation of the information provided by journals themselves about the name and the institutional affiliation of board members was also observed. According to our data, journals from the USA give more complete information than the European ones.

Table 5. Cooperation habits: amount of authors, centres and countries per document

	N	Average	STD	Min	Max	SIG
No. Au/Doc						
Only men	598	1.88	0.90	1.00	7.00	0.000
Only women	32	1.09	0.30	1.00	2.00	
Men and women	103	2.80	1.00	2.00	7.00	
Total	733	1.98	0.97	1.00	7.00	
No. Centre/Doc						
Only men	598	1.68	0.78	1.00	5.00	0.000
Only women	32	1.09	0.30	1.00	2.00	
Men and women	103	2.29	0.91	1.00	5.00	
Total	733	1.74	0.82	1.00	5.00	
No. Countries/Doc						
Only men	598	1.40	0.58	1.00	3.00	0.000
Only women	32	1.06	0.25	1.00	2.00	
Men and women	135	1.56	0.73	1.00	4.00	
Total	733	1.43	0.61	1.00	4.00	

Editorial boards

Our results show that only 4% of the members of the editorial boards of the Maths journals analysed are women. This value is much lower than the percentage of female authors that publish in the journals (10%). We have analysed eight boards, in which there are 21 men per woman. The infra-representation of women in editorial boards has been described also in other areas, such as Medical Science and Life Science (Gilbert, Williams and Lundberg, 1994; Katterman, 1996). Editorial board members form part of the exclusive scientific elite, so this finding indicates the difficulty of women to arrive to the most prestigious positions in science, as has been repeatedly shown through the well known "scissors diagram". Although the percentage of women in the higher education system is around 43% in Europe, it decreases at the highest level of the academic hierarchy, and only 2-18% of full professor are women (She figures, 2006). This percentage is even lower in Mathematics in which less than 5% of full professors were female in 1997 in some European countries (ETAN, 2000).

Impact and collaboration

Although previous studies on scientific activity by gender have discribed that documents published by women are more frequently cited than those published by men (Feller, 2004; Long 1992), inter-gender differences in number of citations per document or in the percentage of non-cited documents were not found in our study. Concerning collaboration, the indicators show low collaboration in Mathematics, which is a characteristic of the field (Bordons and Gómez, 2000). Comparing "only women" with "only men" documents, female scientists show a greater trend to publish single-authored documents.

In summary, our data indicate that women are under-represented in the editorial boards of the Mathematics journals analysed. We think that the participation of women in editorial boards should be increased, at least to assure the parity between the percentages of women on editorial boards of journals and the presence of women as scientists in the area.

References

Bordons, M.; Gomez, I. (2000). Collaboration networks in science. In B. Cronin & H. B. Atkins (Eds.), *The web of knowledge: A festschrift in Honor of Eugene Garfield*, (pp 197-213). Medford, NJ: Information Today.

Evans, J. L., Hsieh, P., & Robinson, D. H. (2005). Women's involvement in educational psychology journals from 1976 to 2004. *Educational Psychology Review*, 17, 263-271.

ETAN expert working group on Women and Science. Science Policy in the European Union. Promoting excellence through mainstreaming gender equality. European Commission. Directorate-General for Research.

Gilbert, J. ; Williams, E.; Lundberg, G. (1994). Is there gender bias in JAMA's peer review process? *JAMA*, 272, 139-142.

Katterman, L. (1996). Women still rarely named to fill top posts at life sciences journals. *The Scientist*. March 4, 1-8.

Long, J. S. (1992). Measures of sex differences in scientific productivity', *Social Forces*, 71, (1), 159-78.

Stark, Barbara L., Katherine A. Spielmann, Brenda Shears, and Michael Ohnersorgen (1997) *'The Gender Effect on Editorial Boards and in Academia.' SAA Bulletin*, 15 (4),p 1-8.

Women and Science Statistics and Indicators, She figures, 2006 Comisión Europea, Dirección General de Investigación.

Zuckerman, H. (1987) Persistence and change in the careers of men and women scientists and engineers: A review of current research', in L.S. Dixon (ed.) Women: Their Under-Representation and Career Differentials in Science and Engineering. Washington: National Technical Information Service, 123-156.

Analysing Influence Over Time: An Historiographic Mapping of the Research of Conrad Hal Waddington (1905-1975)

Katherine W. McCain

kate.mccain@ischool.drexel.edu
College of Information Science & Technology, Drexel University, 3141 Chestnut Street, Philadelphia, PA 19104 (USA)

Abstract
Historiographic mapping is used to study the influence of Conrad Hal Waddington, a developmental biologist/evolutionary geneticist whose publications span the years 1929 – 1977. Data were gathered from the Web of Science for a 60 year period (1945-2004) and analyzed using HistCite. A map of the entire period shows that Waddington is recognized for two major research themes – canalization/genetic assimilation and embryonic induction. When the data set is split into three 20 year periods, canalization/genetic assimilation appears as an active, visible area of contemporary research in all three while embryonic induction is essentially invisible in the middle period (1965-1984). This corresponds to historians' accounts of research trends in that time frame, in which funding and interest shifted from the level of the organism to the molecular level. The emergence of Evolutionary Developmental Biology is marginally visible in the last time period, although work explicitly linking Waddington's two themes is not prominent. A "time slice" approach to historiographic mapping can bring to light research activities and influences that may not be visible in a data set that covers a longer time period.

Keywords
algorithmic historiography; citation networks; evolutionary developmental biology; historiographic mapping; history of science

Introduction

When Eugene Garfield introduced the concept of algorithmic historiography and made his HistCite software available to researchers, he provided a valuable new tool for dissecting and visualizing the flow of ideas in scholarly fields (Garfield, 2004; Garfield, Pudovkin & Istomin, 2003). The data used for this type of analysis are gathered from the Web of Science and the results include lists of highly cited works within and without a central network of highly cited papers indexed by ISI. Historiographs – visualizations of the citation networks – are also available.

The "standard" use of HistCite has been to gather a single multi-year data set from WoS based on subject terms or an author's ISI-indexed *oeuvre* and explore its content and structure. In this paper I present a different use of HistCite – a multi-period "time slice" analysis of the influence of a single author's full published body of work. The author is Conrad Hal Waddington and I examine the citation networks associated with his work over three successive time periods from 1945 to 2004. Waddington is considered to be a foundational author in evolutionary developmental biology (Evo-Devo or EDB). EDB is an emerging interdisciplinary field that rejoins two streams of research—embryology/developmental biology and genetics/evolutionary studies—previously separated since the early 20th century (Gilbert et al. 1996). Waddington is of particular interest because his work spans both research streams and has experienced a resurgence in interest with the rise of EDB.

The results demonstrate the usefulness and limitations of this type of analysis in studying intellectual history. A related poster (if accepted) focuses on Waddington's position in EDB in the most recent decade (1995-2004) using both historiographic mapping and author tri-citation mapping.

Conrad Hal Waddington, 1905-1975

Slack (2002) characterizes Waddington as a "true 20th century polymath." Not only did he publish in geology and many related areas in the biological sciences (experimental embryology and developmental biology, developmental genetics, population genetics, theoretical biology), but he also wrote on operations research in WWII, ethics, and on the relationship between science and modern art.

Waddington is best known today for two areas of research, embryonic induction and canalization/genetic assimilation (Robertson, 1977, Slack, 2002). *Embryonic induction* is the effect of one set of cells (the "organizer") in the developing embryo on a second set – an effect that changes the developmental pathway of the second. In the 1930s, Waddington was able to demonstrate the existence of an organizer in mammals and birds – a region of the developing embryo that, when transplanted, could induce a new complete embryonic structure – and collaborated on studies of the chemical basis of this "organizer." *Genetic assimilation* is the process by which changes in an organism that are originally induced by the environment become expressed even in the absence of the environmental cause—as hidden genetic variation becomes visible after severe stress. *Canalization* (one of many novel terms introduced by Waddington) is "the property of developmental pathways to produce standard phenotypes despite mild environmental or genetic perturbations." (Gilbert 2000). This branching set of paths along which development occurs (via successive stages of induction and change) was described by Waddington as the "epigenetic landscape." While today (see later discussion), these two topic areas may still appear as distinct research topics (outside of EDB), Waddington himself saw genetics, embryology, and evolution as a single, strongly interconnected field – "diachronic biology." (Gilbert, 2000; Hall, 2004; Vijayraghavan, 2006). For more details, and an extensive bibliography of Waddington's publications, see Robertson (1977), Hall (1992) and Slack (2002).

Methods

Data Gathering

Algorithmic historiography / historiographic mapping focus on highly cited works within a document set and the citation links among them (Garfield 2004). In order to look at the influence of Waddington and his coauthors, it was necessary to first build an appropriate document set. The final set included:

- All papers citing anything by Waddington (as first author) were retrieved from the Web of Science (WoS) for the years 1945-2004.
- All papers in the Web of Science written by Waddington as either first or co-author were retrieved for the same time period.
- Pre-1945 papers by Waddington were entered into the set in WoS retrieval format so that they could be included as cited works visible in the citation network. Citations were not included as part of the WoS record.
- Books and other highly cited works by Waddington that were not indexed in the WoS were also entered in WoS retrieval format, again without citations. All editions of the same work were given the same year of publication (the earliest) and the 4 volumes of *Toward a Theoretical Biology* were given the date of volume 1 (1968).
- Cited references to Waddington's works were normalized with a standard generic citation string. This was primarily necessary for the books, since not only did the date of publication vary for different editions but often the citation string included reference to a specific page number.

The data were subdivided in 20 year periods for this report.

Analysis and Post-Processing

The major results of algorithmic historiography are reported as lists of documents within and outside the citation network, rankable both by within network and total citation counts (see Garfield et al., 2003 for screen shots of the various output displays). Graphs of citation links between documents cited above a researcher-selected threshold can also be generated and it is these graphs that constitute historiographic mapping. In order to create a citation linkage network that both contained ID labels for the network nodes (citing and cited works) and that could be manipulated for clarity, I converted the citation matrix for the network being analyzed into a form that could be processed by UCINet and its graphic display component NetDraw.

For each of the time periods, I experimented with a range of citation thresholds to find one producing networks that struck a useful balance between detail and clarity. Because I was interested in identifying significant research themes, I sought solutions with a clear structure (whether showing a

single network or several distinct networks). For legibility, I only include networks of three or more documents in the graphs presented and do not represent total citation counts for the networked articles by varying the size of the nodes in the network.

Within-network research themes were initially identified using NetDraw's Newman-Girvan algorithm. This algorithm (Girvan & Newman, 2002) identifies groups of nodes within a network that are more connected internally than externally. The directional arrows showing citing/cited links have been reversed from the original HistCite convention and indicate information flow from older to more current papers rather than citation pointers from newer works to the older ones they cite. For clarity in presentation, nodes within a network are arranged in temporal order (oldest at the top). Small networks are grouped on the display area. Their placement has no relationship to the temporal arrangement of the nodes in the main network. Citing pairs (dyads) and disconnected cited works (singletons) are not included. Network subgroups as well as separate networks are highlighted by differences in node shape (and color, although this is not useful in photocopies or b/w printing). The two major Waddington-relevant themes—canalization/genetic assimilation (C/GA) and embryonic induction are consistently coded with the same shapes/colors to support discussion of their relative importance in each time period. (Constraints on paper length prevent providing full bibliographic information for the coded nodes in each network. A bibliography for each time period is available from the author.)

Results & Discusión

Overview of Waddington's citation counts
Figure 1 shows the annual count of citing papers for all of Waddington's work (authored and co-authored)

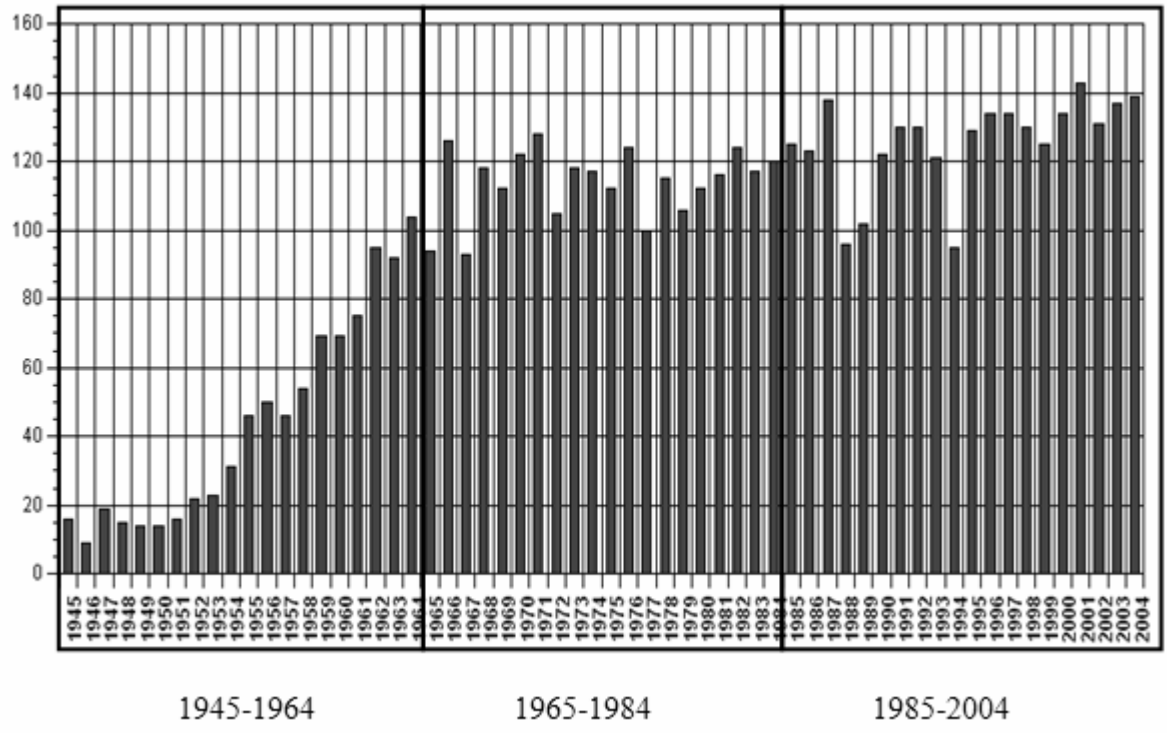


Figure 1. Citation history for Waddington-authored papers, 1945-2004, showing 20 year partitions

Table 1 gives some basic statistics for the different time periods studied. The citation threshold was increased for each successive time period to keep the above-threshold paper sets roughly the same size

Table 1. Basic statistics for 60 year and 20 year data sets

Time Period	#articles*	citation threshold	#above threshold	#3+ networks	# dyads	# singletons
1945-1964	916	9	64	4	1	8
1965-1984	2238	15	63	4	1	19
1985-2004	2515	25	65	1	0	7
1945-2004	5669	35	86	1	0	9

*all time periods also include articles by Waddington and coauthors that predate the time span

Time period network analyses

Pre-1945 publications
The Web of Science does not go back before 1945 and papers in Waddington's pre-war years are included in this study only as cited works. In this time period, after a brief fling with paleontology he was drawn to experimental embryology and in particular the study of induction. He drew on the work of Spemann, who won the Nobel Prize for his studies of induction in amphibians, and collaborated with the Needhams on a search for the chemical cause of induction and cell differentiation in mammals and birds. It is this work that signals later research streams in aspects of embryonic induction. His book *Organisers and Genes*, published in 1940, brings together his ideas from experimental embryology and genetics, including the idea of epigenetic landscapes and canalization (Stern 2000)—the latter topic being a major focus of his post-war research.

1945-2004
When the data are considered as a 60 year whole, a single network of 77 documents (plus 9 singletons—cited but un-linked works) exists above the 35 citation threshold. Waddington's work is differentially cited by two distinct research communities. On the left (circles)is a single,densely connected subnetwork focusing on C/GA. On the right, there are 4 subnetworks relating to experimental embryology and embryonic induction: amphibian embryology [diamonds], theoretical models and experimental embryology [upward triangles], embryology of *Drosophila* [downward triangles], and embryonic induction [squares]. The two sides are loosely linked by a small number of direct or indirect citations from each to *Organisers & Genes* – Waddington's 1940 book which introduced the idea that "development and evolution can be viewed as a succession of relatively stable states, separated by periods of instability and change" (Stern 2000) and *Introduction to Embryology,* his major embryology text .

An author-focused analysis over a 60 year period naturally favors the author's work and older works in general – and a high proportion of the works in Figure 1 (below) are by Waddington and his colleagues. For a more detailed examination of Waddingtonian research themes over time, the data were divided into three 20-year time periods. Visual coding of network elements in the remaining maps use the same node shapes—circles for canalization/genetic assimilation and squares for embryonic induction—to highlight Waddington's two main themes. Coding is based on the presence in the network of two or more of these key works:

- Canalization/genetic assimilation (circles):
    - 1942c. Canalization of development and the inheritance of acquired characters. *Nature* 150:563
    - 1953a. Genetic assimilation of an acquired character. *Evolution* 7(2):118-126
    - 1956a. Genetic assimilation of the bithorax phenotype. *Evolution* 10(2):1-13
    - 1957. *The Strategy of the Genes* London: Allen & Unwin

- Embryonic induction (squares)
    - 1932. Experiments on the development of chick and duck embryos, cultivated in vitro. *Phil Trans Roy Soc B*. 221:179-230
    - 1933a. Induction by the primitive streak and its derivatives in the chick. *Journal of Experimental Biology* 10:38-46
    - 1933b Induction by the endoderm in birds. *Wm Roux Arch Entw Mech Org* 128:502-521

Organisers and Genes appears to function as a "bridging" work connecting the two themes. Its coding is dependent on the network structure in a given time period.

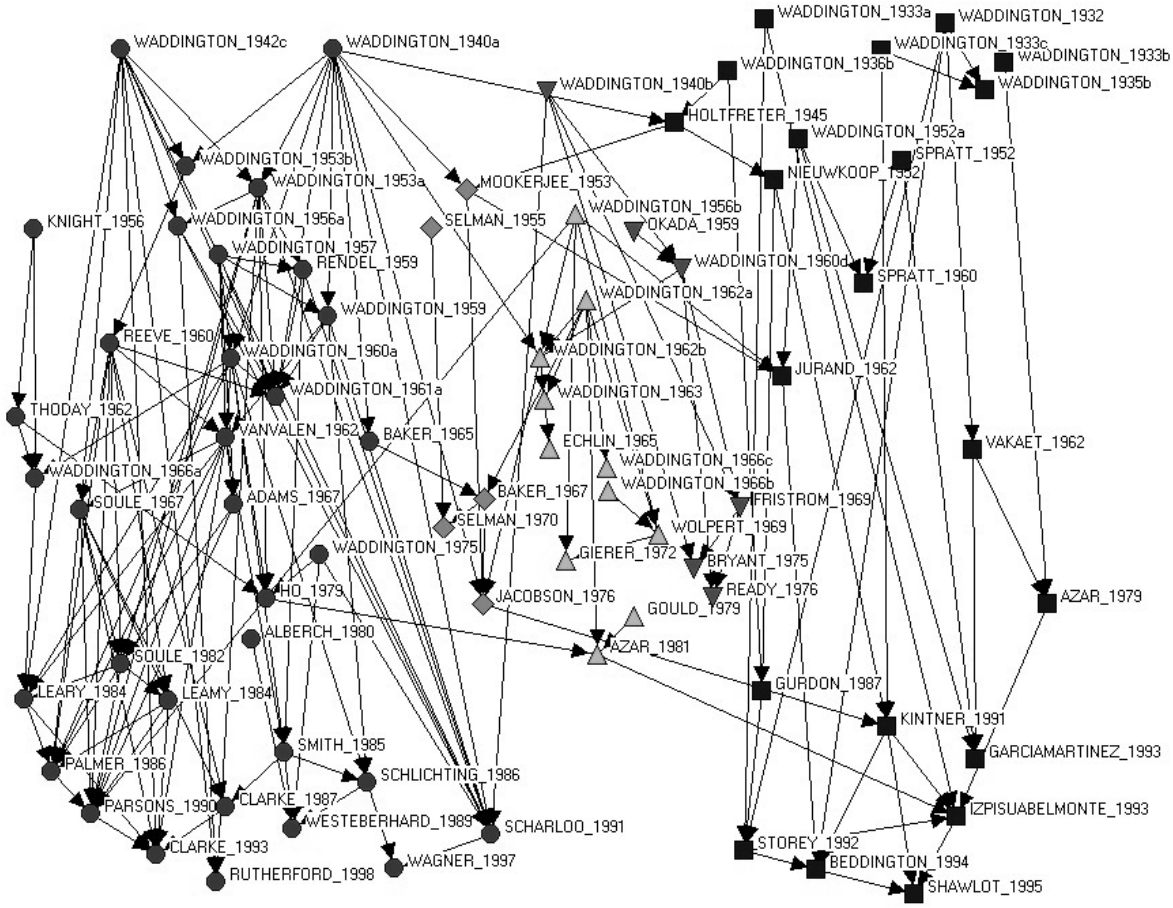


Figure 2. 1945-2004 network with citation threshold = 35+

1945-64

This 20 year period encompasses most of Waddington's post-war research. He is an active participant in the research streams and only his pre-1945 publications are entered without citations. Using a citation threshold of 9+ citations, the citing literature breaks out into 1 large and three smaller networks with 3 or more components.. Both of Waddington's major themes—canalization/genetic assimilation and embryonic induction—are visible in the largest network, connected by a subset of papers focusing on the induction of the nervous system [upward triangles] and another that deals with amino acid uptake and metabolism in the developing embryo [downward triangles].

His most highly cited book, *The Strategy of the Genes*, is barely visible, being cited only by one contemporary paper. The individual smaller networks focus on cellular differentiation [double triangle], the teratogenic effects of trypan blue [box] and the mechanics of cell cleavage [diamonds].

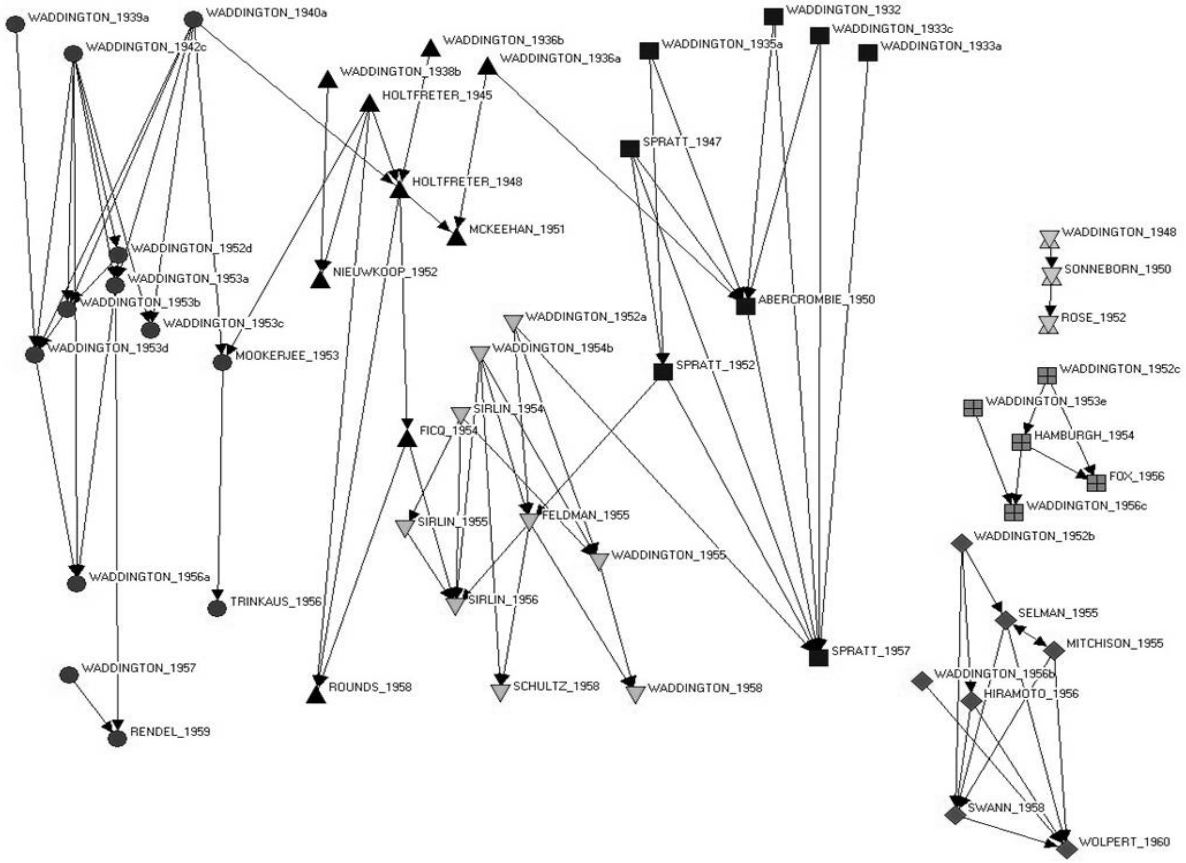


Figure 3. 1945-1964 networks with citation threshold = 9+

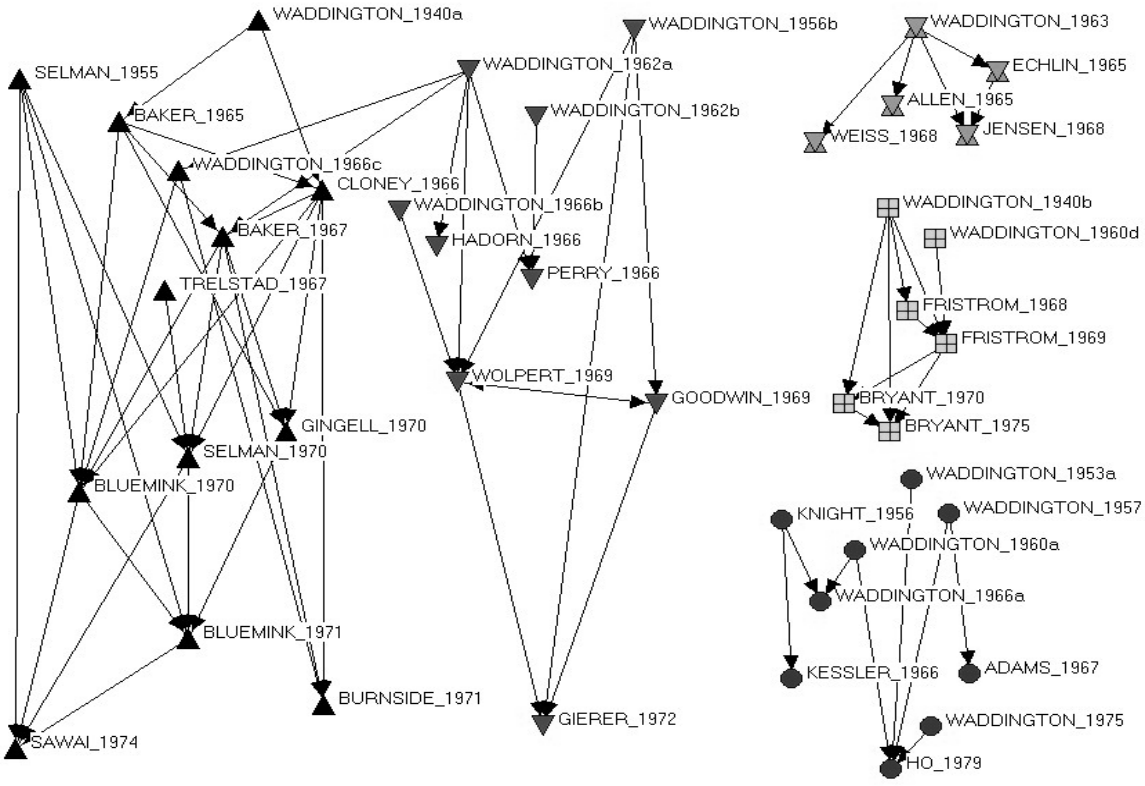


Figure 4. 1965-1984 networks with citation threshold = 15+

1965-84

The 1965-84 period (Figure 4 above) includes Waddington's work in later years as well as work by his students. The number of papers has more than doubled and the threshold has been raised to 15. There is one large and three smaller networks—the largest network contained paper subsets on cell structure and cytokinesis [upward triangles] and cellular development and differentiation [downward triangles]. The smaller networks represent polyribosome structure and configuration [double triangles], wing development in *Drosophila* [boxes] and C/GA (circles). Waddington's second major theme – embryonic induction – is not visible in this time period as an area of active research.

1985-2004

This last 20 year period spans the emergence of Evolutionary Developmental Biology and the reinvigoration of interest in Waddington's work on embryonic induction. While the number of citing papers is roughly the same as in the previous time period, the citation threshold has been raised to 25, producing a single network with several components.

As in the 60-year map, he major network is loosely split between topics in C/GA and Embryonic Induction, with the link between the two again made by citations to *Organisers and Genes* [Waddington 1940_a]. To the left are subnetworks dealing with inheritance of epigenetic defects (double triangles), C/GA focusing on developmental stability/fluctuating asymmetry (circles) and phenotypic plasticity (the ability of organisms to change their appearance based on changes in the environment –e.g. workers, soldiers and queen bees) (downward triangles), genetic regulation (circle in square), and genetic control of wing development in *Drosophila*. *The Strategy of the Gene* is cited by current papers in the latter two subnetworks. Embryonic induction reappears, connected to the left-hand networks via citations to *Organisers and Genes* and having a tightly linked set of current papers citing key Waddington work from the 1930s. The emergence of Evolutionary Developmental Biology is hinted at by the linkage of Gilbert et al (1996) to Waddington's text *Principles of Development and Differentiation* [Waddington_1966b] and *Organisers and Genes.*

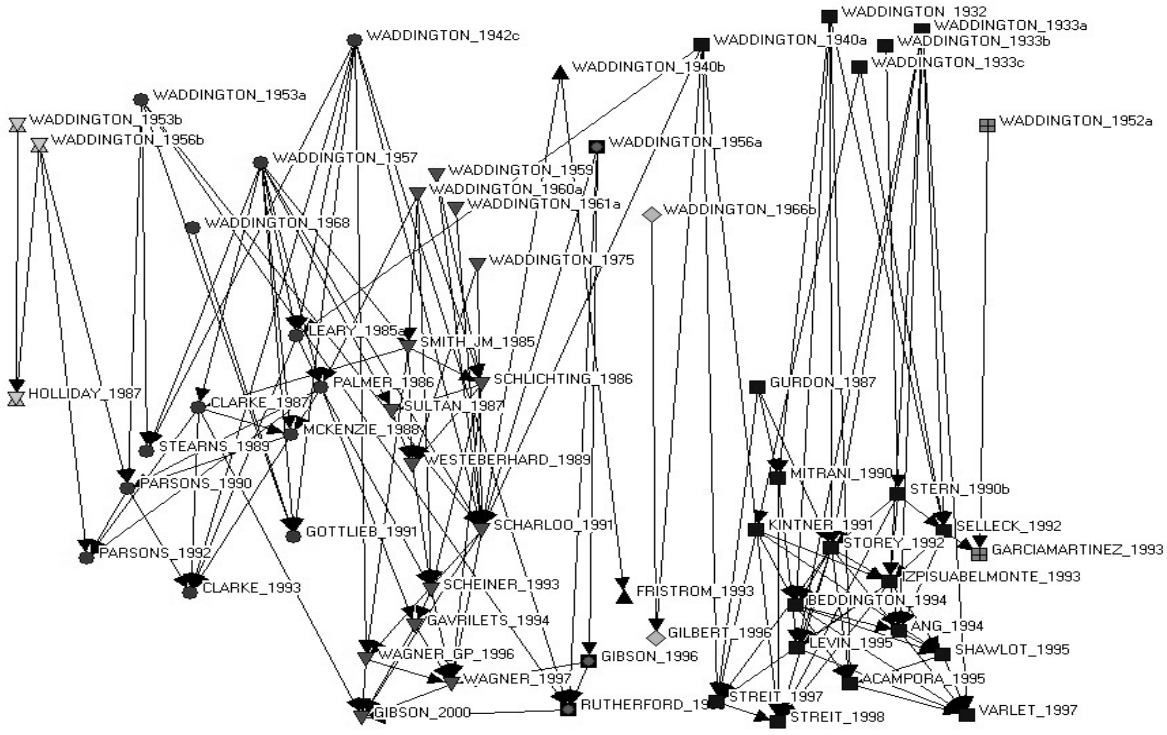


Figure 5. 1985-2004 networks with citation threshold = 24+

Assessing subject themes and change over time

When looking at the networks in Figures 2-5, it's important to remember that each contemporary paper (that is, a paper published in the selected time period) appears in a network for two reasons. First, because it is sufficiently highly cited within its time period to rise above the established citation threshold and, second, because it cites some published work by Waddington and his students or colleagues that also received sufficient citations to be potentially visible in the network. Papers published early in the time period have a better chance of making the cut than do papers published toward the end. Works by Waddington that are highly cited in general, but not by highly cited contemporary papers, appear as singletons—as do contemporary papers that cite less-well-referenced Waddington works. (A good example of the latter is Gould and Lewontin, 1979, which cites a Waddington paper that is sufficiently obscure to be missing from Robertson's (1977) extensive bibliography.) Singletons and dyads represent nascent networks that might (but are not guaranteed to) emerge were the citation threshold to be lowered.

So the networks are, in a real sense, the tip of icebergs whose full mass includes many other papers published in that time period and additional works authored or coauthored by Waddington. The appearance of a network of 3 or more papers is a good indication that the subject matter represented by the contemporary papers and the older Waddingtonian works is an active area of fairly extensive research interest in that time period– even if the papers within the network seem to represent an internally coherent and even incestuous author group. The visible portion of the network may contain primarily works by Waddington and his students—but the citing papers include many authors not immediately connected to Waddington.

Each of the 20-year networks includes small, coherent research topics that appear only in that time period. Canalization/genetic assimilation appears to be the only truly persistent research theme that is grounded in Waddington's work. Two of Waddington's works in this area [Waddington 1953a, 1957] appearing in all three 20 year periods and *Organisers and Genes,* a book that is cited by both C/GA and embryonic induction researchers, is also visible throughout the entire 1945-2004 period. What changes over time is the number of papers focusing on C/GA broadly or specifically, the density of citation linkages among the papers, and the visibility of specialized topics within C/GA, such as phenotypic plasticity and fluctuating asymmetry in the years spanning the emergence of Evolutionary Developmental Biology.

Embryonic induction, the second of Waddington's research themes, has a more interesting pattern. It is highly visible in both the earliest (1945-64) and latest (1985-2004) networks but is essentially invisible in the intervening time period – no network for this theme exists in 1965-84. This corresponds to accounts in the history of embryology/developmental biology (Gilbert & Saxen 1993, Gilbert et al. 1996). During this time period, the concept of embryonic induction was being shunted aside in favor of interest in gene regulation. Funding and support for Waddington's epignetics laboratory at Cambridge also supported mostly molecular biology research (Robertson, 1977, p. 583).

Why do older works continue to be cited?.

Cited works are understood to represent concept symbols (Small, 1978) and are valued as those concepts continue to be important to the more recent authors working in that research area. While books are likely to have long citation histories as their ideas diffuse across time and subject areas (see, e.g. Garfield, 1985, on Price's *Little Science, Big Science* and McCain & Salvucci, 2006, on Brooks' *The Mythical Man-Month*), citation histories for journal articles generally follow one of two patterns (Aversa, 1985; McCain & Turner 1989). The majority receive maximum citation counts in years 2-3 after publication and then their citation counts drop precipitously as the findings are incorporated into the body of scientific knowledge. A minority have citation peaks at year 6 or later and may continue to be cited for decades. McCain and Turner found that, at least in molecular biology, the difference seemed to be the role of the article content in subsequent research. Late-peaking, slowly aging papers were likely to be cited for important widely useful methodologies or fruitful, broadly relevant theoretical insights while the early-peaking, quickly-aging papers were being cited primarily for their explicit research results. A citation context analysis is beyond the scope of this study, but would be a

useful next step in testing this characterization for the various papers published by Waddington alone and with colleagues and students (most of the work in experimental embryology/embryonic induction).

Final Comments

Including books, older cited journal articles, and current literature in the same analysis appears to be a useful way of assessing the impact of that older work over time as it is involved in short term and longer term research. Here, we have seen that some of Waddington's earliest work in canalization/genetic assimilation and embryonic induction continued to be cited for more than 50 years after publication while other, frequently co-authored work had a limited life as the specific research topic was addressed, dealt with, and the findings incorporated into the growing base of scientific knowledge. Many of Waddington's books and journal articles were widely cited but not built upon in the systematic fashion that signals research within a specialty community.

The incorporation of Waddington's ideas as part of Evolutionary Developmental Biology is suggested by the appearance of Gilbert et al. (1996) in Figure 4 and in the resurgence of interest in embryonic induction. However, it is not made explicit in the sense of a number of papers co-citing relevant contributions from his two streams of research. Research reported elsewhere (submitted poster) shows that Waddington's name is a useful probe to identify historical and contemporary authors working in this intersection of genetics, evolution, and developmental biology. But his role as a major source of theoretical insights and relevant research findings requires further investigation.

Taking "time-slices" of a 60 year span of the literature allows a more focused analysis of research topics, their change over time, and the influence of key authors than does a single analysis of the time period as a whole. In the latter case, unless some kind of gradient is built into the analysis (such as those used by Chen in his visualizations—see, e.g. Chen, 2006), older, well-cited works are greatly privileged over newer ones, short-term research streams are not visible, and the "above-threshold" data set simply becomes a "greatest hits" list. For the same reason, successive time slices, such as those in this report, aren't completely satisfactory either. Papers published toward the end of the decade are less likely to be able to receive sufficient citations to make the threshold and are also not included in the next period.

References

Aversa, E.S. (1985). Citation patterns of highly cited papers and their relationshop to literature aging: a study of the working literature. *Scientometrics,* 7, 383-389.

Chen, Chaomei. (2006). CiteSpace: Visualizing patterns and trends in scientific literature. http://cluster.cis.drexel.edu/~cchen/citespace/ Accessed 27 November 2006.

Garfield, E. (2004). Historiographic mapping of knowledge domains literature. *Journal of Information Science*, 30, 119-145

Garfield, E. (1985). In tribute to Derek John de Solla Price: A citation analysis of *Little Science, Big Science. Scientometrics,* 7, 487-503.

Garfield, E.; Pudovkin A.I. & Istomin, V.S. (2003). *Journal of the American Society for Information Science and Technology*, 54 (5):400-412

Gilbert, S.F. (2000). Diachronic biology meets Evo-Devo: C.H. Waddington's approach to evolutionary developmental biology. *American Zoologist,* 40,729-737.

Gilbert, S.F (1996). A brief history of premolecular induction studies. *Seminars in Cell and Developmental Biology,*. 7, 67 -76.

Gilbert, S. F. & Saxén, L. (1993). Spemann's Organizer: Models and molecules. *Mechanisms of Development,* 41, 73 - 89.

Gilbert, S.F.; Opitz, J.M. & R.A. Raff. (1996). Resynthesizing evolutionary and developmental biology. *Developmental Biology,* 173, 357-372.

Girvan, M. & Newman, M.E.J. (2002). Community structure in social and biological networks. *PNAS*, 99, 7821-7826.

Gould, S.J. & Lewontin R.C. (1979). The spandrels of San Marco and the Panglossian paradigm. A critique of the adaptionist program. *Proceedings of the Royal Society of London, Series B, Biology,* 205, 581-598.

Hall, B.K. (2004). In search of evolutionary developmental mechanisms: the 30-year gap between 1944 and 1974. *Journal of Experimental Zoology (Molecular Development and Evolution),* 302B, 5-18.

Hall, B.K. 1992. Waddington's legacy in development and evolution. *American Zoologist* 32:113-122.

McCain, K.W & Salvucci, L.J. (2006). How influential is Brooks' Law? A longitudinal citation context analysis of Frederick Brooks' *The Mythical Man-Month Journal of Information Science,* 32, 277-295.

McCain, K.W. & Turner, K. (1989) Citation context analysis and aging patterns of journal articles in molecular genetics. *Scientometrics,* 17, 127-163.

Robertson, A. (1977). Conrad Hall Waddington, 8 November 190 – 26 September 1975. *Biographical Memoirs of Fellows of the Royal Society,* 23, 575-622.

Slack, J.M.W. (2002). Conrad Hal Waddington: the last Renaissance biologist? *Nature Reviews: Genetics,* 3, 889-894.

Small, H.G. (1978). Cited documents as concept symbols. *Social Studies of Science*, 8, 327-340.

Stern, C.D. (2000). Conrad H. Waddington's contributions to avian and mammalian development, 1930-1940. *International Journal of Developmental Biology,* 44, 15-23.

Vijayraghavan, K. (2006). Reading Waddington today. *Journal of Genetics,* 85, 99.

Fusion Approach to Citation-Based Quality Assessment

Lokman Meho and Kiduk Yang

{meho, kiyang}@indiana.edu
School of Library and Information Science, Indiana University 1320 E. 10th Street, LI011,
Bloomington, IN 47401 (USA)

Abstract
Citation analysis is one of the most widely used methods in evaluating the research performance of scholars. Despite the widespread use, critics claim that citation analysis has serious limitations in both data and methodology. The problems reported in literature point to two fundamental shortcomings with the typical citation analysis approach. First, conventional citation analysis methods yield one-dimensional and sometimes misleading evaluation as a result of not taking into account differences in citation quality, not filtering out citation noise such as self-citations, and not considering non-numeric aspects of citations such as language, culture, and time. Second, the coverage of citations in citation databases of today is disjoint and incomplete, which can result in conflicting quality assessment outcomes across different data sources. To address these limitations, we are developing a multi-faceted fusion approach to quality assessment that employs a range of citation-based methods to analyze data from multiple sources. The paper briefly describes a work-in-progress prototype system called CiteSearch, and discusses a citation analysis pilot study, which measures the impact of scholarly publications based on the data mined from Web of Science, Scopus, and Google Scholar.

Keywords
citation analysis; quality assessment; fusion

Introduction

Citation analysis is one of the most widely used methods in evaluating the research performance of scholars (Holden, Rosenberg, & Barker, 2005; Moed, 2005; Lewison, 2001). The basic assumption underlying citation analysis is that citations are a way of giving credit to and recognizing the value, quality, or significance of an author's work (Cronin, 1984; van Raan, 1996). While the proponents have reported the validity of using citation counts for research assessments (Aksnes & Taxt, 2004; Holmes & Oppenheim, 2001; Martin, 1996), critics claim that citation analysis has serious limitations in both data and methodology (MacRoberts & MacRoberts, 1996; Seglen, 1998). The problems reported in literature point to two fundamental shortcomings with the typical citation analysis approach. First, conventional citation analysis methods yield one-dimensional and sometimes misleading evaluation as a result of not taking into account differences in citation quality, not filtering out citation noise such as self-citations, and not considering non-numeric aspects of citations such as language, culture, and time. Second, the coverage of citations in citation databases of today is disjoint and incomplete, which can result in conflicting quality assessment outcomes across different data sources.

To address these limitations and produce a reliable and efficient indicator for assessing the relative impact and quality of scholarly publications, we are developing a multi-faceted fusion approach to information quality assessment that employs a range of citation-based methods to analyze data from multiple sources. The paper briefly describes a work-in-progress prototype system called *CiteSearch*, which will analyze combined data from multiple citation databases to produce citation-based quality evaluation measures such as CiteRank, H-Index, and Mentor-Index, and discusses a citation analysis pilot study, which measures the impact of scholarly publications based on the data mined from Web of Science, Scopus, and Google Scholar.

CiteSearch System

The CiteSearch system, which is being developed by the VCoB project[1], is a Web-based citation search and analysis system that facilitates the citation-based assessment of information by extracting

[1] The Virtual Collection Builder (VCoB) is one of the research project undertaken by the Web Information Discovery Integrated Tool (WIDIT) Laboratory (http://elvis.slis.indiana.edu/) at Indiana University School of Library and Information

and analyzing citation metadata from multiple citation databases. The implementation of CiteSearch prototype is currently under way, so what follows is a general description and brief overview of the system design.

In order to realize the goal of developing a robust citation-based quality assessment method that overcomes the weaknesses of existing citation analysis approaches, the CiteSearch system focuses on two key objectives: integration of citation data from multiple sources to ensure that quality assessment data is both complete and clean (i.e., without noise), and multi-faceted citation analysis to leverage not only multiple sources but also multiple aspects of evidence in the citation data landscape. Integration of citation data from multiple sources involves data mining, filtering, metadata extraction, data normalization and data fusion, while multi-faceted citation analysis involves derivation and fusion of multiple citation-based quality evaluation measures such as CiteRank (a citation propagation measure similar to Google's PageRank), H-Index (an index to quantify an individual's scientific research output), and Mentor-Index (an index to measure mentoring impact by aggregation of students' research impact).

Given a publication title, for example, the *CiteSearch* system will automatically search multiple Web-based citation databases such as *EBSCO*, *Proquest* and *Google Scholar*, and analyze the search results to produce bibliographical metadata of all citations and compute various citation-based quality evaluation measures such as CiteRank, weighted CiteRank, which is CiteRank weighted by source, author, or time of citations. As a pilot study, we are using complete publication lists of full-time faculty members randomly selected from American Library Association-accredited library and information science programs to generate citation metadata with CiteSearch system. The initial citation metadata will then be aggregated and analyzed to produce meta-level citation measures for authors, publications, and schools. Figure 1 displays the overview of the CiteSearch system architecture.

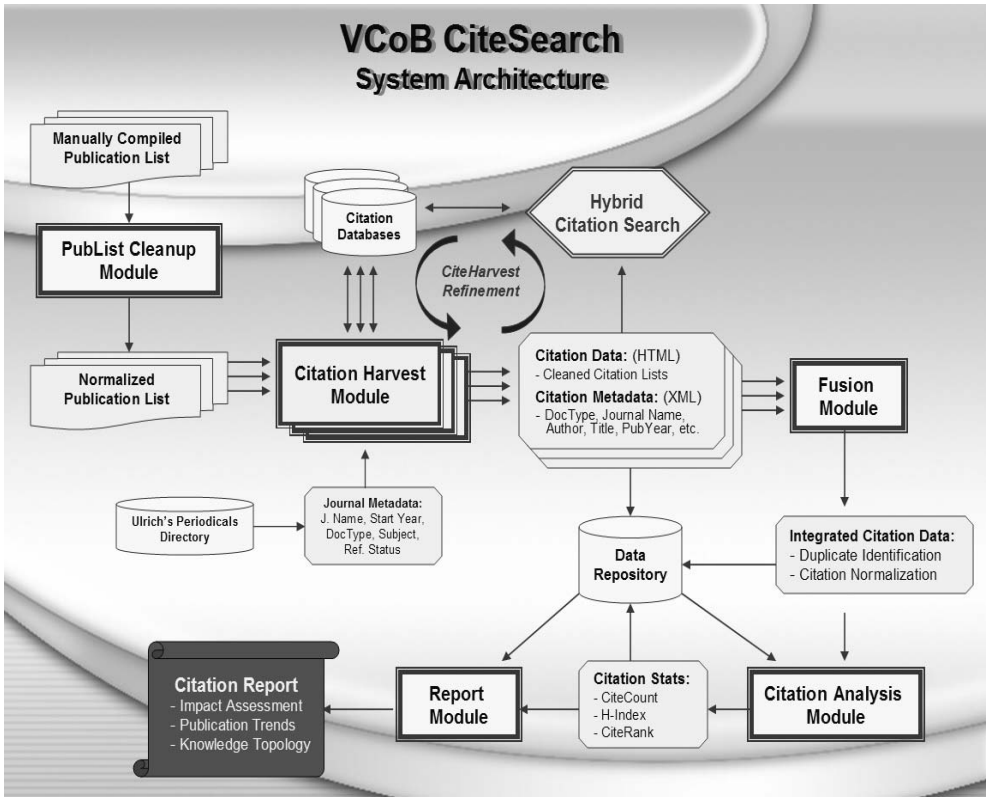


Figure 1. CiteSearch Prototype System Architecture

Science. The aim of the VCoB project is to develop an adaptive, interactive agent for building and maintaining a virtual collection of Web documents.

CiteSearch Study

To explore the existing citation analysis environment for the CiteSearch system, we conducted a pilot study to examine the impact of using *Scopus* and *Google Scholar* on the citation count and citation ranking of LIS faculty members as measured by *Web of Science*. *Web of Science*, which comprises the three ISI citation databases (*Arts & Humanities Citation Index*, *Science Citation Index*, and *Social Sciences Citation Index*), has been the standard tool for a significant portion of citation studies worldwide.

Web of Science's offers powerful features for browsing, searching, sorting and saving functions, as well as exporting to citation management software (e.g., *EndNote* and *RefWorks*). Coverage in *Web of Science* goes back to 1945 for *Science Citation Index*, 1956 for *Social Sciences Citation Index*, and 1975 for *Arts & Humanities Citation Index*. As of October 2006, there were over 36 million records in the database from approximately 8,700 scholarly titles, several hundred conference proceedings and over 190 open access journals (Harnad & Brody, 2004). Subjects covered in *Web of Science* include disciplines found in the curricula of most universities in arts, humanities, sciences, and social sciences.

Similar to ISI, Elsevier, the producer of *Scopus*, offers powerful features for browsing, searching, sorting, and saving functions, as well as exporting to citation management software. Coverage in *Scopus* goes back to 1966 for bibliographic records and abstracts and 1996 for citations. As of October 2006, there were over 28 million records in the database from over 15,000 "peer-reviewed" titles including 500 Open Access journals, 700 conference proceedings, and 600 trade publications (Elsevier, 2006). Subject areas covered in *Scopus* include: Chemistry, Physics, Mathematics, and Engineering, Life and Health Sciences, Social Sciences, Psychology, and Economics, Biological, Agricultural, and Environmental Sciences, and General Sciences (Goodman and Deis, 2005; Jacso, 2005).

In contrast to ISI and Elsevier, Google does not offer a publisher list, title list, document type identification, or any information about the time-span or the refereed status of records in *Google Scholar*. This and other studies, however, have found that *Google Scholar* covers print and electronic journals, conference proceedings, books, theses, dissertations, preprints, abstracts, and technical reports (Bauer & Bakkalbasi, 2005; Gardner & Eng, 2005; Wleklinski, 2005). Although *Google Scholar* does not cover material from all major publishers, it identifies citations to articles from these publishers when documents from other sources cite these articles. As of August 2006, *Google Scholar* provided searching of more than 500 million documents (Giustini, 2006).

Citation Searching Methods

All *Scopus* and *Web of Science* data were manually collected and processed twice by one of the authors (LIM) in October 2005 and again (for accuracy and updating purposes) in March 2006. *Google Scholar* data were harvested by *CiteSearch* system in March 2006; however, identifying their bibliographic information took approximately 3,000 hours of work over a six months period to manually verify, clean, format, standardize, and enter the data into EndNote Libraries and Access databases.

The "Cited Author" feature was used in *Web of Science* to identify citations to each of the 1,093 items published by the 15 faculty members constituting the study sample. The "Cited Author" search feature in *Web of Science* displays all the cited items of an author. The searcher then goes through all the entries and selects the relevant ones based on the information displayed for each entry (e.g., cited author, cited source, publication year of the cited item, and page numbers relevant to the cited item). Citations to items in which the faculty members were not first authors were included in the study.

Unlike *Web of Science*, *Scopus* does not provide the ability to browse the cited authors, cited works, or cited journals indexes or fields. Consequently, all available methods had to be used to accurately locate all potentially relevant citations in the database for each individual faculty member. The methods used were:

- *Author Search*: This feature allowed us to retrieve all articles in the database for each individual faculty member and, subsequently, identify all the records in the database that have cited these articles. Although the majority of citations found through this method overlapped with those found in the Exact Match Searches method described above, this method identified a few unique citations.
- *Exact Match Search*: This method used the title of an item as a search statement to locate an exact match of it in the "References" field, thereby allowing us to identify most of the documents in the database that have cited items published by the study sample. In cases where the title was too short or ambiguous to refer to the item in question, we used additional information as keywords (e.g., the author's last name). In cases where the title was too long, we used the first few words of the title. The Exact Match Searches identified the largest number of relevant citations for our study sample.

Google Scholar was searched for citations using two methods: *Author* search and *Exact Phrase* (or *exact match*) search. The Author search usually retrieves items published by an author and ranks the items by relevance. In most cases, highly cited items rank first, as *Google Scholar* uses Google's crawler to index the content of research materials and automatically extracts and adds citation counts to retrieved documents to raise or lower individual articles in the rankings of a result set. Once the items are retrieved, a click on the "Cited by . . ." link allows the searcher to display the list of citing documents. The "Cited by . . ." link is automatically generated by *Google Scholar* for each cited item. The Exact Match search identified 14.63% (or n=534) additional citations that were missed by the Author search. Upon manual inspection of 1,301 records uniquely found through the Exact Phrase search, only 534 (or 41.0%) were found to be relevant citations. Almost all of the false drops were documents retrieved when searching for citations to short-title items.

A major disadvantage of *Google Scholar* is that its records are retrieved in a way that is very impractical for use with large sets or large numbers of study participants, requiring a very tedious process of manually extracting, cleaning, organizing, classifying, and saving the information into meaningful and useable formats. Unlike *Web of Science* and *Scopus*, *Google Scholar* does not allow re-sorting of the retrieved sets in any way (such as by date, author name, or data source). Other major disadvantages of *Google Scholar* include duplicate citations (i.e., counting a citation published in two different forms, such as preprint and journal article, as two citations) as well as the lack of any information about document type, document length, and the refereed status of the retrieved citations. In many cases, the item for which citations are sought for is retrieved and considered a citation by *Google Scholar*. The presence of all these problems in *Google Scholar* suggest that unless a system is developed that automatically and accurately parses result sets into error-free, meaningful, and usable data, *Google Scholar* will be of limited use for large-scale comparative citation and bibliometric analyses.

To make sure that citations were not overlooked because of searching or indexing errors, we looked for the bibliographic records of all citations that were missed by one or two of the three tools. For each of these missed citations, we examined: (1) why it was not retrieved through the citation search methods described above; and (2) whether it should be counted as a citation. Items that were overlooked due to searching errors (16 in the case of *Web of Science* and 27 in *Scopus*) were counted as citations toward their respective databases; most of the searching errors were due to having missed selecting a relevant entry in the cited references field in *Web of Science* and typographical errors in the case of *Scopus*. Items that were missed due to database/system errors were tallied but were not counted as citations. These included:

- *Web of Science*: 10 citations (0.5% of the database's relevant citations) were missed due to incomplete lists of references.
- *Scopus*: 75 citations were missed due to lack of cited references information and 26 citations due to incomplete lists of references (total 4.4% of its relevant citations).

- *Google Scholar*: missed 501 (12.0% of its relevant citations) due to system errors. Many of the errors in *Google Scholar* were a result of matching errors due to hyphenated words or missing reference section header (e.g., Bibliography, References).

These results suggest that if citation searching of individual LIS scholars were limited to *Scopus*, a searcher would miss an average of 4.4% of the relevant citations due to database errors. In the case of *Google Scholar*, the percentage would be 12.0%; this percentage would increase to 26.6% had we not used the Exact Phrase search approach described earlier. The results also suggest that when using *Google Scholar* one must use both the Author search and Exact Match search methods.

Sample and Units of Analysis

In order to analyze the impact of using additional sources besides *Web of Science* on the citation count and citation ranking of LIS faculty members and be able to generalize the findings to the field, this study explored the difference in citation count and citation ranking *Scopus* and *Google Scholar* make to results from Web of Science for all 15 faculty members of the School of Library and Information Science at Indiana University (SLIS). These faculty members not only cover most if not all of the mainstream LIS research areas as identified by the Association of Library and Information Science Education (ALISE, 2006), they also belong to one of the largest and most published and cited American Library Association accredited LIS programs in North America (Adkins & Budd, 2006; Budd, 2000). From 1970 to December 2005, the 15 SLIS faculty members had published or produced over 1,093 scholarly works, including: 312 refereed journal articles, 12 refereed review articles, 305 conference papers (almost all refereed), 131 chapters (some refereed), 36 books, and 35 edited volumes, among others (Table 1). The inclusion of high and diverse number of faculty members and publications in this study provided us with valuable framework to make citation comparisons between *Web of Science*, *Scopus*, and *Google Scholar*.

Table 1. SLIS Publication Data (1970-2005)*

Document Type	**Count****
Refereed journal articles	322 (312)
Conference papers	313 (305)
Chapters	135 (131)
Non-refereed journal articles	94 (93)
Technical reports / Working papers	84 (83)
Articles in professional journals	59 (59)
Books	36 (36)
Edited volumes	35 (35)
Bibliographies	16 (16)
Review articles	15 (12)
Encyclopedia articles	11 (11)
Total	**1,120 (1,093)**

*Book reviews, meeting abstracts, letters to editors, panels, presentations (invited and otherwise), and so on are excluded from this table.**Figures in parentheses refer to unique records (i.e., after removing duplicates due to co-authorship among SLIS faculty members).

Results And Discussion

The results of this study are presented and discussed in two sections: (1) the effect of using *Scopus* on the citation counts and rankings of SLIS faculty members as measured by *Web of Science*; and (2) the effect of using *Google Scholar* on the citation counts and rankings as measured by *Web of Science* and *Scopus* combined. Because the three tools provide different citation coverage in terms of document type and time period, we limited most of the analysis to citations from types of documents and years

common to all three tools, that is, conference papers and journal items (e.g., journal articles, review articles, editorials, book reviews, and letters to the editor) published between 1996 and 2005.

Effect of Scopus on Citation Count and Citation Ranking of LIS Faculty

To show the difference that *Scopus* makes to the citation counts and rankings of LIS faculty members as measured by *Web of Science*, we compare the number of citations retrieved by both databases, show the increase *Scopus* makes toward the total number of citations of SLIS as a whole and also of individual faculty members, explore the effect *Scopus* has on altering the relative citation ranking of SLIS faculty members, and examine the overlap and unique coverage between the two databases.

As shown in Table 2, *Scopus* includes 278 (or 13.7%) more citations than *Web of Science*, suggesting that *Scopus* provides more comprehensive coverage of the LIS literature than *Web of Science*.[2] Further analysis of the data shows that combining citations from *Scopus* and *Web of Science* increases the number of citations of SLIS as a whole by 35.1% (from 2,023 to 2,733 citations). This means that were only *Web of Science* used to locate citations for LIS faculty members, on average, more than one-third of relevant citations (found in the union of *Web of Science* and *Scopus*) would be missed; the percentage of missed citations would be 18.8% were only *Scopus* used. Data also show that the percentage of increase in citation counts for individual faculty members varies considerably depending on their research areas, ranging from 4.9% to 98.9% (Table 5). These findings not only imply that certain subject areas will benefit more than others from using both *Scopus* and *Web of Science* to identify relevant citations, they also suggest that to generate accurate citation counts for faculty members, and by extension schools, and to accurately compare them to one another, a researcher must use both databases. The importance of using *Scopus* in addition to *Web of Science* is further evidenced by the facts that:

- The relative ranking of faculty members changes in eight out of 15 cases, strikingly so in the cases of faculty members E, F, H, and I (Table 3).
- The overlap of LIS citations between the two databases is relatively low—58.2% (Figure 2) with significant differences from one research area to another ranging from a high 82.0% to a low 41.1% (Table 5).
- The number of unique citations found in *Scopus* is noticeably high in comparison to that of *Web of Science* (26.0% vs. 15.8%, Figure 2). The overlap and uniqueness between the two databases is almost identical to what Whitley (2002) found in her study that compared the duplication (60%) and uniqueness of citing documents in *Chemical Abstracts* (23%) and *Science Citation Index* (17%).

Regarding the type of documents in which the citations were found, the main difference between the two databases is in the coverage of conference proceedings. *Scopus* retrieves considerably more citations from refereed conference papers than *Web of Science* (359 vs. 229, Table 4). Of all 496 citations from conference papers, 53.8% are uniquely found in *Scopus* in comparison to only 27.6% in *Web of Science* (19.6% of citations from conference papers are found in both databases). This could have significant implications for citation analyses and the evaluation of individual scholars, especially when those evaluated include authors who use conferences as a main channel of scholarly communication. Without *Scopus*, authors who communicate extensively through conferences will be at a disadvantage when their citation counts are compared with those who publish primarily in journals due to poor coverage of conference proceedings in *Web of Science*.

[2]Table 2, however, also shows that *Web of Science* includes 391 (or 17.0%) more citations than *Scopus*, when citations from pre-1996 are counted.

Table 2. Citation Count by Year – *Web of Science* and *Scopus*

Years	*Web of Science*	*Scopus*	Union of *Web of Science* and *Scopus*
1971-1975	1	-	1
1976-1980	15	-	15
1981-1985	129	-	129
1986-1990	201	-	201
1991-1995	323	-	323
Subtotal	**669**	**-**	**669**
1996	119	101	140
1997	121	119	144
1998	142	123	167
1999	131	128	164
2000	175	171	219
2001	207	242	278
2002	202	220	271
2003	251	291	348
2004	323	459	510
2005	352	447	492
Subtotal	**2,023**	**2,301**	**2,733**
TOTAL	**2,692**	**2,301**	**3,402**

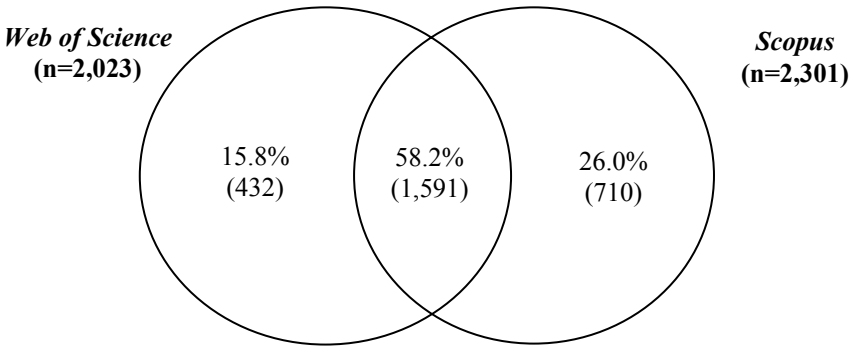


Figure 2. Distribution of unique and overlapping citations in *Web of Science* and *Scopus* (N=2,733)

Table 3. Impact of Adding *Scopus* Citations on the Ranking of Faculty Members

Faculty Members*	*Web of Science*		Union of *Web of Science* and *Scopus*	
	Count	Rank	Count	Rank
A	544	1	853	1
B	508	2	564	2
C	273	3	365	3
D	162	4	188	4
E	123	5	137	8
F	122	6	128	9
G	118	7	154	7
H	115	8	165	6
I	88	9	175	5
J	83	10	93	10
K	35	11	42	12
L	32	12	44	11
M	29	13	31	15
N	28	14T	40	13
O	28	14T	34	14

Table 4. *Web of Science* and *Scopus* Citation Count by Document Type

Document Type	*Web of Science*		*Scopus*		Union	
	Count*	%	Count*	%	Count*	%
Journal articles	1,529	75.6%	1,754	76.2%	1,968	72.0%
Conference papers	229	11.3%	359	15.6%	496	18.1%
Review articles	172	8.5%	147	6.4%	175	6.4%
Editorial materials	63	3.1%	36	1.6%	64	2.3%
Book reviews	17	0.8%	0	0.0%	17	0.6%
Letters to editors	9	0.4%	2	0.1%	9	0.3%
Bibliographic essays	2	0.1%	2	0.1%	2	0.1%
Biographical item	2	0.1%	1	0.0%	2	0.1%
Total	**2,023**	**100.0%**	**2,301**	**100.0%**	**2,733**	**100.0%**
Total from Journals	1,794	88.7%	1,942	84.4%	2,237	81.9%
Total from Conference Papers	229	11.3%	359	15.6%	496	18.1%
Total	**2,023**	**100.0%**	**2,301**	**100.0%**	**2,733**	**100.0%**

*Excludes duplicate citations.

Table 5. Impact of Adding *Scopus* Citations on Faculty and School Citation Counts

Research areas of individual faculty members*	***Web of Science***	***Scopus***	**Union**	**% Increase**	**Overlap**
Human-computer interaction	544	740	853	56.80%	430 (50.4%)
Citation analysis, informetrics, scholarly communication, and strategic intelligence	508	459	564	11.00%	403 (71.5%)
Computer-mediated communication, gender and information technology, and discourse analysis	273	313	365	33.70%	221 (60.5%)
E-commerce, information architecture, information policy and electronic networking	162	168	188	16.00%	142 (75.5%)
Bibliometrics, Collection development and management, evaluation of library sources and services, and serials	123	108	137	11.40%	94 (68.6%)
Information seeking and use, design and impact of electronic information sources, and informetrics	122	111	128	4.90%	105 (82.0%)
Information visualization, data mining, and data modeling	118	129	154	30.50%	83 (53.9%)
Intelligent interfaces for information retrieval and filtering, knowledge discovery, and user modeling	115	133	165	43.50%	92 (55.8%)
Communities of practice and social informatics	88	159	175	98.90%	72 (41.1%)
Classification and categorization, ontologies, metadata, and information architecture	83	80	93	12.00%	70 (75.3%)
Critical theory and documentation	35	37	42	20.00%	30 (71.4%)
Computational linguistics, computer-mediated communication, and sociolinguistics and language acquisition	32	38	44	37.50%	26 (59.1%)
Citation analysis, bibliometrics, and data retrieval and integration	29	21	31	6.90%	19 (61.3%)
Information retrieval and data integration	28	32	40	42.90%	20 (50.0%)
Information policy, social and organizational informatics, and research methods	28	31	34	21.40%	25 (73.5%)
Faculty Members Total	**2,288**	**2,559**	**3,013**	31.70%	**1,832 (**60.8%)
School Total*	**2,023**	**2,301**	**2,733**	35.10%	**1,591 (**58.2%)

*Each row in the table represents a single faculty member and the main research topics covered by him/her. It would have been practically impossible to classify citations by individual topics rather than individual faculty members.
**Excludes duplicate citations.

In conclusion, the findings suggest that many of the previous studies that used *Web of Science* exclusively to generate citation data to evaluate and/or rank scholars, journals, programs, and so on have been based on skewed and incomplete data and may have resulted in inaccurate assessments and imprecise rankings. Given the low overlap in citations between the two databases, the findings further suggest that the use of *Scopus* in addition to *Web of Science* may have significant implications on the *h*-index scores of authors and journals (Bornmann & Daniel, 2005, in press; Cronin & Meho, 2006; Hirsch, 2005), journal impact factors (Garfield, 1996, 2006; Nisonger, 2004b; Saha, Saint, & Christakis, 2003), and research assessment exercises as those conducted in the United Kingdom (Oppenheim, 1995; Smith & Eysenck, 2002; Warner, 2000)

.

Effect of Google Scholar on Citation Counts and Rankings of LIS Faculty

Data collected in this study show that, in contrast to *Web of Science* and *Scopus*, which index citations mainly from journal articles and conference papers, citations found through *Google Scholar* come from many different types of documents (Table 6). Results show that *Google Scholar* identifies 1,448 (53.0%) more citations than *Web of Science* and *Scopus* combined (4,181 vs. 2,733) and combining citations from all three sources increases the number of citations by 93.4% (from 2,733 to 5,285 citations). In other words, one would miss over 93.4% of relevant citations if searching were limited to *Web of Science* and *Scopus*.

Table 6. *Google Scholar* Citation Count by Document Type

Document Type	Count	%
Journal Articles	2,215	40.32%
Conference Papers	1,849	33.66%
Doctoral Dissertations	261	4.75%
Master's Theses	243	4.42%
Book Chapters	199	3.62%
Technical Reports	129	2.35%
Reports	110	2.00%
Books	102	1.86%
Review Articles	86	1.57%
Conference Presentations	72	1.31%
Unpublished Papers	65	1.18%
Bachelor's Theses	34	0.62%
Working Papers	31	0.56%
Editorial Materials	25	0.46%
Research Reports	23	0.42%
Workshop Papers	15	0.27%
Doctoral Dissertation Proposals	9	0.16%
Conference Posters	9	0.16%
Book Reviews	3	0.05%
Master's Thesis Proposals	3	0.05%
Preprints	3	0.05%
Conference Paper Proposals	2	0.04%
Government Documents	2	0.04%
Letters to the editor	2	0.04%
Biographical Item	1	0.02%
Total	5,493	100.00%
Total from Journals	2,332	42.45%
Total from Conference Papers	1,849	33.66%
Total from Journals and Conference Papers	4,181	76.12%
Total from Dissertations/Theses	538	9.79%
Total from Books	301	5.48%
Total from Reports	262	4.77%
Total from other document types	211	3.84%
Total	**5,493**	**100.00%**

Despite the large increase in citation counts, adding *Google Scholar*'s unique citations data to those of *Web of Science* and *Scopus* does not significantly alter the relative ranking of faculty members—Spearman Rank Order correlation coefficient = 0.976 at 0.001 level (Table 7). *Google Scholar* also misses 1,104 (40.4%) of the 2,733 citations found by *Web of Science* and *Scopus* (Figure 3). This is

strikingly high, especially given the fact that virtually all citations from *Web of Science* and *Scopus* come from refereed and/or reputable sources.

Given the fact that *Google Scholar* is so cumbersome to use, misses a significant number of citations from refereed sources, and has little or no influence on the relative rankings of scholars, one could conclude that, as far as LIS is concerned, *Google Scholar* is superfluous, especially when the focus of a study is on citations in refereed journals and conference proceedings and when both *Web of Science* and *Scopus* are used to generate citation counts for assessing and comparing scholars, journals, and academic departments to one another. Results of this study also show that the increase in Spearman's *rho* correlation from .874 between *Google Scholar* and *Web of Science* to 0.976 between *Google Scholar* and the Union of *Web of Science* and *Scopus* provide evidence of the importance and value of using *Scopus* in addition to *Web of Science* in: (1) diminishing the value or impact of *Google Scholar* on the relative ranking of individual scholars; and (2) creating a more accurate ranking of scholars.

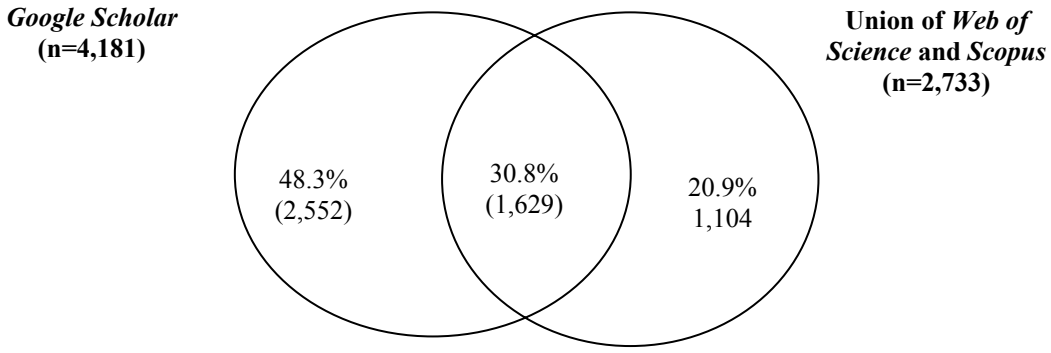


Figure 3. Distribution of citations in *Google Scholar*, *Web of Science* & *Scopus* (N=5,285)

Table 7. Impact of Adding *Google Scholar* Citations on the Ranking of Faculty Members

Faculty Members	**Union of *WoS* & *Scopus***		**Union of *WoS*, *Scopus*, & *Google Scholar***	
	Count	Rank	Count	Rank
A	853	1	2,078	1
B	564	2	802	2
C	365	3	797	3
D	188	4	244	7
F	175	5	402	4
E	165	6	249	6
G	154	7	291	5
I	137	8	169	9
H	128	9	171	8
J	93	10	125	10
K	44	11	92	11
L	42	12	60	12
O	40	13	59	13
M	34	14	42	14
N	31	15	39	15

Google Scholar, however, could be very useful for individual scholars preparing for tenure and promotion and/or for comparative citation analysis studies that attempt to map or visualize scholarly networks (Börner, Chen, & Boyack, 2003; Small, 1999; White, 1997) and/or those studies that try to show one's international impact (Nisonger, 2004a). It should be emphasized here that *Google Scholar* uncovers many unique citations (2,552 or 48.3% of the 5,285 citations found in all three sources) that could be very useful in such studies or for such purposes (Figure 3).

Considering the type of documents in which the citations were found, *Google Scholar* retrieves significantly more (almost four times as many) citations from conference papers than *Web of Science* and *Scopus* combined (1,849 vs. 496). In contrast, *Web of Science* and *Scopus* retrieve almost as many citations from journals as *Google Scholar* does (1,968 vs. 2,332). It should be emphasized here that the relatively poor coverage of conference papers by *Web of Science* and *Scopus,* or the relative good coverage by *Google Scholar* of this document type, has much to do with the fact that many authors make their conference papers available online themselves. Almost half of *Google Scholar*'s unique citations from conference papers and many of its citations from journals were identified through full-text documents made available online by their authors (i.e., self-archived) rather than from the official web sites of the publishers of the conference proceedings and journals.[3] What this kind of findings reveals is that there is a "discernible difference" in terms of the frequency with which the article is cited; there is a dramatic advantage in favor of the articles that their authors made available online.

Conclusions And Implications

This study provides meaningful suggestions for conducting citation analyses and informs of novel ways of identifying citations to scholarly publications. The study found that addition of *Scopus* citations to those of *Web of Science* could significantly alter the ranking of scholars but addition of *Google Scholar* does not alter the rankings despite the fact that citation count almost doubles. Significantly, this study has demonstrated that:

1) *Web of Science* should not be used alone for locating citations to an author or title.
2) *Scopus* provides more comprehensive coverage of LIS and LIS-related literature than *Web of Science*; however, the two databases complement rather than replace each other.
3) *Google Scholar* can be very useful for illustrating one's international impact and impact on students' research; however, *Google Scholar* is not conducive for large-scale comparative citation analyses.
4) Both *Scopus* and *Google Scholar* can help identify a significant number of citations not found in *Web of Science*, but only *Scopus* significantly alters the ranking of scholars as measured by *Web of Science*.
5) All three tools are necessary to develop more accurate maps or visualizations of scholarly networks and impact both within and between disciplines.

This study has significant implications for funding agencies and editors and publishers of journals who may wish to use citation counts and rankings to identify subject experts to review grant applications or submitted manuscripts and to determine the impact of projects and articles they funded or published. The study has also significant implications for the wider scholarly community as researchers begin to adopt the methods and tools such as *CiteSearch* described here to identify sources of citations that may otherwise remain unknown.

We hope that other researchers can use this study as a model for exploring the impact of broadening the sources of citations in other fields. The increasing availability of online information resources and open access journals will make *Google Scholar* very popular among scholars as they try to find citations to their work or to items they are using for research. *Google Scholar*, however, will become useful for extensive citation analyses only if it is made easier to search and work with.

[3]In these cases, we most often had to go to the root web site to identify the full bibliographic information of the citing documents. Most often, the root web sites were the curricula vitae of the authors of the citing documents.

References

Adkins, D., & Budd, J. (2006). Scholarly productivity of U.S. LIS faculty. *Library & Information ScienceResearch*, 28(3), 374-389.

Aksnes, D. W., & Taxt, R. E. (2004). Peer reviews and bibliometric indicators: A comparative study at a Norwegian university. *Research Evaluation*, 13(1), 33-41.

Bauer, K., & Bakkalbasi, N. (2005). An examination of citation counts in a new scholarly communication environment. *D-Lib Magazine*, 11(9). Retrieved March 25, 2006, from http://www.dlib.org/dlib/september05/bauer/09bauer.html.

Börner, K., Chen, C., & Boyack, K. W. (2003). Visualizing knowledge domains. *Annual Review of Information Science and Technology*, 37, 179-255.

Bornmann, L., & Daniel, H-D. (in press). What do we know about the h index? *Journal of the American Society for Information Science and Technology*.

Budd, J. M. (2000). Scholarly productivity of U.S. LIS faculty: An update. *The Library Quarterly*, 70(2), 230-245.

Cronin, B. (1984). *The citation process: The role and significance of citations in scientific communication*. London: Taylor Graham.

Cronin, B., & Meho, L. (2006). Using the *h*-index to rank influential information scientists. *Journal of the American Society for Information Science and Technology*, 57(9), 1275-1278.

Elsevier Science Publishers (2006). *Scopus*. New York: Elsevier. [Electronic database]

Gardner, S., & Eng, S. (2005). Gaga over Google? Scholar in the Social Sciences. *Library Hi Tech News*, 22(8), 42-45.

Garfield, E. (1996). How can impact factors be improved? *British Medical Journal*, 313 (7054), 411-413.

Giustini, D. (2006). Blog posting at UBC Academic Search—Google Scholar Blog. Retrieved September 25, 2006, from http://weblogs.elearning.ubc.ca/googlescholar/archives/025964.html.

Goodman, D., & Deis, L. (2005). Web of Science (2004 version) and Scopus. *The Charleston Advisor*, 6(3). Retrieved March 25, 2006, from http://www.charlestonco.com/dnloads/v6n3.pdf.

Harnad, S., & Brody, T. (2004). Comparing the Impact of Open Access (OA) vs. Non-OA Articles in the Same Journals. *D-Lib Magazine*, 10(6). Retrieved on April 27, 2006 from http://www.dlib.org/dlib/june04/harnad/06harnad.html.

Hirsch, J. E. (November 15, 2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, 102(46), 16569-16572. Retrieved February 15, 2006, from http://www.pnas.org/cgi/reprint/102/46/16569.

Holden, G., Rosenberg, G., & Barker, K. (2005). Bibliometrics: A potential decision making aid in hiring, reappointment, tenure and promotion decisions. *Social Work in Health Care*, 41(3-4), 67-92.

Holmes, A., & Oppenheim, C. (2001). Use of citation analysis to predict the outcome of the 2001 Research Assessment Exercise for Unit of Assessment (UoA) 61: Library and Information Management. *Information Research*, 6(2). Retrieved June 15, 2005, from http://informationr.net/ir/6-2/paper103.html.

Jacsó, P. (2005). As we may search—comparison of major features of the Web of Science, Scopus, and Google Scholar citation-based and citation-enhanced databases. *Current Science*, 89(9), 1537-1547. Retrieved March 15, 2006, from http://www.ias.ac.in/currsci/nov102005/1537.pdf.

Lewison, G. (2001). Evaluation of books as research outputs in history of medicine. *Research Evaluation*, 10(2), 89-95.

MacRoberts, M. H., & MacRoberts, B. R. (1996). Problems of citation analysis. *Scientometrics*, 36(3), 435-444.

Martin, B. R. (1996). The use of multiple indicators in the assessment of basic research. *Scientometrics*, 36(3), 343-362.

Moed, H. F. (2005), *Citation analysis in research evaluation*, Springer, Berlin.

Nisonger, T. E. (2004a). Citation autobiography: An investigation of ISI database coverage in determining author citedness. *College & Research Libraries*, 65(2), 152-163.

Nisonger, T. E. (2004b). The benefits and drawbacks of impact factor for journal collection management in libraries. *The Serials Librarian*, 47(1-2), 57-75.

Oppenheim, C. (1995). The correlation between citation counts and the 1992 Research Assessment Exercise Ratings for British library and information science university departments. *Journal of Documentation*, 51(1), 18-27.

Saha, S., Saint, S., & Christakis, D. A. (2003). Impact factor: A valid measure of journal quality? *Journal of the Medical Library Association*, 91(1), 42-46.

Seglen, P. O. (1998). Citation rates and journal impact factors are not suitable for evaluation of research. *Acta Orthopaedica Scandinavica*, 69(3), 224-229.

Small, H. (1999). Visualizing science by citation mapping. *Journal of the American Society for Information Science*, 50(9), 799-813.

Smith, A. T., & Eysenck, M. (2002). The correlation between RAE ratings and citation counts in psychology. Technical Report, Psychology, University of London, Royal Holloway. Retrieved October 12, 2006, from http://cogprints.org/2749/.

Van Raan, A. F. J. (1996). Advanced bibliometric methods as quantitative core of peer-review based evaluation and foresight exercises. *Scientometrics*, 36(3), 397-420.

Warner, J. (2000). A critical review of the application of citation studies to the Research Assessment Exercises. *Journal of Information Science*, 26(6), 453-459.

White, H. D., & McCain, K. W. (1997). Visualization of Literatures. *Annual Review of Information Science and Technology*, 32, 99-168.

Wleklinski, J. M. (2005). Studying Google Scholar: Wall to wall coverage? *Online*, 29(3), 22-26.

Scientific Output by Gender in Spain (Web of Science, 2004)

Félix Moya-Anegón[*], Zaida Chinchilla-Rodríguez[*], Benjamín Vargas-Quesada[*], Elena Corera-Álvarez[*], Antonio Muñoz-Molina[*], Francisco José Muñoz-Fernández[*] and Rocío Gómez-Crisóstomo[**]

**felix@ugr.es, zchinchi@ugr.es, benjamin@ugr.es, ecorera@ugr.es, agmo@ugr.es, franjo@ugr.es,*
Grupo SCImago, Department of Library and Information Sciences, University of Granada, Campus of Cartuja, 18071 Granada (Spain)

***mrgomcri@alcazaba.unex.es,*
Grupo SCImago, Department of Computer Science, University of Extremadura, Pl. Ibn Marwan, 06005 Badajoz, (Spain)

Abstract
The objective of this study was to obtain bibliometric indicators by gender applied exclusively to scientific publications registered in the Thompson Scientific databases. Aspects related with the volume of production, visibility, patterns of collaboration and networks of coauthorship will be analyzed below. The data are presented broken down by scientific field.

Keywords
bibliometrics; scientific production; gender; Spain

Introduction

There is ever greater interest surrounding the participation of women in science as an indicator of social progress, and of how their presence in academic and scientific circles has increased over the years. For this reason, numerous persons and organizations have embarked on gender studies, focusing mainly on demonstrating how scarce female representation in scientific and technological areas may be, as well as on differentiating the professional categories largely occupied by women versus the achievements of their male peers.

The interest in promoting gender parity in all the realms, and particularly in Science and Technology, began in the United States in the 1970´s (with the foundation of the *Association for Women in Science,* in 1971) and in Europe in the 1980´s[1]. Initiatives on the part of Scandinavian countries and the United Kingdom were followed by a general surge of awareness by the European Community that crested in 1999. As a result of the Conference "Woman and Science" celebrated in Brussels in 1998, the theme of gender was first incorporated into the history of EU research policy. The "Helsinki Group" was created to examine the situation of women scientists in 30 countries. The plan of action —to promote gender equality in Science and Technology— included the elaboration of the ETAN Report, published in the year 2000. Data therein reflected that women researchers were under-represented in the key positions of the 30 countries involved, a discrimination traced to multiple factors. Since then, slow advancements have taken place, yet more in the legislative or normative realm than in social reality overall.

The national Institute of Statistics (*Instituto Nacional de Estadística, INE*) includes among its data some statistics broken down by sex. According to this information, the number of female students registered in Spanish universities comes to some 54% of the total, although this percentage is much lower in the studies of Architecture and Engineering, where women represent only 30% of the total students registered. In Doctoral programs, the percentage of students registered is quite even, with women making up 51%; yet when the dissertations are finally read, only 47% pertain to women students. Women make up just 35% of all Spanish University professors, and only 13% of the Department Heads in these universities are female. According to these data, a high percentage of women scientists are seen to "drop out" of their research endeavours, leaving research as a principally

[1]. The study "Mujer y Ciencia: La situación de las mujeres investigadoras en el sistema español de ciencia y tecnología" Fecyt, 2005, presents a panorama of the gender variable from its origins.

masculine undertaking. At the same time, we can see inequality of gender in those employed in R+D activities in the different sectors of execution: over 60% of contracted personnel are male, their presence being greater in the private sector and lesser in non-profit Public Institutions and Administrations.

Despite these figures, Spain is not one of the worst situated countries insofar as the female involvement in science and technology is concerned: according to the report *She Figures 2006*, based on data from the Office of Community Statistics (Eurostat), the European mean for women researchers revolves around 29%, whereas for Spanish women it reaches 36%, in addition to evidencing higher growth amongst women on the national level (11% in the period 99-03) than internationally (4%). At the European level, half of the persons who work in science- or technology-related activities are women, a figure that has risen 4% in the period 1998-2004, nearly twice the increase undergone by men, at 2.2%. The same report confirms that women make up 43% of the total Ph.D. degree holders in the European Union, although Spain surpasses this figure, with 45%. However, the growth of women in science during the period 1998-2004 was greater on the European level, with a 7% increase, as opposed to the 5% growth seen in Spain; in contrast, the body of male scientists grew 4% nationally, but just 2% at the international level.

Studies on Scientific – Technological Output

Few are the studies on scientific production by gender, mainly due to the lack of availability of data broken down by sex. Aware of this difficulty, the European Union, in 1993, dictated recommendations as to the retrieval and comparison of statistics on gender in science and technology. In 1998 it again insisted on this point, yet it was not until the year 2003 when this type of information finally saw the light in the report "*She Figures"*. Organizations such as the Helsinki Group sought to promote the inclusion of these data in studies carried out on either a national or an international level, in order to compare them and determine the factors influencing under-representation in certain fields, as well as the scarce presence of women in positions of leadership.

At present, desegregation by gender cannot be considered fully extended: for example, some authors underline the lack of attention conceded in bibliometric studies (Webster, 2001). In large bibliographic databases, such as that of Thomson Scientific, the information regarding authors may be incomplete or erroneous, with authors having more than one entry for their persona, or with only the initials of authors´ first names indicated, making gender distinction impossible. In the science database of China, however, 50% of the journals indicate this information, along with the age and the academic degree of the undersigning authors. Other databases, such as Scopus, with a greater coverage than the ISI (Moya, Chinchilla, et al., 2007) feature registers and tools (*Author Profile*) with author information where the full first name and middle initial appear as the norm..

The alternative is to dispose of lists of research communities classified by specialty, institutional affiliation, geographical base, or the means of financing. Thus we encounter studies based on lists of doctoral candidates (Goel, 2002), censuses (Lemoine, 1992) and institutional directories (Gupta, 1999) (Rusell, 2003) (Bordons and Mauleón, 2004). Meanwhile, indicators of technological activity would be easier to obtain, as patents must include the full name of the inventors and the persons who register them, though there are no documented reports of such for Spain. Studies relative to entrepreneurial environs are practically non-existent, except for the ETAN Report and the reflections gathered up in the EU report on "*Women in Industrial Research*".

Whichever the source, studies of gender and scientific output have been denouncing vertical discrimination (Bordons and Mauleón, 2004), horizontal bias (Stack, 2002) and categorical differences in salary. Generally, these studies evidence lower levels of production, giving rise to reports aimed to solve the so-called "puzzle of productivity" or *productivity gap* (Cole and Zuckerman, 1984). Being left out of the mainstream of research (Davenport and Zinder, 1995) can accompany lower levels of participation, position, and recognition (Long and Fox, 1995), limited representation as a human resource in science, and a discrepancy between the number of graduates with scientific degrees and the number of women actually active in research (Prpic, 2002) as well as the orientation of a publication through national networks (Webster, 2001) (Lemoine, 2002). Differences like these are lesser when they are seen in light of other influential factors such as professional category or age.

Some authors point out a "discipline effect" with regards to the relative differences in output between male and female researchers; that is, a horizontal segregation owing to factors such as educational stereotypes. In fields like sociology, with a high proportion of women, the traditional barriers for these researchers to overcome may be weak or non-existent (Snack, 2002). And it would appear that family obligations do not affect yield (Cole, 1979) (Cole and Zuckerman, 1987). The ETAN report affirms (p.42):

> *In this study, there was no relationship between academic productivity and family structure – those women with children did not produce less than their childfree colleagues. Indeed previous studies have shown that married women, as a rule, produce more scientific papers per year than single women - and those with children have equal or higher productivity than those without children (Cole and Zuckerman, 1987; Luukkonen-Gronow and Stolte-Heiskanen, 1983; Kyvik, 1988). Career breaks and childcare are important issues nevertheless and taken up elsewhere in this report.*

Another issue is the apparent finding that women are less productive in terms of the number of publications, but that their works are of high quality (Garfield, 1981) (Garfield, 1983) (Zuckerman, 1997) (Schiebinger, 1999) (Nilsson, 1997; Feller, 2004). Certain studies show women to publish more information per document than men (Long, 1992). While they publish fewer *papers*, they are cited more often (Sonnert and Holten, 1995a, 1995b, 1996). Notwithstanding, other studies do not reflect such differences (Lewison, 2001). In fields such as Astrophysics, citation analysis suggests that a publication is more likely to be cited if the author is a man, with women at a comparative disadvantage for gaining acknowledgement and funding (Rusell, 2003).

Generally speaking, there is little information about the size of research groups of men and women. Long shows that the work position does not depend on output, but rather that output depends on one´s position, from the relationship between the size of research groups and their productivity (Long, 1978). Some studies show that females are more likely to work in collaboration, but not at the international level, which in turn means a reduced participation in international congresses, key for establishing future scientific contacts.

When the predominant factor for seeking a position of influence in the hierarchy of the scientific community is related with high indexes of output in quality journals, the importance of indicators of scientific output according to gender becomes clear. Without such information, any attempt to explain the minor ladder of women in scientific communities will suffer from the lack of substantive foundations, and its validity will be questioned (Rusell, 2003).
In Spain, the *Fundación Española de Ciencia y Tecnología* features a line of work and reflection on "The role of women in scientific and technological activities". Its objective is to "help identify and analyze the factors and circumstances related with their low participation and leadership in the scientific research system, in the development of technology, and industrial innovation". In April of 2005, the monograph "Women and science: the situation of women researchers in the Spanish system of science and technology" (*Mujer y Ciencia: La situación de las mujeres investigadoras en el sistema español de ciencia y tecnología*) came out. Spain´s Higher Council for Scientific Research (*Consejo Superior de Investigación Científica*, CSIC), at the urging of the commission for Women and Science, published a study about the topic. In the *Centro de Información y Documentación Científica* or CINDOC, a project now underway, "Incorporación de la dimensión del género a los estudios bibliométricos", is being financed by the Spanish *Instituto de la Mujer*.

Material and Methods
An *ad hoc* database was created, to which a portion of the documents extracted from the Thomson-ISI database was exported. The selection of the source can be justified on the basis of its use by the National Evaluating Commission (*Comisión Nacional Evaluadora de la Actividad Investigadora*, CNEAI) as the source of reference for the evaluative processes for the concession of research incentives. The documents were selected among the 35,790 documents with international projection

published in the year 2004 and having at least one author belonging to a Spanish institution.

Selection of the sample
In order to attain a significant simple of Spain´s scientific output, we calculated the number of ISI categories in each one of the scientific areas established by the *Agencia Nacional de Evaluación y Prospectiva* (ANEP), previously associated with the different ISI categories in order to determine which area each study pertained to. The selection involved 25% of these, representing 25% of the output for that area. The documents were chosen at random. Then, a table with a register for each author was created, and a field for information on gender was inserted, along with another field for the documents not located. These tables were used to establish the percentage of women per subject matter in each one of the sample categories. Such a representative sample, made up of 9,292 documents (25.96% of the total), may entail bias in very small conglomerates, such as Law or the technological areas, where the data volume may not have been statistically significant.

At this point, each one of the authors of every document was identified. The search process was initiated in different referential sources, and was carried on in sources that give full-text access to the document (ISI, Scopus, Scyrus, Google Scholar, Sciencedirect, etc.) Upon locating each document constituting the sample, the new information was incorporated into the fields mentioned above. There was a small percentage of documents for which one or more authors could not be located despite subjecting them to exhaustive search processes. In these cases the document was extracted from the sample and in its place another study was randomly included. In this manner, after identifying and modifying all the authors of each one of the documents, we proceeded to carry out the corresponding consultation in the database, so as to elaborate tables and graphs for our results.

The indicators analyzed are:

- Total output (ndoc): number of documents
- Primary output (ndocc): only articles appearing in journals, be they with or without Impact Factor, and including contributions in Art and Humanities (A&H)
- Citable output (ncit): articles in journals with an Impact Factor registered in the Journal Citation Report (JCR), but limited to Sciences (SCI) and Social Sciences (SSCI)
- Normalized Impact Factor (FINP): that of the respective journals of publication, with each document "inheriting" the IF corresponding to the journal in which it is published, which is therefore recorded in the Journal Citation Report for the year of publication of the document; and when this information is missing, it is assigned the IF for the nearest year known. Then it is transformed by means of a procedure of normalization based on typification or standardization that allows us to work with it in comparative terms, following other authors (Braun, Glänzel and Schubert, 1985) (Rousseau, 1988), in order to generate IF values that conserve variability, while at the same time making the scales of the different categories compatible and comparable. This normalization process marks a point of reference when situating the position of the domain in question, unlike other calculations in which the resulting value is placed within a range. Thus, the typified impact factor (TIF) is calculated using the formula:

$$tif_{jc} = \frac{if_{jc} - \overline{if_c}}{\sigma if_c}$$

Where *if* stands for the Impact Factor of a journal j, in a category c, of the JCR; and tif is the normalized IF of a journal j in a category c of the JCR. The resulting values can be positive or negative, and allow us to make comparisons among diverse categories. Yet this value can be difficult to comprehend and utilize in a cumulative sense when it is negative; and so, for this purpose, we propose the scale corrector: $fin_{jc} = m+(TIF_{jc} / k)$. Hence, we may adopt m and k as two constant values that give rise to positive numerals, adequate for the objectives of analysis. In our case we used m= 1 and K=3. Thus, we manage to generate values that conserve their variability, are positive, allow for comparison among different categories, and ultimately make it possible to normalize the mean IF of a category having a value of 1, which is therefore assigned to each of the documents therein. However, upon comparison of the impact factors achieved by a given collective

with respect to another (greater) one, within a thematic area that includes several different JCR categories, certain lacks of adjustment may occur as a consequence of the different weights that each category bears in the production of each group, and of the different habits of citation within each category. In order to solve this problem, a weighted normalized impact factor is introduced, to measure the mean weighted citation expected for a set of publications pertaining to a specific community or thematic level of aggregation, and to indirectly indicate the possibility of a greater audience on the part of the scientific community (Moya, et.al., 2004) (Chinchilla y Moya, 2007). Its formula is:

$$finp = \frac{(\sum NDocc * fin)}{\sum NDocc}$$

- Patterns of behaviour: the language of publication and type of document
- Patterns of coauthorship: coauthorship index (number of authors per document) and order of the signing authors and patterns of collaboration in which we distinguish: No Collaboration (documents signed by a single institution, regardless of the number of institutional departments participating in the research; National Collaboration (documents signed by authors from more than one Spanish Institution); Interregional Collaboration (documents signed by researchers from more than one Autonomous Community of Spain); and International Collaboration (documents signed by authors from more than one country). In the computation of documents in collaboration, there may be some overlap due to the fact that a given document may pertain to more than one category (for instance, national collaboration may also entail interregional collaboration).

Results

Whereas over 95% of the output has at least one male author, women are authors or co-authors of just 65% of the publications studied. This is a considerable difference, though more so for those documents where the authors are all of the same sex: more than 30% are signed exclusively by men, and less than 5% of the publications have only female authors.

Table 1. Percentage of Ndoc, Ndocc and NdocCit by subject matter and gender of authors

Thematic Area	Abr.	Ndoc				Ndocc				NdocCit			
		Male	Female	Only Male	Only Female	Male	Female	Only Male	Only Female	Male	Female	Only Male	Only Female
AGRICULTURE	AGR	91,93	78,88	21,12	8,07	91,93	78,88	21,12	8,07	91,93	78,88	21,12	8,07
FOOD SCIENCE AND TECHNOLOGY	ALI	90,26	89,23	10,77	9,74	90,05	90,05	9,95	9,95	90,05	90,05	9,95	9,95
CIVIL ENGINEERING & ARCHITECTURE	CIV	93,94	42,42	57,58	6,06	93,94	42,42	57,58	6,06	93,94	42,42	57,58	6,06
COMPUTER SCIENCE & TECHNOLOGY	COM	98,17	43,22	56,78	1,83	98,48	43,56	56,44	1,52	98,48	43,56	56,44	1,52
SOCIAL SCIENCES	CSS	90,91	51,14	48,86	9,09	93,15	49,32	50,68	6,85	93,15	49,32	50,68	6,85
LAW	DER	100,00	100,00			100,00	100,00			100,00	100,00		
ECONOMY	ECO	82,67	42,67	57,33	17,33	84,72	41,67	58,33	15,28	84,72	41,67	58,33	15,28
ELECTRICAL, ELECTRONIC & AUTOMATED ENGINEERING	ELE	100,00	35,71	64,29		100,00	38,46	61,54		100,00	38,46	61,54	0,00
PHYSIOLOGY & PHARMACOLOGY	FAR	98,23	77,88	22,12	1,77	97,87	82,98	17,02	2,13	97,87	82,98	17,02	2,13
PHILOLOGY & PHILOSOPHY	FIL	70,00	37,50	62,50	30,00	70,00	36,67	63,33	30,00	73,68	36,84	63,16	26,32
PHYSICS & SPACE SCIENCES	FIS	98,94	51,46	48,54	1,06	98,91	51,91	48,09	1,09	98,91	51,91	48,09	1,09
LIVESTOCK & FISHING	GAN	96,77	82,26	17,74	3,23	96,49	82,46	17,54	3,51	96,49	82,46	17,54	3,51
HISTORY & ARTS	HIS	75,00	50,00	50,00	25,00	76,00	52,00	48,00	24,00	80,95	52,38	40,00	19,05
MATERIALS SCIENCE AND TECHNOLOGY	MAR	98,39	72,58	27,42	1,61	98,34	71,82	28,18	1,66	98,34	71,82	28,18	1,66
MATHEMATICS	MAT	95,36	41,06	58,94	4,64	95,21	41,78	58,22	4,79	95,21	41,78	58,22	4,79
MECHANICAL, NAVAL & AERONAUTIC ENGINEERING	MEC	99,07	31,48	68,52	0,93	99,06	32,08	67,92	0,94	99,06	32,08	67,92	0,94
MEDICINE	MED	95,86	77,04	22,96	4,14	96,05	78,20	21,80	3,95	96,05	78,20	21,80	3,95
MOLECULAR & CELLULAR BIOLOGY & GENETICS	MOL	96,64	77,87	22,13	3,36	97,41	77,18	22,82	2,59	97,41	77,18	22,82	2,59
PSYCHOLOGY & EDUCATIONAL SCIENCES	PSI	95,31	76,56	23,44	4,69	94,55	72,73	27,27	5,45	94,55	72,73	27,27	5,45
CHEMISTRY	QUI	94,65	79,68	20,32	5,35	94,39	80,19	19,81	5,61	94,39	80,19	19,81	5,61
ELECTRONIC & TELECOMMUNICATIONS TECHNOLOGY	TEC	100,00	31,71	68,29		100,00	31,58	68,42		100,00	31,58	68,42	
GEOSCIENCES	TIE	96,23	58,16	41,84	3,77	96,44	59,56	40,44	3,56	96,44	59,56	40,44	3,56
CHEMICAL TECHNOLOGY	TQU	93,90	70,73	29,27	6,10	93,83	71,60	28,40	6,17	93,83	71,60	28,40	6,17
PLANT & ANIMAL BIOLOGY, ECOLOGY	VEG	98,20	61,71	38,29	1,80	98,10	62,86	37,14	1,90	98,10	62,86	37,14	1,90

The difference is even greater in certain scientific areas where male participation is seen in over 90% of the documents in all areas (Table 1); only in Philology and Philosophy (FIL), History and Art (HIS), and Economics (ECO) is the percentage lower. On the other hand, female intervention ranges between the 31.48% seen in Mechanical, Naval and Aeronautic Engineering (MEC) and the 89.23% in Food Science and Technology (ALI). In primary production, the results do not vary: with respect to total documents, the publications turned out solely by men range from the 10.77% in Food Science and Technology (ALI) to 70% in the areas of Mechanical, Naval and Aeronautic Engineering (MEC) and Electronic and Communications Technology (TEC). The contributions by women are negligible in

these categories, while they are highest in Philology and Philosophy (FIL) with 30%, and in History and Art, with a 25% share.

Table 5. Type of document and language of publication

Document Type	Male	Female	Only Male	Only Female
Art Exhibit Review		100,00		100,00
Article	95,76	66,39	33,61	4,24
Biographical-Item	100,00		100,00	
Book Review	78,57	21,43	78,57	21,43
Correction	100,00	60,00	40,00	
Editorial Material	87,50	56,25	43,75	12,50
Letter	96,83	60,32	39,68	3,17
Meeting Abstract	98,35	84,30	15,70	1,65
Review	96,88	67,19	32,81	3,13
Theater Review		100,00		100,00

Language	Male	Female	Only Male	Only Female
English	95,80	66,79	33,21	4,20
French	80,00	40,00	60,00	20,00
Spanish	69,57	56,52	43,48	30,43

Whereas 33% of the scientific *articles* published are signed exclusively by men, those authored solely by women stand for less than 5%. However if we calculate the percentage that these represent in view of the total documents signed only by women, they amount to over 83%, thus reducing the differences with respect to articles signed by men with respect to the total (88%).

Meanwhile, publications in the English language that are signed only by women make up under 5%, while those with only male authorship are over 30%. This is due to the fact that 98% of the studies published by men are in English. Of studies by women authors, 93% are in English and around 6% in Spanish; that is, the percentage of articles written in Spanish by women is higher than that of men.

Patterns of Co-authorship

The order of appearance of author names on a document is considered useful for determining their importance, as not all positions of author names have the same value; and their importance could furthermore depend to a great extent on the subject field.

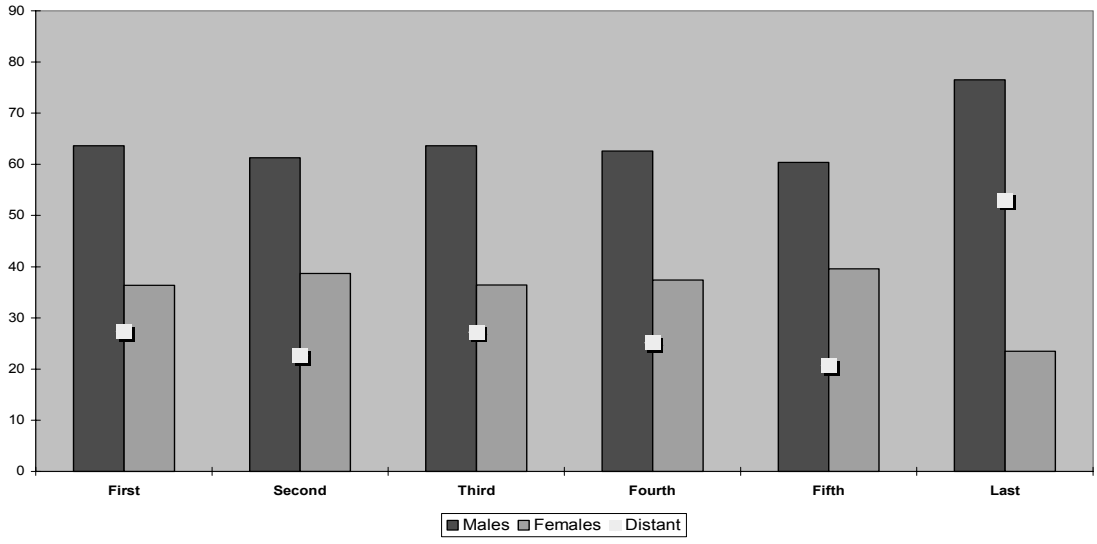


Figure 1. Percentage of documents by order of appearance of author names and gender

The usual finding is that the first and last positions are occupied by male names, related with the understanding that these positions bear a greater relevance on the work. In practice, authors themselves tend to reserve these two positions for contributors wielding a more important role. This hypothesis can be corroborated in the breakdown by gender. Yet the differences in the distances between the percentages reveal that the final position is most affected by this practice. Over 65% of the documents are signed in the first place by a male author (Figure 1), leaving the other 35% first positions for women as principal authors. As the position of the author name advances (to second,

third, fourth or fifth place), the difference in this percentage decreases. However, the distance of the final position is very noteworthy: over 75% of the studies are signed in final place by a male.

With regards to distribution by subject matter, the documents featuring a woman as first signing author are the majority in Food Science and Technology (ALI) and Law (DER), where the importance of female contributions appears to be more relevant than male input, with women signing over 50% of production. Under this figure, yet over 40%, are the areas of Agriculture (AGR), Pharmacology and Pharmacy (FAR), History and Art (HIS), Materials Science and Technology (MAR), Molecular, Cellular and Genetic Biology (MOL), Psychology (PSI), Chemistry (QUI) and Chemical Technology (TQU), wherein the contribution of women scientists is seen to bear great importance. In the final author position, we see that only for the area of Food Science and Technology (ALI) are there female names on over 40% of the studies; and in Agriculture (AGR), Philology and Philosophy (FIL) Livestock and Fishing (GAN), History and Art (HIS) and Psychology (PSI) women authors sign lastly in over 30% of cases. There are areas in which women scientists sign below 30% of documents in first and last positions, and therefore, in these fields the relevance of their scientific contribution would be considered inferior to that of male colleagues.

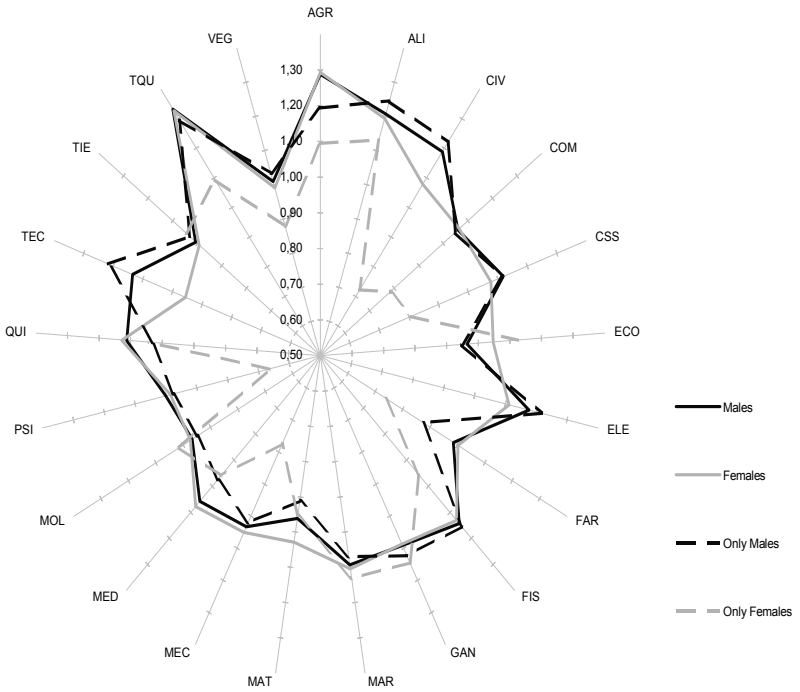


Figure 2. Standardized Impact Factor by ANEP area and gender

Although as a general rule, the documents signed only by male authors appear in journals with a higher Impact Factor, in some areas of science the opposite is true. Such is the case of Economics (ECO), Livestock and Fishing (GAN), Materials Science and Technology (MAR), Molecular, Cellular and Genetic Biology (MOL) and Earth Sciences (TIE).

Meanwhile, we see that of the documents signed only by men, over 80% are coauthored, whereas among women there is a lesser tendency toward coauthorship —some 40% of these are undersigned by a single female author (Figure 3). Out of all the coauthored documents, the ones signed solely by women stand for less than 3%, a very low figure in comparison with the 30% represented by only male coauthorship. This fact is reflected in the low index of collaborative authorship of the documents in whose elaboration only women intervene, with 2.33 authors per study. Yet the highest level of this index (5.19) is seen in conjunction with the documents having at least one woman author, meaning that in studies where men and women work together, the number of total authors is greater than in those produced by men alone. Figure 4 clearly illustrates how the more highly coauthored documents

tend to show female collaboration. The fact that there are more studies involving women where three to four authors participate leads us to surmise that these belong to areas such as Biomedicine, where the pattern of coauthorship is high. Therefore, there would appear to be a more direct relationship between patterns of publication and collaboration depending on the thematic category, than gender bias *per se*.

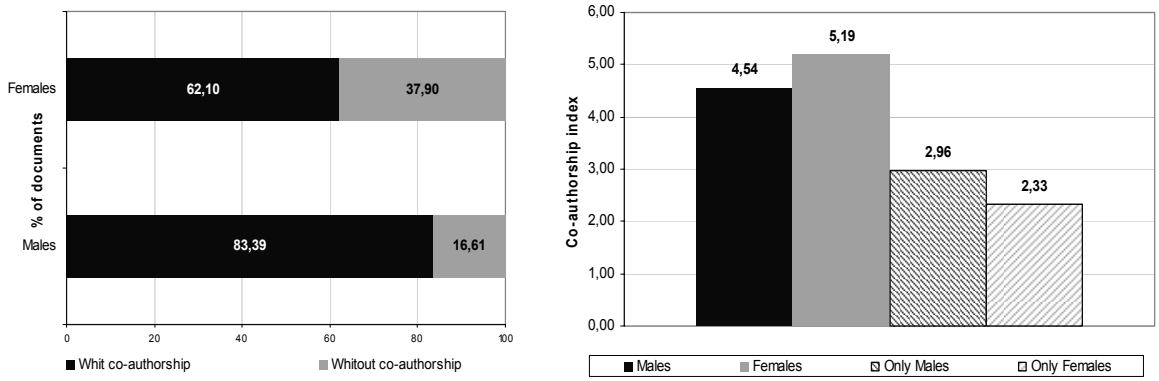


Figure 3. Percentage of documents by number of signing authors, coauthorship index and gender

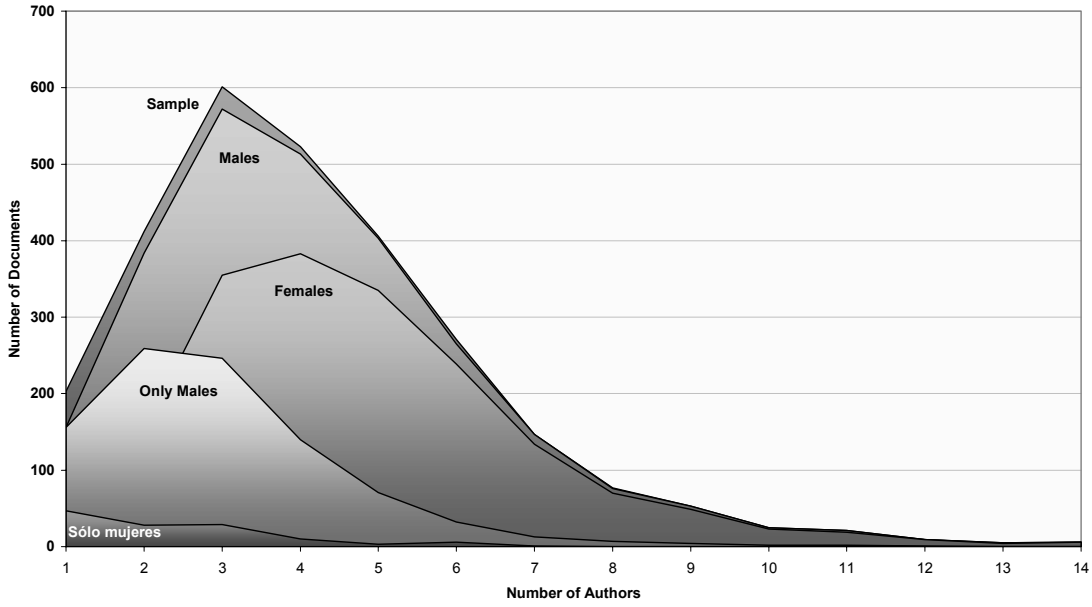


Figure 4. Patterns of coauthorship by gender

Finally, we observed an interesting network of coauthorship in the area of Agriculture within the University of Granada. This network consists of a total of 91 authors (nodes), of which 41 are men and 50 are women, making manifest the superiority (in number of participants) of female research in this category and time period. The relationships (links) between and among authors conform the different authorship networks, allowing us in some cases us to distinguish specialties for each.

From the structural standpoint, we can easily spot components or associations of authors, which implies that research is made up by fourteen nuclei of consolidated areas of investigation (regardless of some overlap). In most of the components there is a symbiosis of authors of both sexes, with a slight female predominance. There are three exceptions: a network of five authors and another of two

authors, situated in the upper and lower right sectors of the figure, respectively made up uniquely by men; and another of three authors, in the lower right section, made up exclusively by women. In the rest of the networks, we can find at least one female author. Noteworthy are those components constituting five or more authors, in that the point of interconnection or *break point* of each is consistently represented by a female researcher. This suggests the significance, capacity of centralization and the prestige of each of the following women scientists: Urbano, G., Barrionuevo, M.,and DeLaSerrana, HLG.

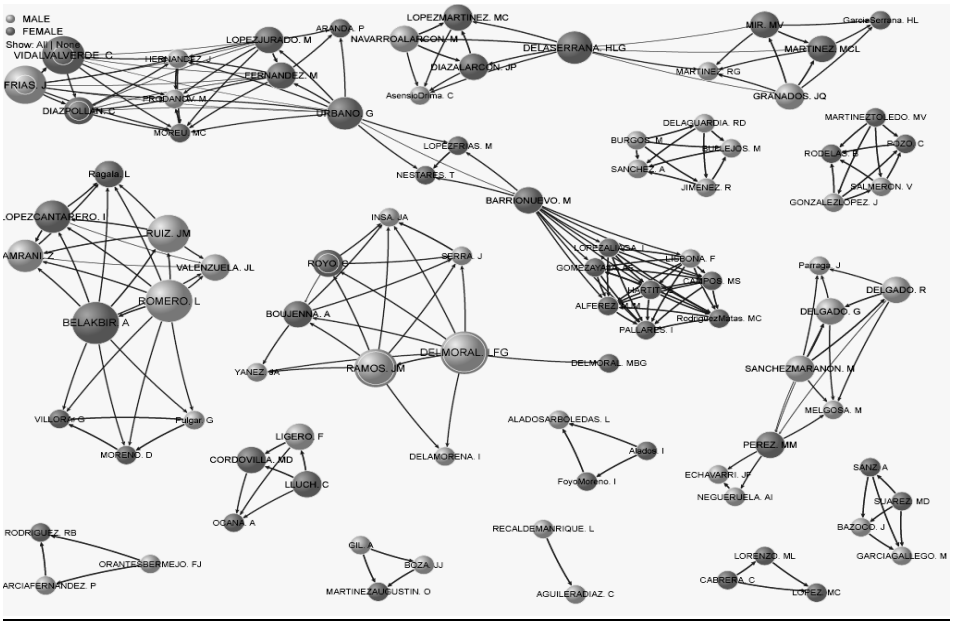


Figure 5. Network of Co-authorship, University of Granada. ANEP area: Agriculture

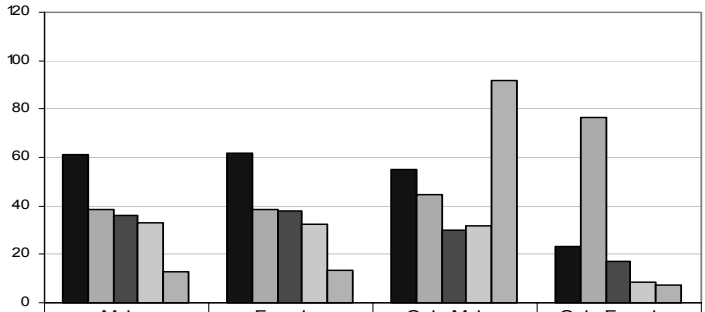
Patterns of institutional collaboration

Documents having exclusively male intervention account for 30% of the total in each type of collaboration, while works undersigned only by women do not surpass 3% in any type of collaboration. The patterns of collaboration (Figure 6) show that for men, the most habitual type of collaboration is interregional (more than 90% of their work), followed by international, at nearly 32%, and national, with just over 30%. On the other hand, as women hold a documental share of over 75% without institutional collaboration, the figures for collaboration are much lower. The 17% seen for national collaboration among women is followed by an 8% for international collaboration, and finally an interregional collaboration rate of 7%.

Conclusions

When considering the identification of authors of scientific articles, the information offered by data sources is very often incomplete, and sometimes erroneous. To alleviate this obstacle, authors should always sign with the exact same name, and whenever possible, Spanish or Hispanic authors should connect their two surnames with a hyphen. Otherwise, foreign language databases are likely to make mistakes when assigning them an entry.

A horizontal segregation can be perceived in our results, owing to the fact that in certain areas, an unequal development by gender might be attributed to stereotypes lodged in formative or professional expectations surrounding women. Yet it is vital to bear in mind that these differences are strongly influenced by the patterns of publication in the different subject areas. Therefore, it may be that gender differences are rooted not only in stereotypes, but also in the dynamics of different areas of scientific activity or research.



	Males	Females	Only Males	Only Females
%col	61,21	61,65	55,31	23,39
Without Institutional Collaboration	38,79	38,35	44,69	76,61
Domestic	36,09	37,76	30,23	16,94
International	33,32	32,38	31,94	8,87
Inter-regional	13,16	13,23	91,94	7,26

Figure 6. Patterns of collaboration

Table 6. Percentage of documents and form of collaboration by areas and gender (*)

Thematic Area	Whitout Institutional Collaboration				Domestic Collaboration				International Collaboration				Inter-regional Collaboration			
	Males	Females	Only Males	Only Females	Males	Females	Only Males	Only Females	Males	Females	Only Males	Only Females	Males	Females	Only Males	Only Females
AGR	41,61	39,13	8,70	6,21	60,25	51,55	14,91	1,86	24,84	17,39	8,07	0,62	12,42	10,56	2,48	0,62
ALI	48,72	52,82	3,59	7,69	64,10	65,64	7,18	1,03	18,97	15,90	4,62	1,54	7,18	6,15	1,03	
CIV	57,58	24,24	39,39	6,06	72,73	36,36	42,42		15,15	12,12	3,03		6,06		6,06	
COM	48,35	17,95	31,50	1,10	69,60	30,04	40,66	0,73	26,01	13,92	12,09		10,99	4,40	7,33	0,73
CSS	40,91	20,45	26,14	5,68	63,64	32,95	36,36	3,41	28,41	14,77	13,64		12,50	6,82	7,95	2,27
DER	100,00	100,00			100,00	100,00				0,00	0,00		0,00			
ECO	44,00	28,00	30,67	14,67	62,67	33,33	44,00	2,67	24,00	5,33	18,67		17,33	6,67	12,00	1,33
ELE	39,29	17,86	21,43		67,86	21,43	46,43		46,43	10,71	35,71		10,71	3,57	7,14	
FAR	29,20	20,35	9,73	0,88	55,75	46,02	11,50		35,40	32,74	3,54	0,88	9,73	3,54	6,19	
FIL	40,00	27,50	37,50	25,00	50,00	27,50	47,50	2,50	12,50	2,50	10,00		5,00		5,00	
FIS	36,34	18,30	18,83	0,80	68,17	35,01	33,95	0,27	42,71	22,02	20,69		12,20	7,69	4,77	0,27
GAN	37,10	32,26	8,06	3,23	58,06	45,16	16,13		38,71	30,65	8,06		19,35	19,35		
HIS	35,71	28,57	28,57	21,43	46,43	32,14	35,71		10,71	3,57	7,14		17,86	14,29	3,57	
MAR	45,16	31,18	15,59	1,61	69,89	51,08	20,43		30,11	24,73	5,38		10,75	7,53	3,23	
MAT	46,36	18,54	31,13	3,31	70,20	31,13	43,05	0,66	30,46	13,91	17,22	0,66	12,58	2,65	10,60	0,66
MEC	49,07	12,04	37,04		75,93	22,22	53,70	0,93	29,63	10,19	19,44		9,26	2,78	7,41	0,93
MED	30,89	25,33	8,28	2,72	50,30	40,36	13,02	1,07	27,34	21,78	6,04	0,47	15,03	11,95	3,31	0,24
MOL	34,78	28,26	9,68	3,16	59,29	47,23	15,22	0,20	34,58	27,47	7,11		12,06	9,68	2,37	
PSI	50,00	42,19	10,94	3,13	70,31	57,81	17,19		26,56	21,88	6,25	1,56	3,13	1,56	1,56	
QUI	43,32	36,90	10,70	4,28	64,35	54,01	15,15	0,53	27,45	22,10	6,24	0,89	10,70	9,80	1,07	0,18
TEC	56,10	19,51	36,59		87,80	29,27	58,54		34,15	9,76	24,39		2,44		2,44	
TIE	29,71	17,57	14,64	2,51	67,36	38,49	31,38	1,26	48,54	28,87	20,08	0,42	15,06	9,21	6,69	0,84
TQU	41,46	30,49	14,63	3,66	63,41	48,78	20,73		26,83	20,73	8,54	2,44	12,20	9,76	2,44	
VEG	30,63	18,92	13,51	1,80	65,32	40,09	27,03		45,50	29,28	16,22		14,41	8,56	5,86	

(*)Percentages calculated with respect to total documents signed by men, women, only men or only women.

Similarly, the patterns of coauthorship seen here are more closely tied to the subject area than to author gender; and even though women are unevenly distributed in this sense, there are also areas with a high participation in terms of authors and institutions. For instance, Biomedicine has a very noteworthy female presence, whereas in the Humanities, the norm is solitary output, traditionally at the hands of more women than men. We cannot, then, state that women participate less in coauthored works. Simply, it depends on the specific area analyzed. Likewise, we are not able to conclude here, at least, that female researchers tend to resort to national journals for publication rather than international ones. The fact that the woman scientist publishes a greater percentage of studies in the Spanish language has more to do with the area of study, as female participation is traditionally greater in the humanities and social sciences.

Finally, this study will lead us in the near future to enlarge the sample, to dispose of data for total production in view of gender and elaborate in greater depth an analysis of the different subject areas, at different levels of thematic grouping, where bibliometric arguments and the weight of social structures can be evoked to help us comprehend the true role of women in science in each particular

discipline. At the same time, further study will help us avoid the types of bias that may have been introduced here due to the limited volume of the data conglomerates.

References

Bordons, M. & Muñoz, A. (2004). *Estudio bibliométrico sobre mujer y ciencia: segundo informe para el grupo de expertas Mujer y Ciencia*. Madrid: Fecyt.

Bordons, M. & Mauleón, E. (2003a). *Indicadores bibliométricos por género aplicados al estudio del CSIC: Informe preliminar para el grupo de expertas Mujer y Ciencia*. Madrid: Fecyt.

Bordons, M., Morillo, F., Fernández, M.T., & Gómez, I. (2003b). One step further in the production of bibliometric indicadors at the micro level: Differences by gender and professional. *Scientometrics*, 57, 159-173.

Bordons, M., & Mauleón, E. (2006). Indicadores bibliométricos por género en tres áreas del Consejo Superior de Investigaciones Científicas (CSIC). *Proceedings of the I International Conference on Multidisciplinary Information Sciences & Technologies*. Mérida: , Spain, 25-28 October.

Braun, T.; Glänzel, W., & Schubert, A. (1985) *Scientometric Indicators: A 32-Country Comparative Evaluation of Publishing Performance and Citation Impact.* Philadelphia: World Scientific; 1985.

Chinchilla Rodríguez, Z. & Moya Anegón, F. (2007) *La investigación científica española (1995-2002): una aproximación métrica*. Granada: Universidad de Granada.

Cole, J. R. & Zuckerman H. (1987). Marriage, motherhood and research performance in science. *Sci.Amer*. 256, 119-25.

Cole, J. R. (1979). *Fair science: women in the scientific community*. New York: Free Press.

Davenport, E., & Snyder, H. (1995). Who cites women? Whom do women cite? An exploration of gender and scholarly citation in sociology. *Journal of Documentation*, 51, 4004-410.

Garfield E. (1980). The 1,000 contemporary scientists most-cited 1965-1978. Part 1. The basic list and introduction. *Current Contents*, 41, 5-14

Garfield E. (1983). Why aren't there more women in science? *Currents Contents*, 17, 5-12, 26 April 1982. (Reprinted in: Garfield E. Essays of an information Scientist, Philadelphia: ISI Press, 5, 498-505

Goel, K. (2002). Gender differences in publication productivity in psychology in India. *Scientometrics*, 55, 243-258

Gupta, B.M., Kumar, S., Aggarwal, B.S. (1999). A comparison of productivity of male and female scientists of CSIR. *Scientometrics*, 45, 269-289.

Helsinki Group. Retrieved October 10, 2006 from: http://ec.europa.eu/research/science-society/page_en.cfm?id=2906

Instituto Nacional de Estadística INEbase (2006). Retrieved October 18, 2006 from: http://www.ine.es/inebase/index.html

Lewison, Grant (2001). The quantity and quality of female researchers: a bibliometric study of Iceland *Scientometrics*, 52, 29-43.

Long J S. (1992). Measures of sex differences in scientific productivity. *Sot. Forces* 7, 159-78.

Long, J. S., Fox, M. F. (1995). Scientific careers: universalism and particularism. *Annual Review of Scociology*, 21, 45-71.

Mauleón, E., & Bordons, M. (2005). *Indicadores bibliométricos por género en el CSIC: estudio del área de Ciencia y Tecnología Físicas.* Madrid: Cindoc-Csic.

Moya Anegón, F., Chinchilla Rodríguez, Z., Corera Álvarez, E., Herrero Solana, V., Muñoz Fernández, F.J., & Vargas Quesada, B. (2004). *Indicadores bibliométricos de la actividad científica española*. Madrid: Fecyt.

Moya Anegón, F.; Chinchilla Rodríguez, Z., Vargas Quesada, B., Corera Álvarez, E., Muñoz Fernández, F. J., González Molina, A., & Herrero Solana, V. (2007). Coverage analysis of Scopus: a metric approach. *Scientometrics*, 73, (in press)

Oficina de Estadística Comunitaria Eurostat. (2006). Retrieved October 20, 2006 from: http://epp.eurostat.ec.europa.eu/portal/page?_pageid=1090,30070682,1090_33076576&_dad=portal&_schema=PORTAL

Prpic, F (2002). Gender and productivity differentials in science. *Scientometrics*, 55, 27-58.

Rousseau, R. (1988) Citation Distribution of Pure Mathematics Journals. Informetrics 87/88. *Select Proceedings of the First International Conference on Bibliometrics and Theoretical Aspects of Information Retrieval*; Diepenbeek, Belgium. Elsevier Science Publishers, 249-261.

Rusell, J. (2003). Los indicadores de producción científica por género: un caso especial. *Tercer Taller de Obtención de Indicadores Bibliométricos*. Madrid: Red Iberoamericana de Indicadores de Ciencia y Tecnología.

Stack, S. (2002). Gender and scholarly productivity: 1970-2000. *Sociological Focus*, 35, 285-296.

UNESCO (1999). Ciencia para el siglo XXI: por un nuevo compromiso. *World Conference on Science, 1999*, Budapest: Unesco. Retrieved October 5, 2006 from: http://www.unesco.org/science/wcs/esp/ declaracion_s. htm

Webster, B. M. (2001) Polish women in science: a bibliometric analysis of Polish science and its publication, 1980-1999. *Research Evaluation*, 10, 185-194.

Emergence of a New Discipline in the Earth Sciences: Bibliometric Analysis of Photogrammetry and Remote Sensing Literature

Ahat A. Nabiullin

ahatnabi@poi.dvo.ru
Laboratory of Scientific Information, Department of Informational Technology, V.I. Il'ichev Pacific Oceanological Institute, Vladivostok, 690041 (Russia)

Abstract
Remote sensing science is a rapidly growing field of the Earth sciences. Since its emergence and to the present day an extensive volume of literature has been published which traces wide application of remote sensing in human activities. According to the ISI Web of Science in 1965-2005 more than 19,000 papers were published on remote sensing. A number of papers grew exponentially with a doubling period of about 6 years. Notwithstanding all specialized journals there are a lot more remote sensing papers published in a vast list of source titles (up to 350 journals). Only 25% of the retrieved papers are published in 10 journals which ISI is assigned to a subject category of remote sensing. The time line of remote sensing periodicals issued in 1908-1995 shows an exponential growth with a doubling period of about 15 years. After 1995 there is a prominent deviation from the exponential curve which shows the demand saturation for specialized journals. The revealed features are discussed in terms of dynamics and impact of remote sensing in the current Earth sciences development.

Keywords
remote sensing literature; earth sciences

Introduction

Remote sensing is a youngest discipline in the Earth sciences. The term "remote sensing" was first coined by E. L. Pruitt and W. H. Bailey in the early 1960s. Early in the 1970s the first remote-sensing satellites were launched in the United States and in the U.S.S.R. (Space exploration, 2006). Satellite remote sensing quickly evolved from the sphere of pure research to that of the worldwide day-to-day application. Now the remote sensing is gathering force in ecology and environmental sciences.

Remote sensing science is a rapidly growing field of the Earth sciences (English & Miller, 1989). Since its emergence and to the present day an extensive volume of literature has been published which traces the wide application of remote sensing in human activities. The aim of the report is to discuss this progress in terms of dynamics and impact of remote sensing in current Earth sciences development.

Data Acquisition and Processing

The main bibliographic database used in the study was Thomson-ISI Web of Science (WoS). Searches were performed using "remote sens*" or "photogrammetr*" keywords in the SCI-EXP database in the 1900-2005 time span. More than 19,000 papers were published during the given time in the ISI-covered scientific journals worldwide. To estimate number of papers published in the Earth sciences we used the Georef database, produced by the AGI. It contains more the 2.3 million papers of the geoscience literature of the world.

For evaluation of numbers of journals issued on remote sensing the Ulrich's Periodicals Directory was used. Data retrieval was performed with "remote sensing" keyword. More than 50 currently issued journals were found in this database. Number of the Earth sciences journals was estimated using the Ulrich's Periodicals Directory as well. In all cases only "Active" and "Academic/Scholarly" journals were taken into account.

To calculate the growth rates of the literature the semi-logarithmic plots of cumulated numbers of the papers published or journals issued were used (Figure 1). Assuming exponential growth of those data is correct, we may calculate the doubling periods ***D*** by the slopes of the linear trends.

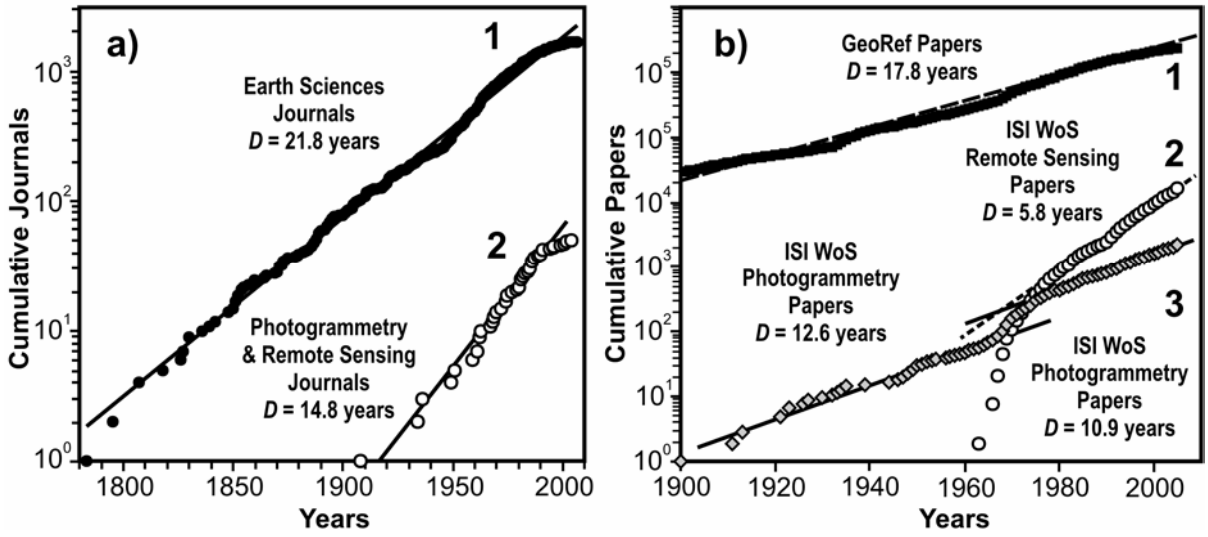


Figure 1. Growth of photogrammetry and remote sensing literature, 1900-2005: **a**) Earth sciences (1) and Photogrammetry and remote sensing (2) journals. Data from the Ulrich's Periodicals Directory, 2006; **b**) GeoRef papers (1), Remote sensing (2) and photogrammetry (3) papers. Data from GeoRef, 1800-2005 and the WoS, 1900-2005.

Results and Discussion

According to Derek Price (Price, 1963) to measure a progress of science or some scientific discipline one can evaluate the number of people involved or their scientific outputs (books, journals, papers, patents etc.) as well as material inputs (for example, GERD or university monetary funds). In remote sensing science we have tried to estimate a growth of scientific outputs as dynamics of periodicals issued and scientific papers appeared in the ISI-covered journals.

Exponential Growth of Remote Sensing Literature

Figure 1 shows the increase of remote sensing literature in terms of journals issued (a) and articles published (b). Figure 1a depicts the growth of the Earth sciences journals and papers during last one and a half century as well. Doubling period ***D*** of this exponential growth is 21.8 while a number of the remote sensing journals increased with a doubling period of 14.8 years between 1908 and 1995. Last decade noticeable decrease of a number of newly issued journals is observed.

Data of the GeoRef database show the constant rate of the serial papers with doubling period ***D*** = 17.8 years (Figure 1b). The search in the WoS database revealed an exponential growth of the remote sensing papers published in the ISI-covered journals in the 1975-2005 time span. Doubling period of this increase is 5.8 years. However, the growth of photogrammetry papers has a more complicated shape. Since the beginning of the last century till mid-1960s and from mid-1970s to our time the cumulative number of photogrammetry papers increased at approximately the same rates (average doubling period is 11.7 ± 0.9 years). But in the decade between 1965 and 1975 the rate of the growth was as great as 3-fold (***D*** ~ 3.5 years). This period coincides with the period of rapid growth of remote sensing papers as well. There was the time of the first remote sensing observations from the space and rapid advances in digital image processing (Campbell, 2002).

Diffusion of Remote Sensing Knowledge

Analysis of the remote sensing articles retrieved from the ISI database showed wide spreading of the papers between scientific periodicals and disciplines. Notwithstanding all specialized journals, there are a lot more remote sensing papers published in a vast list of source titles (up to 350 journals). Only 25% of the retrieved papers are published in 10 journals which the ISI assigns to the subject category of remote sensing. That is the reason why the growth rate of remote sensing literature is 2.5 times greater than that of photogrammetry and remote sensing journals.

In the 1975-2005 time span there is a prominent increase of the remote sensing applications in ecology and environmental sciences, meteorology and oceanography (Figure 2). On the other hand, the trend is toward decrease of share of publications in geochemistry/geophysics as well as in optics and imaging science. Other remote sensing applications account for *cca* 40% of the papers published in 1975-2005 (Figure 2).

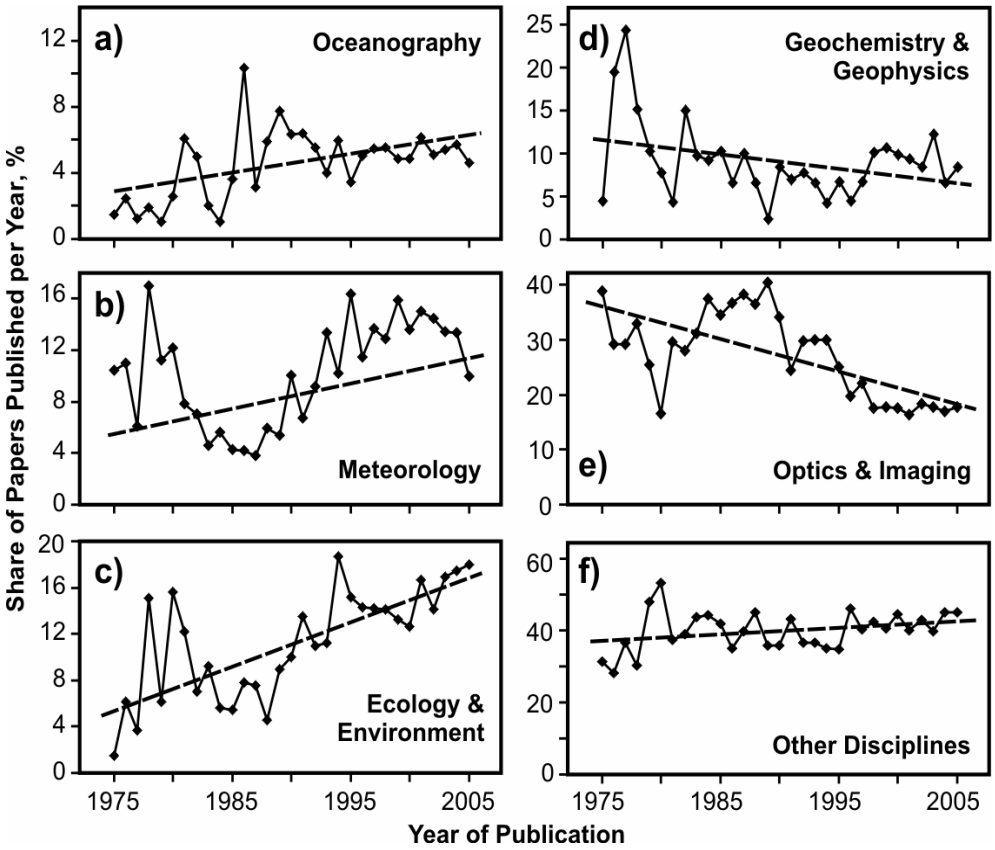


Figure 2. Shares of some scientific disciplines in remote sensing papers, 1975-2005. Data from the Wows.

Citation Delay in Remote Sensing Science

To answer a question about the use of scientific literature on remote sensing we analyzed reference lists in all papers of the remote sensing journals published in 2005. Figure 3a shows this citation delay curve. Obviously the citation curve has two regions: first one, 1960-2005 of "current" literature with delay half-life of 6.2 years and second, 1820-1960 of "classic" (Price, 1965) citations with delay of 20 years. First delay half-life is almost equal to the doubling period of the periodical papers growth and is slightly lower than the average cited half-life of the remote sensing journals (7.1, Table 1). So aging of the literature in remote sensing science is very close to that in technical literature (Burton & Kebler, 1960).

It is notable that according to Figure 3 "current" literature in remote sensing is all publications issued in 1960-2005. In other words, 99% of the published papers still used in remote sensing belong to "current" literature domain and only 1% of cited works is older than 50 years. This is obviously because the remote sensing science is still young. Figure 3b also depicts the citation delay of the papers published in 2005 journals on neural networks, another young science emerged in the middle of the last century. The trend of this curve is the same and citation half-life of the "current" (1950-2005) literature is 5.8 years.

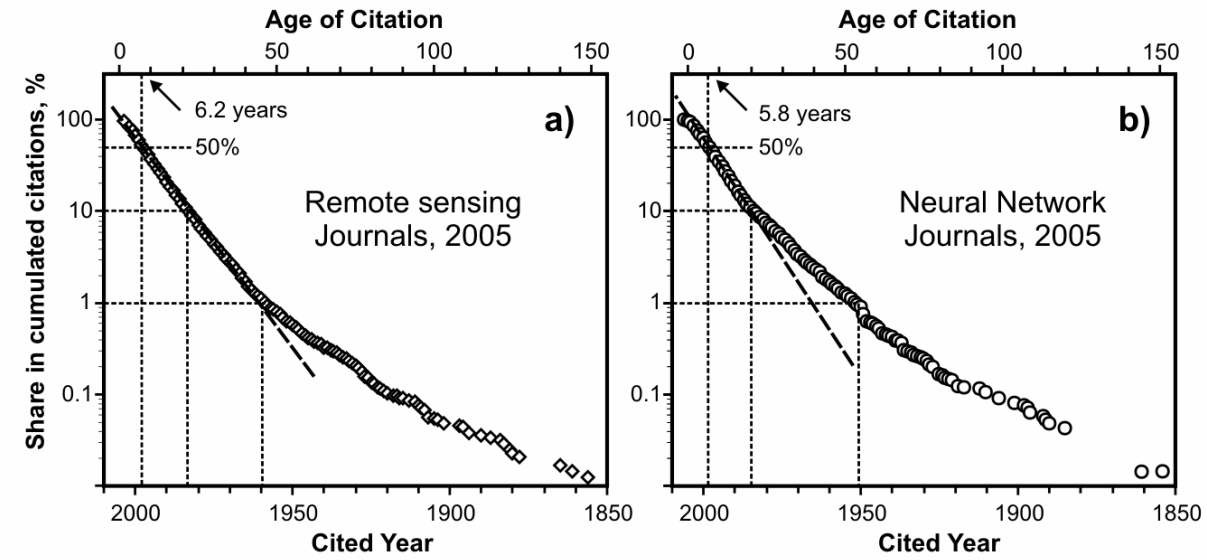


Figure 3. Citation delay of references in remote sensing (**a**) and neural networks (**b**) journals published in 2005. Data from the WoS.

Impact of Remote Sensing Journals

The ISI Journal Citation Report (JCR), 2005 assigns 10 journals to subject category of remote sensing (Table 1). These journals annually publish 20-370 citable items. Half of them are small enough (less than 100 papers per year); three journals are large (more than 200 issued papers annually). In average, these journals publish by 35% more papers per year as compared to the average of all Earth sciences journals (Table 1). But they receive citations by only 5% more, and their average impact factor is slightly lower than that of all geosciences journals. An average immediacy index of remote sensing journals is less than that of all geojournals by 50%. It means that the "instantaneous" citations (citing papers published in the same year) are infrequent in remote sensing. Probably, it is due to the wide scattering of remote sensing papers discussed above.

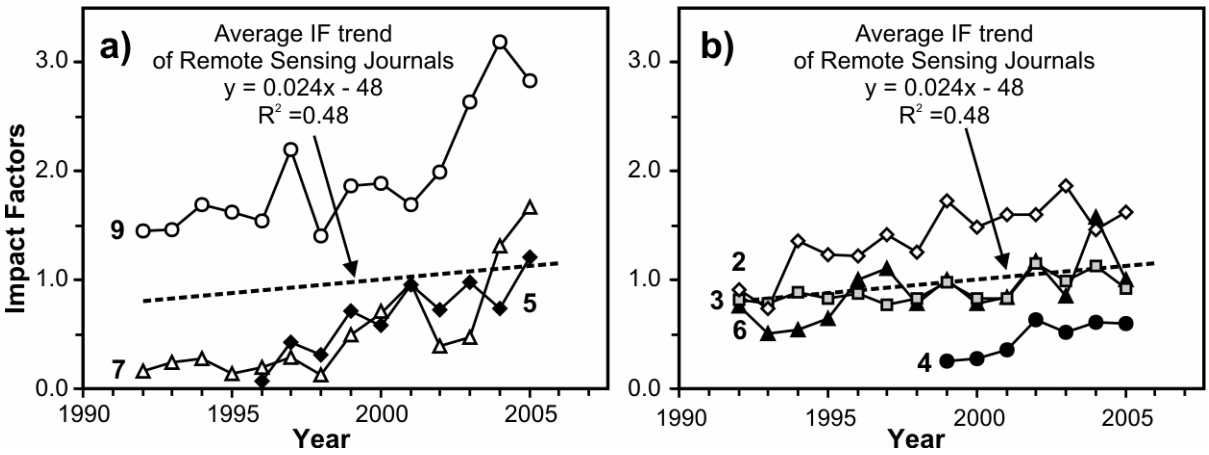


Figure 4. Impact factors of some remote sensing journals, 1992-2005 (numbering see in Table 1). Dotted line is a trend of average impact factors of all ISI-covered remote sensing journals. Data from JCR, 1992-2005.

Basic publication and citation data of geosciences journals evaluated from the JCR, 1998 (Ren and Rouseau, 2002) show the same features in comparison of remote sensing journals with all geosciences journals. These parameters increased by 50-70% in 2005. In Figure 4 the trend of average impact factors (1992-2005) of remote sensing journals is shown along with the changes of impact factors of some journals listed in Table 1. This trend demonstrates ~2.4% per year increase of the impact of

those journals. The dramatic change of the impact factor of *Remote Sensing of Environment* journal favours the view that in recent years remote sensing applications in environmental studies increase both in number (see Figure 2c) and impact.

Table 1. Basic publication and citation data of remote sensing journals. Data from the JCR, 2005. Start years are from the Ulrich's Periodicals Directory.

No	Abbreviated Journal Title	ISSN	Start Year	Total Cites	Impact Factor	Immediacy Index	Source Items	Half-life
1	CAN J REMOTE SENS	0703-8992	1975	537	0.862		41	4.9
2	IEEE T GEOSCI REMOTE	0196-2892	1963	5613	1.627	0.303	290	6.5
3	INT J REMOTE SENS	0143-1161	1980	5309	0.925	0.182	369	7.4
4	ISPRS J PHOTOGRAMM	0924-2716	1949	447	1.674	0.000	22	6.4
5	J GEODESY	0949-7714	1927	605	1.205	0.160	75	4.7
6	PHOTOGRAMM ENG REM S	0099-1112	1934	2402	1.000	0.159	107	9.7
7	PHOTOGRAMM REC	0031-868x	1953	142	0.594	0.235	17	7.2
8	RADIO SCI	0048-6604	1966	2391	0.951	0.191	110	>10.0
9	REMOTE SENS ENVIRON	0034-4257	1968	6543	2.833	0.346	231	6.9
10	SURV REV	0039-6265	1931	68	0.137	0.000	29	
Average of 10 remote sensing journals				2406	1.181	0.175	139	7.1
Average of 689 Earth Sciences journals				2292	1.296	0.269	102	7.2

Conclusions

The bibliometric analysis of remote sensing literature shows it is still young science with the growth rate higher than that of other Earth sciences. Some applications of remote sensing (for example in ecology and environmental studies) shoot up even greater.

Currently, we are experiencing an information boom in remote sensing. The literature on the subject is quite extensive, widely spreaded between source titles and scientific disciplines. That is why the scientists have no time to read and/or to cite latest papers in their publications. All one need to do is to use the abstract services or bibliographic databases to find all publications of interest.

References

Burton, R.E. & Kebler, R.W. (1960) The "Half-Life" of some scientific and technical literature. *American Documentation*, 11, 18-22.

Campbell, J.B. (2002) *Introduction to Remote Sensing*. London: Taylor & Francis. p. 6.

English, P.W. & Miller, J.A. (1989) *World Regional Geography*. New York: John Wiley & Sons.

Price, D.J. (1963) *Little Science, Big Science*. New York: Columbia University Press.

Price, D.J. (1965) Networks of Scientific Papers. *Science*, 149, 510-515.

Ren, S. & Rousseau, R. (2002) A citation data analysis of JCR-covered journals in geosciences. *Journal of Library and Information Science*, 28, 4-13.

Space exploration (2006) In *Encyclopædia Britannica*. Retrieved September 13, 2006, from Encyclopædia Britannica Online: http://www.search.eb.com/eb/article-237076.

Applying Bibliometric Mapping in a High Level Science Policy Context. Mapping the Research Areas of Three Dutch Universities of Technology [1]

Ed C.M. Noyons and Clara Calero-Medina

noyons@cwts.leidenuniv.nl, clara@cwts.nl
Centre for Science and Technology Studies (CWTS), Leiden University, PO Box 9555, 2300 RB, Leiden (The Netherlands)

Abstract
Bibliometric maps have evolved to useful tools for science policy issues. The complexity of the structures, however, makes it often very difficult to interpret the results. Moreover, interactive tools to explore the maps and results complicates the matter of interpretation and the drawing of unique conclusions. In this study, we present a case study in which we use the bibliometric mapping results to address a high level science policy issue of research efficiency. By revealing the results in a alternative way, we hope to increase the utility of biblometric mapping within the science policy context.

Keywords
science policy; bibliometric mapping; collaboration network analysis; research profile analysis

Introduction

For the past ten years bibliometric mapping applications have been developed to be used to provide possible answers to policy related questions like:

-What does a particular field look like?
-Who are prominent actors in that field?
-What is the expertise of a particular actor?
-What is the excellence of a particular actor?

These mapping techniques were developed for this purpose, because we presume mapping constitutes a proper platform to provide valuable answers. However, we have to admit that it appears difficult in many cases to understand them and to interpret the results and as such to actually use them to provide useful answers. In other words we have not yet been able to constitute an interface between results of bibliometric mapping and the user at science policy level.

Since the graphical interfaces have become an important tool to explore information, interactive maps have been developed to be used to provide answers to the policy related questions as the ones listed above. However, we may wonder whether this is the right approach to equip policy makers. An interface enables the user to extract a variety of answers depending on particular parameters, while a policy maker tends to need one answer or result.

Maps in this context are developed to structure science fields and thus to sophisticate results. It seems however that it is often very difficult to understand or interpret this structure. Experts are often needed to interpret the results and to create a reference to the actual situation. In the study presented in this paper, we propose a way of presenting the results of a bibliometric mapping approach in such a way that the user does not need to interpret the structure and is still able to get an answer to a high level science policy related question.

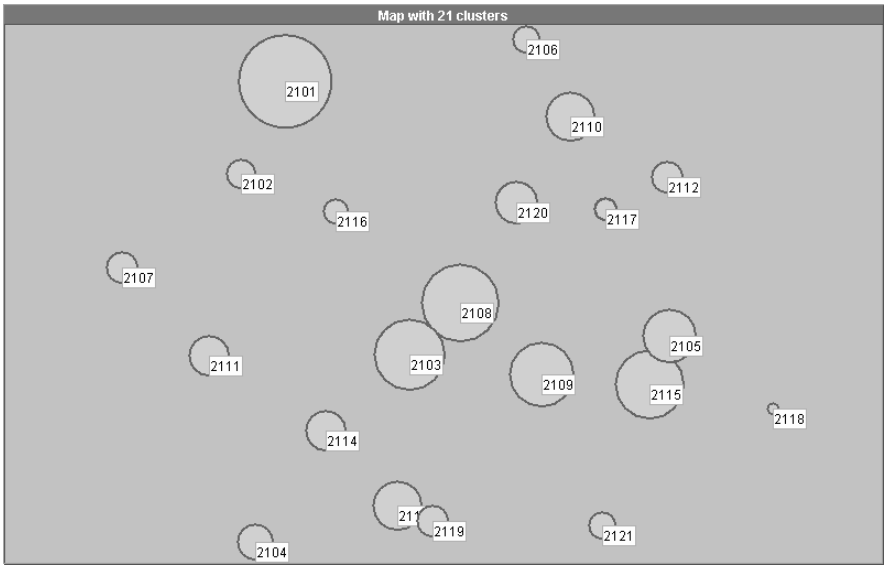
[1]. This study was conducted within the contract of 'The 3TU landscape' for the 3TU federation in close collaboration with the 3TU/IST board. We wish to thank in particular the three deans involved Prof. KCAM Luyben (TUD), Prof. A Bliek (UT), Prof. WMJ de Jonge (TUe) as well as the 3TU/IST secretary ANM Langendorff.

Background of the study

The preparation of the federation of the three Dutch technology universities (3TU) was initiated in 2003 to improve their contribution to the knowledge economy by an increased collaboration platform. In 2007 this federation should be created. The three universities are:

-Delft University of Technology;
-Eindhoven University of Technology; and
-University of Twente.

An important issue at the start-up op the federation is the picture of the landscape of the research at the three different universities. Therefore, the Institute of Science and Technology of the 3TU Federation, decided to use bibliometrics to create this landscape of 3TU on the basis of the 3TU publication output. This landscape was meant partly to assess the potential of the three universities and partly as a requirement of the Dutch government to determine the amount of overlap of research done at the different universities, and thus to be able to manage the research more efficiently.Normally a landscape would be a 2-dimensional representation of the research done at the three universities involved. This landscape would be created with the standard CWTS bibliometric mapping analysis, in which the map would contain sub-domains, defined by clusters of topics. These clusters of topics would be determined by co-word analysis. A typical map is shown below. The circles represent the sub-domains (topic clusters), while their position represents the cognitive orientation of these sub-domains in relation to each other. Sub-domains with a similar cognitive orientation in relation to all other sub-domains are in each other's vicinity, while sub-domains with a different orientation are distant from each other. The cognitive orientation is determined by the overlapping publications and topics.



Typical bibliometric map to represent a science field structure

In view of these requirements, the main objective of the study was to create the map of the universities involved the 3TU landscape. On of the problems, however, was: would the user be able to understand the cognitive maps that we usually provide in such studies. An additional problem we encountered was that it appeared that the scope of 3TU research so broad that the map would probably become an atlas. Together with this atlas we would need a set of experts to interpret the maps and results. At an early stage of the project we came to the conclusion that the cognitive maps would not be such a good idea as a result of the study. In other words the landscape would not be a landscape of topics. Instead, we decided in close collaboration with 3TU/IST to create a map not with the usual sub-domains but with elements that were easy to understand and referred to 'the real world'.

Data

The data provided by the three universities for 2000-2004, were extracted from the national output management system METIS, by each individual faculty, in a user-friendly well-structured format. This structured format with tagged fields allowed us to process the data in a relatively easy way and thus to identify the relevant bibliographic fields for this study:

-authors
-source type
-source title
-other source information (volume and page numbers)
-publication year
-title
-keywords and abstract

Only for one faculty we received the data as a list of references. These data were processed with an in-house tool to identify the relevant information.
If possible, the additional information was added to the original bibliographies by matching the journal articles to the in-house database, built upon Thomson Scientific's Web-of-Science, formerly the ISI indexes. We will refer to those publications as the ISI (covered) publications. By using the additional information from the in-house database we were able to:

-identify double publications (within one faculty or across faculties)
-add additional address data (not covered by the METIS system properly)
-add keyword and abstract information
-add citation data to conduct a performance analysis (not discussed in this paper)

For the non-ISI publications we collected the available information and used that in our study. With this information we also were able to identify double publications, i.e., publications entered into the system twice within one faculty or by different faculties. This repository should represent the entire output of the faculties of the three universities.

Approach of the study

The map had become the holy grail of the study. The landscape was expected to reveal the overlap and full potential of the three universities. However, we concluded that the standard bibliometric map, with sub-domains as clusters of topics, would not be appropriate to reveal this overlap. The maps should not be a standard bibliometric or co-word map (Noyons, 1999) with clusters of topics representing cognitive sub-domains of the entire field, in this case, covering the research done at the three TU's.

Roughly, there were two reasons not adopt this standard approach. The first one was the fact that the research scope of the three TU's was too broad to be represented in a comprehensive way. The second was the fact that the results should be discussed with people at the highest level of S&T policy while these maps should be interpreted on the level of the field experts. Moreover, we argue that in view of the main objective of this study - being to determine the amount of overlap of research done at the different TU's - the standard cognitive oriented map would not be necessary.

The map we created contains items that refer to something that is recognized by this kind of user and is fit to be used to determine the amount of overlap. The items we used to create the maps are *faculties*. Actually, we would have preferred to use departments but the structure of the METIS data did not allow us to break down to that level. The faculties per university we used to create the maps are:

Delft University of Technology

BK	Architecture
CITG	Civil engineering & Geosciences
EWI	Electrical Engineering, Mathematics & Computer science
IRI	Interfacultary Reactor Institute
LR	Aerospace Engineering
OCP	Industrial Design Engineering
OTB	Research Institute for Housing, Urban and Mobility Studies
TBM	Technology, Policy & Management
TNW	Applied sciences

Eindhoven University of Technology

BK	Architecture, Building and Planning
EWI	Electricla Engineering, Mathematics and Computer Science
ID	Industrial Design
ST+TN	Chemical engineeruing, Chemistry and Applied Sciences
TM	Technology Management
WBMT	Mechanical Engineering & Biomedical Engineering

University of Twente

BK	Public Administration and Public Policy
CTW	Engineering Technology
EWI	Electrical Engineering, Mathematics and Computer Science
GW	Behavioural Sciences
TM	Business Administration
TNW	Science & Technology

To determine the overlap we used the research profile of each individual faculty. These profiles were compared and similarity was determined. The maps were created using the similarity of the faculties' research (or cognitive) profile. The profile were created on the basis of the keywords attached to the ISI publications as well as on the keywords resulting from a sophisticated analysis of the titles of the non-ISI publications. For these additional publications we determined per first-author-oeuvre, i.e., the set of publications with the same first author, the most relevant keywords on the basis of the TF/IDF measure (Salton & McGill, 1983).

The combination of keywords identified for the ISI publications on the one hand and the keywords from the additional papers on the other, allowed us to create the profile using multi-language keywords. This was considered important because a significant amount of authors published their work in Dutch and German and outside the scope of the Web-of-Science.The profiles of the individual faculties were input a similarity analysis in order to define the amount of overlap. The similarity was calculated using the Cosine vector.

$$sim(x,y) = \frac{\sum_i (x_i y_i)}{\sqrt{((\sum_i x_i^2)(\sum_i y_i^2))}}$$

This measure ranges between 0 and 1, where zero indicates that two profiles are completely different while 1 indicates that they are exactly the same, in the sense that two faculties focus on exactly the

same topics. In other words, we were able to determine the research similarity between two faculties based on all their publications. This similarity would be characterized with a measure between 0 and 1. In addition to this profile similarity study, we mapped the collaboration of the faculties in order to investigate the existing communication. We consider the collaboration results important to enhance the profile information. It reveals a more complete picture of the actual situation and the potential to join forces. Both types of maps (cognitive orientation and collaboration) were created with a software package UCINET to analyse social networks (Wassermann and Faust, 1994 and Borgatti et al., 2002).

Results

The landscape we created was based on the entire output 2000-2004. We used a threshold of 0.11 to represent the similarity between the 21 faculties involved. With this threshold, we covered around 55% of the possible similarities. This threshold is relatively low, also because we wanted as much as possible faculties to be represented in the network map. The map below thus shows the similarity of these faculties in terms of profile overlap. The thickness of the connecting lines indicates the similarity as measured by the cosine (0.11<threshold <0.6).

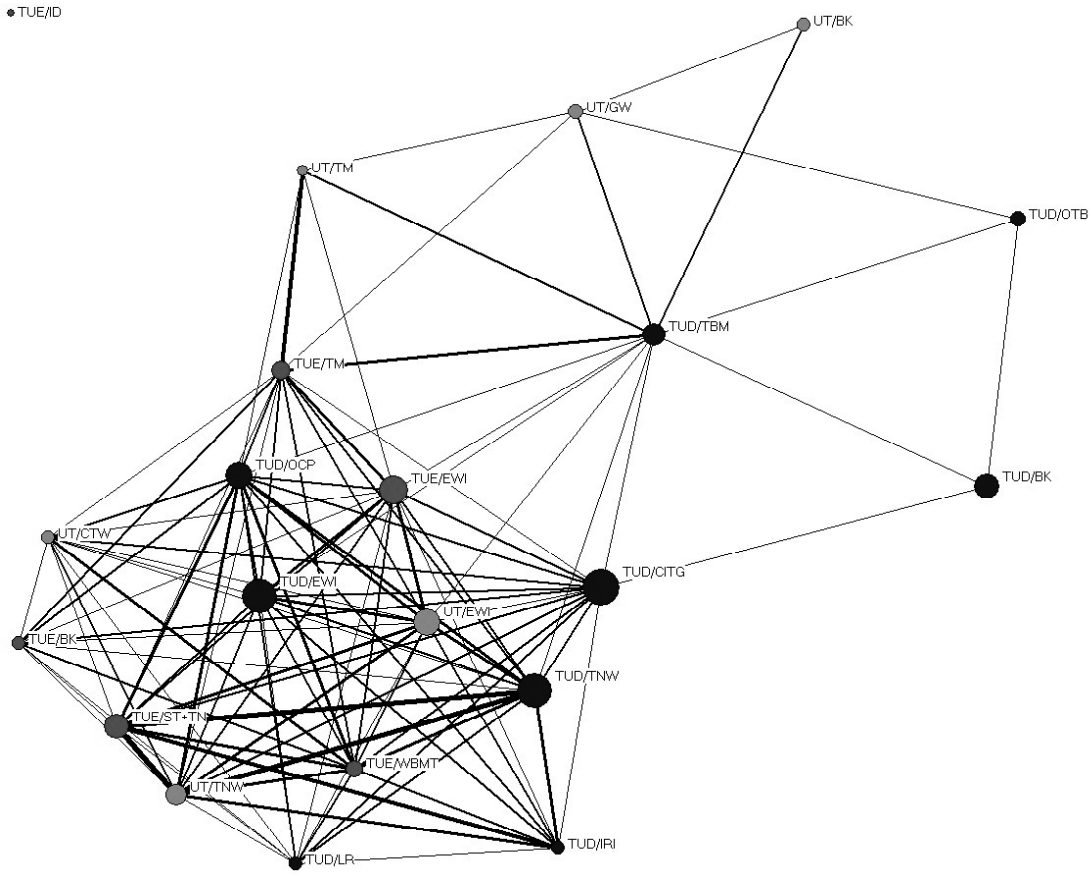


Figure 1. Thematic orientation of 21 faculties of the three universities of technology (2000-2004). The size of the circles represent the amount of publications of a faculty, between 200 and 5000. The colors indicate the origin (TUD: blue, TUE: green and UT: orange). The thickness of the connecting lines represents the measure of similarity en (between 0.11 and 0.6). The positions are defined by all similarities above 0.11.

The map indicates that there is a coherence of research done the three universities. There is only one faculty (in Eindhoven) that does not meet the similarity threshold of 0.11 with any of the others. It was well recognized that the research at Industrial Design, does not match the research at the other faculties. In that sense, the research at TUE/ID remains outside the core research at 3TU.Furthermore, the map shows that TUE/TM, TUD/TBM and UT/TM function

as a sort of bridge between the 'hard' science represented in the lower area of the map and the 'softer' research as done at the faculties in the upper part of the map.

At the area with the 'hard' faculties, there is a clear coherent research orientation, as shown by the many connecting lines between the different faculties, but there are only a few that have a strong similarity. Only the three core faculties, TUD/TNW, UT/TNW and TUE/STTN show a similarity of > 0.6. But as these faculties represent the basic research of each of the universities, we should expect such an 'overlap'.

If we accept the faculty as the aggregation level to investigate the overlap, what are the conclusions we can draw from this map with respect to the overlap issue? In other words: is a similarity of 0.6 much? To investigate this, we compared in more detail the contents of the research profiles of TUD/TNW, UT/TNW and TUE/STTN. We compiled an overview of most prominent topics in publications by these three universities. In the figure below, we listed the keywords most often used by these faculties. The proportion of publications for each of these three is indicated by a colored bar in the chart. Moreover, we drew a line at the overall proportion of the individual faculties as compared to the entire set of publications of these three. These lines thus indicate the expected value of a faculty for each of the topics listed. Each topic is assigned to a faculty if its proportion is significantly above its expected value. The topics are grouped depending on the universities they are assigned to.

In the profile overview, we see a clear distribution of related topics over faculties, or actually over universities. Most topics are shared by all three universities but it seems there are always at least two universities with an increased interest. There are only a few topics with a dominance of one university only. This would explain, from a bibliometric perspective, the increased similarity of their profiles. On the other hand, if we look at the clusters of topics assigned to one or two universities, being the ones with a significant focus, we can see that these clusters of topics are associated to each other: yeast research in Delft, polymers in Eindhoven.In addition, we compiled a similar list for faculties that have a less similar research profile. In the case of TUD/TM, TUE/TBM and UT/TM, we clearly see a different picture.

In this profile overview, we see a dominance of colors in separate parts of the chart. This indicates that there is a much clearer distribution of topics over universities, a more separated research focus, which would explain that their similarity is much lower than in the first example.

And because at this level of similarity (~0.5) the topics differ already so much from each other, we can conclude that the overall picture reveals no alarming overlap between any of the faculties.But if their research profiles differ so much, is there still a proper basis for a federation. In order to investigate that, we created the social network graphs to illustrate collaboration between the faculties.We were able to derive from the information of double publications from different faculties. This approach enabled us to identify collaboration both with ISI and non-ISI publications. In general, the information in METIS does not contain the addresses of the authors involved and as such, does not allow a proper assessment of collaboration by measuring co-authorships.

First of all the map shows that there is a strong network of collaboration between the faculties in general. There is no faculty that remains outside the network and the density is relative high. Furthermore, the map shows that the collaboration tends to take place within the same university. In particular in Twente the collaboration remains primarily on campus.Finally, we can see that Eindhoven seems to play a bridging role between Delft and Twente. The latter two have very little collaboration with each other. The lack of cognitive overlap on the one hand and the density of collaboration on the other, does also support each other in the sense that if the faculties would be doing exactly the same, they probably would not collaborate but rather act as each other's competitors.

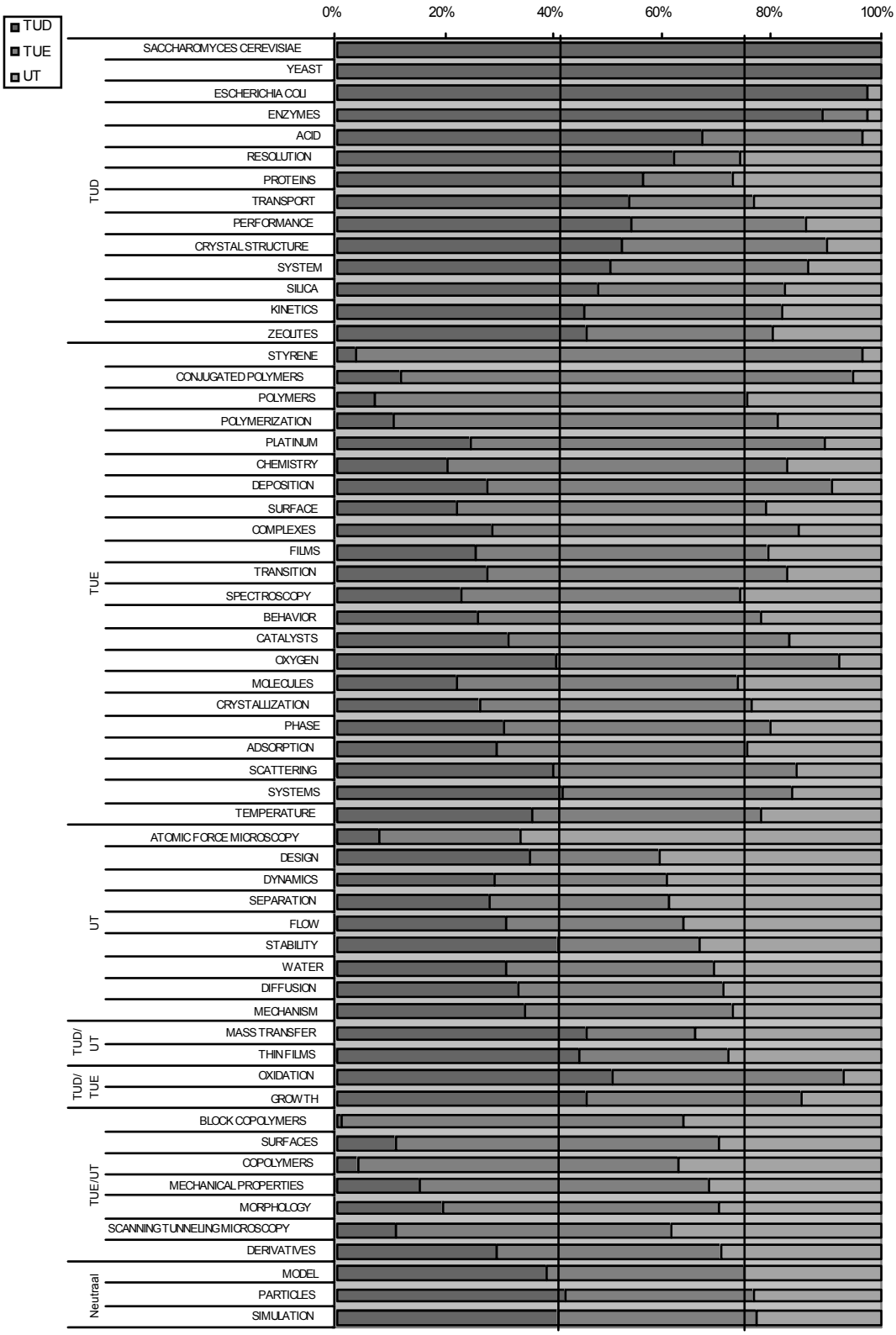


Figure 2. Comparison of the research profile of three faculties based on the most prominent topics in their publications (2000-2004)

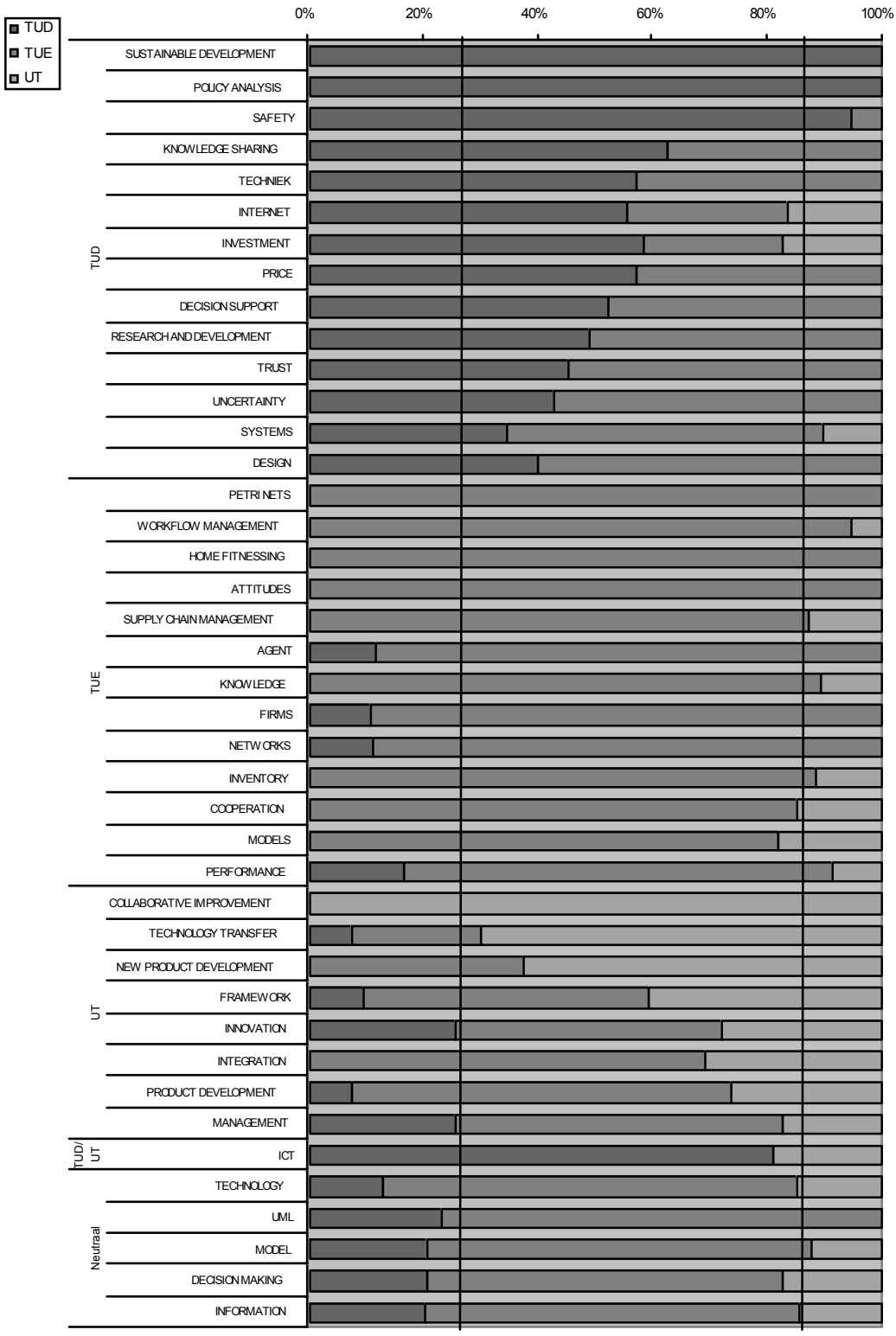


Figure 3. Comparison of the research profile of three Technology Management related faculties based on the most prominent topics in their publications (2000-2004)

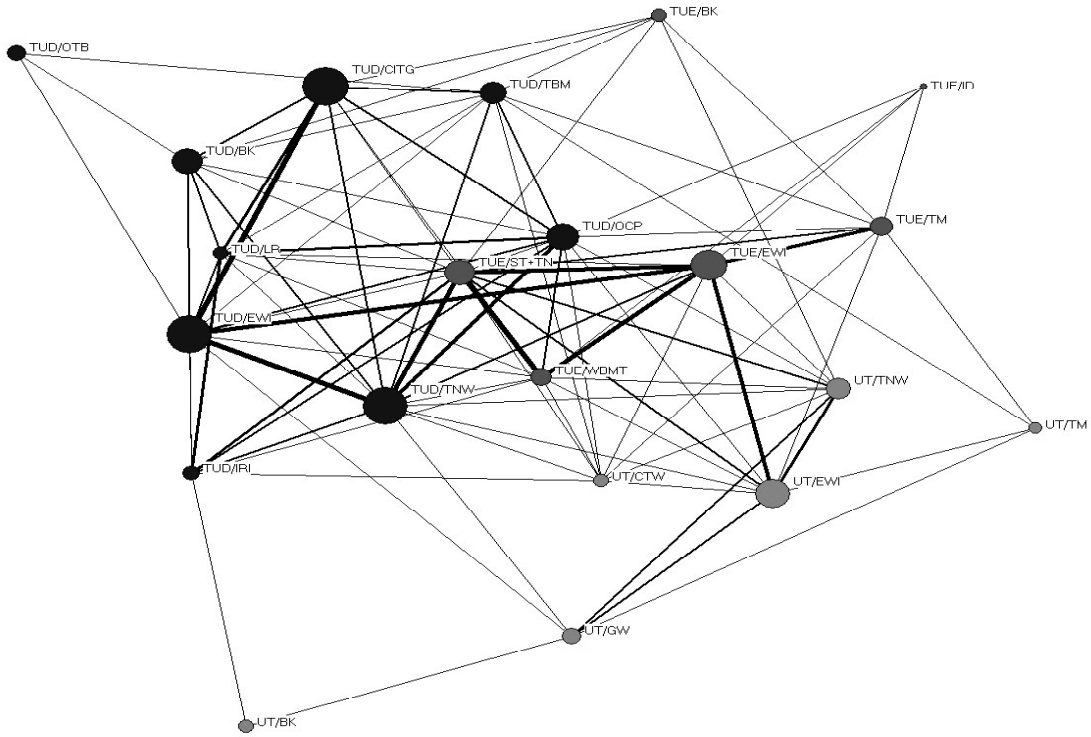


Figure 4. Collaboration between 21 faculties of the three universities of technology (2000-2004). The size of the circles represent the production of the faculties (number of papers). The colors refer to theoriginating university (see Figure 1). The thichness of the connecting lines indicate the amount of co-authored papers between 1 and 50

In other words the limited overlap and strong collaboration network, indicate a sound basis for an increased integration using the individual strengths within the context of a federation as far as research is concerned.

Conclusions

In this paper, we showed an applied mapping methodology to address a high level policy issue. By using bibliometric mapping but not exposing it in the traditional way, we were able to use the sophisticated but complicated results of a cognitive profile analysis to address a set of policy related issues regarding the initiation of a federation of the three universities of technology in the Netherlands. The conclusions of the bibliometric analysis are presented in a network map with elements relevant and comprehensive for the policy makers to interpret. The faculties refer to entities of the 'real world' and as such subjects of research to determine overlap. The publication keywords, the topics, are a means to determine the amount of overlap and as such not relevant primarily to be part of the results. We found a way to conceal the complexity of the cognitive maps and still use them to investigate the research profiles of the three universities of technology in the Netherlands.

Moreover by a comprehensive overview of the research profiles of individual faculties we were able to illustrate the background of the conclusive results as found by the profile analysis.

References

Borgatti, S. P., Everett, M. G. & Freeman, L. C. (2002), *Ucinet for Windows: Software for Social Network Analysis*. Analytic Technologies, Harvard.

Noyons, ECM (1999). *Bibliometric mapping as a science policy and research management tool*. DSWO Press, Leiden. 1999.

Salton, G. and McGill, M. J. 1983 *Introduction to modern information retrieval*. McGraw-Hill, ISBN 0070544840.

Wasserman, S., and Faust K. (1994), *Social Network Analysis*. Cambridge University Press, Cambridge.

Linear Analysis of Cybermetric Data: Quantifying The European University Web Space

José Luis Ortega and Isidro Aguillo

{jortega; isidro}@cindoc.csic.es
Cybermetrics Research Group, CINDOC-CSIC, Joaquín Costa, 22. 28002 Madrid. Spain

Abstract
The aim of this communication is to analyze the European academic web space using the linear regression. The objective is to describe and understand how each country receive links from others and to explain the relationships and characteristics of the whole European university web space. Several linear regression models were applied to webometric data to show the relationships between several variables. The results show that there are two variables that explain the attraction of European links: number of web pages, and language of these pages. Pages that use the English language are more likely to get links than the non-English ones, because this language is the lingua franca in the European academic web space.

Keywords
webometrics; cybermetrics; linear regression; european universities.

Introduction

Jeong, Neda and Barabasi (2003) argue that there is a preferential attachment phenomenon when it comes to gain links in evolving free-scale networks as the Web. Hence, the most connected nodes are more likely to get links than nodes less connected. However, Pennock, Flake, Lawrence, Glover, and Giles (2002) showed that this preferential attachment behavior changes according to different content categories in the Web. Payne and Thelwall (2005) think that there is a second law that will be able to explain the link attraction. Thelwall's group has found positive correlations between research measures and inlink counts in United Kingdom. In other study (Cothey, 2005) it was found that the web contents are the main reason to attract links from the European Union Research Area (EURA).
However, the scalar properties of the Web cause several problems to determine and estimate scientific indicators. Payne and Thelwall (2005) found that there was a linear relationship between web site size and research productivity. But in the other hand, Katz (2000) and Katz and Cothey (2006) have observed that many scientometric and webometric variables follow a power law relationship so it is necessary to develop scale-independent indicators to measure these relationships.

Objectives

The aim of this work is to study the relationship between several web variables in order to answer the following questions:
Is it possible to apply the linear regression models to webometric data?
What variables can estimate and explain the amount of incoming links that a country receives?
Can the linear regression model explain the European Academic web space relationships?

Methodology

535 universities of 14 European countries (EU except Luxembourg) in 2004 were selected from *Webometrics Ranking of World Universities* (www.webometrics.org) This site ranks 3,000 universities according to two main criteria: size (number of pages and rich files) and visibility (number of incoming links) (Aguillo, Granadino, Ortega, & Prieto, 2005). *Yahoo! Search* was used to obtain the outcoming links data with the following query (Aguillo, Granadino, Ortega, & Prieto, 2006):
+site:{university domainA} +linkdomain:{university domainB}
and *Google* was used to retrieve the total number of pages of each university:
+site: {university domain}
These data were obtained in August of 2005 and they were aggregated at the country level.

Being the Web a scale-free network, where a few nodes attract or make a huge amount of links, and the majority just attract or receive a few ones. The distribution of out- and in-links follows a power

law behavior. Hence, the relationships between variables are scalar and not linear (Katz, 2000; Katz et al., 2006). A power law relationship between two scalar variables *x* and *y* can them be represented as:

$$y = ax^k$$

where *a* (the constant of proportionality) and *k* (the exponent of the power law) are constants. Power laws can be seen as a straight line on a log-log graph since, transforming the variables to a logarithmic scale converts the above equation into:

$$\log(y) = k\log(x) + \log(a)$$

which has the same form as the equation for a line.

$$y = kx + a$$

Thus, the scalar relationships between variables, when transformed to a logarithmic scale, can be represented through a linear regression. This transformations determine to the coefficients to be interpreted in a percentage terms. A scatter plot of the residual values and predicted values was carried out (Figure 1) to test the normality assumption. Graphs show that there is no relationship, so there is no residual information on the transformed variables. To develop these statistical analyses MS Excel 2000 and SPSS 13.0 were used.

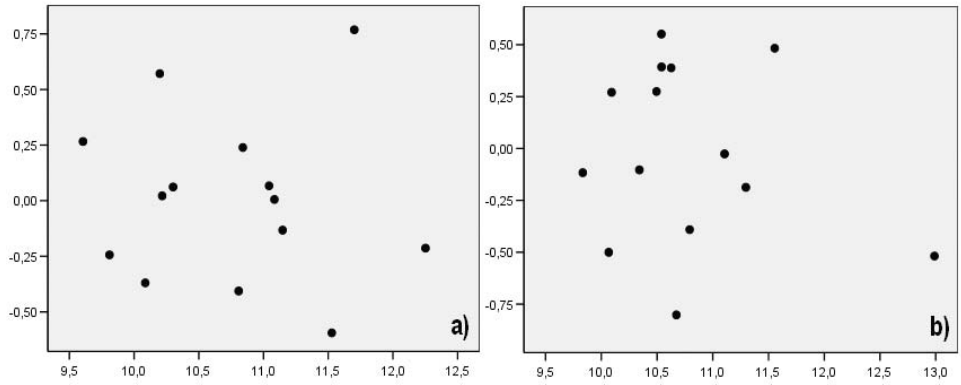


Figure 1. scatter plot of the residual values and predicted values.

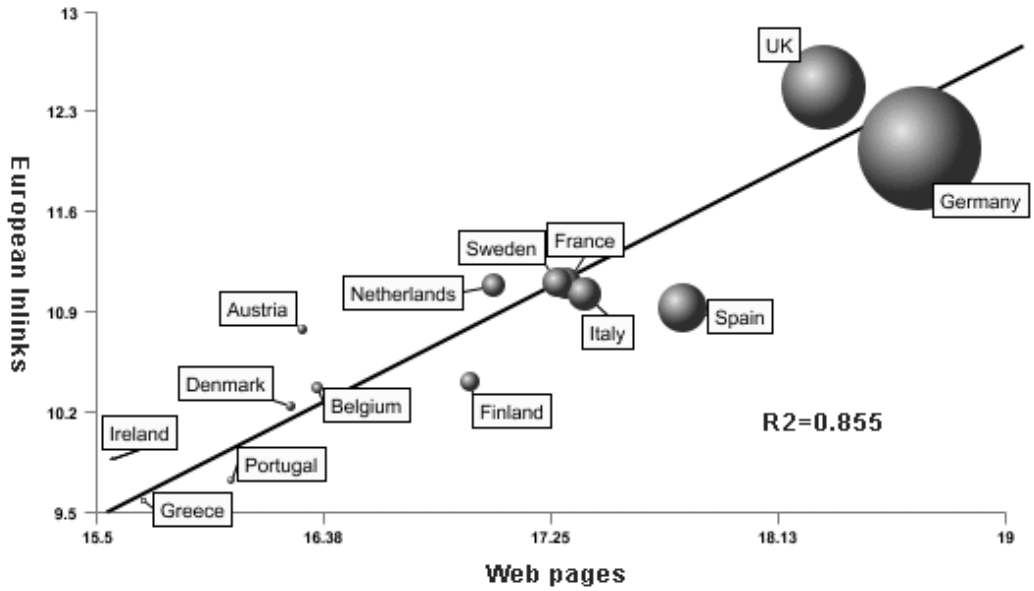


Figure 2. European inlinks and pages Scatter plot (log. scale).

Table 1. European Inlinks and Pages Regression analysis.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.925[a]	.855	.843	.32703

a. Predictors: (Constant), Pages_log

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-2.939	1.631		-1.801	0.096
	Pages_log	.809	.096	.924	8.409	.000

a. Dependent Variable: European_Inlinks_log

The data in Figure 2 and Table 1 show the relationship between the number of pages in each country and the number of incoming links that they receive. There is a good linear fit (R^2=0.855) between the inlinks and the number of pages. Table 2 also shows a good significance of the response variable (t=8.41). Thus, we can estimate that a 10% perceptual increase of web pages causes a 8.1% growth of European inlinks. This let us to argue that the more amount of pages a country makes available on the web, the more links receives from other countries. Nevertheless, there are still countries that are above or below the linear fit predicted. Austria attracts a disproportionate number of inlinks in proportion to the pages that it hosts. Perhaps, because the Austrian network is strongly joined to the German network and many of the inlinks come from Germany. This might also be the case of Ireland according to UK (Ortega, Aguillo, Cothey, & Scharnhorst, 2007). However, other countries with a medium size web space such as Netherlands, Sweden or Belgium show more incoming links than the linear adjustment, and countries with a large size web space such as Germany and Spain show less inlinks that they should to have. So it is possible that there is another variable that explains the number of inlinks that a country receives. Hence, a second question arises. Is there any relationship between the incoming links and the amount of published pages in English language? Perhaps, if we consider the English language as an international *lingua franca*, it would cause the attraction of links from abroad because these contents are more comprehensible for a wide international community.

Table 2. Total pages and English pages by country.

	Total Pages	**Pages In English**	**%**
UK	88,783,300	85,000,776	95.74
Ireland	5,742,000	4,656,000	81.09
Sweden	31,618,000	11,807,167	37.34
Netherlands	24,997,700	9,261,499	37.05
Austria	11,925,000	4,064,750	34.09
Greece	6,448,390	1,936,655	30.03
Denmark	11,359,800	3,133,250	27.58
Finland	22,677,000	5,755,441	25.38
Belgium	12,668,000	2,350,900	18.56
Portugal	9,042,000	1,506,850	16.67
Germany	127,821,190	19,747,708	15.45
Italy	35,225,400	4,935,679	14.01
France	33,080,700	4,446,066	13.44
Spain	51,796,200	4,380,785	8.46
Total	473,184.680	162,983,528	34.44

Table 2 shows the percentage of web pages in English over the total amount of published pages on the Web in each country. These data were obtained from *Google* search engine through the *Advanced Search* page in August of 2005. A computing routine was designed to retrieve the total number of web pages publicly available and the number of web pages only in English from the 535 European universities. Table 2 obviously shows a high percentage of English pages for UK (95.74%) and Ireland (81.09%), followed by Sweden (37.34%) and the Netherlands (37.05%).

Figure 3 shows a linear fit between the amount of incoming links and the number of English pages in each country. Table 3 shows the Regression analysis between both variables, obtaining a good significance of the response variable (t=6.98). The resulting analysis lets us to estimate that a 10% perceptual increase of web pages causes a 8.1% growth of European inlinks a 10% increase of web pages in English would produce a 7.12% growth of European inlinks. With these results we can argue that countries with a high production of English pages attract more European inlinks, and this is because the English is considered an international language in academia.

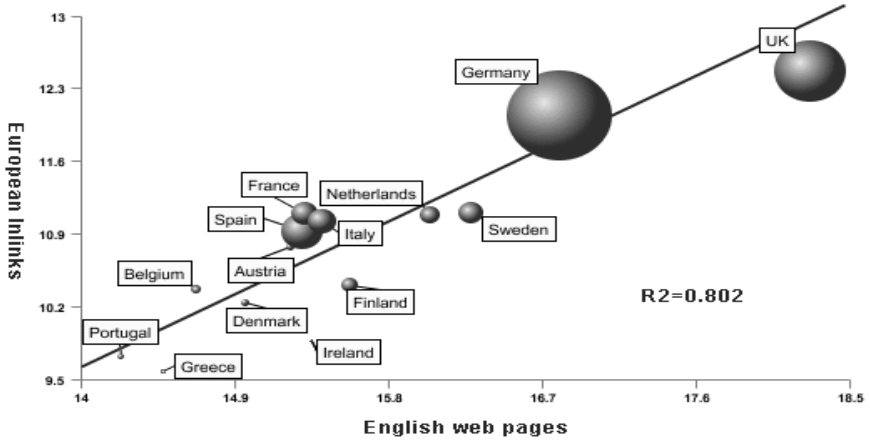


Figure 3. European Inlinks and English pages Scatter plot (log. scale).

Table 3. European Inlinks and English pages Regression analysis.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.896[a]	.802	.786	.38170

a. Predictors: (Constant), English_Pages_log

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-.321	1.591		-.202	.844
	English_ Pages_log	.712	.102	.896	6.980	.000

a. Dependent Variable: European_Inlinks_log

Discussion

Regression Analysis has been a suitable statistical tool to achieve the description of the variables that can explain the inlinks that a country attracts. Mainly, two variables can explain the web visibility: number of web pages and the number of that pages are in English. Unlike another studies, we not only found these relationships, but also we have explained to what extent the inlinks are due to the web pages and English pages production and we have estimated these relationships. Thus, a 10%

perceptual increase of web pages causes a 8.1% growth of European inlinks and an increase of 10% of web pages in English would produce a 7.12% growth of European inlinks.
However, Thelwall, Tang and Price (2003) found that the English language pages contained more international European links than any other language, but they do not prove mathematically this relationship. Thelwall (2002) also applied a multiple regression model to explain the relationship between inlinks and academic indicators, but he does not explain the relationship with web pages. Nevertheless, Thelwall and Harries (2004) found relationship between inlinks and the number of web pages of web sites with a high RAE assessment, although they do not quantify this relationship.
Two separated bivariate regressions were used because the English pages variable are included into web pages variable. Hence, both variables cause colinearity and they were separately studied.

Conclusion

Our results let us to state that the publication of web content is a key factor to obtain links from other countries and, thus, to increase the visibility in the Web. The more pages that a country hosts on the Web more links from other countries will point to this country, simply, because there are contents to link. This results fit in with the previous results of Cothey's (2005) characterization of the EURA.
Thus, many of that inlinks come from English language web pages as well. This allows us to confirm the English as *lingua franca* of the European academic web space. The use of this language increases the visibility of one country and many of the links between countries come from and point to English language web pages.
Nevertheless, the weight of the English language in the European academic web space favors the English native countries. This is the case of Ireland which receives more European inlink than the expected and its visibility is overestimated. This also happen in less extent with UK which is the country with most European inlink but it has less contents than Germany. On the contrary, countries with less use of English have less visibility though they have more web pages such as Germany and Spain (Figure 2).

References

Aguillo, I.F., Granadino, B., Ortega, J.L. & Prieto, J.A. (2005). What the Internet Says About Science. *Scientist*, 19(14), 10

Cothey, V. (2005). Some preliminary results from a link-crawl of the European Union Research Area Web. In: P. Ingwersen & B. Larsen (Eds.), *Proceeding of the 10th International Conference of the International Society for Scientometrics and Informetrics*.

Jeong, H., Neda, Z. & Barabasi, A.L. (2003). Measuring preferential attachment in evolving networks. *Europhysics Letters*, 61(4), 567-572.

Katz, J.S. & Cothey, V. (2006). Web indicators for complex innovation systems. *Research Evaluation*, 15(2), 85-95.

Katz, S. (2000). Scale Independent Indicators and Research Evaluation. *Science & Public Policy*, 27(1), 23-36.

Payne, N. & Thelwall, M. (2005). Mathematical Models for Academic Webs: Linear Relationship or Non-Linear Power Law? *Information Processing & Management*, 41(6), 1495-1510.

Pennock, D.M., Flake, G.W., Lawrence, S., Glover, E.J. & Giles, C.L. (2002). Winners Don't Take All: Characterizing the Competition for Links on the Web. *Proceedings of the National Academy of Sciences of the United States of America*, 99(8), 5207-5211.

Thelwall, M. (2001). Extracting macroscopic information from Web links. *Journal of the American Society for Information Science and Technology*, 52(13), 1157-1168.

Thelwall, M. (2002). A research and institutional size based model for national university web site interlinking. *Journal of Documentation*, 58(6), 683-

Thelwall, M. (2002). Conceptualizing documentation on the Web: An evaluation of different heuristic-based models for counting links between university Web sites. *Journal of the American Society for Information Science and Technology*, 53(12), 995-1005.

Thelwall, M., & Harries, G. (2004). Do better scholars' web publications have significantly higher online impact? *Journal of the American Society for Information Science and Technology*, 55(2), 149-159.

Thelwall, M., Tang, R. & Price, L. (2003). Linguistic Patterns of Academic Web Use in Western Europe. *Scientometrics*, 56(3), 417-432.

Hypergraph Modelling and Graph Clustering Process Applied to Co-word Analysis

Xavier Polanco[*] and Eric Sanjuan [**]

[*]*xavier.polanco@lip6.fr;@inist.fr*
LIP6, Université Pierre et Marie Curie, 104 avenue du Pte. Kennedy, 75016 Paris (France)
SRDI, INIST- CNRS, 2 allée du Parc de Brabois, 54514 Vandoeuvre-lès-Nancy (France)

[**]*eric.sanjuan@iniv-avignon.fr*
LIA, Université d'Avignon, 339 chemins des Meinajaries, 84911Avigon (France)

Abstract
We argue that any document set can be modelled as a hypergraph, and we apply a graph clustering process as a way of analysis. A variant of the single link clustering is presented, and we argue that it is better suited to extract interesting clusters formed along easily interpretable paths of associated items than algorithms based on detecting high density regions. We propose a methodology that involves the extraction of similarity graphs from the indexed-dataset represented as a hypergraph. The mining of informative short paths or geodesics in the graphs follows a graph reduction process. An application for testing this methodology is briefly exposed. We close this paper indicating the future work.

Keywords
co-word analysis; graph clustering; hypergraph modelling

Introduction

The main objective of this paper is to propose a graph clustering methodology including hypergraph modelling of raw data. The position that we support here can be summarized in these three proposals: [1] the hypergraph modelling constitutes an uniform data model that allows us to generalize the method of co-word analysis as we try in this paper, and it can be compared to Galois lattice modelling used by more symbolic methods as formal concept analysis (FCA) (Wille, 1985). This last point we will not treat it here. [2] On the basis of what has been observed in co-word analysis (Courtial, 1990), short paths of strong associations can reveal potential new connections between separated sectors of the word network. Thus, the graph clustering method that we propose does not focus on homogeneous clusters, but highlight some heterogeneous clusters formed along short path of strong associations. [3] Concerning the graph clustering, we support that a variant of the single link clustering (SLC) is better suited to build interesting clusters, formed along easily interpretable paths of associated items, than algorithms based on detecting high density regions.

Section 2 introduces the hypergraph modelling, sections 3 and 4 deal with the graph reduction and clustering process. A real application on SCI dataset try, in section 5, to provide the empirical demonstration of the approach exposed in the former sections.

Hypergraph Modelling

Hypergraph definition (Berge, 1987): A hypergraph is a pair (X, E) where X is a set of elements, called hypernodes or hypervertices, and E is a set of hyperedges; a hyperedge e is a non empty subset of X, $S(X)\backslash\emptyset$. While graph edges are pairs of nodes, hyperedges are arbitrary set of nodes or vertices, because in a hypergraph the edges can connect any number of vertices. A hypergraph is a family of sets drawn from the set X.

Any set of documents can be modelled as a hypergraph H. Each of the documents can be represented by a hypervertex, and the document elements as for instance authors, keywords, and citations are represented each one by the subset of documents sharing these elements. These subsets constitute the hyperedges of the hypergraph. In the sequel, we shall identify each keyword w with the subset of documents $\{d_1, \ldots, d_p\}$ indexed by w. According to the type of information that we want to analyse, intersection graphs that we shall call co-occurrence graphs, can be derived from different subparts of

H. The vertices are the selected hyperedges meanwhile an edge is drawn among two vertices (v_i,v_j) whenever they have a non empty intersection. For example, if we select as subpart of *H*, the hypervertex representing keywords or authors or citations, the resulting intersection graph is the co-keyword or co-author or co-citation graph. If we select both author and keyword hyperedges, we obtain a graph of associations between keywords and authors; if we select both citation and keyword hyperedges, we obtain a graph of associations among keywords and citations.

Graph reduction process

The next consists of applying a graph clustering process from the co-occurrence data matrix, where the data can be keywords, authors or citations. Here, we limit to treat the co-word matrix. This is the input matrix of the graph of co-occurrences $Go(V,E)$.

Many similarities between keywords can be defined based on the cardinality of $w_i \cap w_j$ (Van Cutsem, 1994), in the present case the intersection graph $Go(V,E)$ of *H* defines the co-word graph having as many edges as there are non null values in the similarity matrix. The set *E* of edges of $Go(V,E)$ is the set of pairs of keywords $\{w_i,w_j\}$ where $w_i \cap w_j \neq \emptyset$. But the associations between keywords cannot be considered in a crisp binary way. The co-occurrence frequency alone is not enough to measure the strength of associations, because it favours high-frequency couples compared to those with low frequency. Then, it is necessary to use some normalizing coefficient. For this task, we apply the "equivalent coefficient" (as originally defined in Michelet, 1988) based on the product of conditional probabilities of appearance of a term knowing the presence of the other one. The equivalence coefficient that we note $\sigma(w_i,w_j) = |w_i \cap w_j|^2 / (|w_i| \times |w_j|)$ allows to avoid weak relations and to normalise frequency of keywords. This coefficient also has an easy interpretation in terms of probability theory, since it is the product of conditional probabilities of finding one item knowing the presence of the other. This coefficient is maximized by pairs of items that are in the same closed sets. We denote $G(A) = (V,E,\sigma)$ the weighted graph of associations. Usually, when a co-occurrence matrix is used, a threshold is set on the keyword frequency in order to obtain a less sparse matrix. On the other hand, setting also a threshold on association value $\sigma(w_i,w_j)$ and not only on keyword frequency $|w_i|$ is better suited to form clusters that are closed. Consequently, every value *s* in]0,1[induces a sub-graph $G(A>s) = (V,E_s,\sigma)$ where E_s is the set of pairs of vertices (w_i,w_j) such that $\sigma(i,j) > s$.

Graph clustering

We apply a variant algorithm of the single link clustering (SLC) called CPCL (Classification by Preferential Clustered Link), originally introduced in Ibekwe-SanJuan (1998), here we use its optimised version (SanJuan et al, 2005). This algorithm is applied on the graph $G(A>s)$ building clusters of keywords related by geodesic paths that are constituted of relatively strong associations. Any variant of single link clustering that reduces its chain effect can produce interesting results in this context, since they naturally form clusters along short geodesics of maximal weight. The CPCL algorithm merges iteratively clusters of keywords related by an association strongest than any other in the external neighbourhood. In other words, CPCL works on local maximal edges instead of absolute maximal values like the standard SLC. We refer the reader to Berry et al. (2004) for a detailed description of the algorithm in the graph formalism.

Experimentation

We test this graph clustering process on a corpus of 5,795 bibliographic data on "data and text mining", extracted from the SCI database, over period 2000-2006. The different components of the bibliographic data are catalogued according to 39 fields coded by a combination of two capitals as specified by SCI. These 5,795 data characterized by these 39 fields are submitted to hypergraph modelling. The hypergraph, denoted *H*, is store on the form of a ternary relation i.e. an identification number, a code field and a keyword. The coded fields form the vertices of the hypergraph *H*. From this general hypergraph with 165,850 vertices and 5,795 edges, by selecting a subset of vertices a sub-hypergraph is extracted. In this case we select the set of keywords (SCI field denoted DE). The result is a sub-hypergraph *K* with 8,040 vertice and 3,171 edges, the vertices are the author keywords. The degree of hyperedge corresponds to the number of keywords indexing a reference. Its minimal value is 1, the mean 5.11 and the maximal 35. The degree of vertices is the number of references indexed by a

keyword. On this corpus, its minimal value is 2, its mean 6.35 and its maximal value 1,722 for keyword "data mining". The frequency of all other keywords is lower than 200.

The hypergraph modelling allows introducing a new data analysis instrument: the minimal transversal. Given a subset S of vertices, where $H(S)$ is the set of hyperedges that intersect S, the subsets S of vertices such that for any subset P strictly included in S we have $H(P)$ strictly included in $H(S)$ are the minimal traversals. Consequently, the minimal transversals correspond to closed item-sets in data mining (Zaki, 2004), and to formal concepts in Formal Concept Analysis (Ganter et al, 2005; 1989). The number of hyperedges in which S is included is the support of S. On K there are 2,526 minimal transversals with at most 5 elements and a support greater that 0.001. Table 1 gives some examples of minimal traversals with 5 elements having a largest support.

Table 1. Minimal traversals of the hypergraph K.

Support					
0.00421	bioinformatics	cancer	data mining	genomics	proteomics
0.00316	genomics	machine learning	microarray	proteomics	text mining
0.00316	data mining	dimensionality reduction	feature extraction	neural network	pattern recognition
0.00316	clustering	machine learning	microarray	proteomics	text mining
0.00316	bioinformatics	genomics	machine learning	proteomics	text mining
...	…	…	…	…	…

Computing such sets allow detecting subsets of frequently associated items. However, computing all minimal transversal is intractable since the number of such sets can be exponential on the number of items, because the numerous overlaps between minimal transversal.

The graph of weighted associations $G(A)$ constitutes the intersection graph of the hypergraph K with 8,037 vertices and 34,375 weighted edges. The $G(A)$ graph is small world graph (SWG) since its average clustering measure is 0.47. This value is far from the expected value for a random graph having the same average degree which is the average degree over the number of edges 4.43/2,335. As in random graphs, the average path length is low 2.31. These are the two conditions usually considered to characterise SWGs (Watts, 1999). The SWGs are compact graphs with a high number of simplicial vertices that are vertices whose neighborhood forms a complete graph. Setting a low threshold on association values (0.001 here) is enough to drastically reduce the number of edges and reduces its clustering measure (0.35). However, we do not loose the SW property since 0.35 is much greater than the mean degree 3.28 over the 1,057 number of edges, and the average path length is low 4.12.

The graph $G(A>s)$ is induced from $G(A)$ fixing the threshold s at 0,001, and the co-occurrence frequency at 2. The graph $G(A>0.001)$ has 447 vertices and 1,017 edges, and it is easier to handle. $G(A>0.001)$ presents a central dense connected component of 187 vertices and 369 edges. All other components are small (less than 10% of the total number of vertices). We focus our experiment on this component to observe if it allows detecting complex associations as the one pointed out by minimal transversals. Applying the CPCL algorithm on $G(A>0.001)$ we obtain a clustered graph, G(CPCL), with 187 vertices and 738 edges, each vertex is a cluster. By definition, CPCL output tries to highlight a disjoint family of clusters formed on geodesics of relatively high associations. We have experimentally checked that most of these clusters are minimal transversals. The biggest one has 10 keywords, and the mean size is 4. 76% of these clusters are minimal transversals of the hypergraph K. Table 2 shows three examples of clusters of medium size which are minimal transversal.

Table 2. Clusters in *G*(CPCL) graph

Label	vertex 1	vertex 2	vertex 3	vertex 4
Ontology	thesaurus	query processing	ontology	knowledge mining
OLAP	rule bases reasoning	OLAP	data interchange	XML
Intrusion detection	computer security	intrusion detection	anomaly detection	user profiling

They are labeled by the vertex having in the graph the highest betweenness valued, and the elements are enumerated following the geodesic that cross the edges with a highest value. This is the way that the clusters are extracted.

A numerical notation of the keywords (i.e. co-words) of the graph *G*(*A*>0.001), and also of the clusters of the graph *G*(CPCL), gives an idea of their positions in the structure of the graphs. The structural properties considered are centrality, density, betweenness, degree, and weighted betweenness. Centrality (Centrl) gives the sum of association values involving the vertex, density (Dens) gives the number of edges in the neighborhood over the maximal theoretic number, betweenness (Betw) gives the number of geodesics crossing the vertex, degree gives the number of vertices in the neighborhood, and weighted betweenness (w_Betw) gives the number of geodesics that maximize the sum of association values. Tables 3 and 4 summarize the first five items ordered by decreasing betweenness (Betw). Keywords having the higher betweenness centrality point out keywords shared by the longest minimal transversals showed in table 1. This observation deserves certainly a thorough analysis.

Table 3. Keyword vertices in *G*(*A*>0.001) graph

Vertex	Centrl	Dens	Betw	Degree	w_Betw
text mining	0.07	0.01	14,573	44	11,264
machine learning	0.08	0.03	10,243	30	12,664
bioinformatics	0.13	0.03	8,113	34	3,549
classification	0.1	0.03	7,376	23	4,377
pattern recognition	0.22	0.05	7,286	22	7,111
...	…	…	…	…	…

Table 4 Cluster vertices in *G*(CPCL)

Vertex	Centrl.	Dens.	Betw.	Degre	w_Betw
machine learning	0.035	0.003	7,224	35	7,845
text mining	0.055	0.005	4,854	26	3,867
pattern recognition	0.196	0.064	3,141	18	3,069
classification	0.059	0.002	2,902	17	1,318
bioinformatic	0.083	0.020	2,607	16	1,370
...	…	…	…	…	…

The graphs *G*(*A*>*s*) and *G*(CPCL) constitute data analysis levels. For example, using the AiSee interactive interface (http://www.aisee.com), we can visualized the graph of clusters *G*(CPCL) that reveal the keywords that have the highest score of betweenness centrality since they are used as cluster labels. Opening the clusters, we access to the main pair of non central concepts related by geodesic paths that cross the label of the clusters suggesting potential new interactions between concepts. The higher betwenness and degree values (see tables 4 and 5) allow characterizing hubs and crossroad vertices in the graph structure. In the Appendix, by way of example two excerpts of *G*(CPCL) are presented where clusters within circles -text mining, bioinformatics, proteomics, machine learning, and classification- are hubs and crossroad at the same time in the global structure of the clustered graph. In graphs *G*(*A*>*s* = 0.001) and *G*(CPCL), the data and text mining domain appears for the period observed strongly related to bioinformatics.

Conclusion and Future Work

In this article we have proposed a graph clustering methodology for mining the topic structure of an unordered set of textual data, which also can be browsed. The aim is to discover useful knowledge from an unordered dataset, i.e. useful for watching a research domain. We believe that the hypergraph modelling and the introduction of minimal traversal in the data analysis constitute something new in the use of the graph theory in informetric studies. The studies which we know concentrate principally on citations and then working with directed graph models, and do not refer any to hypergraph modelling; in addition co-word analysis is omitted.

The advantages of the hypergraph theory widely are that it constitutes a uniform data model to generalise co-word analysis, allowing to cover the entire model called relational (Callon et al, 1993). On the other hand, we shall compare it to more symbolic methods as formal concept analysis (FCA) using the Galois lattice model (Ganter et al, 2005; Ganter & Wille, 1999). Note that the number of possible closed sets is exponential on the number of attributes, and the generation of the whole Galois lattice is a NP hard problem. We are interested in clustering methods that naturally highlight particular small closed sets of attributes in linear time. Our intention is to try to point out the theoretical requirements of clustering algorithm in linear or quadratic time to maximize the probability of extracting closed sets.

A tree can be viewed as a graph with no circle of length higher than 2, then a natural generalisation of trees are graphs with no circle longer than 3. This class of graphs is called chordal. We experimentally observed that all subgraphs with less than 30 elements of a co-author graph are chordal, meanwhile the whole graph is not. On the contrary, the co-word graph is more general and does not seem to have special chordal properties. In future work we shall study the chordal properties in the purpose of improving the graph visualisation.

To know that a graph is SWG allows to consider a "graph decomposition process" (Berri et al, 2004), using group of vertices as "clique minimal separators" aiming to split the SWG into not disjoint minimal unites; we are working on this.

References

Berge C. (1987). *Hypergraphes*, Paris: Gauthier-Villars.

Berry A., Kaba B., Nadif M., SanJuan E., Sigayret A. (2004). Classification et désarticulation de graphes de termes. In JADT 2004 Proceedings (pp. 160-170). Leuven.

Callon M., Courtial J-P., Penan H. (1993). *La Scientométrie.* Paris: Presses Universitaires de France, (coll. Que sais-je? Vol. 2727).

Courtial J-P. (1990). *Introduction à la scientométrie*. Paris: Anthropos – Economica.

Ganter B., Stummed G., Wille R. (Eds) (2005). Formal Concept Analysis: Foundations and Applications. Lecture Notes in Artificial Intelligence 3626. Berlin: Springer.

Ganter B., Wille, R., (1999). Formal Concept Analysis: Mathematical Foundations. Berlin: Springer.

Ibekwe-SanJuan F. (1998). A linguistic and mathematical method for mapping thematic trends from texts. In *Proceedings of the 13th European Conference on Artificial Intelligence* (pp. 170-174). Brighton, U.K.

SanJuan E., Dowdall J., Ibekwe-SanJuan F., Rinaldi F. (2005). A symbolic approach to automatic multiword term structuring. *Computer Speech and Language*, 19 (4), 524-542.

Van Cutsem B. (Ed.) (1994). *Classification and Dissimilarity Analysis.* (LNS 93). Berlin: Springer.

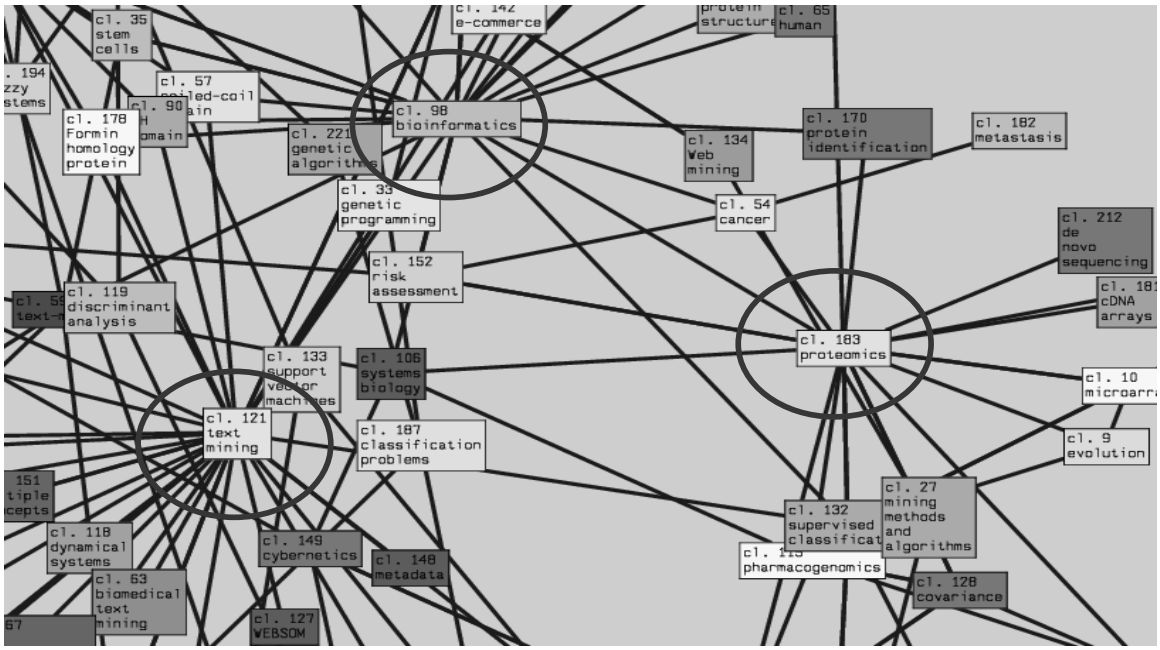
Watts D.J. (1999). Small Worlds. Princeton University Press.

Wille R. (1982). Restructuring lattice theory: an approach based on hierarchies of concepts. *Ordered Sets*, 83, 445-470.

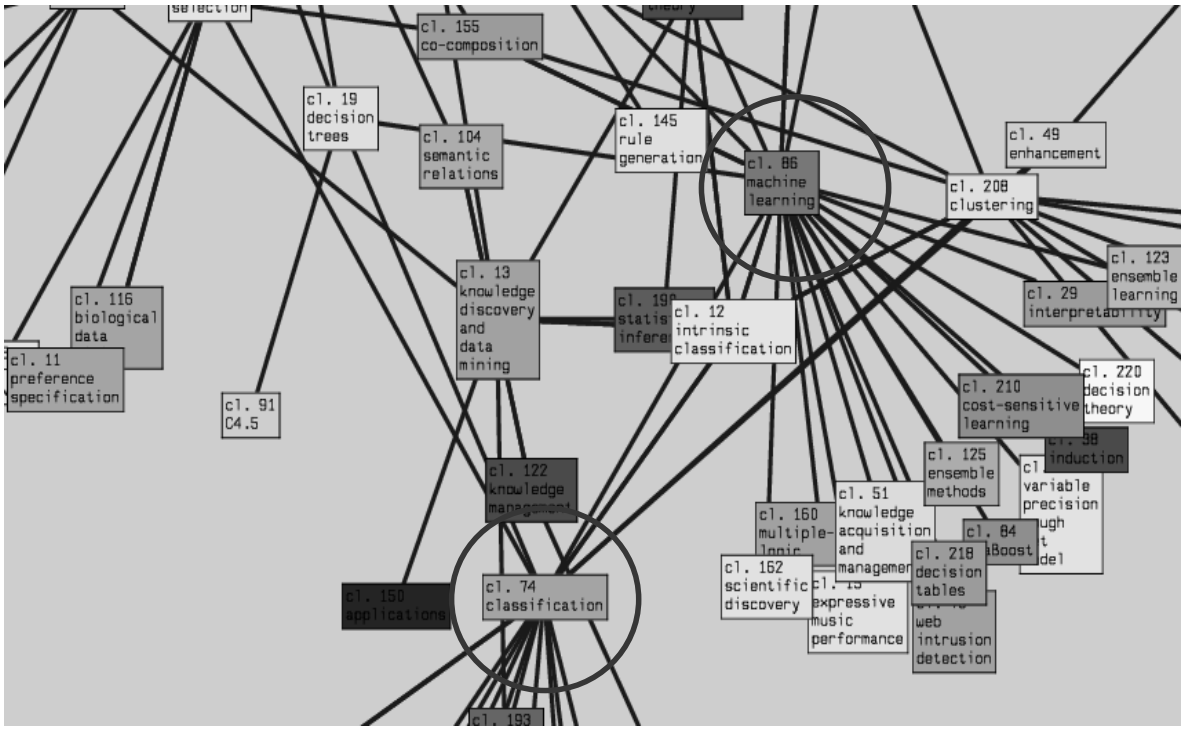
Zaki M. (2004). Mining non-redundant association rules. *Data Mining and Knowledge Discovery*, 9 (3), 223-248.

Appendix

(A) Text Mining, Bioinformatics, Proteomics



(B) Machine learning, Classification



The State of Science and Technology in Africa (2000-2004): A Scientometric Assessment

Anastassios Pouris[*] and Anthipi Pouris[**]

[*]*Anastassios.pouris@up.ac.za*
Institute for Technological Innovation, University of Pretoria, Pretoria 0002 (South Africa)

[**]*apouris@nrf.ac.za*
Institutional Capacity Development Grants, National Research Foundation, PO Box 2600, Pretoria 0001 (South Africa)

Abstract
This article reports for first time the state of science and technology in the African Continent on the basis of two scientometric indicators - number of research publications and number of patents awarded. We suggest that the effort covers partially the need for monitoring indicators for the Continent.Our analysis shows that Africa produced 68 945 publications over the 2000-2004 period or 1.8% of the World's publications. In comparison India produced 2.4% and Latin America 3.5% of the World's research. More detailed analysis reveals that research in Africa is concentrated in just two countries – South Africa and Egypt. These two counties produce just above 50% of the Continent's publications and the top 8 countries produce above 80% of the Continent's research. Disciplinary analysis reveals that few African countries have the minimum number of scientists required for the functioning of a scientific discipline. Examination of the Continent's inventive profile, as manifested in patents, indicates that Africa produces less than one thousand of the world's inventions. Furthermore 88% of the Continent's inventive activity in concentrated in South Africa. On the basis of the recent declarations on the importance of science and technology for development we suggest that the African Governments should pay particular attention in developing their national research systems.

Keywords
Africa; science; technology; research; patents; publications

Introduction

Monitoring and evaluating the various facets of the scientific enterprise is a necessary and integral part of science policy. Rising costs of research and development and competing disciplinary claims for financial resources require intelligent allocation of resources, which presupposes knowledge of the activities and performance of the innovation system.

The development of science, technology and innovation (STI) indicators has grown substantially internationally during the last twenty years. Their ability to enlighten political choices, by informing and allowing decisions to be objective and depersonalized, has been the main force behind their popularity.Additional factors enhancing the importance of STI indicators have been the central role assumed by science and technology as a social resource; the continuing expansion of knowledge and globalisation; and the magnitude of the challenges undertaken, in terms of research, technology and investment.

Many countries and international organizations have established bodies specifically dedicated to the analysis and production of information on, and the development of, indicators for science, technology and innovation. Examples of this are the OECD supported by the Network of Experts on Science and Technology Indicators (NESTI); the European Union's Eurostat; UNESCO's Institute of Statistics; the Network on Science and Technology Indicators (RICyT) in the Ibero-America and the Asia-Pacific Economic Cooperation and Pacific Economic Cooperation Council for the Asia-Pacific countries. Those agencies provide services and conduct studies requiring the coordination of interdisciplinary teams under conditions encouraging a flexible mode of operations, the necessary academic independence, and the capacity for planning medium- and long-term activities.

The importance of indicators has been recognised also in the African Continent in the first NEPAD Ministerial Conference on Science and Technology.The NEPAD Declaration (NEPAD, 2003) commits to "develop and adopt common sets of indicators to benchmark our national and regional systems of innovation" (paragraph 12).The Declaration suggests that the system of indicators will "track the development and functioning of the African national systems of innovation" and it will constitute the mainstay for the production of the "Annual African Innovation Outlook". The Outlook "will report on the developments in science, technology and innovation in Africa at national, regional and continental level and will be produced by the NEPAD Secretariat under the mandate of the Steering Committee and transmitted to the Ministers Council for consideration".Furthermore, the indicators will inform the African Peer Review Mechanism which among others will monitor the success of the African countries in the "effort to commit 1% of Gross Domestic Product (GDP) in public funds to research and development". However, up to date the African continent is the only Continent without an organisation/agency responsible to promote and assist in the development of STI indicators and with scant scientometric coverage in the open literature. Lack of scientometrics expertise, disinterested science authorities and lack of financial resources may account for this phenomenon.

This document aims to report the state of science and technology in the African Continent as it is represented by two of the commonest used scientometric indicators internationally - research publications and awarded patents. These indicators are probably the first to be included in an African monitoring system as they are easily and reliably available and they do not require expertise in the African Continent.

Approach

One of the most efficient and objective methods of assessing research and innovation performance is through scientometric indicators. An indicator is defined (DHEW, 1970) as "statistics of direct normative interest which facilitates concise, comprehensive and balanced judgments about the condition of major aspects of a society. It is in all cases a direct measure of welfare and is subject to the interpretation that, if it changes in the "right" direction, while other things remain equal, things have gotten better or people better off."

Scientometric analysis, the quantitative study of the innovation system, is based mainly on bibliometric and patent indicators. In bibliometrics the number of publications in a field is considered as an indicator of research activity. Similarly in patent analysis the number of patents awarded to an institution or a country is used as an indicator of technological activity. Patent indicators – within the science and technology (S&T) context – are used to measure inventive performance, diffusion of knowledge and internationalization of innovative activities – across countries, firms, industries, technology areas, etc. The philosophy underlying the use of bibliometric indicators as performance measures has been summarized in De Solla Price's statement that "for those who are working at the research front, publication is not just an indicator but, in a very strong sense, the end product of their creative effort." (De Solla Price, 1975)

It is natural that there are many trained scientists who are not required to publish. They may perform managerial or administrative functions, teach available knowledge apply existing knowledge in making new products and in providing services. However, the common characteristic of all these scientists is that they are far away from the research front. They provide the infrastructure for the producers of knowledge and they exploit the end results of research and development. In any case, however, they cannot be considered as "knowledge" producers.

The importance of research publications lies among others in the fact that they are proxies of scientific manpower available in different countries. (Schubert A. and A. Telcs 1986). This is a particular useful characteristic for African countries which are lacking relative monitoring mechanisms of science, technology and innovation.The same way, in which scientific articles are accepted as a legitimate reflection of scientific research, patents are accepted as a reflection of technological achievements. Griliches (1990) has pointed out that "Patent statistics remain a unique source for the analysis of the

process of technical change. Nothing else even comes close in the quantity of available data, accessibility, and potential industrial, organizational, and technological detail."

Bibliometric indicators and patent analysis possess a number of strengths that facilitates their universal use. They are highly reliable because they are well defined and unambiguous. They facilitate detailed categorisation and hence make possible the study of scientific and technological fields and sub-fields and they make possible international comparisons. In the context of developing countries scientometric indicators can provide unique insides of the strengths and weaknesses of the STI systems as other sources of information are usually lacking.Scientometric indicators are used internationally for monitoring purposes. In the USA the National Science Foundation (NSB, 2004) is using bibliometrics and patent analysis to monitor the health of American science and technology on a continuous basis; in Europe the European Commission (EC, 1997) is using similar approaches in order to monitor the health of the European innovation system and the OECD (2003) (is using the indicators for monitoring and comparative purposes).

For the bibliometric analysis we utilise the databases of the Institute for Scientific Information (ISI) – Science Citation Index, Social Science Citation Index and Arts and Humanities Citation Index. ISI currently indexes more than 6000 of the world's leading scholarly scientific and technical journals, approximately 1800 social sciences journals, and 1150 titles from the arts and humanities journals. All journals indexed by ISI are peer reviewed. As a group, the ISI indexed set of journals represents an elite body of internationally influential research publications. In these databases, ISI includes articles, notes, reviews and proceeding papers but not items such as editorials, letters, corrections and abstracts. There is a substantial number of articles using the databases. (Braun, T., Schubert, S., Zsindely, S. 1997 and King, DA, 2004). An important advantage of the ISI databases is that they provide information (affiliation) for all co-authors in the indexed articles. Hence, papers can be attributed to all co-authors equally. A paper is attributed to a country if it carries at least one author address of that country.

For the purposes of the patent analysis we are utilising patents awarded by the United States Patents and Trade Office (USPTO) to residents of African countries. Although most countries in the world have their own patent authorities, the use of the USPTO provides a number of advantages. First in the majority of the patent offices, patents are not examined for originality, usefulness and novelty. Consequently counting and comparing patents awarded by different patent offices in different countries may be misleading because of differences in the criteria used and the easiness of awarding patents, bias towards local patents etc. The obvious solution in order to avoid the above-mentioned shortcomings is to use a common denominator such as an external patent system with an objective approach in its awarding patents approach (i.e. the USPTO). The USPTO examines claims according to a number of criteria. These are (Fordis, B.J. and Sung, M.L., 1995):

- *Subject matter:* an invention must fall into one of the categories the patent law divides patentable subject matter into.
- *Utility:* An invention must fulfil the substantive requirement of "utility". An invention must perform a designed function or achieve some minimum human purpose.
- *Novelty:* an invention has to be novel.
- *Non-obviousness:* the knowledge in the technological field at the time of invention must not make the invention obvious to one of ordinary skill in that area.
-*Definiteness:* one skilled in the art must understand the limits of the invention based on the claim language.

Second, the US represents the most important single market for technological sales and hence is a key drawing card for technology-based products. Owners of important commercial inventions will make sure that they are protected in the USA market. Third, the costs involved and the complexity of filing foreign patents in the USA tend to screen out trivial patents. These characteristics make the approach appealing and a large number of investigations are reported internationally annually. (Hicks D., Breitzman T., Olivastro D. and Hamilton K., 2001)

It should be mentioned that although patents facilitate the development of a number of useful indicators they have a number of drawbacks. Patented inventions are not necessarily all the inventions produced in a country or organization. Many inventions are not patented because there are other barriers to entry (e.g. lack of brand names among the competitors), because inventors may undertake other measures of protection (e.g. the encapsulation of products in epoxy resin to deter imitation) or because inventors consider that the invention will be profitable even if imitators may appear in the foreseeable future. Similarly high costs for application or monitoring infringement as well as lack of appreciation are additional reasons that may limit the number of patents from a particular country or organisation.

The State of S&T in Africa

Table 1. Number and share of world publications-Selected regions 2000-2004

Country	Number of Publications	World share of publications
European Union	1,461,813	38.8%
USA	1,267,203	33.6%
China	175,522	4.7%
Latin America	130,569	3.5%
India	89,976	2.4%
Africa	68,945	1.8%
Other	574,406	15.2
World total	**3,768,434**	**100.0%**

Table 1 shows the number of publications[1] produced by researchers in different regions of the world and the relevant shares during the five year period 2000-2004. The European Union and USA appear at the top of the list producing 38.8% and 33.6% of the world publications respectively. Africa produced during the period 68.945 publications or 1.8% of the world total. Africa produces half as many publications as Latin America and substantially less than what India is publishing.Table 2 shows the number of publications and country shares in the Continent.

Table 2. Number and share of publications of the most prolific African countries: 2000-2004

Country	Number of Publications	Share in Continents' publications	Cumulative share
South Africa	20,762	30.1%	30.1%
Egypt	13,942	20.2%	50.3%
Morocco	5463	7.9%	58.2%
Nigeria	4040	5.9%	64.1%
Tunisia	3930	5.7%	69.8%
Kenya	3231	4.7%	74.5%
Algeria	2766	4.0%	78.5%
Tanzania	1368	2.0%	80.5%
Ethiopia	1321	1.9%	82.4%
Cameroon	1301	1.9%	84.3%

[1] The National Science Indicators database of the Institute for Scientific Information (ISI) is used for this purpose. The database covers comprehensively the most important journals in the World and it provides the standards in the field of science monitoring.

South Africa and Egypt appear on the top of the list with 20.762 and 13.942 publications respectively. The last column in the table shows that the top two counties in the Continent produce just above 50% of the Continent's publications and that the top 8 countries produce above 80% of the Continent's research.
Table 3 shows the number of publications per discipline produced by the various African countries. If we assume a-priori[2] that a country needs approximately 300 publications over a 5-year period in order to be considered as having some minimal expertise in the field (i.e. approximately 50 to 60 publishing researchers), it becomes apparent that the majority of the African countries are well below the set standards. The vertical sums in the table identify the disciplinary concentration of research in the Continent. The disciplines of clinical medicine, plant and animal studies and chemistry are those with the largest number of publications - 12.998, 8.546 and 8.317 publications respectively.

In the field of ecology (a discipline necessary for environmentally friendly and sustainable development) only four countries (South Africa, Egypt, Nigeria and Kenya) produce 300 or more publications. On the other hand "land and primary resources" sciences (i.e. agriculture, ecology, geosciences and plant and animal sciences) occupy 26.4% of the research in Africa while the relevant figures for USA and India are 13.5% and 19.5% respectively. The challenge appears to be size of the research system and not disciplinary adjustments. This conclusion is supported further by the findings of the Latin American and Caribbean Regional Workshop (UN, 2003) on Science and Technology for Sustainable Development (STSD) which state that "new critical knowledge is needed in all areas….The knowledge required to solve these problems relate to a range of areas and disciplines. In general, this knowledge concerns the study of society-nature interactions" (p15) and that sustainable development requires "methodologies relating to supra-disciplinary approaches" (p18)

Table 4 shows the concentration of disciplinary research in the Continent. South Africa appears to be producing above 70% of the Continent's space science and psychology and above 50% of the Continent's research in education, economic sciences and neurosciences. Egypt is producing above 40% of the Continent's research in chemistry, and materials science and just below forty percent of the Continent's research in engineering. Nigeria is the main producer of agricultural research in the Continent(18.9%).

[2] This figure is well below a theoretical "threshold" level that would trigger a virtuous interaction between science and technology (Bernardes, A. and Albuquerque, E. (2003) Cross-over, thresholds and the interactions between science and technology: lessons for less-developed countries. *Research Policy*, **32(**5**):**867-887). This "threshold level" for 1998 data was in the neighbourhood of 150 papers per million people and is increasing.

Table 3. Number of publications per discipline produced by the various African countries

Field	Agricultural Sciences	Biology & Biochemistry	Chemistry	Clinical Medicine	Computer Sciences	Ecology/Environment	Economics & Business	Education	Engineering	Geosciences	Immunology	Law
South Africa	500	1,112	1,744	3,684	122	1,187	257	151	915	1,302	298	5
Egypt	477	455	3,579	1,542	92	316	17	3	1,812	363	55	1
Morocco	188	160	1,023	878	23	157	9	0	422	238	21	1
Nigeria	665	183	192	846	11	293	24	8	234	84	18	11
Tunisia	87	238	640	753	58	62	29	0	376	98	34	0
Kenya	306	114	50	769	1	315	42	7	69	61	227	1
Algeria	65	57	614	85	28	56	1	0	412	102	1	0
Tanzania	88	42	11	501	0	100	5	2	31	69	69	1
Ethiopia	139	33	35	397	0	70	32	2	13	73	56	1
Cameroon	141	50	48	264	5	65	8	0	64	62	60	0
Zimbabwe	128	40	16	253	1	98	18	2	41	96	64	0
Uganda	75	29	1	406	0	79	4	6	3	10	131	0
Ghana	120	31	25	309	3	58	10	4	21	20	54	1
Senegal	68	49	51	307	0	85	1	1	16	24	88	0
Cote Ivoire	59	45	46	252	0	32	8	0	10	17	60	0
Malwi	48	15	1	371	0	8	5	3	9	7	71	0
Botswana	54	7	69	55	3	46	12	18	24	76	9	1
Burkina Faso	41	27	14	184	0	40	4	0	7	15	32	0
Sudan	53	13	17	154	0	17	1	1	14	3	30	0
Zambia	36	7	3	150	0	18	3	0	4	21	51	0
Gambia	7	14	0	214	0	6	1	1	0	2	78	0
Benin	42	15	11	61	0	24	2	0	11	4	8	0
Bagon	13	11	2	119	0	16	1	0	1	4	55	0
Libya	5	9	48	57	0	11	2	0	52	18	3	0
Namibia	0	6	4	8	1	38	2	3	3	31	1	0
Mauritius	13	23	31	17	0	25	4		29	4	2	0
Nigeria	43	14	6	52	1	18	1	1	2	21	3	0
Mozambique	12	4	0	71	1	10	4	0	5	7	7	0
Congo peoples rep	12	7	4	73	0	12	0	1	0	6	5	0
Eritrea	12	0	1	10	0	5	2	0	4	26	0	0
Congo Democratic Republic	4	2	7	40	0	2	0	1	1	3	6	0
Mauritania	9	3	9	15	0	6	0	0	5	3	0	0
Cent Afr replublic	1	0	2	38	0	3	0	0	0	1	10	0
Swaziland	4	1	2	8	2	4	4	0	4	3	0	0
Angola	1	2	6	15	0	0	0	0	0	11	2	0
Rwanda	1	0	1	18	0	2	0	1	1	0	7	1
Lesotho	0	1	2	6	0	1	1	2	0	1	0	0
Burundi	1	0	2	6	0	1	0	0	1	1	1	0
Liberia	0	0	0	7	0	1	0	0	1	0	1	0
Somalia	0	0	0	3	0	0	0	0	0	0	0	0
Senegambia	0	0	0	0	0	0	0	0	0	0	0	0
TOTAL	**3,518**	**2,819**	**8,317**	**12,998**	**352**	**3,287**	**514**	**218**	**4,617**	**2,887**	**1,618**	**24**

Field	Materials Science	Mathematics	Microbiology	Molecular Biology & Genetics	Multidisciplinary	Neurosciences & Behaviour	Pharmacology	Physics	Plant and Animal Sciences	Psychology/ Psychiatry	Social Sciences, General	Space Science	Grand Total
South Africa	402	275	545	328	659	334	364	1,101	3,621	440	989	427	20,762
Egypt	1,193	177	202	63	279	56	538	1,827	705	35	93	62	13,942
Morocco	249	345	70	50	56	73	132	1,010	301	8	22	27	5,463
Nigeria	77	35	88	42	103	21	229	80	537	49	200	10	4,040
Tunisia	189	249	102	61	22	32	40	623	224	1	12	2	3,932
Kenya	12	0	139	56	77	18	53	36	700	17	161	0	3,231
Algeria	252	111	16	27	14	10	25	786	91	1	6	6	2,766
Tanzania	5	4	49	11	12	11	24	8	223	10	90	2	1,368
Ethiopia	5	0	37	25	9	6	19	15	298	12	43	1	1,321
Cameroon	11	16	78	13	34	5	54	86	190	0	47	0	1,301
Zimbabwe	6	5	36	4	14	10	27	7	232	11	94	1	1,204
Uganda	4	2	53	7	6	6	11	4	157	11	89	1	1,095
Ghana	8	0	43	23	10	4	27	27	116	4	86	1	1,005
Senegal	16	9	63	10	1	6	6	28	133	1	40	0	1,003
Cote Ivoire	7	8	41	6	2	8	26	8	105	2	17	2	761
Malwi	0	0	28	2	3	0	1	1	47	0	25	0	645
Botswana	2	21	14	1	11	0	10	28	90	7	74	0	632
Burkina Faso	4	4	29	13	1	3	15	11	88	1	19	0	552
Sudan	4	1	31	4	11	1	6	21	79	2	5	4	472
Zambia	0	2	17	4	2	1	1	4	62	5	41	3	435
Gambia	0	0	13	5	1	1	2	0	34	0	12	0	391
Benin	0	8	7	3	2	2	12	21	122	0	8	1	364
Bagon	0	6	44	19	1	3	7	1	25	0	5	0	333
Libya	4	2	7	3	2	2	7	14	14	0	1	1	262
Namibia	1	0	3	1	16	0	0	5	106	0	12	3	244
Mauritius	3	4	3	7	4	2	6	1	31	2	3	14	228
Nigeria	5	1	4	3	1	0	1	7	34	0	8	0	226
Mozambique	0	1	7	0	2	0	0	1	44	0	11	0	187
Congo peoples rep	1	1	4	1	1	0	6	5	40	0	3	0	182
Eritrea	0	0	1	0	2	0	1	5	20	0	7	1	97
Congo Democratic Republic	1	0	3	2	1	1	8	1	6	0	6	0	95
Mauritania	3	8	1	0	0	2	1	2	12	0	1	0	80
Cent Afr replublic	0	0	8	0	0	0	0	4	12	0	0	0	79
Swaziland	0	3	0	0	1	1	1	7	10	0	3	0	58
Angola	0	0	1	0	2	1	0	0	10	1	2	0	54
Rwanda	0	0	2	1	0	0	1	0	9	0	3	0	48
Lesotho	0	0	0	0	1	0	0	2	9	0	6	0	32
Burundi	0	2	0	0	0	2	0	8	9	0	0	0	34
Liberia	0	0	0	0	0	0	1	1	1	0	2	0	15
Somalia	0	0	0	0	0	0	0	0	0	0	1	0	4
Senegambia	0	0	0	0	0	0	0	0	0	0	0	0	0
TOTAL	**2,464**	**1,300**	**1,789**	**795**	**1,363**	**622**	**1,662**	**5,796**	**8,547**	**620**	**2,247**	**569**	**68,943**

Table 4. Concentration of disciplinary research in the African Continent

Field	Agricultural Sciences	Biology & Biochemistry	Chemistry	Clinical Medicine	Computer Sciences	Ecology/ Environment	Economics & Business	Education	Engineering	Geosciences	Immunology	Law
South Africa	14.21%	39.45%	20.97%	28.34%	34.66%	36.11%	50.00%	69.27%	19.82%	45.10%	18.42%	20.83%
Egypt	13.56%	16.14%	43.03%	11.86%	26.14%	9.61%	3.31%	1.38%	39.25%	12.57%	3.40%	4.17%
Morocco	5.34%	5.68%	12.30%	6.75%	6.53%	4.78%	1.75%	0.00%	9.14%	8.24%	1.30%	4.17%
Nigeria	18.90%	6.49%	2.31%	6.51%	3.13%	8.91%	4.67%	3.67%	5.07%	2.91%	1.11%	45.83%
Tunisia	2.47%	8.44%	7.70%	5.79%	16.48%	1.89%	5.64%	0.00%	8.14%	3.39%	2.10%	0.00%
Kenya	8.70%	4.04%	0.60%	5.92%	0.28%	9.58%	8.17%	3.21%	1.49%	2.11%	14.03%	4.17%
Algeria	1.85%	2.02%	7.38%	0.65%	7.95%	1.70%	0.19%	0.00%	8.92%	3.53%	0.06%	0.00%
Tanzania	2.50%	1.49%	0.13%	3.85%	0.00%	3.04%	0.97%	0.92%	0.67%	2.39%	4.26%	4.17%
Ethiopia	3.95%	1.17%	0.42%	3.05%	0.00%	2.13%	6.23%	0.92%	0.28%	2.53%	3.46%	4.17%
Cameroon	4.01%	1.77%	0.58%	2.03%	1.42%	1.98%	1.56%	0.00%	1.39%	2.15%	3.71%	0.00%
Zimbabwe	3.64%	1.42%	0.19%	1.95%	0.28%	2.98%	3.50%	0.92%	0.89%	3.33%	3.96%	0.00%
Uganda	2.13%	1.03%	0.01%	3.12%	0.00%	2.40%	0.78%	2.75%	0.06%	0.35%	8.10%	0.00%
Ghana	3.41%	1.10%	0.30%	2.38%	0.85%	1.76%	1.95%	1.83%	0.45%	0.69%	3.34%	4.17%
Senegal	1.93%	1.74%	0.61%	2.36%	0.00%	2.59%	0.19%	0.46%	0.35%	0.83%	5.44%	0.00%
Cote Ivoire	1.68%	1.60%	0.55%	1.94%	0.00%	0.97%	1.56%	0.00%	0.22%	0.59%	3.71%	0.00%
Malwi	1.36%	0.53%	0.01%	2.85%	0.00%	0.24%	0.97%	1.38%	0.19%	0.24%	4.39%	0.00%
Botswana	1.53%	0.25%	0.83%	0.42%	0.85%	1.40%	2.33%	8.26%	0.52%	2.63%	0.56%	4.17%
Burkina Faso	1.17%	0.96%	0.17%	1.42%	0.00%	1.22%	0.78%	0.00%	0.15%	0.52%	1.98%	0.00%
Sudan	1.51%	0.46%	0.20%	1.18%	0.00%	0.52%	0.19%	0.46%	0.30%	0.10%	1.85%	0.00%
Zambia	1.02%	0.25%	0.04%	1.15%	0.00%	0.55%	0.58%	0.00%	0.09%	0.73%	3.15%	0.00%
Gambia	0.20%	0.50%	0.00%	1.65%	0.00%	0.18%	0.19%	0.46%	0.00%	0.07%	4.82%	0.00%
Benin	1.19%	0.53%	0.13%	0.47%	0.00%	0.73%	0.39%	0.00%	0.24%	0.14%	0.49%	0.00%
Bagon	0.37%	0.39%	0.02%	0.92%	0.00%	0.49%	0.19%	0.00%	0.02%	0.14%	3.40%	0.00%
Libya	0.14%	0.32%	0.58%	0.44%	0.00%	0.33%	0.39%	0.00%	1.13%	0.62%	0.19%	0.00%
Namibia	0.00%	0.21%	0.05%	0.06%	0.28%	1.16%	0.39%	1.38%	0.06%	1.07%	0.06%	0.00%
Mauritius	0.37%	0.82%	0.37%	0.13%	0.00%	0.76%	0.78%	0.00%	0.63%	0.14%	0.12%	0.00%
Nigeria	1.22%	0.50%	0.07%	0.40%	0.28%	0.55%	0.19%	0.46%	0.04%	0.73%	0.19%	0.00%
Mozambique	0.34%	0.14%	0.00%	0.55%	0.28%	0.30%	0.78%	0.00%	0.11%	0.24%	0.43%	0.00%
Congo peoples rep	0.34%	0.25%	0.05%	0.56%	0.00%	0.37%	0.00%	0.46%	0.00%	0.21%	0.31%	0.00%
Eritrea	0.34%	0.00%	0.01%	0.08%	0.00%	0.15%	0.39%	0.00%	0.09%	0.90%	0.00%	0.00%
Congo Democratic Republic	0.11%	0.07%	0.08%	0.31%	0.00%	0.06%	0.00%	0.46%	0.02%	0.10%	0.37%	0.00%
Mauritania	0.26%	0.11%	0.11%	0.12%	0.00%	0.18%	0.00%	0.00%	0.11%	0.10%	0.00%	0.00%
Cent Afr replublic	0.03%	0.00%	0.02%	0.29%	0.00%	0.09%	0.00%	0.00%	0.00%	0.03%	0.62%	0.00%
Swaziland	0.11%	0.04%	0.02%	0.06%	0.57%	0.12%	0.78%	0.00%	0.09%	0.10%	0.00%	0.00%
Angola	0.03%	0.07%	0.07%	0.12%	0.00%	0.00%	0.00%	0.00%	0.00%	0.38%	0.12%	0.00%
Rwanda	0.03%	0.00%	0.01%	0.14%	0.00%	0.06%	0.00%	0.46%	0.02%	0.00%	0.43%	4.17%
Lesotho	0.00%	0.04%	0.02%	0.05%	0.00%	0.03%	0.19%	0.92%	0.00%	0.03%	0.00%	0.00%
Burundi	0.03%	0.00%	0.02%	0.05%	0.00%	0.03%	0.00%	0.00%	0.02%	0.03%	0.06%	0.00%
Liberia	0.00%	0.00%	0.00%	0.05%	0.00%	0.03%	0.00%	0.00%	0.02%	0.00%	0.06%	0.00%
Somalia	0.00%	0.00%	0.00%	0.02%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Senegambia	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

Field	Materials Science	Mathematics	Microbiology	Molecular Biology & Genetics	Multidisciplinary	Neurosciences & Behaviour	Pharmacology	Physics	Plant and Animal Sciences	Psychology/ Psychiatry	Social Sciences, General	Space Science
South Africa	16.31%	21.15%	30.46%	41.26%	48.35%	53.70%	21.90%	19.00%	42.37%	70.97%	44.01%	75.04%
Egypt	48.42%	13.62%	11.29%	7.92%	20.47%	9.00%	32.37%	31.52%	8.25%	5.65%	4.14%	10.90%
Morocco	10.11%	26.54%	3.91%	6.29%	4.11%	11.74%	7.94%	17.43%	3.52%	1.29%	0.98%	4.75%
Nigeria	3.13%	2.69%	4.92%	5.28%	7.56%	3.38%	13.78%	1.38%	6.28%	7.90%	8.90%	1.76%
Tunisia	7.67%	19.15%	5.70%	7.67%	1.61%	5.14%	2.41%	10.75%	2.62%	0.16%	0.53%	0.35%
Kenya	0.49%	0.00%	7.77%	7.04%	5.65%	2.89%	3.19%	0.62%	8.19%	2.74%	7.17%	0.00%
Algeria	10.23%	8.54%	0.89%	3.40%	1.03%	1.61%	1.50%	13.56%	1.06%	0.16%	0.27%	1.05%
Tanzania	0.20%	0.31%	2.74%	1.38%	0.88%	1.77%	1.44%	0.14%	2.61%	1.61%	4.01%	0.35%
Ethiopia	0.20%	0.00%	2.07%	3.14%	0.66%	0.96%	1.14%	0.26%	3.49%	1.94%	1.91%	0.18%
Cameroon	0.45%	1.23%	4.36%	1.64%	2.49%	0.80%	3.25%	1.48%	2.22%	0.00%	2.09%	0.00%
Zimbabwe	0.24%	0.38%	2.01%	0.50%	1.03%	1.61%	1.62%	0.12%	2.71%	1.77%	4.18%	0.18%
Uganda	0.16%	0.15%	2.96%	0.88%	0.44%	0.96%	0.66%	0.07%	1.84%	1.77%	3.96%	0.18%
Ghana	0.32%	0.00%	2.40%	2.89%	0.73%	0.64%	1.62%	0.47%	1.36%	0.65%	3.83%	0.18%
Senegal	0.65%	0.69%	3.52%	1.26%	0.07%	0.96%	0.36%	0.48%	1.56%	0.16%	1.78%	0.00%
Cote Ivoire	0.28%	0.62%	2.29%	0.75%	0.15%	1.29%	1.56%	0.14%	1.23%	0.32%	0.76%	0.35%
Malwi	0.00%	0.00%	1.57%	0.25%	0.22%	0.00%	0.06%	0.02%	0.55%	0.00%	1.11%	0.00%
Botswana	0.08%	1.62%	0.78%	0.13%	0.81%	0.00%	0.60%	0.48%	1.05%	1.13%	3.29%	0.00%
Burkina Faso	0.16%	0.31%	1.62%	1.64%	0.07%	0.48%	0.90%	0.19%	1.03%	0.16%	0.85%	0.00%
Sudan	0.16%	0.08%	1.73%	0.50%	0.81%	0.16%	0.36%	0.36%	0.92%	0.32%	0.22%	0.70%
Zambia	0.00%	0.15%	0.95%	0.50%	0.15%	0.16%	0.06%	0.07%	0.73%	0.81%	1.82%	0.53%
Gambia	0.00%	0.00%	0.73%	0.63%	0.07%	0.16%	0.12%	0.00%	0.40%	0.00%	0.53%	0.00%
Benin	0.00%	0.62%	0.39%	0.38%	0.15%	0.32%	0.72%	0.36%	1.43%	0.00%	0.36%	0.18%
Bagon	0.00%	0.46%	2.46%	2.39%	0.07%	0.48%	0.42%	0.02%	0.29%	0.00%	0.22%	0.00%
Libya	0.16%	0.15%	0.39%	0.38%	0.15%	0.32%	0.42%	0.24%	0.16%	0.00%	0.04%	0.18%
Namibia	0.04%	0.00%	0.17%	0.13%	1.17%	0.00%	0.00%	0.09%	1.24%	0.00%	0.53%	0.53%
Mauritius	0.12%	0.31%	0.17%	0.88%	0.29%	0.32%	0.36%	0.02%	0.36%	0.32%	0.13%	2.46%
Nigeria	0.20%	0.08%	0.22%	0.38%	0.07%	0.00%	0.06%	0.12%	0.40%	0.00%	0.36%	0.00%
Mozambique	0.00%	0.08%	0.39%	0.00%	0.15%	0.00%	0.00%	0.02%	0.51%	0.00%	0.49%	0.00%
Congo peoples rep	0.04%	0.08%	0.22%	0.13%	0.07%	0.00%	0.36%	0.09%	0.47%	0.00%	0.13%	0.00%
Eritrea	0.00%	0.00%	0.06%	0.00%	0.15%	0.00%	0.06%	0.09%	0.23%	0.00%	0.31%	0.18%
Congo Democratic Republic	0.04%	0.00%	0.17%	0.25%	0.07%	0.16%	0.48%	0.02%	0.07%	0.00%	0.27%	0.00%
Mauritania	0.12%	0.62%	0.06%	0.00%	0.00%	0.32%	0.06%	0.03%	0.14%	0.00%	0.04%	0.00%
Cent Afr replublic	0.00%	0.00%	0.45%	0.00%	0.00%	0.00%	0.00%	0.07%	0.14%	0.00%	0.00%	0.00%
Swaziland	0.00%	0.23%	0.00%	0.00%	0.07%	0.16%	0.06%	0.12%	0.12%	0.00%	0.13%	0.00%
Angola	0.00%	0.00%	0.06%	0.00%	0.15%	0.16%	0.00%	0.00%	0.12%	0.16%	0.09%	0.00%
Rwanda	0.00%	0.00%	0.11%	0.13%	0.00%	0.00%	0.06%	0.00%	0.11%	0.00%	0.13%	0.00%
Lesotho	0.00%	0.00%	0.00%	0.00%	0.07%	0.00%	0.00%	0.03%	0.11%	0.00%	0.27%	0.00%
Burundi	0.00%	0.15%	0.00%	0.00%	0.00%	0.32%	0.00%	0.14%	0.11%	0.00%	0.00%	0.00%
Liberia	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.06%	0.02%	0.01%	0.00%	0.09%	0.00%
Somalia	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.04%	0.00%
Senegambia	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

Table 5 shows the number of patents awarded to inventors with African addresses during the 2000-2004 period. The whole Continent produced 633 patents during the period. In comparison the rest of the World received 817197 patents. The African Continent produces less than one per thousand of the inventions which are protected in the largest economy of the World. It is interesting that while the Continent produces 1.8% of the World's knowledge (as it is manifested in research publications) it produces less than O.1% of the World's inventions.

Table 5. Patents Awarded to African Inventors by USPTO - 2000-2004

Country	Utility Patents	Design Patents	Plant Patents	Reissue Patents
Algeria	2			
Angola		1		
Egypt	29			
Ivory Coast	1			
Jordan	1			
Kenya	17	12		
Morocco	4	1		
Namibia	1			
Nigeria	12			
South Africa	557	61	10	2
Tanzania	2	1		
Tunisia	1			
Zimbabwe	4			
Uganda	2			
Africa Total	633	76	10	2
World Total	**817,197**	**82,006**	**4,275**	**2,184**

Table 5 further shows that 557 out of the 633 utility patents (88%) are produced by South Africa. South Africa produces 30% of the Continents publications and 88% of the Continents patents.

Discussion

This article reports the results of an effort to develop the scientometric indicators –research publications and patents- of the African Continent, compare them with those of other regions and discuss their implications.

Identification of the world share of publications by different regions shows that Africa produced only 68 945 publications over the 2000-2004 period or 1.8% of the World's publications. In comparison India produced 2.4% and Latin America 3.5% of the World's research. More detailed analysis reveals that research in Africa is concentrated in just two countries – South Africa and Egypt. These two counties produce just above 50% of the Continent's publications and the top 8 countries produce above 80% of the Continent's research.

Disciplinary analysis reveals that few African countries have the minimum number of scientists that may be required for the functioning of a scientific discipline. For example, in the field of ecology (a discipline necessary for environmentally friendly and sustainable development) only four countries (South Africa, Egypt, Nigeria and Kenya) produce 300 or more publications. Furthermore it is identified that African research over-emphasises "land

and primary resources" sciences (i.e. agriculture, ecology, geosciences and plant and animal sciences). While "land and primary resources" research occupies 26.4% of the research effort in Africa, the relevant figures for USA and India are 13.5% and 19.5% respectively.

Examination of the Continent's inventive profile, as manifested in patents, indicates that Africa produces less than one thousand of the world's inventions. Furthermore 88% of the Continent's inventive activity in concentrated in South Africa.

It may be interesting to speculate on the forces creating the big difference between research publications and patent contribution of the African Continent. International collaboration may be one such force. International donors and collaborators may be prepared to work with their African colleagues when the output will in the public domain but may refrain to do so when the output of the collaboration is proprietary knowledge. Innovation (patent) take-off may require a minimum capability threshold in research; high costs of patenting in USA and Europe and isolationism (the "we don't need to" syndrome) may be additional factors.

The African Continent is lagging substantially behind the rest of the World both in science and technology. However, a number of developmental reports argue that "science, technology and innovation underpin every one of the Millennium Development Goals" (UN Millennium Project, 2005) (p16) and hence science, technology and innovation become prerequisites for development.

In the same vein with the findings of the Latin American and Caribbean Workshop on STSD we argue that the main concern in the Continent should be how to develop research capacity across the board (and not in particular disciplines) and we suggest that African countries have to do two things simultaneously if they wish to do so. More research-trained students have to be graduated and graduates have to be enticed to remain in their countries.

Finally we suggest that it is time that the African Continent joins the rest of the world through the production of national and continental science and technology indicators reports. Such reports will be produced regularly and they will include additional indicators (e.g. educational statistics) monitoring progress in science and technology in the Continent. NEPAD and UNESCO may be the natural candidates to undertake such an effort.

References

Braun, T., Schubert S., Zsindely, S. (1997), "Nanoscience and nanotechnology on the balance", *Scientometrics* 38(2): 321-325

De Solla Price D (1975) "The Productivity of Research Scientists" In *Yearbook of Science and the Future, Encyclopaedia Britannica Inc*., University of Chicago, Chicago

DHEW (1970) "*Towards a Social Report*" Department of Health, Education and Welfare, University of Michigan Press, Ann Arbor

EC (1997) "*Second European Report on S&T Indicators 1997*" European Commission, Directorate General XII. Science, Research and Development, Brussels.

Fordis, B.J. and Sung, M.L. (1995) "How to avoid patent rejection", *Bio/Technology* 13, 42-43

Griliches, Z. (1990) "*Patent Statistics as Economic Indicators: A Survey*." Journal of Economic Literature, 28 :1661–1707 p1702

Hicks D., Breitzman T., Olivastro D. and Hamilton K. (2001) "The changing composition of innovative activity in the U.S. — a portrait based on patent analysis". *Research Policy,* **30**, 681–703.

King, D.A. (2004) "The scientific impact of nations", *Nature*, 430: 311-316.

NEPAD (2003) "Declaration of the First NEPAD Ministerial Conference on Science and Technology" 7th November 2003, Johannesburg, South Africa

NSB (2004) "*Science and Engineering Indicators-2004*", National Science Board, Arlington, VA: National Science Foundation.

OECD (2003), "*Main Science and Technology Indicators*", Organisation for Economic Cooperation and Development, Paris

Schubert A. and A. Telcs (1986) "*Publication Potential – An Indicator of Scientific Strength for Cross-National Comparisons*" Scientometrics **9**(5-6)**:**231-238

UN Millennium Project (2005) *"Innovation: Applying Knowledge in Development"* Task Force on Science, Technology and Innovation, EarthScan, London, Sterling VA

United Nations (2003) "*Science and Technology for Sustainable Development: A Latin American and Caribbean Perspective*" Latin American and Caribbean Regional Workshop on Science and Technology for Sustainable Development, Santiago, Chile, 5-8 March 2002, UN, Sustainable Development and Human Settlements Division

Diversity Measures and Network Centralities as Indicators of Interdisciplinarity: Case Studies in BionanoScience[1]

Ismael Rafols[*] and Martin Meyer[**]

[*] *i.rafols@sussex.ac.uk*
SPRU, University of Sussex, Brighton (United Kingdom)

[**] *m.s.meyer@sussex.ac.uk*
SPRU, University of Sussex, Brighton, United Kingdom; Steunpunt O&O Statistieken, Katholieke Universiteit Leuven, Belgium; Helsinki University of Technology, Institute of Strategy and International Business (Finland)

Abstract
Mapping and evaluating interdisciplinarity poses major challenges, due to its multidimensional character and its inherent conflict with categorisation methods. Here we develop indicators of interdisciplinarity as *cognitive diversity* that are not dependent on categorisation. To do so, we integrate in a conceptual framework measures of diversity used in ecology and economics, the measures of similarity predominant in bibliometrics and recently proposed indicators of interdisciplinarity based on social network analysis. We carry out two case studies in bionanoscience which illustrate how these indicators do capture the diversity of research topics engaged by an author or a publication, in contrast to category-based indicators that fail to do so. We suggest that these simple and ready-to-use indicators of cognitive diversity may be of potential importance in comparative studies of emergent scientific and technological fields, where claims of novelty and interdisciplinarity are rife but not always justified.

Keywords
interdisciplinary research; nanotechnology; nanoscience; diversity; indicators; social network analysis.

Introduction

In the policy discourse interdisciplinarity is often perceived as a mark of 'good' research: interdisciplinary research is seen as more successful at making breakthroughs and generating more relevant outcomes, be it in terms of innovation for economic growth or for social needs. This belief has led to the design of policies aimed at fostering interdisciplinarity, particularly in those fields, such as biotechnologies or nanotechnologies that are regarded as emerging through technological convergence.

However, the concept of interdisciplinarity and its variants (multi-, trans- and cross-) is problematic in many respects, if not plainly controversial (Weingart and Stehr, 2000). In the first place, because, given its polysemous and multidimensional nature (Sanz-Menéndez et al., 2001), there has been no agreement so far concerning its most pertinent indicators, or the appropriateness of categorisation methods based on disciplines (van Raan, 2000; Bordons, 2004, Rafols and Meyer, 2007). Second, although the etymology of inter-, multi-, trans- and cross-disciplinarity suggests that this is a property of research lying between, beyond or across various disciplines, interdisciplinarity is currently widely (and ambiguously) used to mean research spanning over a variety of areas, whether the areas are academic disciplines, technological fields and/or even industrial sectors. Given this use beyond the disciplinary boundaries, Glaser (2006) suggested that 'interdisciplinarity' as employed in science policy is a misnomer: *cognitive diversity* (in relation to either disciplines, specialties, technologies, industries, stakeholders, research fronts or specialties, etc.) would be a more appropriate label.

As a result of the problems highlighted above, policies fostering interdisciplinarity appear to be based more on conventional wisdom rather than on an empirical analysis of research practices. The aim of

[1] This research is supported by an EU postdoctoral Marie-Curie Fellowship to IR and the Daiwa Anglo-Japanese Foundation. We have benefited from comments by J . Gläser, L.Leydesdorff and various of our colleagues at SPRU, in particular B.R. Martin, A. Stirling, S. Katz and E. Steinmueller.

this investigation is to inform policy-making on the dynamics of emerging fields by providing simple measures that can help capture the intensity of interdisciplinarity in the wider sense of *cognitive diversity*.

Diversity is a concept used across a range of scientific fields, from ecology to economics and cultural studies, to refer to three different attributes of a system composed of different elements and/or categories (Stirling, 1998, 2007; Purvis and Hector, 2000):

- Variety: number of distinctive categories.
- Balance: evenness of the distribution of categories.
- Disparity: degree to which the categories examined are different from each other.

Table 1. Measures of diversity.

Notation	
Proportion of elements in category i:	p_i
Distance between categories i and j:	d_{ij}
Similarity between categories i and j:	s_{ij}
Indices	
Shannon-Wiener	$\sum_i p_i \ln p_i$
Herfindahl-Hirschmann[2]	$\sum_i p_i^2$
Disparity (or similarity)	$\sum_{ij(i\neq j)} d_{ij}$ (or $\sum_{ij(i\neq j)} s_{ij}$)
Stirling	$\sum_{ij(i\neq j)} d_{ij}\, p_i\, p_j$

Many bibliometric studies of interdisciplinarity have examined the *variety* and *balance* of disciplines as indicators of degree of diversity using pre-existing categories (see review by Bordons et al., 2004). Many other studies have used *similarity* measures between in order to visualise the relative 'position' of different scientific disciplines and fields (see review by Noyons, 2004; Boyack et al. 2005). Thus, different bibliometric studies on interdisciplinarity have indeed already looked into the three different aspects of diversity listed above. Other attempts to measure interdisciplinarity include: (i) van den Besselar and Heimeriks (2001), who used factor analysis on similarity measures to discriminate the interdisciplinary elements within a set (journals in their case) as those elements that do not fit into the latent classes represented by eigenvectors; (ii) very recently Leydesdorff (2007) who incorporated measures of centrality in social network analysis. The current investigation builds on previous research on mapping using similarity distances defined by co-occurrence (Noyons 2004) and Leydesdorff's introduction of social network analysis.

Here, we use Stirling's measure of diversity (see table 1) (Stirling, 1998, 2007) but later simplify it to disparity to avoid imposing categorisation. This measure is a sum over the *distances* between each pair of categories of a set with a weight proportional to the product of their *shares*. This form incorporates the various properties inherent in the concept of diversity: variety, balance and disparity. In order to apply this diversity measure, we need two previous requirements which will determine the final measure: (i) a categorisation method for partitioning the set; (ii) a definition of distance between categories.

Since our aim is to find a measure of cognitive diversity for a given set of publications, we would need in the first place to assign each publication to a cognitive category (e.g. a discipline or research specialty). However, as outlined above, there are several reasons for avoiding categorisation. In the first place, the only multidisciplinary and most widely used categorisation system is the one provided by ISI, which assigns each journal to one or more disciplines. This assignation may work at the journal level, but is very problematic at the paper level given the heterogeneous contents of many journals. In

[2] Herfindahl-Hirschman's index is equivalent to Simpson's index, used in biodiversity.

consequence, our first intention was to conduct cluster analysis in order to create self-organised categories for each bibliometric set, following the approach pursued by Schmidt and collaborators' (2006). However, clustering exercises result in extremely skewed size distributions (few large clusters, many smalls ones, see Schmidt, 2006, p.9), as one might expect from the power-law distributions of citation networks (Price, 1965). Due to these skewed distributions, the clustering approach is only feasible for large data sets and even then it is very sensitive on the cutting threshold set. The alternative we pursue here is to take each paper as its own category, with all the categories having the same weight p_i. This approach simplifies the measure of diversity to a measure of disparity or dissimilarity, i.e. of how different the various elements are: $\sum_{ij(i\neq j)} d_{ij}$, with *i,j* now being the indices for all the elements of the set. Although this description is mathematically more straightforward, we should like to emphasize that it is not a simplification. On the contrary, we may argue that it is a more complete description as it adds to the diversity measure *between* clusters we had, a measure of the diversity *within* clusters –which was assumed to be zero, but is not negligible in skewed distributions that include macro-clusters.

As it has been explained in many clustering exercises (e.g. Boyack et al., 2005, p.354) in order to measure the distances or the similarities between a given set of elements we need to make some choices. First, we must select a *context* (the e.g. a wider set of papers), that will provide the properties or attributes used to measure the distances. Second, we must choose which *attributes* will be used to compare the units and third we have to adopt a *functional form* for the similarity or dissimilarity matrix.

Network centralities as measures of diversity

Since in this study we are interested in mapping the breadth of knowledge sources a set of papers (e.g. publications by an author during a period), we select the co-occurrences of references as the *attribute* for comparison (bibliographic coupling). This choice implies that we select the references included in our initial set as the *context* of comparison. Then, we compute the similarity among all the papers of the context using as *functional form* the Salton's cosine of the bibliographic coupling (s_{ij} = number of references shared by papers *i* and *j* divided by the geometric mean of references in *i* and *j*). Finally, the distance between two papers is estimated using centrality measures of social network analysis. For an explanation of centrality measures in terms of bibliometrics, we refer to the recent article by Leydesdorff (2007).

In a first instance, we use directly the Salton's cosine value of bibliographic coupling as the similarity between two papers, which a number of studies have claimed to be preferable to others (e.g. see a detailed comparison in Boyack et al. 2005). The similarity measure over the whole network is then the sum of the similarity matrix, which is also the sum of the valued degree centrality for all vertices (S_{degree})[3].

$S_{degree} = (N*(N-1))^{-1} \sum_{ij(i\neq j)} s_{ij,}$ with s_{ij} = Salton's cosine of reference vectors

This measure has a notable downside: if two papers do not share any reference, their similarity is set to zero, irrespectively of their location in the wider context provided. In other words: if they don't share a reference, the distance between a paper in biophysics and one in biochemistry is assumed to be the same as between one in biophysics and one in sociology.

In order to take a wider context into account, we propose a second measure based on the length of the geodesic (the shortest path). Here the length of the geodesic d_{ij} is the minimum number of links or edges that are needed to connect papers *i* and *j* within the network of papers. Two papers are considered to be connected when their similarity value s_{ij} exceed a chosen threshold s_{min}[4]. The

[3] The measure gives us the network similarity instead of the diversity, because we sum over element-to-element similarities instead of distances.

[4] Notice that this procedure converts the s_{ij} matrix into a binary matrix, thus losing information about element-to-element distances.

diversity-dissimilarity measure is then the sum over all the length of geodesics for the connected papers. After normalisation this yields the average distance between vertices (papers) in the network. Given that this is approximately the inverse of the mean over all the vertices of **the closeness centrality,** we will choose this well-established index of social network analysis as the second indicator of similarity (S_{close})[5]:

$S_{close} = (N*(N-1))^{-1} \sum_{ij\ (i\neq j)} 1/d_{ij}$ with d_{ij} = Number of edges crossed to connect *i* and *j*

Here we set s_{min}= 0.05= 1/20, so that the measure can retain the connection between two short research papers with only 20 references each that share one (a typical case in *Nature*, *Science* letters), but at the same time dismiss as irrelevant a small number of shared references between review papers (typically with long reference lists). The normalisation of these indices of similarity/diversity to account for size effects is a complex issue that we aim to address in a future investigation (see discussion). For the purposes of this study, we have normalised the indices so that they take a value in the [0,1] interval and size does not have an effect on the minimum and maximum values.

Table 2. Schematic representation of the the theoretical framework used in this investigation.

Research tradition	Ecology →	Bibliometrics →	Social Network Analysis
Main concept	Diversity →	Similarity →	Centrality measures
Main formulations	$\sum_{ij(i\neq j)} d_{ij}\ p_i\ p_j$	$\sum_{ij(i\neq j)} s_{ij}$ where $s_{ij}=1/d_{ij}$ is Salton's cosine	Valued Degree Centrality $\sum_j s_{ij}$
			Closeness Centrality $\sum_j (N-1)/d_{ij}$

Here we present two case studies to test these measures against other measures of diversity used in previous studies. In particular we compare S_{degree} and S_{close}, with measures of diversity based on the number of categories, first using the journals that appear in the references (N_{jou}=number of journals) and second the widely used ISI subject categories (N_{cat}=number of categories). For these two categorisation methods, we compute the numerical richness (N_{cat}/N), a normalised Shannon-Wiener index[6] and the Herfindahl-Hirschmann index. When feasible, we also compute the betweenness centrality, which was proposed on empirical grounds by Leydesdorff (2007).

These two studies within the emerging field of bionanotechnology take two different units of analysis: first we compare single papers; second, the publication record of a researcher over various periods. We downloaded sets of papers from searches of ISI Web of Knowledge, computed the Salton's cosine of similarities with our own software, then visualised results and computed centrality measures using Pajek (Batagelj and Mvar, 2006), Bibexcel (Persson, 2006) and the statistical packet R (2006) (all freeware). Since the two cases presented are pilot studies based on previous interviews to the authors, they are made of very small publications sets (between 21 and 44 for case 1, and between 8 and 28 for case 2) –results need to be interpreted accordingly as indicative or exploratory. For detailed narratives of the case studies we refer to our previous publication (Rafols and Meyer, 2007).

[5]Closeness centrality for each vertex is defined as $C_i= (N-1)/\sum_j d_{ij}$, with d_{ij} being the length of the geodesic between *i* and *j*. By summing over i and dividing by N, we obtain the mean of closeness centrality, S_{close}.

[6] The Shannon index presented here is normalised with the logarithm of the number of papers in a set. This normalized index reflects the evenness or balance of the distribution.

Table 3. Measures of diversity in the reference set of publications.

		Salton's cosine similarity			Journal category			ISI subject category		
Paper	**N**	**$1/S_{degree}$**	**$1/S_{close}$**	**Betw**	**N_{Jou}/N**	**Shan**	**1/Herf**	**N_{Cat}/N**	**Shan**	**1/Herf**
Funatsu 1995	23	17.6	1.7	0.14	0.57	0.66	4.60	0.26	0.43	3.05
Kojima 1997	24	13.0	1.4	0.07	0.50	0.68	6.40	0.17	0.41	3.43
Ishijima 1998	51	23.5	1.8	0.04	0.37	0.61	6.19	0.12	0.36	3.43
Noji 1997	*21*	*39.1*	*2.3*	*0.35*	*0.52*	*0.72*	*7.41*	*0.24*	*0.45*	*3.60*
Yasuda 1998	31	24.8	1.8	0.07	0.45	0.69	8.66	0.13	0.38	3.33
Okada 1999	29	9.1	1.3	0.01	0.34	0.60	6.23	0.14	0.36	3.06
kkawa 2001	45	13.6	1.5	0.01	0.36	0.63	8.13	0.18	0.41	3.94
Sakakibara 1999	*26*	*32.7*	*2.1*	*0.11*	*0.46*	*0.67*	*6.58*	*0.23*	*0.46*	*4.02*
Burgess 2003	35	19.5	1.6	0.04	0.57	0.78	13.46	0.26	0.46	3.94
Tomishige 2000	44	9.4	1.4	0.01	0.36	0.65	9.20	0.14	0.39	3.82
Tomishige 2001	17	8.3	1.2	0.00	0.71	0.83	8.76	0.29	0.49	3.50
Yildiz 2004	19	14.6	1.6	0.02	0.63	0.79	8.80	0.53	0.67	5.88

Legend: N: Number of references in paper. Betw: Betweenness centrality. Shan : Shannon diversity. Herf : Herfindahl-Hirschmann diversity measure. $N_{jou:}$ Number or journals in references. $N_{cat:}$ Number of ISI categories in references. S_{degree} and S_{close} are the similarity measures as defined above. The form of the indices in the table is presented so that it increases its value with increasing diversity. The two papers highlighted are those presenting a higher diversity according to the centrality measures. Papers within the same box belong to the same project.

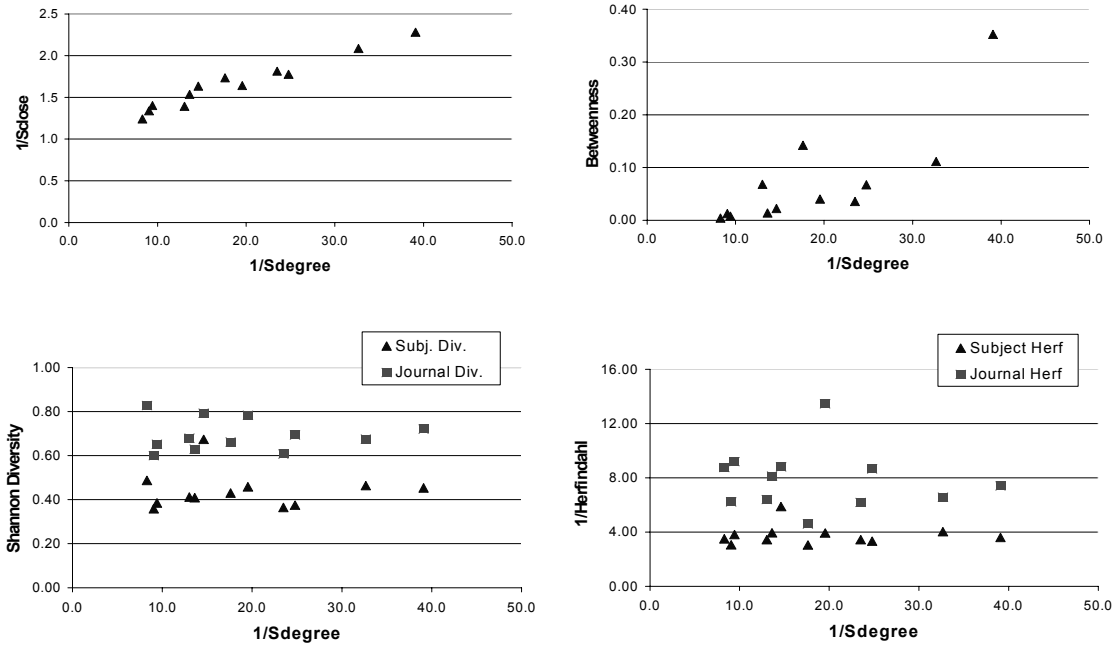


Figure 1 . Relation between the various diversity indices and $1/S_{degree}$ (based on data shown in Table 1). (We observe that there is a linear relation between $1/S_{close}$ and $1/S_{degree}$ and a positive but irregular correlation for Betweenness Centrality and $1/S_{degree}$. However, the more widely used Shannon (entropy) and Herfindalh diversity measures do not show a correlation with $1/S_{degree}$ –neither if computed from ISI subject categories, or using each journal as a category)

Case study 1: Diversity of single articles on molecular motors

In this case, we are interested in tracking the interdisciplinarity of single contributions in molecular motors research. We build on a previous paper investigation that carried out detailed cases studies on interdisciplinary practices in five research projects (Rafols and Meyer, 2007). It emerged from interviews that while all the cases used techniques and concepts from a variety of disciplines (they were similarly interdisciplinary in this respect), in some cases the project was the continuation of a well-established research tradition, while in others the project brought together different research traditions (the later are more interdisciplinary in terms of knowledge integration). The analysis here shows that while the new measures of diversity S_{degree} and S_{close} can track this difference in knowledge integration, category-based measures Shannon and Herfindahl fail to do so.

We assess the diversity of knowledge sources of one article by using its reference as the set of papers for comparison. This implies making two steps back: take an article, look the references, compute the similarities using the references of the references. The threshold for the computation of S_{close} was set at s_{min}=0.05 for all cases except Noji 1997, where in order to avoid the network splitting into two, it was lowered to 0.025. The results are presented in Table 3 and Figure 1. While we see an agreement between those measures of diversity based on reference similarity, there is no correlation between these measures and those based on journal and subject categorisation.

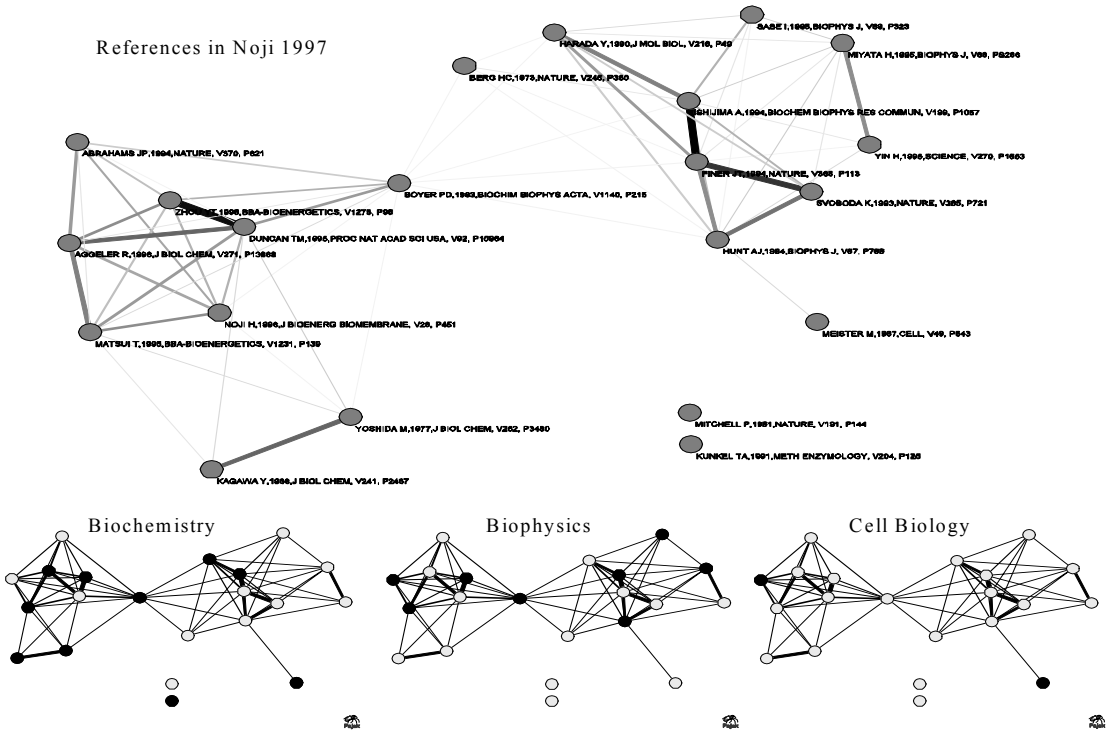


Figure 2. Bibliographic coupling of the reference set in Noji 1997 (threshold=0.025). (The graph illustrates that this seminal paper was based on contribution from two research communities. On the right hand side, the researchers working on linear molecular motors (myosin and kinesin). On the left, those working on bio-energetics of rotary motors (F_1-ATPase). The black circles in the figures at the bottom show the papers classified in different disciplines according to ISI subject classification. Both the left and the right clusters have papers from different disciplines)

In order to investigate this disagreement, we have first plotted the network of references (see figure 2 through figure 5). The measures of diversity S_{degree} and S_{close} can be understood by looking at the number and structure of the clusters formed through similarity links: the two clusters in Noji 1997's mean higher diversity than the only cluster in Ishijima 1998's. However, disciplinary categories are not correlated with the clustering: biochemistry and biophysics are equally found in the two clusters in

Noji 1997 or in the one cluster in Ishijima 1998. As a result, diversity measures based on ISI subject category do not correlate with number of clusters.

Second, we have also compared S_{degree} and S_{close} with betweenness centrality. Here we measured the betweenness centrality by placing the examined paper among the network formed by its own references. Figure 1 shows that these measures are only roughly correlated. Figure 5 compares two cases with high diversity in terms of $1/S_{degree}$ and $1/S_{close}$ but different betweenness. The comparison shows that betweenness is very sensitive to the overall structure of the network. If, as in the case of Sakakibara 1999 (right), the reference set is diverse but there are already other links between groups, then betweeness centrality has a much lower value.

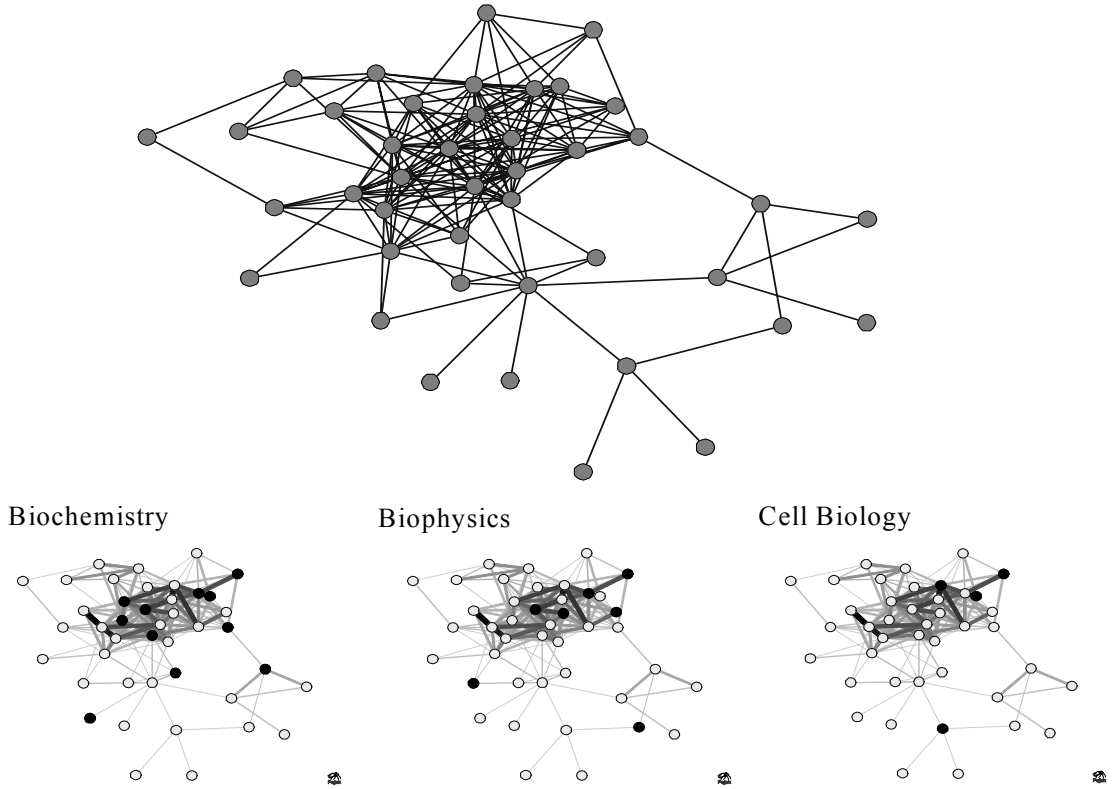


Figure 3. Bibliographic coupling of the reference set in Ishijima 1997 (threshold = 0.05). (This is a case of research related to various disciplines but belonging to a well-established line of research. Therefore, the vast majority of its references are related to each other, but they form a cluster where the incumbent disciplines are mixed. Black circles in the figures at the bottom show the papers classified in different disciplines according to ISI subject classification. Most of those not marked in any category fall into the ISI category of 'Multidisciplinary' journals)

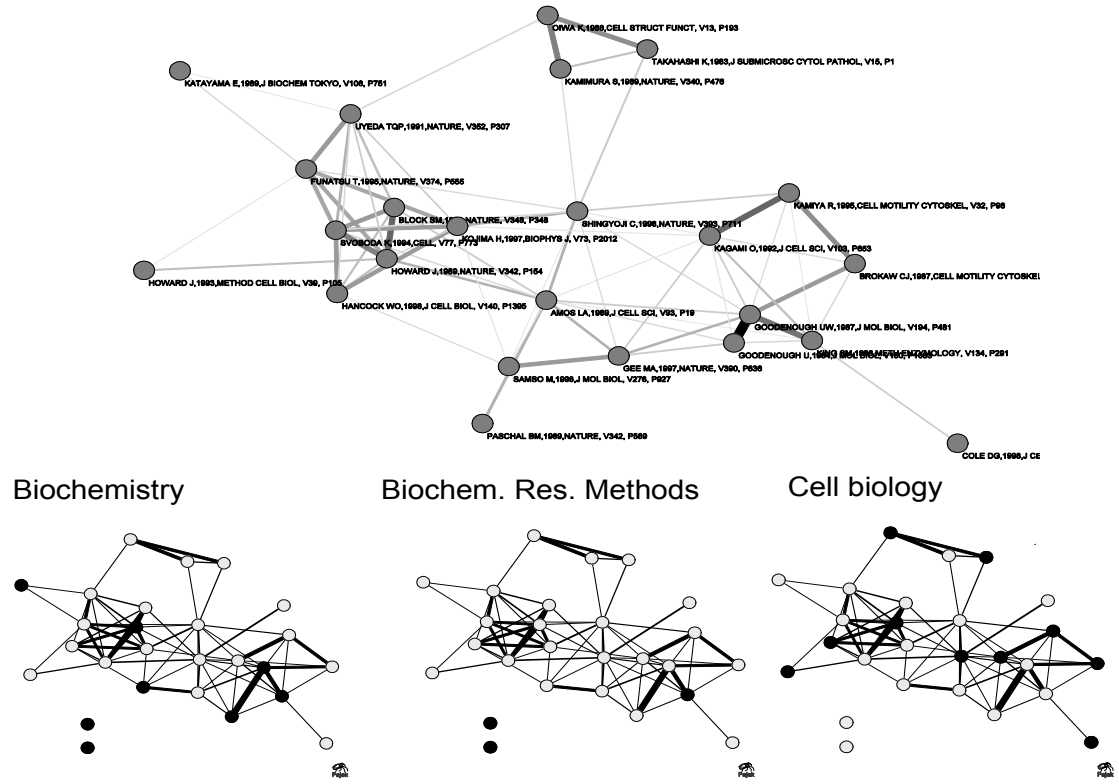


Figure 4. Bibliographic coupling of the reference set in Sakakibara 1999 (threshold = 0.05). (This set integrates literature from the single molecule detection and manipulation based on kinesin and myosin (left cluster), with studies on dynein (right) and dynein role in the axoneme (top). Rather than clear-cut clusters, the network has three or four fuzzy groups, what explains its diversity. The figures at the bottom, show, again, that these groups do not coincide with established categories. The black circles show the papers classified in different disciplines according to ISI subject classification)

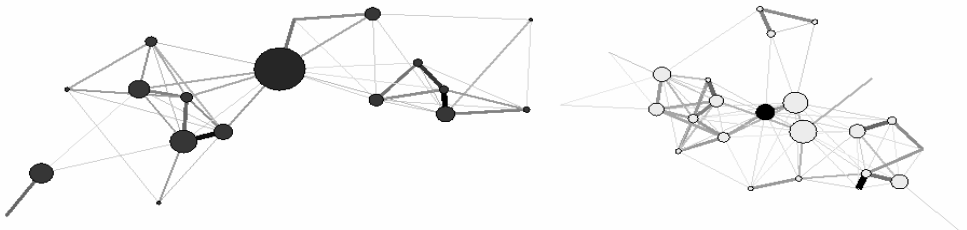


Figure 5. Betweenness centrality (size of nodes) of each of references in Noji 1997 (left) and Sakakibara 1999 (right), with the lines representing the bibliographic coupling (threshold = 0.05). (To the reference set, we have added a black circle in each figure representing the position of the article studied within their own reference set. We can see that this betweenness centrality (of the article) takes a very high value when the article is the only link between clusters (see left: Betw=0.35), but it quickly decreases if there are other links that were already making a brokerage role (see right: Betw=0.11))

Table 4. Measures of diversity in the publications of an author (TQP Uyeda).

Diversity		**Centrality measures**		**Journal category**			**ISI subject category**		
Period	**N**	**$1/S_{degree}$**	**$1/S_{close}$**	**N_{Jou}/N**	**Shan**	**1/Herf**	**N_{Cat}/N**	**Shan**	**1/Herf**
2001-2006	*28*	*23.4*	*2.9*	*0.50*	*0.75*	*10.89*	*0.29*	*0.53*	*4.77*
1996-2000	12	7.1	1.2	0.83	0.89	8.00	0.67	0.79	6.40
1991-1995	8	6.8	1.3	0.88	0.92	6.40	0.63	0.70	3.86
1983-1990	12	11.6	2.0	0.75	0.84	7.20	0.67	0.67	3.65
1983-2006	60	24.4	2.4	0.47	0.76	18.37	0.27	0.53	5.88

Legend: N: Number of references in paper. Betw: Betweenness centrality. Shan : Shannon diversity. Herf : Herfindahl-Hirschmann diversity measure. $N_{jou:}$ Number or journals in references. $N_{cat:}$ Number of ISI categories in references. S_{degree} and S_{close} are the similarity measures as defined above. The form of the indices in the table is presented so that it increases its value with increasing diversity. The two papers highlighted are those presenting a higher diversity according to the centrality measures. Papers within the same box belong to the same project.

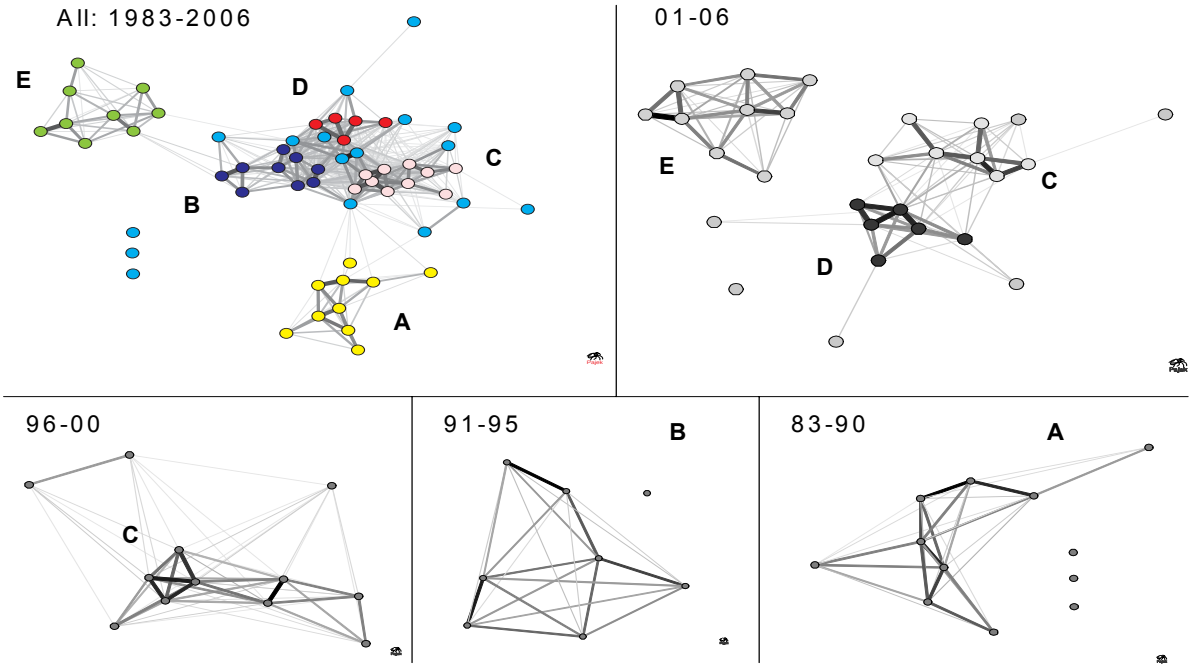


Figure 6. Bibliographic coupling of the publications of TQP Uyeda (1983-2006) (threshold = 0.05). (Islands were created using the island partition algorithm in Pajek (min-size =5, max-size= 10). In the periods 83-90, 91-95 and 96-00 the researcher focused mainly on one topic. However, since the late 1990s he starts branching into two topics of molecular motors (C: role of Myosin II in cytokinesis; D: conformational change in myosin) and later in 2000's into investigations of bioengineering applications of molecular motors (E))

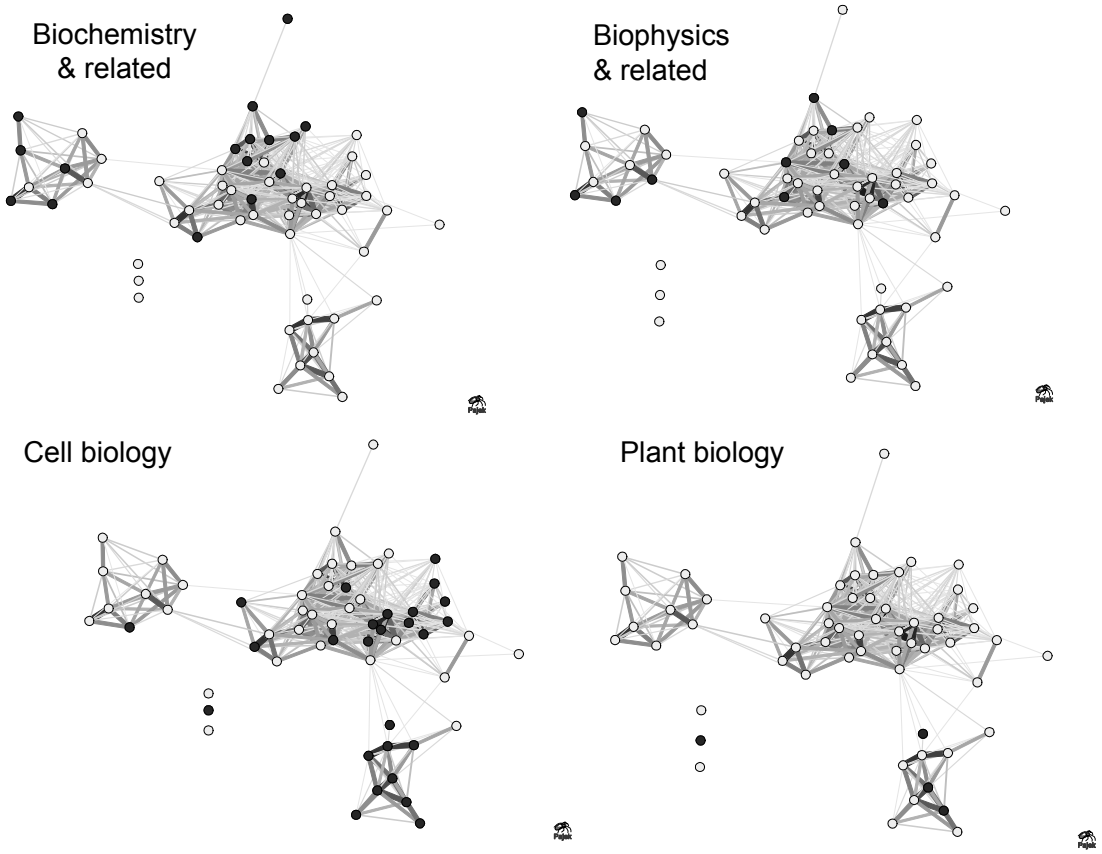


Figure 7. Disciplinary ascription of all TQP Uyeda publications according to ISI subject categories. (The links between nodes represent bibliographic coupling. Black circles correspond to publications in journals of a specific ISI subject category. Whereas the initial publications of Uyeda focused on cell biology and plant biology (bottom cluster), his latest contributions (left) and his work in the last 15 years (centre) can not be distinguished in terms of their disciplinary mix: both comprise some cell biology, and well as biophysics and biochemistry related papers. But the existence of two clearly differentiated clusters indicates that they are different cognitive territories –a result that was confirmed by an interview)

Case study 2: Diversity in the publication set of an author

In a second case studied, we aim to trace how diverse is the career of a given researcher. Here we traced the publications of TQP Uyeda, a researcher in molecular motors, from 1983 to 2006 and computed S_{degree}, S_{close} and Shannon and Herfindahl's diversities at different periods of his career[7] (see Table 3 and Figure 6). In each of the first three periods, Uyeda's research was linked through a common topic, which changed as he moved to new labs. However since 2000, he kept the focus he had since the mid-90's (on the role of myosin, a molecular motor, in cytokinesis), while combining in with papers in a related topic (conformational changes in myosin) and started a new line of research in the bioengineering side (applications of molecular protein on nanofabricated devices).

The indices S_{degree} and S_{close} are able to track this increase in diversity in the last period, whereas Shannon and Herfindahl's can not do it, because the categories used (journals and subject category) cannot make the distinction between distinct clusters that have similar journal or disciplinary mix. The relatively high values of $1/S_{degree}$ and $1/S_{close}$ are attributable to the fact that 3 out of 12 publications were completely unrelated to the others. Comparison of Figure 6 and 7 shows that there is a correlation between research topics and disciplinary categories: in two clusters cell biology is predominant, while in other two we observe a combination of biophysics and biochemistry.

[7] Since in this instance the object of study (the set of publications) is also the *context*, we could not think of a simple and meaningful procedure to compute, for a set, betweenness centrality.

Discussion and conclusions

In this paper we have presented a conceptual development which we think helps to understand the relation between measures of diversity used in ecology and economics (Shannon, Herfindahl and Stirling's; see Stirling, 1998, 2007), the measures of similarity predominant in bibliometrics (van Raan, 2000; Boyack et al., 2005) and recently proposed measures of interdisciplinarity based on social network analysis (Leydesdorff, 2007).

Building on these previous contributions, we have used network centrality measures S_{degree} and S_{close} as indicators of cognitive diversity for two case studies in bionanoscience. The case studies illustrate that the proposed S_{degree} and S_{close} do capture the variety of research topics in the references of a publication or pursued by an author, whereas category-based indicators fail to do so. This result is consistent with the conceptual insights on the dynamics of science in the 1970's (Mulkay, 1974; Small, 1977) which showed that science develops more in terms of ever-changing research fronts and specialties than as a structure of well established disciplines. As a consequence indicators based on categories cannot adequately map emergent fields.

Several aspects of the proposed indicators need further exploration. First, the scalability and field dependence of the diversity measures need to be investigated in order to make meaningful comparisons between sets of different size or field. Trial comparisons between similar sets of different sizes (not shown here) suggest that size dependence is very strong and needs to be taken into account. Moreover, we need extend the measure to other fields in order to have values of reference for what it is a high or low level of diversity. Second, the current method to construct the indicators captures the difference between publication of related research, but cannot make the distinction between: (i) elements that are relatively close but not directly related (e.g. two papers in different research specialties within the big area of biochemistry) and (ii) elements that very far apart (e.g. papers on protein conformational changes and Adorno's social thought). In order to measure this long-range distances, it would be necessary to extend the *context* for their computation to science-wide maps of the type devised by Boyack et al. (2005).

We suggest that the simple and ready-to-use indicators of cognitive diversity presented here may be of special use in comparative studies for evaluation and mapping of emergent fields, where claims of novelty and interdisciplinarity are rife and not always justified. In the policy making process of emerging fields, which by definition are out-of-discipline/field and attract special funding, indicators of interdisciplinarity such as the ones we have developed will be potentially important in order to make the distinction between established developments, conceivable future visions, and the politics and/or rhetoric of technological promise.

References

Batagelj, V. and Mrvar, A. (2006) *Pajek. Program forLarge Network Analysi.* http://vlado.fmf.uni-lj.si/pub/networks/pajek/ Accessed 13-11-2006.

Bordons, M., Morillo, F. and Gómez, I. (2004) Analysis of cross-disciplinary research through bibliometric tools. In Moed, H.F., Glänzel, W. and Schmoch, U. (Eds.) *Handbook of quantitative science and technology research.* Kluwer, Dordrecht, 437-456.

Boyack, K.W., Klavans, R. and Börner, K. (2005) Mapping the backbone of science. *Scientometrics* **64**(3), 351-374.

Glaser (2006) personal communication.

Mulkay, M. (1974) Conceptual displacement and migration in science: a prefatory paper. *Science Studies* **4**, 205-234.

Leydesdorff, L. (2007) "Betweenness Centrality" as an Indicator of the "Interdisciplinarity" of Scientific Journals, *Journal of the American Society for Information Science and Technology* (forthcoming).

Noyons, C. M. (2004) Science maps within a science policy context. In Moed, H. F. et al (eds), *Handbook of Quantitative Science and Technology Research*, Kluwer, Dordrecht, 237-255.

Purvis, A. and Hector, A. (2000) Getting the measure of biodiversity. *Nature* **405**, 212-219.

R Project for Statistical Computing (2006) http://www.r-project.org/. Accessed 13-11-2006.

Persson, O. (2006) Bibexcel. A tool-box programme for bibliometric analysis. http://www.umu.se/inforsk/Bibexcel/ Accessed 13-11-2006.

Price, D.J. de S. (1965). Networks of scientific papers. *Science* 149, 510–515.

Rafols, I. and Meyer, M. (2007) How cross-disciplinary is bionanotechnology? Explorations in the specialty of molecular motors. *Scientometrics* **70**(3), 633-650.

Sanz-Menéndez, L., Bordons and M. Zulueta, M.A. (2001) Interdisciplinarity as a multidimensional concept: its measure in three different research areas. *Research Evaluation* **10**(1), 47-58.

Schmidt, M., Glaser, J., Havemann, F. and Michael, H. (2006) A Methodological Study for Measuring the Diversity of Science. In *Proceedings International Workshop on Webometrics, Informetrics and Scientometrics & Seventh COLLNET Meeting*, Nancy (France).

Small, H.G. (1977) A co-citation model of a scientific specialty: a longitudinal study of collagen research. *Social studies of science* **7**(2), 139-166.

Stirling, A. (1998) On the economics and analysis of diversity. SPRU Electronic Working Paper. http://www.sussex.ac.uk/Units/spru/publications/imprint/sewps/sewp28/sewp28.pdf Accessed 01-04-2006.

Stirling, A. (2007) A General Framework for Analyzing Diversity in Science, Technology and Society. J. R. Soc. Interface. *Forthcoming*. Available at SPRU Electronic Working Paper, 156. http://www.sussex.ac.uk/spru/documents/sewp156.pdf Accessed 25-02-2007.

Van den Besselaar, P., and Heimeriks, G. (2001). Disciplinary, Multidisciplinary, Interdisciplinary: Concepts and Indicators. In Proceedings of the 8th International Conference on Scientometrics and Informetrics - ISSI2001, M. Davis & C. S. Wilson (Eds.). Sydney: University of New South-Wales, pp. 705-716.

van Raan, A.F.J (2000) The interdisciplinary nature of science: theoretical framework and bibliometric-empirical approach. In Weingart, P. and Stehr, N. (Eds), *Practising Interdisciplinarity,* University of Toronto Press, Toronto, pp. 66-78.

Weingart, P. and Stehr, N. (Eds). (2000) *Practising interdisciplinarity*. University of Toronto Press, Toronto.

Aggregation Properties of Relative Impact and other Classical Indicators: Convexity Issues and the Yule-Simpson Paradox

Suzy Ramanana-Rahary[*], Michel Zitt[**] and Ronald Rousseau[***]

[*]*suzy.ramanana@obs-ost.fr*
Observatoire des Sciences et des Techniques (OST), Paris (France)
[**]INRA-Lereco, Nantes (France)
[***]KHBO - Industrial Sciences and Technology Zeedijk 101, B-8400 Oostend (Belgium)

Abstract
Among classical bibliometric indicators, direct and relative impact measures for countries or other players in science are appealing and standard. Yet, as shown in this article, they may exhibit undesirable statistical properties, or at least ones that pose questions of interpretation in evaluation and benchmarking contexts. In this article, we address two such properties namely sensitivity to the Yule-Simpson effect, and a problem related to convexity. The Yule-Simpson effect can occur for direct impacts and, in a variant form, for relative impact, causing an apparent incoherence between field values and the aggregate (all-fields) value. For relative impacts, it may result in a severe form of 'out-range' of aggregate values, where a player's relative impact shifts from 'good' to 'bad', or conversely. Out-range and lack of convexity in general are typical of relative impact indicators. Using empirical data, we suggest that, for relative impact measures, 'out-range' due to lack of convexity is not exceptional. The Yule-Simpson effect is less frequent, and especially occurs for small players with particular specialisation profiles.

Keywords
Yule-Simpson Paradox; convexity; impact measures; research evaluation

Introduction

Typical dimensions of scientometric indicators are the geographic, the thematic and the time dimension. A simple indicator of publication or citation will be presented as an n-tuple, for instance: measure, player, field, period, etc. Similarly, a 'geographically bidimensional' indicator (such as co-authorship) will be represented by an n-tuple: measure, player, player, field, period, etc.

A global set, say all (i.e. "total world") publications collected over a certain period in a given database, can be partitioned according to the same criteria as used to construct the above-mentioned n-tuples. Typical examples are geographic and thematic breakdowns. Assuming counting methods that do not produce overlaps, for example fractional counting, the data can be displayed in a matrix in the form of a contingency table, player (i.e. region, or country) x field (or theme) in our example.

If [P] denotes the publication matrix then p_{ij} denotes the number of publications of country i in domain j. Similarly c_{ij} denotes the number of citations received by country i for publications in domain j. We further assume that [P] and [C] are M x N matrices (M rows, N columns), with the structure of a contingency table.

A number of classical bibliometric indicators are defined on this twofold data structure: direct indicators such as publication volume, citation volume, and the ratio of the latter to the former (direct impact); relative indicators such as world share of publication, world share of citation, and the ratio of the latter to the former (relative impact).

In this article we address some properties of classic bibliometric indicators which create difficulties of interpretation in an evaluation context. The two issues concern aggregation properties. The first one is the Yule-Simpson effect (or Simpson's paradox), a well-known issue in statistics when aggregating indicators in the form of 'mediant fractions'. The second issue is the general question of convexity, in particular in relation to relative impacts. Both issues can combine in the case of relative impacts, resulting in a severe form of non-convexity, with a reversal of results for field values and aggregate

values for a player, an apparently unsatisfactory and confusing outcome, particularly in an evaluation context.

The first section is devoted to the definition of some indicators. In the second one, we recall the Yule-Simpson paradox and show consequences of this effect on direct and relative impact measures, in the realm of classic citation indicators pioneered by Garfield (1967). In the next section, we address the property of convexity at the player-level. It is shown that the resulting 'out-range' of aggregate values can create interpretation difficulties. Section four reports some empirical examples of both effects. The article ends with a conclusion.

Classical indicators and impact measures

First, we introduce the publication matrix.

Table 1. A publication matrix [P]

Field Player	field j	all fields
player i	p_{ij}	p_{i+}
all players	p_{+j}	p_{++}

The total number of publications of country *i* is denoted as $p_{i+} = \sum_{j=1}^{N} p_{ij}$.

The total number of publications in field *j* is denoted as $p_{+j} = \sum_{i=1}^{M} p_{ij}$.

The total number of publications in the system under consideration is denoted as p_{++} with $p_{++} = \sum_{i=1}^{M} \sum_{j=1}^{N} p_{ij}$.

A similar notation is used for the corresponding citations and citation matrix: c_{ij} denotes the citations received by country i in field j using a consistent type of counting. The corresponding citation matrix [C] is shown in Table 2.

Table 2. The citation matrix [C] corresponding to the publication matrix shown in Table 1.

Field Player	field j	all fields
player i	c_{ij}	c_{i+}
all players	c_{+j}	c_{++}

$$c_{i+} = \sum_{j=1}^{N} c_{ij} \; ; \; c_{+j} = \sum_{i=1}^{M} c_{ij} \; ; \; c_{++} = \sum_{i=1}^{M} \sum_{j=1}^{N} c_{ij} \; .$$

Elementary forms of relative bibliometric indicators, i.e. those comparing individual data with the reference aggregate, are based on this structure. We will here choose the macro-level for our examples. We will consider the level player = 'country' and all players = 'world' for the geographic breakdown: the field = 'broad discipline' and all fields = 'all science' for the thematic breakdown.

A common type of relative measure is expressed in terms of *shares* over the aggregate of reference, here the world. The world share of publications of a country in a field, easily interpretable as 'market' share, is a quite common relative indicator. Denoting the share of country i in field j as $v_i(j)$ yields: $v_i(j) = p_{ij} / p_{+j}$ in the field j. We further denote p_{i+} / p_{++} by v_i **.** and observe that this is not equal to $v_{i,+} = \sum_{j=1}^{M} v_i(j) = \sum_{j=1}^{M} \frac{p_{ij}}{p_{+j}}$. The notation v_i**.** refers to the share of all publications of country i in the

total (world) number of publications. Table 3 shows the matrix V with the players' market shares as elements.

Table 3. Matrix [V] of players' market shares (publications) - by field

Field Player	field j	all fields
player i	$v_i(j)$	$v_{i\bullet}$
all players	1	1

A similar matrix for citation shares by field is denoted as [W]: its elements are $w_i(j)$.

Symmetrically, the all-fields share of publication of field j, for a particular player k, is: $x_j(k) = p_{kj} / p_{k+}$. The ratio p_{+j} / p_{++} is denoted as $x_{\bullet j}$. This leads to the matrix X as shown in Table 4. A similar matrix [Y] is drawn for the corresponding citations.

Table 4. Matrix [X] of a field's market shares in all science (publications) - by player

Field Player	field j	all fields
player k	$x_j(k)$	1
all players	$x_{\bullet j}$	1

Specialisation studies make use of various deviation/concentration measures. A well-known cell-level index, used in economics and scientometrics, is the Balassa index, also known under such names and variants as the revealed advantage index, the specialisation index or, in bibliometrics, the specialisation index (Balassa, 1965; Brusoni & Geuna, 2004; Grupp, 1998). In its simplest form, this index refers to the observed/expected (under independence hypothesis) production of country i in field j (by cell); or to the observed/expected share of the field j in the production of country i (by row); or else to the observed/ expected share of country i in the field j (by column). The following three forms are clearly equivalent:

$$a_{ij} = p_{ij} / (p_{+j}\, p_{i+} / p_{++}) = p_{ij}\, p_{++} / (p_{+j}\, p_{i+}), \text{ or}$$
$$a_{ij} = (p_{ij} / p_{+j}) / (p_{i+} / p_{++}), \text{ or}$$
$$a_{ij} = (p_{ij} / p_{i+}) / (p_{+j} / p_{++}).$$

The (a_{ij})-values can be brought together in a matrix [A]. Then we may note that using weighted averages leads to row and column sums equal to one.

$$a_{i\bullet} = \sum_j \frac{p_{+j}}{p_{++}}\, a_{ij} = \sum_j \left(\frac{p_{+j}}{p_{++}} \cdot \frac{p_{ij} \cdot p_{++}}{p_{i+} \cdot p_{+j}} \right) = \sum_j \frac{p_{ij}}{p_{i+}} = 1$$

$$a_{\bullet j} = \sum_i \frac{p_{i+}}{p_{++}}\, a_{ij} = \sum_i \left(\frac{p_{i+}}{p_{++}} \cdot \frac{p_{ij} \cdot p_{++}}{p_{i+} \cdot p_{+j}} \right) = \sum_i \frac{p_{ij}}{p_{+j}} = 1$$

We note that the $a_{i\bullet}$ and $a_{\bullet j}$ are written as convex combinations of the a_{ij}. The weighting scheme, based on a field's weight, is identical for all players, which is not the case for impact measures below. The scheme is analogous column-wise.

Table 5. Matrix [A] of activity index (publications)

Field Player	**field j**	**all fields (weighted averages)**
player i	a_{ij}	$a_{i\bullet}=1$
all players (weighted averages)	$a_{\bullet j}=1$	

These measures, or similar ones, play an important role in scientometric studies, and a similar index can be used on the matrix of citations ("attractivity index"). Formally equivalent ratios are used for studying probabilistic affinities on square matrixes of co-publication (Luukkonen et al., 1993; Tijssen et al., 1987; Zitt et al., 2000, 2003, 2005; see also Glänzel and De Lange, 1997).

Impact measures involve term to term relations between the two matrixes [C] and [P]. We define $s_{ij} = c_{ij} / p_{ij}$ as the direct impact of country i in discipline j. It is not field-normalised and completely depends on citation behaviour in the discipline, as mentioned before.

Table 6. Matrix [S] of direct impact

Field Player	**field j**	**all fields**
player i	s_{ij}	$s_{i\bullet}$
all players	$s_{\bullet j}$	$s_{\bullet\bullet}$

The marginal values in Table 6 are not defined as row-wise or column-wise sums of s_{ij} – values but as ratios of marginal values of C and P (see e.g. (Egghe & Rousseau, 1996)):

$$s_{i\bullet} = c_{i+} / p_{i+}$$
$$s_{\bullet j} = c_{+j} / p_{+j}$$

We observe that $s_{i\bullet}$ and $s_{\bullet j}$ are actually weighted averages:

- all fields impact for player i: the weight is the share of publication by field, in the actors' row:
$s_{i\bullet} = \sum_j c_{ij} / \sum_k p_{ik} = \sum_j (p_{ij}\, s_{ij}) / \sum_k p_{ik} = \sum_j (x_j(i)^* s_{ij})$.

- all actors' impact for field j: the weight is the share of publication by actor, in the field column:
$s_{\bullet j} = \sum_i c_{ij} / \sum_k p_{kj} = \sum_i (p_{ij}\, s_{ij}) / \sum_k p_{kj} = \sum_i (v_i(j)^* s_{ij})$.

A shortcoming of direct impact is its field-sensitivity due to variations of citation behaviour amongst fields, as observed in a vast amount of literature (Murugesan and Moravcsik, 1978, see also Garfield, 1998). A normalised form of impact is often used: the ratio to a field average for all players, usually termed relative impact. It alleviates the problem at the aggregation level considered, though sensitivity to heterogeneity at inferior levels remains. Equivalently, relative impact is the ratio of market share of citations and market share of publications.

We consider here the set of all world players (other sets of reference could be considered).

$$r_{ij} = (c_{ij} / p_{ij}) / (c_{+j} / p_{+j}) = s_{ij} / s_{\bullet j}$$
or
$$r_{ij} = (c_{ij} / c_{+j}) / (p_{ij} / p_{+j}) = w_j(i) / v_j(i)$$

For the world (all players) $r_{+j} = 1$.

The definition of r_{ij} holds at the aggregate level, all fields together:

$$r_{i\bullet} = (c_{i+} / c_{++}) / (p_{i+} / p_{++})$$

The indicator r_{ij} is sometimes termed relative impact (Egghe & Rousseau, 2002), but other terminology is used as well such as: relative citation rate (not to be confused with relative citation ratio RCR) or citation impact (Braun et al., 1985; Schubert and Braun, 1986; May, 1997). Similar forms are discussed by Van Raan (2004) and Vinkler (2003).

Table 7. Matrix [R] of relative impact

Field Player	**field j**	**all fields**
player i	r_{ij}	$r_{i\bullet}$
all players	$r_{\bullet j}$	$r_{\bullet\bullet}$

For further use we define the notion of being out of range.

Definition
A variable x obtained as a function of a set S of values $(x_k)_k$ is out of range for S if $x \notin [\min(x_k), \max(x_k)]$. In this case x can never be written as convex combination of the $(x_k)_k$.

The Yule-Simpson (in short: YS) paradox and impact measures
The Yule-Simpson paradox (usually termed Simpson's paradox) (David & Edwards, 2001) emphasizes counter-intuitive results in statistical aggregations. The paradox refers to the fact that results of comparisons between groups are reversed when groups are combined. Let us consider the following classical example[1], (Table 8) where Ann and Bob are in a competition involving two tests. The outcome in each cell is the number of successes (numerator) over the number of cases (denominator).

Table 8. Ann versus Bob

	Test 1	**Test 2**	**Total**
Player 1 (Ann)	60 / 100	1 / 10	61 / 110
Player 2 (Bob)	9 / 10	30 / 100	39 / 110

It appears that the two sets of data separately support a certain hypothesis, but, considered together, support the opposite hypothesis. In the first test, Ann ($S_A(1) = 60\%$) does not reach Bob's success rate of 90% (= $S_B(1)$). In the second test Ann managed 10% ($S_A(2)$) while Bob achieved 30% ($S_B(2)$). On both occasions Bob is more successful than Ann. But if we combine the two tests, we see that Ann ($S_A = 61/110 = 55\%$) outperforms Bob ($S_B = 39/110 = 35\%$)!

The comment about this example makes it clear that this so-called 'paradox' has simple mathematical roots. If $S_B(1) > S_A(1)$ and $S_B(2) > S_A(2)$ our intuition tells us that S_B must be greater than S_A. However, if different weights are used to form the overall score then this intuitive feeling may be incorrect. Here the first test is weighted 100/110 for Ann and 10/110 for Bob while the weights are reversed on the second test:

$$S_A = 100/110\ S_A(1) + 10/110\ S_A(2)$$
$$S_B = 10/110\ S_B(1) + 100/110\ S_B(2)$$

In other words, the global score of A is a weighted average of his score in the two tests, similarly for B. The implicit weighting scheme however, is different among players. Real and fictitious examples are reported in many domains of hypothesis testing or score aggregation ("batting paradox" on base-ball scores), some showing that the Yule-Simpson effect may have particularly severe consequences.

Mediant fractions and graphical interpretation

[1] Found in several sources online, e.g. Chemical Daily Encyclopedia

The outcome of the Yule-Simpson paradox is related to aggregation in the form of so-called mediant fractions. Mediant fractions are aggregates from component fractions obtained by respectively summing numerators and denominators of each component, which is the scheme applied in the situation described above. For two component fractions a/b, c/d, as in the Ann-Bob case, the mediant fraction is defined as (a+c) / (b+d). The mediant-fraction aggregation has a simple graphical interpretation in terms of vectors sums. Fig.1 illustrates the Ann-Bob example.

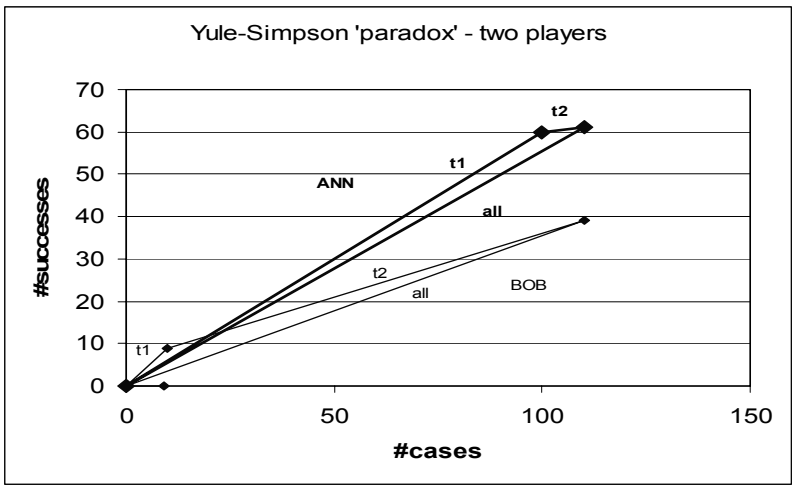


Figure 1. Yule-Simpson between two players - the lower triangle represents Bob, the higher triangle Ann. The indicator is featured by the slope of segments representing the tests t1 and t2, and the sum value. While slopes are superior for Bob both on t1 and t2, the slope for (t1+ t2) is superior for Ann.

Application to direct impact

The scheme above can be extended to a bibliometric indicator exhibiting a similar quotient form such as direct impact. Aggregate direct impacts are defined from impacts of individual articles, allowing any intermediary grouping, for example at the discipline level:

$$s_{i\bullet} = c_{i+} / p_{i+} = (\sum_j c_{ij}) / (\sum_j p_{ij}).$$

The boxes in table 9 represent the direct impact measure, with citation volume in the numerator and publication volume in the denominator. Research, involving individual or institutional players, replaces Ann and Bob, and disciplines 1 and 2 stand for tests 1 and 2. We add a row for 'All players' (here the two players) with the same definition of aggregate impact.

Table 9. Direct impact measure with citation volume in the numerator and publication volume in the denominator

Direct impact s_{ij} (fractions)	**Discipline 1**	**Discipline 2**	**Total all disciplines**
Player 1	60 / 100 = 0.60	1 / 10 = 0.10	61 / 110 = 0.55
Player 2	9 / 10 = 0.90	30 / 100 = 0.30	39 / 110 = 0.35
All Players	69 / 110 = 0.63	31 / 110 = 0.28	100 / 220 = 0.45

Application to relative impacts

The example above also produces an inversion for relative impacts. If, instead of comparing Player 1 and Player 2, we compare each player separately to their aggregate 'All players', we see that in the above example: Player 1 score < All players score on Disciplines 1 and 2, i.e. the world relative impact of Player 1 is inferior to unity, but Player 1 > All players for All disc, i.e. its world relative impact is superior to unity, and conversely for Player 2. Numerical values of relative impacts are given in table 10.

Table 10. Relative impacts expressed as fractions.

Relative impact r_{ij}	**Discipline 1**	**Discipline 2**	**All disciplines**
Player 1	(60/100) / (69/110) = 0.96	(1/10) / (31/110) = 0.35	(61/110) / (100/220) = 1.22
Player 2	(9/10) / (69/110) = 1.43	(30/100) / (31/110) = 1.06	(39/110) / (100/220) = 0.78
Two players	(69/110) / (69/110) = 1.00	(31/110) / (31/110) = 1.00	(100/220) / (100/220) = 1.00

Let us put this in a general framework. The Yule-Simpson paradox occurs if Table 11 is given, together with the requirements that A/U < C/W and B/V < D/X while (A+B)/ (U+V) > (C+D)/ (W+X).

Table 11. General framework for the Yule-Simpson paradox

	Discipline 1	**Discipline 2**	**All disciplines**
Player 1	A/U	B/V	(A+B)/(U+V)
Player 2	C/W	D/X	(C+D)/(W+X)
All players	(A+C)/(U+W)	(B+D)/(V+X)	(A+B+C+D)/(U+V+W+X)

Note that A, B, C, D, U, V, W and X are given. Not just the fractions, as otherwise by bringing the fractions to a common denominator, the paradox could never happen. Concretely: if U = V = W = X and A/U < C/W (= C/U) and B/V (= B/U) < D/X (= D/U) then (A+B)/(U+V) = (A+B)/2U < (C+D)/(W+X) = (C+D)/2U (and not the opposite inequality!).

Proposition 1
The YS effect is present for the pair (Player 1, Player 2) for direct impact, if and only if it is also present for relative impact.
The proof follows immediately from the fact that if in an equality both sides are multiplied or divided by the same positive number, the inequality stays invariant.

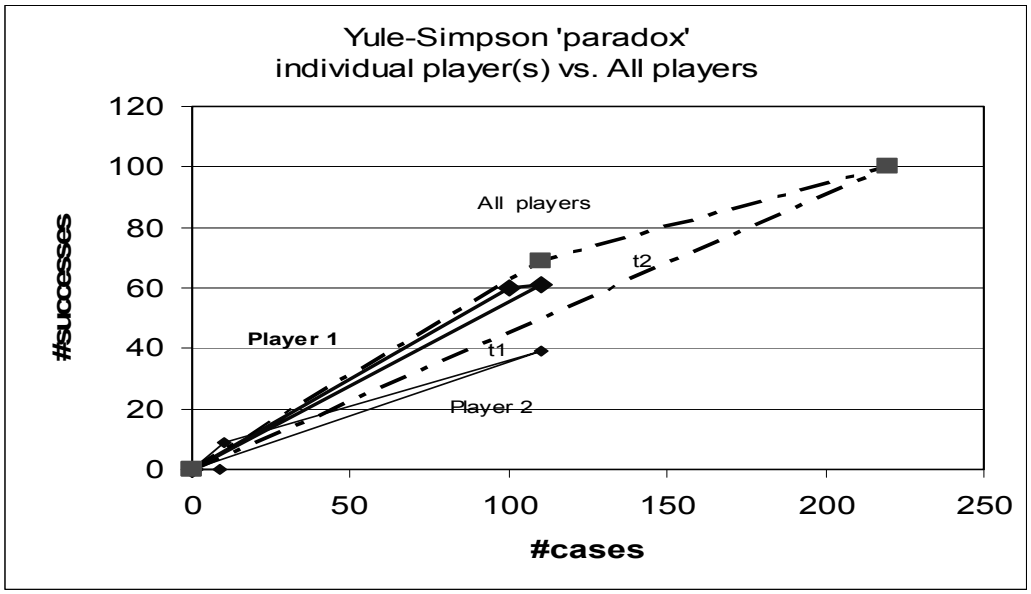


Figure 2. Yule-Simpson between individual player(s) and the sum All players. (Disciplinary values of impact (slope of segments) are superior to the reference 'All Players' for Player 1 (Ann) while all-disciplines value is inferior. In other words, disciplinary values of relative impact are larger than 1 for Player 1, and all-disciplines value is smaller than 1. The situation is reversed for Player 2 (Bob))

Proposition 2
The YS effect for the pair (Player1, Player2) is present if and only if it is present for the pair (Player1, Both Players).
This follows immediately from the fact that, for positive A, C, U and W

$$\frac{A}{U} < \frac{A+C}{U+W} \Leftrightarrow \frac{A}{U} < \frac{C}{W}$$

Proposition 3
If, in a general scheme with n players (n ≥ 2) the YS effect is present for the pair (Player 1, 'All players') on direct impact and hence on relative impact (by Proposition 1), then for Player 1 the aggregate value of relative impact 'All disc.' is out of the range of the disciplinary values (leading to a manifestation of non-convexity). Moreover the unit value lies between the global value and the group of individual disciplinary values. Proof is immediate: the relative impact of 'All players' (quotient of direct impact of Player 1 to direct impact of 'All players') is by construction equal to one. Then if YS effects exist for direct impact:
- either disciplinary values of relative impact of A are below 1 and the 'All disc.' value above 1
- either disciplinary values of relative impact of A are above 1 and the 'All disc.' value below 1.
In other words, the field-level results and the aggregate-level results are opposite with respect to unity. 'Good' indicators (relative impact above 1) in every field are translated into 'bad' indicators (relative impact below 1) at the global level, and conversely.

Note 1
The reciprocal statement of proposition 3 is not true. For a player, the 'out-range' of aggregate value may happen in absence of YS effect between Player 1 and 'All players'. What the YS effect produces is a severe form of out-range resulting in a particularly uncomfortable outcome, with disciplinary and aggregate value on each side of the unit value. In the following we establish the convexity of direct impact measures and the non-convexity of relative impact measures.

Note 2
Indicators using the same weighting scheme for all players are not subject to theYS effect.

Convexity of impact measures

Direct impact
Convexity is a straightforward issue for direct impact. As shown above, the direct impact at the 'All Disciplines' level may be written in the form of a weighted average of all direct impacts at the disciplinary level. This satisfies convexity at the series for this particular player.

$$s_{i\bullet} = \sum_j x_j(i).s_{ij} \text{ where } \sum_j x_j(i) = \sum_j \frac{p_{ij}}{p_{i+}} = 1$$

Similarly, for a discipline, 'All players' impact may be written in the form of a weighted average of direct impacts of each player, and convexity is satisfied.

$$s_{\bullet j} = \sum_i v_i(j).s_{ij} \text{ where } \sum_i v_i(j) = \sum_i \frac{p_{ij}}{p_{+j}} = 1$$

Relative impact

We consider the ratios of a country's share in citations over this country's share in publications, considered per domain and with respect to this domain, i.e. the ratio r_{ij} defined above. Mathematically, we want to consider the following ratios (with country *a* fixed):

$$r_{aj} = \frac{y(j \mid i=a)}{x(j \mid i=a)} = \frac{\frac{c_{aj}}{c_{a+}}}{\frac{p_{aj}}{p_{a+}}} = \frac{\frac{c_{aj}}{\sum_{k=1}^{N} c_{ak}}}{\frac{p_{aj}}{\sum_{k=1}^{N} p_{ak}}}$$

and would like to obtain a convex decomposition of $r_{a.}$ as a function of the r_{aj}. We were, however, not able to find such a convex combination, although a general linear combination with non-negative coefficients is possible.

Proposition

$$r_{a.} = \sum_{j=1}^{N} f_{aj}.r_{aj} \quad where\ f_{aj} \geq 0$$

Proof.

$$r_{a+} = \frac{w_{a.}}{v_{a.}} = \frac{\frac{c_{a+}}{c_{++}}}{v_{a.}} = \sum_{j=1}^{N}\left(\frac{1}{v_{a.}}.\frac{c_{aj}}{c_{++}}\right) = \sum_{j=1}^{N}\left(\frac{c_{+j}}{c_{++}}.\frac{1}{v_{a.}}.\frac{c_{aj}}{c_{+j}}\right) = \sum_{j=1}^{N}\left(y_{.j}.\frac{1}{v_{a.}}.w_{aj}\right)$$

$$r_{a+} = \sum_{j=1}^{N}\left(y_{.j}.\frac{1}{v_{a.}}.\frac{v_{aj}}{v_{aj}}.w_{aj}\right) = \sum_{j=1}^{N}\left(y_{.j}.\frac{v_{aj}}{v_{a.}}.\frac{w_{aj}}{v_{aj}}\right) = \sum_{j=1}^{N}\left(y_{.j}.\frac{v_{aj}}{v_{a.}}.r_{aj}\right) = \sum_{j=1}^{N}\left(y_{.j}.a_{aj}.r_{aj}\right) = \sum_{j=1}^{N}\left(f_{aj}.r_{aj}\right)$$

where, clearly, the coefficients

$$f_{aj} = y_{.j}.a_{aj} = \frac{c_{+j}}{c_{++}}.\frac{v_{aj}}{v_{a.}} = \frac{c_{+j}}{c_{++}}.\frac{p_{aj}/p_{+j}}{p_{a+}/p_{++}} \geq 0$$

Note that $\sum_{j=1}^{N} f_{aj} = \sum_{j=1}^{N} \frac{c_{+j}}{c_{++}}.\frac{p_{aj}/p_{+j}}{p_{a+}/p_{++}}$ is, in general, not equal to 1.

Hence, 'out-range' of aggregate values may happen. A special case of out-range, as seen in the previous section, takes place when an YS effect exists for direct impact between a player and the All-players aggregate.

Other classic indicators

World shares. If we take these basic indicator matrixes, the relation between individual cell values and marginal values is generally straightforward from the definition. In the original matrix [P] (and similarly for [C]) the marginal values p_{+j} , p_{i+} , p_{++} are by definition straight sums over cells. The same is true by definition, column-wise, for players' shares in fields; or row-wise, for field shares for players.

Activity indexes. As seen above, activity indexes for a player (rows) are in the form of weighted averages, with the same scheme for all players. Convexity is satisfied. The same is true column-wise (in a given field).

Empirical evidence

One thing is the theoretical possibility of out-range, another thing is the practical incidence of such situations. Do they frequently occur in "real indicators"? It appears that cases where the property of non-convexity results in 'out-range' of indicators values are not exceptional.

For example, using a two disciplinary classification (8 large academic disciplines, then aggregated into 3 macro-disciplines (biosciences/ physical sciences/ engineering-mathematics[2]) we analysed all countries appearing in the WoS (SCI expanded, OST version) in the period 1999-2004. We limited our test to immediate impact (2-years including publication year) and medium-term impact (5-years, incomplete window for the last period), calculated on a fractional counting option. Incomplete patterns, countries not present in all disciplines for a particular year and/or with zero citations in one or several disciplines, were not considered.

We observed the occurrences of out-range situations (each country vs. world average) and YS effects (country pairs).

Table 12. Relative impact: out-range of the aggregate ('All Disc.') value

disciplines	Citation window	# couples*	Out-range cases	Inversion cases
3 macro disc	Immediate (2 y)	563	73 (13%)	1 (0.2%)
	Med.-term (5y)	642	90 (14%)	0
8 disc	Immediate (2 y)	450	2 (0.4%)	0
	Med.-term (5y)	392	2 (0.4%)	0

* Country-year couples present in all disciplines at the level considered, and with non zero citations in every discipline on the citation window considered

Let us take the example of the three macro-disciplines breakdown, with immediate citation window: the out-range effect was found in 73 year-country couples. Large mainstream countries are not present, but smaller ones like Finland, Korea and Sweden are present in some years. Most countries are third world or emerging countries presenting strong degrees of specialization, a situation favourable to out-range configuration (see Fig.3).

Severe out-range: inversion of performance

At the macro-level, this situation, which corresponds to a Yule-Simpson effect on the Player-All players pair, reveals itself to be rather exceptional, and refers only to a single couple country-year (Panama, 2002) for the shorter citation window. With incomplete patterns (e.g. only two macro-disciplines), a few more cases occur, for very small countries, both on short and long citation windows. On the 8-disciplines breakdown, no case is found with complete patterns, very few again with incomplete patterns.

Out-range and Specialization

We calculated standard deviations of renormalized activity index for each country along disciplines, as a measure of disciplinary “specialization”: the specialization will be very low for a country evenly present in all disciplines, and very high for a country heavily focusing on a few disciplines. In Figure 3, specialization is ranked by classes (approximately deciles) in descending order from left to right. It appears that the more specialised the country, the higher the chances of out-ranging. As specialization tends to grow inversely with the output of systems, out-ranging tends to be found in smaller units as observed above.

[2] The 'life science' category groups fundamental biology, medical research, applied biology; the 'physical sciences' category groups chemistry, physics and earth & space.

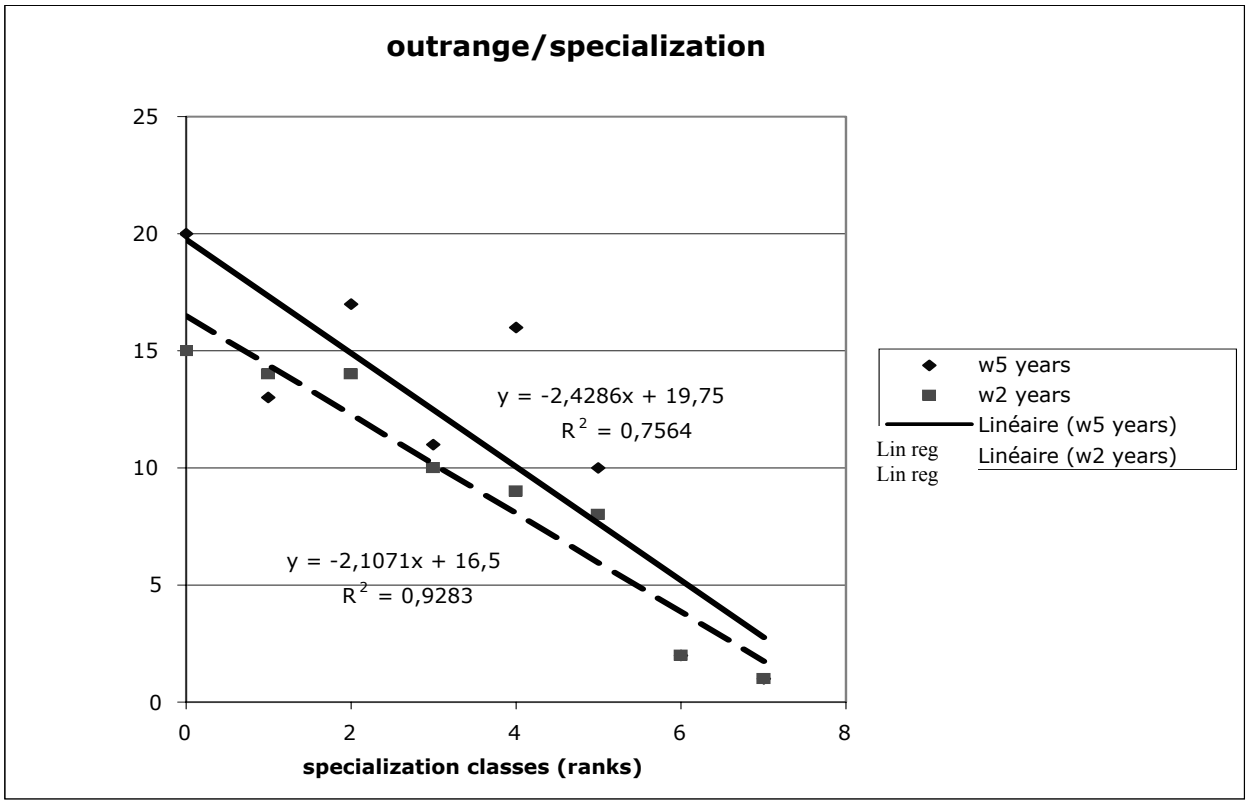


Figure 3. Relation between the out of range phenomenon and specialization

Yule-Simpson between pairs of players
547 cases of YS effect (triplets year country-country) were recorded. Again the probability for a country to be associated with a YS effect is strongly dependent on its specialisation level.

Conclusion
We have shown that some classical bibliometric indicators, direct and relative impact measures for countries or other players in science, may exhibit undesirable statistical properties, or at least properties that pose questions of interpretation in evaluation and benchmarking contexts. We have addressed two such properties, namely the 'Simpson's paradox', and the problem of convexity. The Y-S effect can occur for direct impacts and, in a variant form, for relative impact, causing an apparent incoherence between field values and the aggregate (all-fields) value. For relative impacts, it may result in a severe form of 'out-range' of aggregate values, where a player's relative impact shifts from 'good' (> 1) to 'bad' (< 1), or conversely. Out-range and lack of convexity in general are typical of relative impact forms. Using empirical data, we suggest that, for relative impact measures, 'out-range' due to lack of convexity is not exceptional. The Yule-Simpson effect is less frequent, and occurs especially for small players with particular specialisation profiles.

References
Balassa, B. (1965). Trade Liberalization and 'Revealed' Comparative Advantage. *Manchester School, 33*, 99-123.
Braun, T., Glänzel, W. & Schubert, A. (1985). Scientometric indicators. A 32 country comparison of publication productivity and citation impact. Singapore: World Scientific Publishing Co.
Brusoni, S. & Geuna, A. (2004). Specialisation and Integration: Combining patents and publications data to map the 'structure' of specialised knowledge. In: W. Glänzel, H. Moed & U. Schmoch (Eds.), *Handbook of Quantitative Science and Technology Research* (pp. 733-758). Dordrecht: Kluwer Academic Publishers.
David, H. A. & Edwards, A. F. W. (2001). Yule's Paradox (Simpson's paradox). In *Annotated Readings in the History of Statistics* (pp. 137-140): Springer.
Egghe, L. & Rousseau, R. (1996). Average and global impact of a set of journals. *Scientometrics*, 36, 97-107.
Egghe, L. & Rousseau, R. (2002). A general frame-work for relative impact indicators. *Canadian Journal of Information and Library Science- Revue canadienne des Sciences de l'Information et de Bibliothéconomie, 27*, 29-48.

Garfield, E. (1967). Primordial Concepts, Citation Indexing and Historio-bibliography. *Journal Library History, 2*, 235-249.

Garfield, E. (1998). Random thoughts on citationology. Its theory and practice - Comments on theories of citation? *Scientometrics, 43*, 69-76.

Glänzel, W. & De Lange, C. (1997). Modelling and measuring multilateral co-authorship in international scientific collaboration. Part I. development of a new model using a series expansion approach. *Scientometrics, 40*, 594-604. Part II. A comparative study on the extent and change of international scientific collaboration links. *Scientometrics, 40*, 605-626.

Grupp, H. (1998). *Foundations of the economics of innovation: Theory, measurement and practice*. Cheltenham: Elgar.

Luukkonen, T., Tijssen, R. J. W., Persson, O. & Sivertsen, G. (1993). The measurement of international scientific collaboration. *Scientometrics, 28*, 15-36.

May, R. M. (1997). The scientific wealth of nations. *Science, 275*, 793-796.

Murugesan, P. & Moravcsik, M. J. (1978). Variation of the nature of citation measures with journal and scientific specialties. *Journal of the American Society for Information Science, 29*, 141-155.

Schubert, A. & Braun, T. (1986). Relative indicators and relational charts for comparative assessment of publication output and citation impact. *Scientometrics, 9*, 281-291.

Tijssen, R. J. W., De Leeuw, J. & van Raan, A. F. J. (1987). Quasi-correspondence analysis on scientometric transaction matrices. *Scientometrics, 11*, 351-366.

van Raan, A. F. J. (2004). Measuring Science. In: W. Glänzel, H. Moed & U. Schmoch (Eds.), *Handbook of Quantitative Science and Technology Research* (pp. 19-50). Dordrecht: Kluwer Academic Publishers.

Vinkler, P. (2003). Relations of relative scientometric indicators. *Scientometrics, 58*(3), 687-694.

Zitt, M., Bassecoulard, E. & Okubo, Y. (2000). Shadows of the past in international cooperation: Collaboration profiles of the top five producers of science. *Scientometrics, 47*, 627-657.

Zitt, M., Ramanana-Rahary, S. & Bassecoulard, E. (2003). Correcting glasses help fair comparisons in international science landscape: Country indicators as a function of ISI database delineation. *Scientometrics, 56*, 259-282.

Zitt, M., Ramanana-Rahary, S. & Bassecoulard, E. (2005). Relativity of citation performance and excellence measures: From cross-field to cross-scale effects of field-normalisation. *Scientometrics, 63*, 373-401.

Distributions of Hirsch-Index and G-index: An Empirical Study

I.K. Ravichandra Rao*

*ravichandra.rao@gmail.com
Documentation Research and Training Centre, Indian Statistical Institute, Bangalore 560 059 (India)

Abstract
Empirical distributions of h-index and g-index are studied. It has been observed that exponential distribution fits very well for the observed data on h and g index. It has also been observed that there exists a linear relation between the g-index and the total number of citations received.

Keywords
H-index; G-index; exponential distribution function

Introduction

Hirsch in his paper entitled "An index to quantify an individual's scientific research output" (http://arxiv.org/abs/physics/?0508025) proposed an index called h-index. It is defined as the number of papers (h) with citation number higher or equal to h. i.e., "a scientist has index h, if h of his/her N papers have at least h citations each and the other (N-h) papers have no more than h citations." He has further argued that h is preferable to other single parameter criteria commonly used to evaluate scientific output of a researcher – such as, total number of papers, total number of citations, citations per paper, number of significant papers, number of citations to each of q most cited papers, etc. Each of these parameters has its merits and demerits. The h-index on the other hand measures the broad impact of individual's work. He has also further argued "for a given individual, h increases linearly with time".

Since then many have worked in this area. For instance, Braun, Glanzel and Schubert (http://www.the_scientist.com/article/display/15890/) applied this concept of computing h-index for journals. They compared the ranks of journals based on impact factors with that based on h-index. They suggested that an h-index would advantageously supplement journal impact factors at least in two aspects – it is robust and it combines the effect of quality and quantity. In their study Bornmann and Daniel (2006) observed: "on an average the h-index for successful applications for post-doctoral research fellowships was consistently higher than for non successful applicants". Batista and others (2006) argued that "although h-index is robust in several ways but yet strongly dependent on the research field." They then proposed a complementary index $h_1 = h^2 / N_a$, where N_a is the total number of authors in the considered h papers. van Raan (2005) in his study on "Comparison of the h-index with standard bibliometric indicators and with peer judgment for 147 chemistry research groups" presented characteristics of the statistical correlation between the h-index and several standard bibliometric indicators (SBI) as well as results of peer review judgment. He observed that both h-index and SBI relate in a comparable way with peer judgment. However, he also found that for smaller groups with fewer citations, the SBI appear to be a more appropriate measure of research performance.

Egghe and Rousseau (2006) in their paper on " An informetric model for the h-index", extended the definition of h-index to the general framework of Information Production Processes", using a source-item terminology. They have shown that in each practical situation an IPP always has a unique h-index. In the context of Lotkain model, they have also highlighted the relation between h and the total number of items ($h = T^{1/\alpha}$); if $\alpha = 2$, it coincides with Price's square root law! Egghe (2006) in another paper on 'Theory and practice of the g-index", introduced an index called g-index to measure the global citation performance of a set of articles. It is defined as the largest number such that the top g articles receive at least g^2 citation. He has proved the unique existence of g for any set of articles and g >= h. By definition, g is less than or equal to T_p (total number of papers.) In an e-mail communication Egghe also suggested that g can be greater than T_p, in which case ($g-T_p$) papers are with zero citations each. The index based on this principle is referred to as g'-index in this paper. For instance in row 2 of

Table 1, we may note that g is 22 (it is equivalent to the total number of papers (T_p) and g' is 50; that means there are 28 papers with zero citations each and hence the total number of papers (T_p) becomes 50. It thus inflates the actual value of T_p. The g'^2 is however still less than the total number of citations (T_c). Although g' inflates T_p, it takes into account the high citations received by the top few papers. In this sense, g' is a better indicator than g. Rousseau (to be published) in his article on "Simple models and the corresponding h- and g-index" discussed the relation between h and g in the context of simple models. Egghe (to be published in 2007) in his paper on "Distributions of the h-index and the g-index" discussed the problem of determining the distribution of these indexes. He has applied linear three-dimensional Lotkaian informetrics to determine the distributions. Under the following two assumptions:

1. Article-citation Information Production Process (IPP) to be Lotkaian with exponent α, and
2. Scientist-article IPP to be Lotkaian with exponent α^*

Egghe has shown that both h-index IPP and g-index IPP are Lotkaian with exponent $\alpha\alpha^*$, for both discrete and continuous cases. He further pointed out that both the distributions are highly skewed, due to the high value of $\alpha\alpha^*$!

Objectives of the study

The objectives of this study are to identify

- Distribution of h and g-index
- Relation between h (or g or g') and the total number of papers (T_p)
- Relation between h (or g or g') and the total number of citations received (T_c)
- Relation between h and g

Data collection

Two hundred fifty two references from a bibliography on "optical flow estimation" (1) were considered in this study. 168 authors contributed these references. For each of these authors, a list of their publications was compiled using Web of Science; for each of these papers, the number of citations received was obtained again using the Web of Science. We thus collected

- The number of papers published (T_p) by each of the 168 authors
- The number of citations received by each of the papers for every author (in a set of 168 authors)
- Total number of citations (T_c) received by each of the authors.

We then computed h, g and g', based on the definition given by Hirsch and Egghe (2006). Frequency distributions of h, g, and g' are given in Tables 1-3.

Table 1. The empirical distribution of h.

h	f(h)	h	f(h)	h	f(h)	h	f(h)
0	6	4	23	8	6	13	1
1	39	5	7	9	2	14	1
2	37	6	7	10	4	18	1
3	22	7	8	12	3	25	1
Total		168		Variance		11.9365	
Mean		3.6667		St. Dev.		3.4549	

Table 2. The empirical distribution of g.

g	f(g)	g	f(g)	g	f(g)	g	f(g)
0	6	6	9	12	2	20	1
1	21	7	7	13	4	21	1
2	35	8	7	14	2	22	1
3	19	9	6	15	3	25	1
4	13	10	4	16	1	35	2
5	15	11	4	17	3	45	1
Total		168		Variance		38.04928	
Mean		5.82143		St. Dev.		6.1684	

Table 3. The empirical distribution of g'

g'	f(g')	g'	f(g')	g'	f(g')	g'	f(g')
0	6	8	9	16	3	26	1
1	7	9	10	17	1	27	1
2	26	10	5	18	4	28	1
3	14	11	4	19	2	29	1
4	10	12	4	20	1	38	1
5	15	13	7	21	3	40	1
6	10	14	3	23	5	45	1
7	9	15	1	25	1	46	1
Total		168		Variance		75.56618	
Mean		8.7976		St. Dev.		8.692881	

Data Analysis

Data in Tables 1-3 suggest that the empirical distributions of h, g and g' are highly skewed. Egghe (7) also argued that the distributions of h and g are highly skewed, due to high values of $\alpha\alpha^*$, where α and α^* are the Lotkaian exponents in the distributions of articles vs. citations and scientists vs. articles respectively. The averages ($\bar{x}$) of the distributions of h, g, and g' are 3,67, 5.82, and 8.797 respectively. The standard deviation (s) of h, g, and g' are 3.45, 6.17 and 8.69 respectively. For all the three distributions, $\bar{x} < s^2$; in such circumstances, negative binomial distribution will fit the empirical data fairly well. The values of mean and standard deviations indicate that the empirical distributions of h, g, and g' are likely to follow an exponential distribution, since they are approximately equal to each other. We therefore attempted to fit exponential, negative binomial and also lognormal distributions to the observed set of data.

- The exponential distribution fits the observed data for all the three distributions of h, g and g' very well.
- The negative binomial distribution fits equally well for the empirical distributions of g and g'. The negative binomial distribution even fits the empirical distribution of h at 0.01 level.
- The lognormal distribution fits the observed data only for the distribution of g'.

It has been observed that D_{max} is least for exponential distribution as compared to the lognormal and negative binomial distributions. The D_{max} values are 0.0954, 0.0813, and 0.0575 for distributions of h, g and g' respectively. The critical value at 0.05 levels is 0.1049. We may therefore believe that the exponential distribution is a suitable model to explain all the three distributions of h, g and g'. However, theoretical explanations are yet to be explored in this regard. An attempt was also made to

identify a relation between (T_p & h), (T_p & g), (T_p & g'), (T_c & h), (T_c & g) and (T_c & g'), for all the three cases – linear scale, semi-log scale and log-log scale. For this purpose, correlation coefficients (r) were appropriately computed.

- The values of r for (T_p & h), (T_p & g), (T_p & g') are 0,813, 0.801, and 0.584 respectively. These values indicate that there exists only a weak linear relation between the total number of papers and h- or g- index.
- The values of r for (T_c & g) and (log (T_c) & log (g')) are 0.937 and 0.991 respectively, indicating that there exists a linear relation between the total number of citations and g or g' index. g' as an index is excellent indicator since it takes into account the large citations received by the top few papers as compared to h.
- The value of r between T_c & h (r = 0.791) and also between log (T_c) and log (h) (r = 0.8186) are very low. Even the values of r for (T_c & h), (T_c & g) and (T_c & g') in semi-log scale are very low.

In order to study whether or not there exists a correlation between h and g, the correlation coefficient is computed. Its value is 0.9678. (R^2 is 0.9366; i.e., 93% of the variation in g is due to the variation in h.) It indicates that there exists a strong linear relation between h and g.

Concluding remark

It has been observed that the distributions of h, g and g' follow an exponential distribution. It is much better than negative binomial, lognormal or Lotka type models. Although the methodology of computing g' inflates the size of total number of papers, there exists a linear relation between log (T_c) & log (g'). Since g' takes into account of high citations to the top few papers and since there exists a linear relation between T_c & g' both in linear and log-log scales, g' be used as an indicator of scientific productivity of individuals. It has been further observed that the correlation between (T_p & h) is very poor – only 66% of the variation in h is due to the variation in p. It clearly indicates that h is unlikely to increase linearly with p.

References

A bibliography on "optical floe estimation." Retrieved from: http://itb1.biologie.hu-erlin.de/~wiskott/ Bibliographies/ FlowEstimation.htma=ml

Barista, Pablo, D and Others. (2006). Is it possible to compare researchers with different scientific interests? *Scientometrics*; 68 (1), 179-89

Bornmann, Lutz and Daniel, Hans-Dieter. (2006). Does the h-index for ranking of scientists really work? *Scientometrics*, 65(3) 391-2.

Braun, Tabor, Glaze, Wolfgang, and Schubert, Andres. A Hirsch type index for journals. *The Scientist*, 19 (22), 8.

Egghe, L. (2006). Theory and practice of the g-index. *Scientometrics*, 69 (1), 131-52.

Egghe, L. and Rousseau, R. (2006). In informetric model for the Hirsch-index. *Scientometrics*, 69 (1), 121-9.

Egghe, L. "Distributions of the h-index and the g-index" (to be published.)

Hirsch, J. E. "An index to quantify an individual's scientific research output." Retrieved from http://arxiv.org/abs/physics/?0508025

Rousseau, R. Simple models and the corresponding h- and g-index. (Submitted to Science and Technology Development.)

van Raan, Anthony, F. J. (2005). Comparison of the Hirsch-type index with standard bibliometric indicators and with peer judgment for 147 research groups. Retrieved from http://arXiv:physics/0511206

On Material Transfer Agreements and Visibility of Researchers in Biotechnology

Victor Rodriguez, Frizo Janssens, Koenraad Debackere and Bart De Moor

victor.rodriguez@econ.kuleuven.be
Katholieke Universiteit Leuven (Belgium)

Abstract
When carrying out a research project, some materials may not be available in-house. Thus, investigators resort to external providers for conducting their research. To that end, the exchange may be formalised through material transfer agreements. In this context, industry, government, and academia have their own specific expectations regarding compensation for the help they provide when transferring research material. This paper assesses whether these contracts have an impact on visibility of researchers. Visibility is thereby operationalized on the basis of a bibliometric approach. In the sample, researchers that used these contracts were more visible compared to those who did not use them. Nonetheless, among these contract users, there was no gain in visibility after using them. Providers and receivers could not be distinguished by using these contracts but by research sector and co-authorship.

Keywords
research material; scientific reputation; H-index; biotechnology; material transfer agreement

Introduction

Industry, government, and academia have their own specific expectations regarding "compensation" for the help they provide when transferring research materials. A broad definition of research materials includes cell lines, monoclonal antibodies, reagents, animal models, combinatorial chemistry libraries, clones and cloning tools, databases, and software (under some circumstances). As material transfer agreements (MTAs) are used for the supply of research materials between laboratories (Rodriguez, 2005; Rodriguez et al., 2007), the provider can expect a reward in exchange. This compensation could vary depending on the provider's institutional sector. In academia, the desired reward for providing the research material may range from acknowledgment or reference in a publication to co-authorship of articles or co-assignment of patents made with the aid of the supplied material, depending on the purpose of the material provider. In industry, the provider could ask the receiver to present the results of using the material or any other research results that could be of interest. The rationale of the existence of governmental laboratories differs from academic ones, their expectations are more related to technology transfer and in that sense it could follow the same logic of industrial laboratories.

Being a provider or receiver might be the reflection of differences between scientific leaders and followers in the social structure of science. The role of scientific leader is acquired by the creation of the research material and the subsequent requests for it by the scientific community. The relationship between material owners and would-be users could be catalogued as trade-off of expectations. The provider expects to be rewarded by the receiver for enabling its use and the receiver expects to gain access to the material to carry out the research. Scientists tend to cite contributions that are useful for their own research. In the gift-swap model of scientific exchange, individual scientists contribute their findings to the scientific community and in return can expect to receive various forms of recognition from their peers (Hagstrom, 1965). Greater visibility is one of them.

Due to the role of scientific leader, a provider's visibility could differ from that of receivers after using MTAs because the receiver may monopolise a material that is requested by various researchers. They might reward the provider via citations. This practice might be used as evidence of providers' supposedly greater visibility. The construct researcher's visibility can be operationalized using the h-index, developed by Jorge Hirsch (2005), which is based on the citations each article of an author gets. Substantial research effort has been devoted to studies on author citations (Braun & Glänzel, 2000; Garfield, 1963; Glänzel, 2000; Kostoff, 1998; Leydesdorff, 1998; Moed, 2000; Peritz, 1992),

stratification system within the science institution (Allison, Long & Krauze, 1982; Cole & Cole, 1967; Cole & Cole, 1974; Evans, 2005), Nobel laureates (Crawford, 1998; Garfield & Welljams-Dorof, 1992; Zuckerman, 1967; Zuckerman, 1996) and star scientists (Zucker & Darby, 1996; Zucker & Darby, 2001). However, there has been relatively little analysis of the visibility acquired through the exchange of research materials (Walsh et al., 2003; Walsh et al., 2005).

This paper assesses whether MTAs have an impact on visibility of researchers. In other words, it examines whether involvement in MTA activity positively or negatively affects the visibility of researchers. Visibility is thereby operationalized on the basis of a bibliometric approach. The paper begins with sectoral rewarding of researchers. A concept of visibility and the means to operationalize it is elaborated. The data and methodology respective sections are presented, followed by the results. Finally, some concluding remarks are made.

Rewarding in government, industry, and academia

The labels of basic and applied research are changing with regard to biosciences and biotechnologies. A priori definitions of polar ideal types are vague, imprecise and awkward for empirical operationalisation (Callon, 1997; Kidd, 1965; Latour, 1993). The commercial development of basic research discoveries and commercial interest in academic research in its early stages speed up the applicability of basic research. Thus, the distinction between basic and applied research is difficult to maintain. Nonetheless, what is clear is the research setting where this research is conducted and the nature of the compensation for exchanging research material. Researchers from industry, government, and academia could expect to be rewarded differently for research materials they provide to extramural colleagues.

Researchers in government comprise a significant component of the same scientific labour force as that of scientists in academia. Even though the reason or purpose of the existence of governmental laboratories differs from academic ones, they have had some common educative functions. These arise from their research activities that involve collaboration with the higher education sector and include the supervision of students, lecturing in universities, and training research staff. This role of providing future gatekeepers and linking individuals into wider technological communities (Debackere and Rappa, 1994) is significant in facilitating technology transfer.

Industrial researchers make situational adjustments to their other professional orientations (e.g. communality and autonomy) that must be taken into account when referring to their organisational affiliation (Barnes, 1971). In particular, many science graduates are trained in the pure science ethos but quickly set it aside after a period in an industrial laboratory (Avery, 1960). Most industrial scientists accept the framework of industrial science, do not object to commercial restrictions on publications, are not publishing to try to develop professional reputations beyond their home organisations, and are satisfied with organisational status and recognition, which is accomplished by developing processes and products that are patentable (Ellis, 1972). As the industrial research laboratory is subject to the controls emanating from the business goals of the organisation, a different orientation from academia and government tends to be inculcated (Bailyn, 1985).

Priority of discovery is the basis for a scientist's reputation and this prestige for contributions to the literature is the fundamental reward in the community of academic scientists (Dasgupta & David, 1994). The race for priority serves two very important purposes: it hastens discoveries and it accelerates their subsequent disclosure through publications (Merton, 1973). Scientists are socialized in this norm of openness during their doctoral studies, where the importance of advancing the body of knowledge through publishing articles is always emphasised. However, many doctoral students in biology, chemistry, and other biosciences find themselves working for companies that may not share this desire for openness (McMillan & Deeds, 1998). This privatization of science has been also detected in academia and government since the Bayh-Dole Act and similar mirroring regimes around the world. Lawyers and economists often assume that the alternative to exclusive rights is secrecy, while sociologists tend to equate exclusive rights with secrecy and take the alternative to be open access.

Secrecy and exclusive rights tend to be equated because in most cases the only way to preserve exclusivity in an idea or discovery is to keep it secret. When patent protection extends into fields that are of interest to governmental, industrial, and academic researcher, neither of these assumptions about the relationship between exclusive rights and secrecy is quite right (Eisenberg, 1989).

Visibility

Although a large majority of articles, letters, notes, and reviews are never cited, some are intensely referred to (Aksnes & Siversten, 2004). When a paper is highly cited, even more people become aware of it and its visibility increases the chances of getting more citations (Aksnes, 2003). In fact, high citation scores are the results of many researchers' decisions to cite a particular paper. The construct visibility is usually operationalized using citation counts to gauge the overall impact of a scientist's research output on the scientific community and are generally held to measure quality (Cole & Cole, 1967). An average citation per paper gives an indication of the aggregate level of influence, while highly cited papers reflect the more important contributions to the field. The evidence that citation signals quality is indicated by the literature. Garfield (1970) studied the work of Nobel Prize winners and found that they were among the top 0.1% most cited authors. Zuckerman (1996) found that publication counts, citation counts and peer ratings were intercorrelated. Finkenstaedt and Fries (1978) found that citation counts correlate highly with other measures of quality such as employment in prestigious universities, listing in important bibliographies of scientists, and scientific award and recognition from colleagues.

Research on citation has shown highly skewed distributions, with most articles having few citations, while a handful has exceedingly large numbers. Lotka (1926) and de Solla Price (1965, 1980) showed that in degree distributions of citation networks, the proportions of nodes with degree k varies as a function of $1/k^{\alpha}$, that is, by the inverse power law $P(k) \cong 1/k^{\alpha}$, where α is the power coefficient. Thus, nodes with unusually large number of ties or edges produce network hub phenomena. Recently, more detailed examination has refined the picture. In biosciences and biotechnologies, h-indices tend to be higher than those in physics (Hirsch, 2005), that is to say they are discipline based. Bornmann and Daniel (2005) found that on average the h-index for successful applicants for post-doctoral research fellowships was consistently higher than that for non-successful applicants. Van Raan (2006), dealing with research groups rather than individual scientists, showed that the h-index and several standard bibliometric indicators both correlate in a quite comparable way with peer judgements.

Visibility studies are closely related to the analysis of elite researchers. These "stars" are eminent scientists or technologists who have been accorded recognition in their discipline, who are visible in their community, who are great talents, who are highly productive, who achieve excellence, and who usually choose careers that give them a great deal of freedom and personal independence (Merton, 1968). In particular, they are individuals who are involved in many aspects of scientific or technological life. Once members of distinguished scientific or technologic organisations (e.g. the Royal Society), elite researchers rise quickly to positions of prominence and are called upon to participate actively (i.e. hold office, preside over meetings, chair special interest groups etc.). They are intimately involved in the communication system of their profession by serving as editors and referees of journals (Crane, 1967). Their expertise attracts many requests to share their knowledge by guest lecturing. All of these dimensions of academic involvement are a form of service of a scientifically useful nature (Shilling and Bernard, 1964).

Data

Publications have been selected from the core biotechnology database created by Glänzel et al. (2003), which included patents applied for at the European Patent Office. Surprisingly, none of the sampled patents in industry and government were cited. The absence of forward citation could occur for two reasons. First, the patented technology was so applied that it was located in the frontier of knowledge. Second, other organisations were not able to absorb the patented technology. Thus, the analysis was restricted to four types of documents, namely: articles, letters, notes, and reviews, because they have references.

Researchers working for industry, government, and academia in Belgium authored the biotechnology publications. To study the effect of MTAs on visibility, representatives from the organisations appearing as institutional affiliations distinguished whether or not the documents have used materials received through MTAs. The publication year of MTA-related publication was used as a proxy for MTA year because, in fact, we were not able to determine when the MTA was signed due to the confidentiality clause inherent in this contract. Consequently, MTA year refers to the publication year of the research project that handled external materials formalised in an MTA. A content analysis of representative MTAs in the survey was not possible because of the confidentiality clauses. For that reason it cannot be determined how many of them explicitly required that receivers should acknowledge providers in relevant publications, or that receivers should not mention providers at all for whatever reasons.

The timeline of events per author is shown in Figure 1. Pre-MTA refers to the period of time between the year of first publication done by the author and the proxied MTA year inclusive. Post-MTA denotes a period of time after the proxied MTA year till 2004 inclusive.

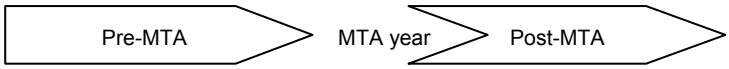


Figure 1. Publication timeline of an author that used an MTA

The receivers and providers of research materials were listed from the set of MTA-related documents. Then, the dataset was compiled by retrieving information concerning publications of receivers and providers in all languages in journals recorded in the three databases of the *Web of Science* of Thomson Scientific (viz. natural sciences, social sciences, arts and humanities). The documents were collected for each author up to 2004 inclusive by using the four steps refinement feature—last name, first initial, middle initial; supplementary initials; research field; institutional affiliation—upgraded by the *Web of Science* in July 2006 to reduce the homonym bias. The total citation window ran from the year of publication to the end of 2004. Our study involved a total of 5194 publications and 68874 citations according to the *Science Citation Index*. Visibility was then operationalized using the h-index. Scientists have an h-index equals to h if h of their N_p publications have at least h citations each, and the rest (N_p - h) have fewer than h citations each (Hirsch, 2005).

The h-index is the highest number of papers a scientist has, published over n years, which have each received at least that number of citations. Thus, someone with an h-index of 50 has written 50 papers that have each had at least 50 citations (Ball, 2005). If a scientist has 21 papers, 20 of which are cited 20 times, and the 21st is cited 21 times, there are 20 papers—including the one with 21 citations—having at least 20 citations, and the remaining papers has no more than 20 citations (Van Raan, 2006). Therefore, the scientist's h-index equals 20. Despite the fact that some researchers in the sample were retired in 2004, the h-index remains useful as a measure of cumulative achievement. Visibility may continue to increase over time, even after the scientist has stopped publishing (Hirsch, 2005). We calculated h-indices for each researcher before and after the MTA year. While other studies used a fixed citation window (i.e. a shorter period of time after the publication year to measure recent performance), we counted citations from the year of publication to the end of the proxied MTA year and after the proxied MTA year to end of 2004. While publications are counted for an entire period of time in a total block analysis, in our study the publications were counted from the first year that the researcher has published till 2004 (i.e. total time length). Total time length represents the researcher's seniority. When plotting total time length with total h-index, we obtained an intercept of 3.23 and a slope of 0.49 (Figure 2). Then, publications were allotted before and after MTA year.

In our sample, all documents were multi-authored. Thus, scientists with high h-indices achieved mostly through publications with many co-authors would be treated overly kindly by their h-indices. As we found large differences in the number of co-authors for each paper in our dataset, we used a factor Ω to capture co-authorship in order to compare different individuals. This factor can be specified as $\Omega = 1/P_i \sum 1/A_j$, where P_i represents total number of publications of each author i and A_i

stands for the number of authors of each paper *j* of author *i*. The absence of co-authors is expressed as Ω = 1. When plotting total inverse omega, 1/Ω, with total h-index (Figure 3), we obtained an intercept of 24.16 and a slope of -2.16. The total inverse omega was then computed before and after the MTA year.

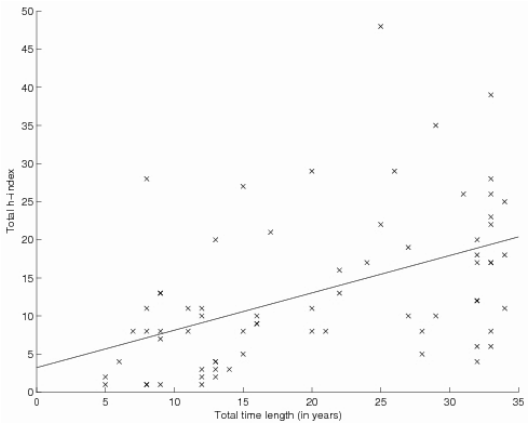


Figure 2. Plotting total time length with total h-index

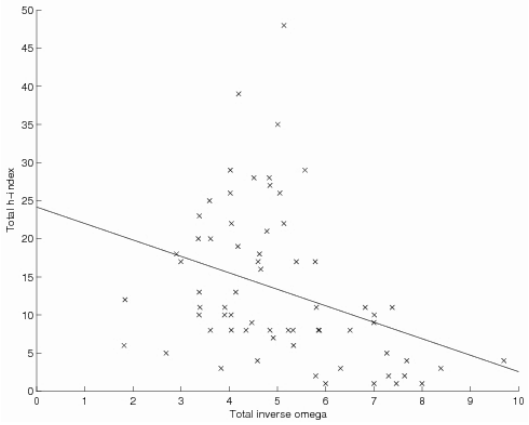


Figure 3. Plotting total inverse Ω with total h-index

Methodology

The sampled researchers were grouped according to some features, such as use of MTA in the period 1992–2000, role-played and sector of activity when they used an MTA. Group membership was assumed to be mutually exclusive (i.e. no case belongs to more than one group) and collectively exhaustive (i.e. all cases are members of a group). Regarding the coding scheme for the qualitative data, we dichotomized the levels such that *k*-1 zero-one dummy variables were created for each variable having *k*-levels instead of assigning numerical values to the levels of variables. When a researcher had more than one institutional affiliation in the publication we chose the first one listed by the author in order to get exclusive sectoral dedication.

In our group variables, *MTA* is a dummy that equals 1 if the researcher used an MTA in the period 1992–2000, 0 otherwise; *Provider* is a dummy variable that equals 1 if the researcher is a provider, 0 otherwise; *Academia* is a dummy variable that equals 1 if the researcher is in academia, 0 otherwise; *Government* is a dummy variable that equals 1 if the researcher is in government, 0 otherwise.

As the aim is to assess the MTA's impact on visibility of researchers, the first empirical strategy was to carry out a one-way analysis of variance to test the null hypothesis that the visibility of the typical researcher that used MTA is equal to that of the one who did not. The quantitative variables employed in the analysis were: *c* for all citations obtained by each sampled researcher till 2004; Ω for the omega

factor of each researcher till 2004; *h* for the h-index scored by each sampled researcher till 2004; and *t* for the seniority of each sampled researcher. The descriptive statistics are shown in Table 1.

Table 1. Group statistics between MTA and non-MTA group

	MTA	N	Mean	Std. deviation	Std. error mean
c	1	70	983.91	1420.679	169.804
	0	69	416,38	855,337	102,970
Ω	1	70	.223132	.0915903	.0109471
	0	69	1.00000	.0000000	.0000000
h	1	70	13.14	9.993	1.194
	0	69	6.59	7.009	.844
t	1	70	18.51	10.796	1.290
	0	69	12.16	5.677	.683

To test if MTAs improved researcher's visibility, the second empirical strategy was to compare the visibility before (pre-MTA) and after (post-MTA) the researchers used an MTA. The construct visibility was operationalized using h-indices and citations counts. The paired-samples t-test procedure is used to test the hypothesis of no difference between visibility before and after MTAs. This procedure compares the means of two variables for a single group. For that reason, we split up the single factor variables before and after MTA year. Thus, the descriptive statistics for these quantitative variables are shown in Table 2, where the post-MTA scores are higher.

Table 2. Paired samples statistics

Pair 1	Mean	N	Std. deviation	Std. error mean
h Pre-MTA	6.81	70	7.490	.895
h Post-MTA	7.33	70	6.569	.785
c Pre-MTA	624.10	70	1113.530	133.092
c Post-MTA	359.81	70	553.438	66.149

Next, the correlation between the two variables for each operationalisation is shown in Table 3. There is a strong positive correlation not only between *h Pre-MTA* and *h Post-MTA*, but also between *c Pre-MTA* and *c Post-MTA*. Researchers who were visible before MTA also were visible afterwards.

Table 3. Paired samples correlations

Pair 1	N	Correlation	Sig.
h Pre-MTA and *h Post-MTA*	70	.477	.000
c Pre-MTA and *c Post-MTA*	70	.383	.001

Finally, the last empirical strategy was to detect which variables discriminated best among providers and receivers using the dummy variables *Academia* and *Government*, and the following quantitative variables: *t Pre-MTA* for the time length (in years) till the MTA year inclusive; *t Post-MTA* for the time length after the MTA year till 2004; *1/*Ω *Pre-MTA* for co-authorship till the MTA year inclusive;

1/Ω Post-MTA for co-authorship after the MTA year till 2004; *h Pre-MTA* for the h-index till the MTA year inclusive; *h Post-MTA* for the h-index after the MTA year till 2004; *c Pre-MTA* for the citations till the MTA year inclusive; and *c Post-MTA* for the citations after the MTA year till 2004. The descriptive statistics are shown in Table 4.

Table 4. Group statistics

Provider	**Variables**	**Mean**	**Std. deviation**	**N**
0	*Academia*	.568966	.4995461	58
	Government	.155172	.3652312	58
	t Pre-MTA	11.206897	9.1895414	58
	t Post-MTA	6.017241	3.6152279	58
	1/Ω Pre-MTA	4.746461	2.0028993	58
	1/Ω Post-MTA	5.083801	2.0851841	58
	h Pre-MTA	6.965517	7.8337227	58
	h Post-MTA	7.275862	6.6088540	58
	c Pre-MTA	685.120690	1203.8892395	58
	c Post-MTA	368.793103	564.9490199	58
1	*Academia*	.333333	.4923660	12
	Government	.666667	.4923660	12
	t Pre-MTA	17.166667	9.0937874	12
	t Post-MTA	7.583333	2.3143164	12
	1/Ω Pre-MTA	2.881189	1.4832684	12
	1/Ω Post-MTA	5.228183	1.8923709	12
	h Pre-MTA	6.083333	5.7597085	12
	h Post-MTA	7.583333	6.6532061	12
	c Pre-MTA	329.166667	391.2397157	12
	c Post-MTA	316.416667	514.7335513	12

Results

For assessing the MTA's impact on visibility of researchers, the first empirical strategy was to carry out a one-way analysis of variance to test the null hypothesis that the visibility of the typical researcher that used MTA is equal to that of the one who did not. As the effects are found to be significant (Table 5) using the above procedure, it implies that the means differ more than would be expected by chance alone. In terms of the above experiment, it would mean that MTA-users were more visible than non-MTA users.

The results of the paired-samples t-test, which is based on the difference between the two variables, are shown in Table 6. The descriptive statistics for the difference between the two variables are displayed under "Paired differences." When the construct visibility was operationalized using h-indices, the significance value greater than 0.05 for change in h-index level shows that MTA involvement did not significantly increase their h-indices. By contrast, when citation counts were used, a significant difference was obtained between pre and post-MTA mean citations. Accordingly, since the significance value for change in citation is less than 0.05, we can conclude that the average loss of 264.286 citations per MTA-user is not due to chance variation, and can be attributed to MTA involvement. Thus MTAs did not help to increase visibility measured by citations counts.

The results of the discriminant analysis are shown in Table 7. Both the dummy *Government* and the variable *1/Ω Pre-MTA* were good discriminant factors between providers and receivers.

Table 5. Analysis of variance

		Sum of squares	df	Mean square	F	Sig.
MTA	Between groups	10.802	30	.360	1.624	.037
	Within groups	23.946	108	.222		
	Total	34.748	138			
t	Between groups	6483.741	30	216.125	4.530	.000
	Within groups	5152.273	108	47.706		
	Total	11636.014	138			
Ω	Between groups	6.728	30	.224	1.634	.036
	Within groups	14.823	108	.137		
	Total	21.550	138			
c	Between groups	197544726.721	30	6584824.224	267.231	.000
	Within groups	2661224.416	108	24640.967		
	Total	200205951.137	138			

Table 6. Paired samples test

Pair 1	Paired differences					t	df	Sig. (2-tailed)
	Mean	Std. deviation	Std. error mean	95% Confidence interval of the difference				
				Lower	Upper			
h Pre-MTA - h Post-MTA	-.514	7.233	.864	-2.239	1.210	-.595	69	.554
c Pre-MTA - c Post-MTA	264.286	1036.415	123.875	17.161	511.410	2.133	69	.036

Table 7. Stepwise statistics for dummy variables: Variable entered [a, b, c, d]

Step	Entered	Wilks' Lambda							
		Statistic	df1	df2	df3	Exact F			
						Statistic	df1	df2	Sig.
1	*Government*	.798	1	1	68.000	17.224	1	68.000	.000
2	*1/Ω Pre-MTA*	.738	2	1	68.000	11.895	2	67.000	.000

At each step, the variable that minimizes the overall Wilks' Lambda is entered.
[a] Maximum number of steps is 20 ; [b] Minimum partial F to enter is 3.84. ; [c] Maximum partial F to remove is 2.71.; [d] F level, tolerance, or VIN insufficient for further computation.

All in all, the t-test, the one-way analysis of variance and the discriminant function analysis are mathematically equivalent but are used to answer different questions. In the sample, researchers that used MTAs were more visible compared to those who did not use MTAs. Nonetheless, among those MTA-users, there was no gain in visibility after using MTAs. Providers and receivers could not be distinguished by using MTA but by research sector and co-authorship before MTA.

Discussion

In the sample, being an MTA-user might or not might, to some extent, be the reflection of systematic differences in the stratification of science based on visibility. Seniority and co-authorship helped to gain visibility as well. In an effort to determine if the sample in analysis of variance was biased we have chosen to perform the two-sample Kolmogorov-Smirnov test (Table 8). This is a general test that detects differences in both the locations (or central tendencies) and the shapes of distributions. The probability of the Kolmogorov-Smirnov Z statistic falls well below 0.05. By that standard, the distributions are significantly different from each other. From Table 8, we can conclude that the significance of the difference between the two distributions is due only to their different locations on the scale, not to any differences in shape.

Table 8. Two-sample Kolmogorov-Smirnov test statistics [a]

		c	***t***	**Ω**	***h***
Most extreme differences	Absolute	.297	.357	1.000	.410
	Positive	.085	.057	1.000	.000
	Negative	-.297	-.357	.000	-.410
Kolmogorov-Smirnov Z		1.753	2.102	5.895	2.416
Asymp. sig. (2-tailed)		.004	.000	.000	.000

[a] Grouping variable: *MTA*.

For validating the pair samples test, we firstly carried out a sign test (Table 9). Apart from that, the paired-samples t-test is appropriate whenever two related sample means are to be compared. The difference scores are assumed to follow a reasonably normal distribution, especially with respect to skewness. The procedure used to test the assumption of normality was a one-sample Kolmogorov-Smirnov test (Table 10).

Table 9. Sign test statistics

	c Post-MTA - c Pre-MTA
Z	-2.032
Asymp. sig. (2-tailed)	.042

Table 10. One-sample Kolmogorov-Smirnov test

		c Pre-MTA	***c Post-MTA***
N		70	70
Normal parameters [a,b]	Mean	624.10	359.81
	Std. deviation	1113.530	553.438
Most extreme differences	Absolute	.288	.285
	Positive	.281	.285
	Negative	-.288	-.258
Kolmogorov-Smirnov Z		2.406	2.384
Asymp. sig. (2-tailed)		.000	.000

[a] Test distribution is normal ; [b] Calculated from data.

In addition, the Wilcoxon signed-rank test was executed (Table 11) as complementary to the paired t-test. Such differences in visibility might exist in a different sample but drawing inferences from them is not as easy as it may appear. In fact, the difference cannot be explained by the use of MTAs.

Table 11. Wilcoxon signed ranks test statistics

	c Post-MTA - c Pre-MTA
Z	-1.870 [a]
Asymp. sig. (2-tailed)	.062

[a] Based on positive ranks.

For assessing the validity of the discriminant analysis, the total sample was randomly dived into an analysis sample consisting of around 70% of the entire sample (50 observations) and a around 30% holdout sample (20 observations). Since the holdout sample was small, it is possible that there might be a good deal of variability simply due to the reduced number of observations. Therefore, a more conservative test was performed in which the discriminant coefficients were derived using all of the available sample observations (analysis plus hold-out samples).

The classification function coefficients based on absolute magnitudes for both the analysis and total samples are given in Table 4. It is worth noting that there are no sign reversals when the same variables were entered by the stepwise. The table clearly indicates that for the dummy variables *Provider* there is no discrepancy in the number of discriminant factors between the analysis and total sample. Accordingly, those researchers that belonged to wealthy research environments, such as governmental or industrial research laboratories, might be distinguished as providers or receivers. Co-authorship before MTAs worked also well as a discriminant factor between providers and receivers.

Table 12. Classification function coefficients

Group	**Discriminant**	**Analysis sample (ca. 70%)**		**Complete sample (100%)**	
		0	1	0	1
Provider	*Government*	1.641	4.149	1.708	4.869
	1/Ω Pre-MTA	1.403	0.943	1.312	.877

Fisher's linear discriminant functions.

The t-test deals with the problems associated with inference based on small samples. An extension of the two-sample t-test is the one-way analysis of variance for a quantitative dependent variable by a single factor (independent) variable. Analysis of variance is used to test the hypothesis that several means are equal. However, it should be clear that, if the means for a variable are significantly different in different groups, then we could say that this variable discriminates between the groups. In other words, in addition to determining that differences exist among the means, it is interesting also to know which means differ. Thereby the basic idea underlying discriminant function analysis is to determine whether groups differ with regard to the mean of a variable. Discriminant analysis may act as a univariate regression and is also related to analysis of variance (Wesolowsky, 1976). The relationship to analysis of variance is such that discriminant analysis may be considered as a multivariate version of analysis of variance.

The h-index after MTAs was higher that that before MTAs for the typical provider and receiver, as well as for the usual non-academic researcher as shown in Table 13. The industrial researcher visibility remained unchanged after MTAs. This could be explained by the fact that they do not fit the traditional mold of the academic reward. Oddly, though, the post-MTA visibility of academic researchers was, on average, not larger than that of the pre-MTA period. This result is not consistent with the predictions of the theory of open science conduct. In fact, this contradicts the reason why

academia relies more in a system of scientific communication related to the reward system. Some studies have looked beyond the researcher's organisational affiliation and have shown that other orientations come into play, in interaction with the organisation setting. Cotgrove and Box (1970) revealed that the interplay between scientific roles and identities could be the basis for the development of a topology of scientists that is not tied to the organisational setting.

Table 13. Group h-indices

Group	Mean *h Pre-MTA*	Mean *h Post-MTA*
Provider	6.08	7.58
Receiver	6.97	7.28
Academia	7.41	6.92
Non-academia	6.15	7.79
Government	7.65	9.76
Industry	6.55	6.55

Upon becoming involved in MTAs, it was assumed that the h-index of providers would be higher than that of receivers. Being a provider was considered a factor for acquiring scientific leadership and gaining greater visibility than a receiver. This was assumed simply because the provider could be rewarded with more citations and thereby increase their h-index. The fact of creating research material that is requested by members of the scientific or technologic community was not generating differences in visibility in our sample. We could not reject the null hypothesis that the mean h-index or citation after MTA was the same between providers and receivers.

Concluding remarks

A crucial challenge in the literature of science and technology studies is to detect the effect of MTA involvement on researchers' visibility. In this paper, this issue was tackled exploiting a natural experiment. When carrying out a research project, some materials may not be available in-house. Thus, investigators resort to external providers for conducting their research. To that end, the exchange is formalised through MTAs. In this context, the provider supplied the material; receivers used it and published their results between 1992 and 2000.

To study the effect of MTAs on visibility of researchers, we were conscious that MTA year is antecedent to the publication year of research results. Usually, the difference between them is more than one year. The problem was that we were not able to determine the date of the MTA signature due to confidentiality clauses included in the contract itself. That is why we proxied MTA year as publication year of research results that used external material received through MTAs.

Bibliometric data of researchers involved in MTAs was collected and a control group of researchers that did not use them between 1992 and 2000 was created. In the sample, researchers that used MTAs were more visible compared to those who did not use MTAs. Nonetheless, among MTA-users, there was no gain in visibility after using MTAs. Providers and receivers could not be distinguished by using MTA but by research sector and co-authorship before MTA.

Even if some connections to visibility were found as suggested in the results, does it mean that using MTAs modifies the social structure of science and technology? This study shows how difficult it is to discriminate between h-indices and other dummies, such as provider, sector, etc.

Some last remarks with implications for citation studies are worth making. Material receivers identified the provider either in the acknowledgments section, in the main text or in the reference list of the publication. Thereby the role of supplier was identifiable in the scientific publication. In rare cases, the provider was not indicated as such in the publication but appeared as a co-author, a situation only distinguishable by the authors themselves and not by bibliometricians. Apart from that,

some providers could ask receivers not to mention them in publications because that citation could be deemed as prior art for patent purposes. So, citation counts might not reflect visibility in cases of omission for several reasons. Only by taking these comments into account can research on visibility make unequivocal conclusions when vested interests are at stake.

References

Aksnes, D., 2003. Characteristics of highly cited papers. *Research Evaluation* 12: 159-170.

Aksnes, D., and G. Siversten, 2004. The effect of highly cited papers on national citation indicators. *Scientometrics* 59: 213–224.

Allison, P., J. Long, and T. Krauze, 1982. Cumulative advantage and inequality in science. *American Sociological Review* 47: 615-625.

Avery, R., 1960. Enculturation in industrial research. *Transactions on Engineering Management* 7: 20-24.

Bailyn, L., 1985. Autonomy in the industrial R&D lab. *Human Resource Management* 24: 129-146.

Ball, P., 2005. Index aims for fair ranking of scientists. *Nature* 436: 900.

Barnes, S., 1971. Making out in industrial research. *Science Studies* 1: 157-175.

Bornmann, L., and H. Daniel, 2005. Does the h-index for ranking of scientists really work? *Scientometrics* 65: 391-392.

Braun, T., and W. Glänzel, 2000. Chemistry research in Eastern Central Europe (1992-1997): Facts and figures on publication output and citation impact for 13 countries. *Scientometrics* 49: 187-214.

Callon, M., 1997. Actor-network theory: The market test. *Actor Network and After Workshop*. Centre for Social Theory and Technology, Keele University, UK.

Cole, S., and J. Cole, 1967. Scientific output and recognition: A study in the operation of the reward system in science. *American Sociological Review* 32: 377-390.

Cole, S., and J. Cole, 1974. Social stratification in science. *American Journal of Physics* 42: 923-924.

Cotgrove, S., and S. Box, 1970. *Science, industry and society*. New York, NY: Barnes and Noble.

Crane, D., 1967. The gatekeepers of science: Some factors affecting the selection of articles for scientific journals. *American Sociology Review* 32: 195-201.

Crawford, E., 1998. Nobel: Always the winners, never the losers. *Science* 282: 1256-1257.

Dasgupta, P., and P. David, 1994. Toward a new economics of science. *Research Policy* 23: 487-521.

Debackere, K., and M. Rappa, 1994. Technological communities and the diffusion of knowledge: A replication and validation. *R&D Management* 24: 355-371.

de Solla Price, D., 1965. Networks of scientific papers. *Science* 149: 510-515.

de Solla Price, D., 1980. A general theory of bibliometrics and other cumulative advantage processes. *Journal of the American Informatics Society* 27: 292-306.

Eisenberg, R., 1989. Patents and the progress of science: Exclusive rights and experimental use. *University of Chicago Law Review* 56: 1017-1086.

Ellis, N., 1972. The occupation of science. In: B. Barnes (Ed.), *Sociology of science*. Baltimore, MD: Penguin Books.

Evans, J., 2005. Stratification in knowledge production: Author prestige and the influence of an American academic debate. *Poetics* 33: 111-133.

Finkenstaedt, T., and M. Fries, 1978. Zur forschungsmessung in den geisteswissenschaften. *Ad Acta* 3: 110-164.

Garfield, E., 1963. Citation indexes in sociological and historical research. *American Documentation* 14: 289-291.

Garfield, E., 1970. Citation indexing for studying science. *Nature* 227: 669-671.

Garfield, E., and A. Welljams-Dorof, 1992. Of Nobel class: A citation perspective on high impact research authors. *Theoretical Medicine and Bioethics* 13: 117-135.

Glänzel, W., 2000. Science in Scandinavia: A bibliometric approach. *Scientometrics* 48: 121-150.

Glänzel, W., M. Meyer, B. Schlemmer, M. Du Plessis, B. Thijs, T. Magerman, K. Debackere, and R. Veugelers, 2003. Biotechnology: An analysis based on publications and patents. Leuven: Steunpunt O&O Statistieken.

Hagstrom, W., 1965. The scientific community. New York, NY: Basic Books.

Hirsch, J., 2005. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences* 102: 16569-16572.

Kidd, C., 1965. Basic research: Description versus definition. In: Kaplan, N. (Ed.), *Science and society*. Chicago, IL: Rand McNally.

Kostoff, R., 1998. The use and misuse of citation analysis in research evaluation: Comments on theories of citation. *Scientometrics* 43: 27-43.

Latour, B., 1993. We have never been modern. London, UK: Harvester Wheatsheaf.

Leydesdorff, L. 1998. Theories of citation. *Scientometrics* 43: 5-25.

Lotka, A., 1926. The frequency distribution of scientific productivity. *Journal of the Washington Academy of Science* 16: 317-323.

McMillan, G., and D. Deeds, 1998. The role of reputation in the recruitment of scientists. *R&D Management* 28: 299-278.

Merton, R., 1968. The Matthew effect in science. *Science* 159: 59.

Merton, R., 1973. *The Sociology of science: Theoretical and empirical investigations*. Chicago, IL: University of Chicago Press.

Moed, H., 2000. Bibliometric indicators reflect publication and management strategies. *Scientometrics* 47: 323-346.

Peritz, B., 1992. On the objectives of citation analysis: Problems of theory and method. *Journal of the American Society for Information Science* 43: 448-451.

Rodriguez, V., 2005. Material transfer agreements: Open science vs. proprietary claims. *Nature Biotechnology* 23: 489-491.

Rodriguez, V., F. Janssens, K. Debackere, and B. De Moor, 2007. Do material transfer agreements affect the choice of research agendas? The case of biotechnology in Belgium. *Scientometrics* 71.

Seglen, P., 1992. The skewness of science. *Journal of the American Society for Information Science* 43: 628-638.

Shilling, C., and J. Bernard, 1964. *Informal communication among bioscientists*. Washington, DC: George Washington University.

Van Raan, A., 2006. Comparison of the Hirsch-index with standard bibliometric indicators and with peer judgement for 147 chemistry research groups. *Scientometrics* 67: 491-502.

Walsh, J., A. Arora, and W. Cohen, 2003. Research tool patenting and licensing and biomedical innovation. In: Cohen, W., Merrill, S. (Eds.), *Patents in the knowledge-based economy*. Washington, DC: National Academies Press.

Walsh, J., C. Cho, and W. Cohen, 2005. *Patents, material transfers, and access to research inputs in biomedical research: Final report to the National Academy of Sciences' Committee Intellectual Property Rights in Genomic and Protein-Related Inventions*. Chicago, IL: University of Illinois at Chicago.

Wesolowsky, G., 1976. *Multiple regression and analysis of variance*. New York, NY: John Wiley & Sons.

Zucker, L., and M. Darby, 1996. Star scientists and institutional transformation: Patterns of invention and innovation in the formation of the biotechnology industry. *Proceedings of the National Academy of Sciences of the United States of America* 93: 12709-12716.

Zucker, L., and M. Darby, 2001. Capturing technological opportunity via Japan's star scientists: Evidence from Japanese firm's biotech patents and products. *Journal of Technology Transfer* 26: 37-58.

Zuckerman, H., 1967. Nobel laureates in science: Patterns of productivity, collaboration, and authorship. *American Sociological Review* 32: 391-403.

Zuckerman, H., 1996. *Scientific elite: Nobel laureates in the United States*. New Brunswick, NJ: Transaction Publishers.

Evaluation of Scientific Publications in the Humanities

Adelaida Román, Mª-Dolores Alcain and Elea Giménez

adelaida@cindoc.csic.es, mdalcain@cindoc.csic.es, elea@cindoc.csic.es
CINDOC (Centre for Scientific Information and Documentation), CSIC (Spanish Council for Scientific Research) Pinar St., 25, 3º. 28006 Madrid (Spain)

Abstract
This paper describes the evolution and development of the research on the evaluation of scientific publications in the social sciences and humanities in Spain. Considering specific characteristics of the research in Humanities an specific model of scientific publications evaluation is proposed. This model is based on several indicators as editorial quality, international visibility, peer review and impact –obtained through the citation analysis- and generating an integral evaluation of Spanish scientific journals on these subjects. Due to the differentiated characteristics and working habits, levels of obsolescence, types of documents cited in the Humanities, the evaluation of other documents as books is being studied. The quality indicators that will be used in the monographs evaluation process are presented.

Keywords
scientific journal evaluation; book evaluation; quality indicators; peer review; citation analysis; humanities.

Background: evaluation of scientific journals.
Research on the evaluation of scientific publications in the humanities and social sciences began a little over a decade ago. The most interesting lines of work have revolved around the design of systems for the evaluation of scientific journals. In the last few years, universities, research centres and evaluation agencies have made a variety of proposals which are today being debated. These range from following the ISI model, which is well suited to the experimental sciences, to various attempts at models better suited to the situation and nature of scientific journals in these fields, approaching quality evaluation by means of a combination of indicators offering different approximations to it from complementary points of view.

This paper describes the evolution and development of the authors' research work on the evaluation of scientific publications in the social sciences and humanities.

Starting out from the specific characteristics of the research in these areas, and in particular in the humanities, our work has focused on applying a model for the evaluation of scientific journals (the first means of dissemination of research results examined) from the various aspects of quality.

Firsts steps on journal evaluation. Results.
This work has progressed in parallel along two paths: evaluation of formal quality aspects (the degree to which the journals conform to international publishing standards) and the quality of the editorial processes (system for evaluating manuscripts, publication mainly of original research, etc.) on the one side, and evaluation of the scientific content of the journals by means of a far-reaching consultation with university lecturers in the various areas of knowledge, on the other.
In this way we have obtained individualised data on the situation of around 2000 social sciences and humanities journals.

Moreover, to complete the scoring of each journal in the context of its field of specialisation, it was necessary to undertake an analysis of the degree of use and the actual respect accorded to the journals evaluated, by using citation analysis: the result could reveal the extent of the agreement between the assessments by university lecturers in the surveys and the degree of use (measured by citations) that the journals receive from these lecturers, in their role as the authors of the articles analysed.

Taking as a basis the results of the surveys and selecting in each speciality the journals that were given the best assessment by university teaching staff, the evaluation data were contrasted with other aspects,

such as regularity of publication and the consolidation of each journal in terms of the number of years of continuous publication, and based on this information, the source journals or citing journals were selected for the citation analysis.

From this analysis, as well as the central objective of obtaining data on the use, and the prestige and influence of each journal, significant data were obtained regarding the characteristics and scientific habits of research in each of the disciplines for which journals were analysed: types of sources used in each area for scientific work, type of documents cited, preferred working languages, types of audits, average lifetime and as a result, the degree of obsolescence of the scientific literature in each discipline, etc. These data highlighted that although the social sciences and humanities are normally treated as a unified set with similar characteristics, in reality they constitute a conglomerate of disciplines each with their own differentiated characteristics and working habits, levels of obsolescence, types of documents cited, etc.

The table below, which gives the significant data on the average lifetime of the documents cited and the distribution of citations according to the type of document, illustrates some of these assertions. The data were obtained from the analysis of the humanities journals published in 2000 and 2001 and analysed in our project as citing journals:

Table 1. Average lifetime and type of documents cited in the humanities

Disciplines	Average life of cited literature (years)	Total citations analysed 2000-2001	Journal citations	Citations to other documents
Anthropology	9	2.229	21,31	78,69
Archaeology and Prehistory	12	8.573	41,2	58,8
Greek and Latin Philology	12	4.675	29,95	70,05
Spanish Philology (excluding Literature)	14	4.675	23,91	76,09
Spanish Literature	15	5.760	16,46	83,54
Geography	8	10.613	27,14	72,86
Philosophy	6	4.603	21,81	78,19
Art	22	5.360	25,97	74,03
Arab and Hebraic Studies	17	6.993	32,46	67,54
Philology of Modern Languages	9	8,198	25,24	74,76
History	13,8	25.218	28,93	71,07

Work deriving from the results obtained in the first stage

The set of information obtained in this process was decisive for the focus of the research currently underway. Some of the observed characteristics that have had greatest influence on the decision regarding the activities developing nowadays are the following:

- Average lifetimes that range between 8 years and 24 years, depending on the discipline.
- Long publication intervals: annual publication was chosen by approximately 50% of the scientific journals published.
- Preference for use of monographs as a means of publishing research findings, and also to lay the foundations of scientific work: the majority of the cited literature is in the form of monographs.
- Uneven degree of standardisation, i.e. different levels of compliance with international publication standards are observed.
- Relatively high level dissemination in prestigious international specialised databases.
- Highly consolidated journals that are well regarded in the scientific community, in many of the humanities disciplines and in some social science specialities (Jurisprudence, for example).

For all these reasons, the research group working on publications assessment decided to:

- Propose an *integrated evaluation* model that binds together various indicators for humanities and social sciences journals which is suited to the special characteristics of journals in these areas and which takes into account aspects of editorial quality, dissemination and content quality

simultaneously, the latter being measured by both specialists' assessments and usage and influence indicators drawing upon citations analysis.
- Undertake a study of the indicators for the evaluation of the quality of monographs.

Journals integrated evaluation
Integrated evaluation is understood to mean a system for the evaluation of scientific journals based on the application of indicators expressing different aspects of journal quality. In this context, the following quality indicators are being applied on a regular basis:

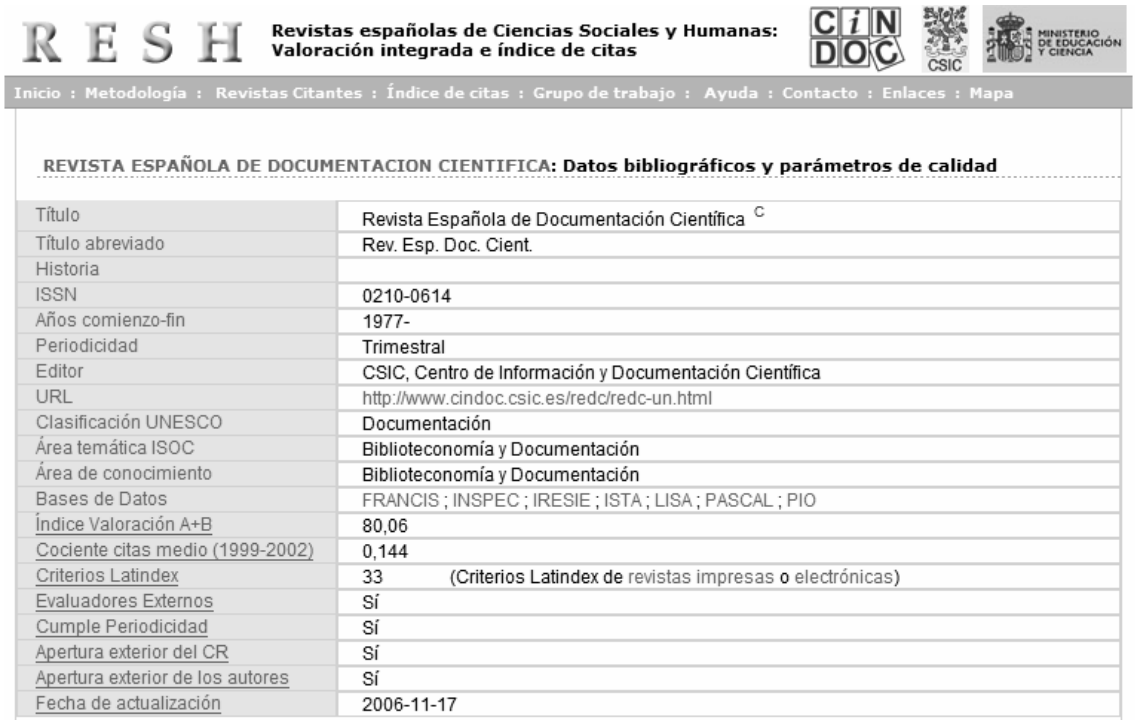
- The evaluation index (Iv A+B), representing the assessment of the quality of each journal's content made by lecturers and researchers.
- The number of *Latindex* criteria complied with, out of a total of 33 referring to different formal aspects of editorial quality and visibility.
- Existence of external referees for the selection of manuscripts submitted to the journal.
- Compliance with the stated publication frequency.
- Opening up of the editorial board to the outside or number of external members of the publishing body, so as to ensure a degree of the plurality in the editorial policy and management of the journal.
- Opening up to the outside of the authors or number of authors outside the publishing body.
- Visibility in international databases.
- The Average Impact Index (IIM). It is an indicator of use and influence for journals and is obtained by calculating the average of impact indexes obtained in the last five years. Up to know citation indexes for 1999, 2000, 2001, 2002 and 2003 have been constructed, and IIM for 2003 has been calculated. IIM is an indirect indicator of the quality of the content, and it will complete and nuance the results of the evaluations by lecturers. It should be noted that the citation ratios or impact indices are calculated based on a five-year window and not the two-year window normally used in the traditional impact factor formula. The calculation of this variant of the impact factor is consistent with the data observed: the average lifetime of the literature cited in the humanities is much longer than in other disciplines and this makes it necessary to work with longer periods when analysing citations. Researchers in the humanities do not cite literature as recent as researchers in science and technology disciplines.

Results from the journals integrated evaluation
One of the more encompassing results of the work done so far is RESH (http://resh.cindoc.csic.es), a website created in 2005 and in constant evolution, compiling all the data on the quality of journals in the social sciences and humanities (integrated evaluation):

The constant review and updating of these parameters in the case of Spanish journals are permanent goals of the research group, given the usefulness of RESH as a tool to support evaluators of scientific activity, researchers and publishers.

RESH allows users to look up the evaluation of all the journals cited in the 1999, 2000, 2001, 2002 and 2003 citations indices, and for each (around 1,300 titles), it gives the details of compliance with the parameters analysed.

Designing indicators for monographs evaluation
If the results of the citations analysis conducted so far are as expressive of the importance of monographs in the humanities, it would seem logical to study a system with which to evaluate them. To do so, the data already collected for the citations analysis will be utilised. In this new stage of the analysis it will be possible to group together monographs according to certain variables, such as the publisher, country or language. This first approach will be completed with a number of elements that are currently being debated, not only in the context of this research group, but also in that of other groups the main focus of whose work is the evaluation of research activity in this area of knowledge. Some of the indicators being considered and which would be used to prepare an initial experimental model are:



RESH Revistas españolas de Ciencias Sociales y Humanas: Valoración integrada e índice de citas

Inicio : Metodología : Revistas Citantes : Índice de citas : Grupo de trabajo : Ayuda : Contacto : Enlaces : Mapa

REVISTA ESPAÑOLA DE DOCUMENTACION CIENTIFICA: Datos bibliográficos y parámetros de calidad

Título	Revista Española de Documentación Científica [C]
Título abreviado	Rev. Esp. Doc. Cient.
Historia	
ISSN	0210-0614
Años comienzo-fin	1977-
Periodicidad	Trimestral
Editor	CSIC, Centro de Información y Documentación Científica
URL	http://www.cindoc.csic.es/redc/redc-un.html
Clasificación UNESCO	Documentación
Área temática ISOC	Biblioteconomía y Documentación
Área de conocimiento	Biblioteconomía y Documentación
Bases de Datos	FRANCIS ; INSPEC ; IRESIE ; ISTA ; LISA ; PASCAL ; PIO
Índice Valoración A+B	80,06
Cociente citas medio (1999-2002)	0,144
Criterios Latindex	33 (Criterios Latindex de revistas impresas o electrónicas)
Evaluadores Externos	Sí
Cumple Periodicidad	Sí
Apertura exterior del CR	Sí
Apertura exterior de los autores	Sí
Fecha de actualización	2006-11-17

Figure 1. Spanish journals on Social Sciences and Humanities system (RESH)

- Presence of monographs, or where appropriate, publishers in general and specialist research library catalogues, and in both national and international bibliographic databases.
- Summaries of monographs in scientific journals in the relevant field of knowledge.
- Citations received by the monographs. In order to obtain this information we will start, as mentioned, with the information collected and processed so far (bibliographies included in articles in Spanish humanities journals between 1999 and 2003) and this will be completed as the project advances.

The main difficulty involved in this work is the enormous quantity of monographs emerging from the journal citations analysis already carried out. It is possible that the complexity and volume of monographs the research group encounters will make it advisable to tackle the evaluation of publishers first and then proceed to evaluate each monograph later. It may be envisaged that the model of evaluation proposed will be applied to some of the humanities disciplines, which in the light of the results, have shown themselves to have a longer average life, and perhaps as a result, a higher rate of citations of other types of document (other than journals).

This line of research would represent one of the first attempts at evaluating monographs in Spain, notwithstanding the work already begun by the research group at the Documentation Faculty of the University of Barcelona (Borrego, 2004).

Future research lines

Foreseen activities for current year are:

- Preparation of the citations indices for 2004 and 2005.
- Redefining the international dissemination indicator based on the presence of journals in international databases, taking into account the criteria applied by each database when deciding whether to accept journals. We also propose to establish a classification of databases and giving weightings to the presence of journals based on this classification.
- Designing a number of quality indicators for the evaluation of monographs, as mentioned previously, in order to establish an initial experimental model that will allow progress to be made towards evaluating this type of text, given that monographs represent the publication channel for 60% of research in the humanities in Spain.

References

Borrego, A. (2004). *La evaluación de revistas en procesos de evaluación de la investigación española en humanidades y ciencias sociales: desarrollo de un modelo de experimentación en el área de la Filología Hispánica.* Report. Retrieved May 16, 2006 from: http://wwwn.mec.es/univ/html/informes/estudios_analisis/resultados_2004/ea0025/ea0025.pdf

European Science Foundation. Standing Committee for the Humanities. (2001). *Abstracts from Workshop on the Evaluation of the Scientific Production in Social Sciences and Humanities*.

Giménez, E, Román, A. & Alcain, M.D. (in evaluation process). From experimentation to coordination in the evaluation of Spanish scientific journals in the humanities and social sciences. *Research evaluation*

omán, A. Elaboración de una propuesta integrada de categorización de las revistas españolas de humanidades, con la incorporación del índice de citación recibido por cada revista en los años 2000, 2001 y 2002, como indicador del uso y el prestigio. (2005) Report. Retrieved November 20, 2006 from http://www.mec.es/univ/proyectos2005/EA2005-0012.pdf

White paper on research in the Humanities (2006). Madrid: FECYT.

Internationalisation of Patents by Public Research Organisations from a Historical and an Economic Perspective[1]

Ana Romero-de-Pablos[*] and Joaquín M. Azagra-Caro[**]

[*]*anaromero@ifs.csic.es*
Departamento de Ciencia, Tecnología y Sociedad, Instituto de Filosofía, CSIC, Pinar 25, 28006 Madrid (Spain)

[**]*joaquin-maria.azagra-caro@ec.europa.eu*
INGENIO (CSIC-UPV) and IPTS (European Commission - Joint Research Centre), Edificio Expo, Inca Garcilaso s/n, E-41092 Sevilla (Spain)

Abstract
Concerns on the organisation of science about how academics generate patents tend to focus on a single set of patents, either national or international. The main objective of this research is to study both in order to understand their differences. We adopt historical and economic perspectives to address the issue, with data from the Spanish National Research Council (CSIC), the largest PRO in Spain. We distinguish three periods in the history of CSIC, according to the political context: dictatorship (1939-1975), transition to democracy (1976-1986) and democracy (1987-to date). The legal and institutional framework conditioned the evolution of patenting by CSIC along these periods, finally characterised by a strong internationalisation of patents. For the late period, we explore the trends of some other economic inputs of patents and we show that not all of them lead to every kind of patents. We conclude that the political and normative context may shape the culture of internationalisation of patents at PRO like CSIC and that increasing technological cooperation has supported this internationalisation. However, very often foreign partners are included in the application to extend protection abroad for commercial reasons, so their number may not be a good indicator of inventive activity.

Keywords
academic patents; internationalisation; PRO

Introduction

University patents have become an important object of analysis because they reflect changes in knowledge production and diffusion; they are an indicator of research results that may be taken into account in the evaluation of merits and a potential instrument for technology transfer. The studies carried out so far tend to apply a single disciplinary perspective, e.g. historical (Mowery and Sampat, 2001; Metlay, 2006), economic (Feller, 1990; Pavitt, 1998), scientometric, etc.

Somewhere else we have shown that similar debates should expand to the other main knowledge producer in the public R&D system, namely Public Research Organisations (PRO), and that a multidisciplinary approach allows for a richer interpretation of the results (Azagra and Romero, 2006). In this paper we apply this philosophy to the study of a question that has been scarcely examined by the literature: what are the factors leading academic institutions to patent abroad? While most studies focus their attention on one set of patents only, one of us has shown that differences among sets of patents matter in the analysis of their generation, e.g. national vs. international patents, since normative conclusions can be raised (Azagra et al., 2006).

Our object of study will be the Spanish Council for Scientific Research (CSIC), the largest PRO in Spain according to its historical trajectory, its resources, its wide regional implantation and its multidisciplinarity across research units. It is also very active on patenting, holding more than 10 percent of Spanish patent applications nowadays.

Processes of patent internationalisation

Sovereign states legislate on intellectual property rights. Inventors can apply for a patent through their respective national patent offices. Most developed countries have eventually allowed universities to

[1] This work was supported by CSIC's Science, Technology and Innovation (CTI) Network.

apply for patents and license them. The traditional way to protect knowledge abroad is to apply for a patent in the foreign country, which requires a new administrative procedure. In order to simplify it, states may agree to sign international treaties to allow other participants to apply for a patent in their territory through a common procedure. Citizens of the participating countries can designate in which of these countries they want protection, including their own.

A country's decision to further engage into internationalisation processes through a treaty is a particular regulatory change which deserves attention. It aims not to increase the number of patents per se but to facilitate patenting abroad. However, this possibility may make it more interesting for applicants to patent, thus increasing the overall number of patents.Most European countries participate in two main pacts. First, the Patent Cooperation Treaty (PCT) (Washington, 1970), that gave rise to the so-called PCT patent. Second, the European Patent Agreement (EPA) (Munich, 1973), that gave rise to the so-called European Patent[2]. Application takes place through the International Office of the World Industrial Property Organisation and the European Patent Office (EPO), respectively, but national patent offices may also start the procedure.[3] However, we should clarify that, unlike the EPA, the PCT system does not end with the granting or refusal of a patent but is limited to processing and handling applications.

From dictatorship to democracy in Spain: milestones in the history of internationalisation of CSIC

The CSIC was created after the civil war in 1939. The new leaders felt that the best way to control and organise research activity was to group or divide it into boards. The history of the Board of Technical Research (*Patronato Juan de la Cierva*, hereafter PJC) and therefore the work carried out at its offices cannot be studied separately from the situation in which Spain found itself, not only after the civil war but also after the Second World War. Spain's difficult international position, in addition to its internal need to deny all acknowledgements of what had happened prior to 1936, meant that this external and internal isolation became a liability that took a long time to eradicate. In a post-war situation, the particular interests of companies had to give way to national interest, and this conditioned the structure of CSIC; the aim was to make this body the basis of scientific mobilisation at the service of great national concerns.

1977 was a year of changes for CSIC: a new internal restructuring was put into practice with the first democratic mechanisms. 1986 is the last of the outstanding historical milestones for this article; it is a watershed in Spanish science policy. It is the year of the Act to Promote and Generally Coordinate Scientific and Technical Research, the so-called "Science Act" (Act 13/1986, Official State Bulletin of the 18th April 1986). Although this act must also be placed in relation to Spain joining the European Economic Community, we must not forget that it was the result of the growing awareness of the importance of scientific research for Spain. The new design of organisation and control specified by this act not only aimed to be more effective, capable of managing the new funding, and at the same time capable of handling the start-up of initiatives such as the National Plan for Scientific Research and Technological Development in 1988, but this new act also sought to provide a response to the new situations arising as a result of the arrival of democracy.The engagement in the process of internationalisation of patents also experiences strong regulatory support during this period. In July 1986, although it did not become effective until October 1987, Spain joined the EPA. Another important fact to bear in mind is Spain joining the PCT, which came into force in November 1989. The PCT system normally used by CSIC to place its patents on the international market has advantages as it facilitates the application procedure at an international level.

[2] It should not be confused with the existing project of enabling a single patent for all members of the European Union.

[3] There is a procedure for PCT extensions to designate Europe for protection, that is to say, all countries that signed the EPA. It makes unnecessary to designate each one separately and reduces bureaucracy, since EPO manages the application and not several national offices. This procedure, called PCT-European route, is equivalent to apply for a patent through the European route if no more countries outside the EPA are designated.

Data and methods

The historical approach
The patent data used in this article come from various sources. For the case of patents applied for during the period when CSIC was organised into Boards (1939-1975), the historical database of the Spanish Patents and Trademarks Office (OEPM) has been used. We could only find patents applied for by PJC, despite our exhaustive search of patents applied for by all Boards. Then we completed and compared the database with the PJC Annual Reports. From 1975 and up to 2005, data have been used kept by the Technology Transfer Office (TTO) of CSIC, as well as a database generated based on CSIC Annual Reports, allowing us to associate CSIC patents with specific research units and, through these, to ascribe them to specific scientific areas and regions .

The first patents registered by the OEPM from the PJC are from 1949, the year when the analysis therefore starts. Chronologically, there are three periods established, periods that respond to the already mentioned historical and legislative changes. The first period goes from 1949 to 1975, the year of structural change in CSIC with the disappearance of the boards. The second period, shorter in time, is from 1976 to 1986, the year when Spain joined the EEC and two decisive acts were passed for patents in Spain, the Patent Act and the Science Act. The third period starts in 1987 and lasts until 2005. Within this time framework, the evolution of applications is analysed, as well as the scientific areas registered, their distribution over space, collaboration with other groups or research units and their venture into the external world.

The economic approach[4]
We have more detailed data for recent times, so the question is how to choose the appropriate cutting edge. The historical approach justifies defining the late period, 1987-2005, for the economic analysis, given the more homogeneous legal and institutional framework. This is an example of cross-fertilisation between the two approaches. The historical approach employs data from CSIC TTO in the late period that we also use here. It consists of 2,793 patents applied for by CSIC. We classified them by an average of 95 CSIC research units per year and 19 years (1987-2005), so we built a 1,812-observation panel. As far as we are aware, this is one of the largest databases in Europe for the study of public patents at institutional level.

We applied count data econometric models in order to estimate the determinants of CSIC patents.
Independent variables include those on collaboration patterns from the same TTO database: logarithm of number of co-applicants, by type (other CSIC research units, other PRO, universities[5], business firms and other institutions) and nationality (national and foreign). A second database from CSIC's memories allows us to include other independent variables on the research units' characteristics. These are total staff (in logs) and their composition (proportion of CSIC civil servants, CSIC working staff, university personnel and other personnel –mainly temporary), per capita amount of financial activities (national projects, other projects, contracts, in logs), scientific area, region[6] and being part of a service centre –an institution for the joint management of research units.
Since we are looking at research units, we have to detect the presence of outliers. There is one in the sample, the Institute of Chemistry (ITQ)[7], so we create a dummy variable to allow it to have a separate intercept. Finally, we will control for time differences through the length of each CSIC presidency under each political party in the government. There have been eight CSIC presidents and three political parties in Spain in the period of observation.

[4] In rigour, it is a microeconomic approach. A macroeconomic approach to analyse whether patent internationalisation by PRO brings forth any impact on the overall economy is out of the scope of this paper and could be the object of further research.
[5] According to Spanish law, patents applied for by joint centres have both CSIC and universities as the assignees. We consider that technological cooperation has taken place only when CSIC research units apply for a patent with a university of which they do not form part. Therefore, we checked one by one the cases of co-applications by CSIC and university and counted just those in the latter situation.
[6] We refer to 10 Spanish NUTS-2 regions plus Rome. We take Castile-Leon and Balearic Islands as a benchmark.
[7] ITQ holds 20% of all patent applications. The next research unit *just* holds 8%.

The growth of international patenting by CSIC, 1949-2005

1949-1975: scarce internationalisation
Although only sixteen PJC applications were registered outside Spain in this first period, it is worth paying them some attention. It was in 1971 when Spanish patents were first registered abroad. Places in Europe such as France, Belgium, Great Britain, Germany, Italy, Holland, Poland, Norway, Yugoslavia, Russia and Turkey were chosen. Nevertheless, Spanish interest did not focus solely on Europe; the data indicate that patents were also registered outside the continent: the United States, Canada, Mexico, Brazil, Argentina, Peru, Morocco, Algeria, Japan and Australia are the other countries reached by PJC patents. The aforementioned situation facing Spain after the Second World War, namely that of autarky, the agreements with the United States, all of this undoubtedly affected the country's scientific life. Nevertheless, perhaps this is the right place to mention the Cultural Relations Board, an organisation dependent on the Ministry of Foreign Affairs, as it allows us to tackle an important point when analysing the use of knowledge abroad generated by CSIC, namely the training of Spanish scientists in foreign research centres. This, which had been one of the points forming the basis of the JAE, forms part of the legacy that José Maria Albareda, general secretary/junior minister of CSIC, took on and tried to strengthen (Delgado and Elizalde, 2005). The relations established through the scientists themselves explain this situation.

1976-1986: incipient internationalisation
In this period, between 1976 and 1986, some degree of internationalisation is also present. There are sixteen patent applications for protection outside Spain. Germany, Belgium, France, Holland, Italy and Yugoslavia are the European countries where attempts were made to protect this knowledge. This period was also marked by an outstanding lack of presence of Spanish research abroad. These were years of internal organisation of the body, of starting up new policies and new regulations whose results would be felt in later years.

1987-2005: strong internationalisation
The increase in patent applications, starting in the 80's, continued until the middle of the 90's, when there was another significant jump. Table 1 allows us to evaluate more precisely, what this growth really meant. First are the total patent applications by CSIC throughout this period. In the next three columns, there is a breakdown of this total. The second column includes applications for the year made in Spain. In the third column, we can see foreign applications directly applied for abroad trough national procedures. The last column contains data on international extensions, i.e. families of the two former sets, generated after application for through an international procedure. The data for total applications in this period show a growth at the end of the 80's which, over the years, not only consolidates but even almost triples during the following years. This growth not only justifies the necessary legislation on patents but is also a reflection of the different regularisation initiated during the last few years of the 70's and principally the 80's. This is the time when, within the Ministry of Education and Science, the Secretary of State for Universities and Research and the General Directorate of Scientific Policy were created.

During the late 80's, the number of Spanish applications was almost coincident with the total. Up to 1995, they used to represent more than 50 percent. From 1995 onwards, there was a reversal and international patents (the sum of foreign applications plus international extensions) became more than half of the total. As already mentioned, change occurred after Spain joining the European Community and everything this entailed. Of course, direct foreign applications are rare as compared with international extensions, after the EPA and the PCT came into force.

There is a lot of diverse information that makes it possible to break down these data. The countries that received the greatest number of Spanish patent applications between 1987 and 2005 were the United States (154), Japan (69), Canada (65), the United Kingdom (54) and France (50). Other interesting data are that, during this period, 487 applications were processed in the international market via the PCT treaty and 168 of them were processed via the EPA. These data indicate

institutional strategies but at the same time, and related to the previous point, they also lead us to think some scientific areas are more highly internationalised than others are.An analysis of the families

Table 1. CSIC patent applications (1987-2005)

Years	Total	Spanish applications	Foreign applications	International extensions
1987	9	9	0	0
1988	19	19	0	0
1989	45	43	0	2
1990	62	42	0	20
1991	54	25	1	28
1992	44	27	0	17
1993	68	31	0	37
1994	65	37	0	28
1995	57	37	0	20
1996	102	51	0	51
1997	142	63	1	78
1998	190	63	3	124
1999	243	94	4	145
2000	238	101	4	133
2001	250	98	1	151
2002	342	123	2	217
2003	316	94	3	219
2004	251	102	4	145
2005	296	111	8	177
Total	2793	1170	31	1592

according to scientific area can provide us with an answer to the question of whether there are areas highly accustomed to or traditionally applying for patents abroad. Because the areas of Chemistry and Biology & Biomedicine are those applying for most patents, both nationally and internationally, a difference can be appreciated in the countries designated to obtain protection. In the area Chemistry, it is of interest to protect knowledge most in the United States, then in Japan, Canada, the United Kingdom and France. On the other hand, in the area of Biology & Biomedicine, patents were applied for in countries such as the United States, Canada and Japan and, unlike the previous case, Australia.

Determinants of CSIC international patents by research unit in the late period, 1987-2005

In Table 2 we show our favourite specifications: a zero inflated negative binomial model for Spanish, national, patents and a standard negative binomial regression for international extensions. We did not try to fit a model for direct international patents since their number was so small. According to the estimations, cooperation with other national PRO and national business firms has the highest, positive, effect on patent production[8]. Having other CSIC research units and Spanish universities as co-applicants also enhances national patent production, but to a lesser extent. Cooperation with other institutions does not have a significant effect on the production of CSIC patents. Until now, this is valid for both national and international patents. When it comes to foreign partners, some differences arise. While cooperation with foreign PRO, foreign universities and foreign firms is not significant for increasing national patents, it is positively significant for increasing international patents. Regarding human capital, the higher the number of total staff, the more likely it is that CSIC research units generate national and international patents.

[8] Due to the limited extension of the paper, here we focus the interpretation of the models in the distinction between national patents and international extension. For more discussion on the implications of the models, see *self-quotation #1*.

Tabla 2. Determinants of the number of CSIC patents by research unit and year

	Number of CSIC Spanish applications	**Number of CSIC international extensions**
Best-fit model	Zero Altered Neg. Binomial Regression	Negative Binomial Regression
Number of observations	1,812	1,812
Poisson-associated R^2	0.55	0.76
Log likelihood function	-1,363.13	-1,098.77
Prob[χ^2>value] =	0.00	0.00
	Coeff. (t-ratio)	Coeff. (t-ratio)
Constant	-1.51 (-5.6) **	-10.03 (-11.47) **
Technological cooperation with		
Other CSIC research units	0.74 (5.39) **	0.42 (2.19) *
Other national PRO	1.16 (1.96) *	1.79 (2.66) **
Foreign PRO	1.56 (0.33)	1.02 (3.58) **
National universities	0.78 (6) **	0.51 (2.88) **
Foreign universities	0.7 (1.6)	0.66 (2.06) *
National business firms	1 (4.21) **	1.29 (8.29) **
Foreign business firms	0.13 (0.09)	1.18 (4.73) **
Other institutions	-0.25 (-0.38)	1.3 (1.49)
Human capital		
Total staff	0.33 (6.54) **	1.49 (10.86) **
Proportion of CSIC civil servants	-0.04 (-0.28)	1.79 (3.52) **
Proportion of CSIC working staff	-0.67 (-2.78) **	-3.25 (-3.02) **
Proportion of university personnel	-0.55 (-3.21) **	-0.72 (-1.02)
Financial capital		
National projects	0.06 (2.03) *	-0.09 (-0.79)
Other projects	0.03 (1.46)	0.14 (1.58)
Contracts	0 (0.09)	0.03 (0.55)
Scientific area		
Chemistry	0.41 (5.95) **	0.91 (4.77) **
Food	0.4 (4.95) **	0.63 (2.62) **
Materials	0.35 (4.98) **	-0.09 (-0.45)
Agriculture	0.11 (1.71)	-0.4 (-1.37)
Natural resources	-0.15 (-2.47) **	-1.38 (-3.94) **
Region		
Andalusia	0.16 (1.87)	0.94 (1.82)
Valencian Community	0.13 (1.27)	0.78 (1.42)
Madrid	0.08 (0.98)	0.76 (1.46)
Asturias	0.19 (0.99)	-28.89 (-0.32)
Catalonia	-0.01 (-0.15)	0.15 (0.27)
Galicia	0.04 (0.39)	-1.22 (-1.45)
Aragon	-0.06 (-0.6)	0.61 (1.04)
Murcia	-0.14 (-0.87)	0.65 (0.91)
Canary Islands	-0.23 (-1.51)	0.7 (0.99)
Service centre	0.22 (4) **	0.92 (5.56) **
ITQ	1.82 (12.39) **	3.74 (8.16) **
CSIC president_Party in the government		
Pres1_Party1	-0.08 (-1.2)	-3.46 (-3.8) **
Pres2_Party1	0.08 (1.23)	-0.29 (-1.01)
Pres3_Party1	-0.13 (-1.08)	0.15 (0.29)
Pres4_Party2	0.09 (0.44)	0.06 (0.08)
Pres5_Party2	0.13 (2.34) **	0.86 (4.03) **
Pres6_Party2	0.22 (3.35) **	1.41 (6.12) **
Pres7_Party2	0.14 (1.71)	1.04 (3.37) **
Pres8_Party1	0.17 (1.49)	1.86 (5.07) **
α	0.13 (2.14) *	1.33 (8.68) **
τ	-5.24 (-5.37) **	

** Significant at 1%. * Significant at 5%

Research units with a higher proportion of CSIC civil servants (42% on average) patent through the national route so much as those with a higher proportion of other personnel –the benchmark (38% on average). Research units with a higher proportion of CSIC working staff (10% on average) and university personnel (10% on average) patent significantly less than the former through the national route[9] Having a higher proportion of CSIC civil servants also enhances application for international patent extensions, whereas a higher proportion of CSIC working staff also diminishes it.
Regarding financial capital, the per capita number of national projects leads to a higher number of national patents but not to more extensions. Not any other funding actions are conducive to more patents.

It is important to control for differences across scientific areas, since the propensity of some of them to patent is higher, and this is the ranking: Chemistry and Food patent more than average; Materials has also more national patents but not more international patents than average; Physics, Biology & Biomedicine and Agriculture represent the average; Natural Resources patent less than average.We have also controlled for regional variation, but it does not seem to affect patenting. On the contrary, being part of a service centre is significantly positive for patenting.ITQ is actually an outlier, the propensity of which to apply for patents we must control in order to moderate the impact of other effects. The significant, positive sign indicates that it stands out from other research units in both national and international patents.

The effect of successive CSIC presidencies allows us to distinguish between two sub-periods along the period of observation for international patents: the propensity to internationalise at CSIC research units significantly increased from the fifth president onwards. It corresponds to the shift attributed to internationalisation by the historical approach. The impact is not so clear for national patents, though. There was an upward shift with presidents 5 and 6, which was not sustained thereafter. On turn, notice that under president 4 patents were the same regardless party 1 or 2 was in the government[10]. Moreover, the rise of patents continued at the end of the period no matter party 2 or 1 was in the power. This leads us to think that governmental changes have not been influential.

Conclusions
From the historical point of view, the configuration of an eventually clearer institutional and legal framework has favoured the creation of CSIC international patents. In 1986, the so-called Law of Science came into force, linked to the admission of Spain in the European Union and to the new Constitution of Spain and its territorial and administrative organisation.
The late eighties were also crucial because many other legislative changes took place. Until 1986, patent law in Spain dated back to 1929. The admission into the European Union forced the adaptation to the European legislation. The 1986 Patent Law made it possible for Spain to patent chemical, pharmaceutical and food technologies for the first time. Moreover, in July 1986 Spain joined the EPA, which allowed applying for patents through a centralized system. Soon afterwards, Spain joined the PCT, the system most used by CSIC to internationalise its patents.

Once this general framework was set, the economic approach shows that the effect on the rise on international extensions by CSIC was significant after some years, when the presidential style of CSIC assumed these cultural changes. On turn, the changes were solid enough as to be independent of later rotations of the political party in the government. We may wonder whether this innocuous effect of late governmental policy is due to a rhetoric common to left and right parties that does not encompass different actions regarding PRO patents any more.

[9] Carayol and Matt (2004) find that the combination of permanent and temporary researchers enhances patenting and so does the combination of full time researchers and university professors –at least in their Class 4 labs. Our results lead us to agree with the former but not with the latter. More evidence on the negative trade-off between teaching and research is in *self-quotation#2*, where one of us found that the presence of joint research centres was not enough for universities to patent more. It looks like that mixing research with teaching has advantages and costs and the outcome depends on the context (Martin, 2003).
[10] Pres4_Party1 is the benchmark.

The historical approach detects that technological cooperation has been stimulating for CSIC to engage into patent application. There is a parallelism between the increase of patents and the increase of the number of participants in their production. Moreover, this process goes in parallel with the trend of internationalisation of patents, which it is not a casual coincidence.

The economic approach quantifies the impact of technological cooperation in the late period and finds a significant relation with patenting at CSIC. While domestic partners are more conducive to patent production in general, foreign partners are also advantageous for international extensions. A second reading is that once CSIC has applied for Spanish patents, it searches foreign partners to extend protection abroad[11]. Hence, foreign institutions, especially foreign firms, are expected to provide commercial expertise rather than inventive knowledge. Studies on a single set of patents which predict that a higher number of foreign applicants will increase the number of patents should be cautious about the interpretation of the reasons why it may occur.

References

Azagra-Caro, J. M. & Romero-de-Pablos, A. (2006). Patenting at public research organisations: a historical-economic approach, *DIME-Universities Working Paper Series*. Retrieved from: http://uil.dime-eu.org/views/uil_publications.

Azagra-Caro, J. M., Yegros-Yegros, A. & Archontakis, F. (2006). What do university patent routes indicate at regional level? *Scientometrics* 66 (1): 219-230.

Carayol, N. & Matt, M. (2004). Does research organization influence academic production? Laboratory evidence from a large European university. *Research Policy*, 33, 1081-1102.

Delgado, L. & Elizalde, M. D. (Eds.). (2005). *España y Estados Unidos en el siglo XX*. Madrid: Consejo Superior de Investigaciones Científicas.

Feller, I. (1990). Universities as engines of R&D-based economic growth: They think they can. *Research Policy,* 19, 335-348.

Martin, B. (2003). The changing social contract for science and the evolution of the university. In A. Geuna, A. J. Salter & W. E. Steinmueller (Eds.), *Science and Innovation*. Cheltenham: Edward Elgar.

Metlay, G. (2006). Reconsidering Renormalization: Stability and Change in 20th-Century Views on University Patents. *Social Studies of Science,* 36 (4), 565-597.

Mowery, D. C. & Sampat, B. N. (2001). University Patents and Patent Policy Debates in the USA (1925-1980). *Industrial and Corporate Change*, 10 (3), 781-814.

Pavitt, K. (1998). Do patents reflect the useful research output of universities? *Research Evaluation,* 7, 2, 105-11.

[11] We tried to verify this statement by having a closer look at patent families and found that it is not rare that extensions have more applicants than the mother patent. A CSIC's TTO manager verbally certified that it makes sense since it is part of CSIC's strategy.

Gender, Funding Diversity and Quality of Research

Ulf Sandström[*] and Martin Hällsten[**]

[*]*ulfsa@isak.liu.se*
Department for Studies of Social Change and Culture, Linköping University, SE-58183, Linköping (Sweden)

[**]*martin.hallsten@soc.su.se*
Dept. of Sociology & Swedish Institute for Social Research, Stockholm University, SE-106 91, Stockholm (Sweden)

Abstract
We analyze the relation between funding and output using bibliometric methods with field normalized data. Our approach is to create a connection between bibliometric data at the individual researcher level and data on incoming grant (funding) using the specific personal ID-number (social security code). Data on funding include the person responsible for the grant. All types of research income are considered in the analysis yielding a project database with a high level of precision. Results show that productivity can be explained by background variables, but that quality of research is un-related to background variables. This is a paper in progress: our ambition is to extend it in several ways: theoretically, analytically and empirically.

Keywords
gender; research quality; productivity; performance indicators; citation analysis; field normalized methods.

Introduction

Reviewing the literature on careers and innovation Dietz and Bozeman (2005) remark that there are two almost separated literatures on these issues – the management of innovation studies on the one hand and the study of productivity by economists and sociologist on the other hand. Regarding the impact of university scientists and engineers there is by tradition studies that focus on a single sector, either the industrial track or the academic. For those working on academic productivity social factors have been put in the background, favoring instead measures of institutional and personal prestige (Merton, Coles, Long etc.). A number of integrative approaches like Stephan and Levin (1992) or socially-embedded approaches (Callon) have signaled a change in the discussion.

The literature on faculty research productivity is voluminous. We refer to reviews like Hattie & Marsch (1996), and studies like Fox (1992), Ransom & Megdal (1993), Toutkoushian et al (2003) and Lee and Bozeman (2005) for overviews of different research questions of importance. In this context it is clear that the focus on productivity has created a number of problems. Even if these studies are based on summative indices counting different output they all have disadvantages and are open for criticism. The period of time chosen has to be adjusted for, the prestige of the different publications channels and prestige within the respective channels have to be accounted for, and this leads to methodological insecurity and weaknesses. Variations across disciplines due to the nature of work being performed and conventions for communicating research are at the heart of this debate. Qualities of the research have been one of the issues since Braxton & Bayer (1986), but to some extent straight count of citations add more to the problems than to the solutions.

Another important issue is the level of data gathering and analysis. The institutional or departmental level is often used for these studies (Fox 1992; Porter & Umbach 2001; Toutkoushian et al 1998), e.g. departments are ranked on the basis of their total publications. Our understanding is that the departmental level is an aggregation of several research groups of very different sizes and composition. In practice, some of the groups might consist of people at other departments and other institutions. This creates a summation problem, at least for those who are content with counting publications.

Instead, if we use field normalized citations counts (Moed 2005) as a proxy for research quality we come around a number of problems regarding the counting, the level of analysis, etc. The analytical advantage of field normalization is that number of publications, publication strategy, and channels of

publication becomes more or less irrelevant for the results. By the journals a researchers publishes in he or she points out who are the actual colleagues. The most specific feature with the field normalization method is that a researcher is compared to reference values that are relevant for the actual research done. Another feature of the method is that it is scale-independent; the numbers of publications are necessarily of low importance (above a certain threshold of approx. 5-10 full publications per individual researcher).

Using field normalized methods makes it possible to gather data on research leaders or grant holders only, and using them as quality indicators for their respective groups (even if they are not leaders of the group). Here, in this paper, we build on data on publication output and relate that to input data from the university databases.

Data and Methods

In Sweden, as in many other countries, the university administrations have built large data sets on each department and in some cases, on the level of individual researcher. This specific Swedish university (made anonymous to the reader) has a personnel database that covers all researchers (and all other staff) during the period 2000-2006. Data given is personal ID-code (social security number), status (professor, researcher, associate professor etc.) and department. In some cases, but not all, it is indicated to which research group the person is assigned to. This is a service that will be introduced in 2007 for all researchers.

The university also holds a database for all grants received during the same period of time. Now, remember that the personal ID-code is used also in this database. Among other things this gives information on gender. Every grant that is received goes to a specific researcher responsible for the project. Consequently, it is possible to track all research money coming in to the university to a specific researcher (and his or her group).

The financial database also covers who is financing the project (grant giver), project number, title of the project, funding body, starting year, length of project in time (yrs), and department of the researcher. There is also a classification of the financiers in the university database in 17 classes according to type of organization (foundation, research council, governmental authority etc.).

From these data it is possible to be more precise on different aspects of funding that goes together with grants. Types of financing and differences between organizations in Sweden are quite well researched. Sandström, Hällsten and Heyman (2005) proposed a classification of funding bodies according to a theoretical analysis based on Stokes book *Pasteur's Quadrant* (1997). Interestingly, they come up with the idea of splitting the financing landscape into three broad categories:

1) Organizations focused mainly on basic research, (fundamental and no obvious application)
2) Organizations focused mainly on strategic research (fundamental but with applications for industry in the future
3) Organizations mainly focused on user-need research (applied and with applications).

The categorization of contracts should not be done in a mechanical manner. It is obvious that some of the larger granting bodies have contracts of several types. Therefore, every contract has to considered by its own and conclusions not to be automatically drawn from the classification scheme. It gives a hint on the position of the funding body, but each contract can be of a very different type. Figure 1 illustrates the Swedish research landscape in FY 2003 using this categorization.

Inspired by this we have classified every financing body for the university to one of these three categories. We have 1135 contracts in the database covering the period of 2001-2005. The distribution of contracts was skewed: 13 per cent from basic research bodies, 43 per cent from strategic and 44 percent of contracts from user-need funding bodies. Counted as proportion of research money per year the basic funds accounted for 14 percent, strategic for 54 percent and user-need for 31 percent. One

conclusion is that this university is to a large extent dependent on contracts with strategic research bodies, and to a less extent dependent of research money from basic research councils.

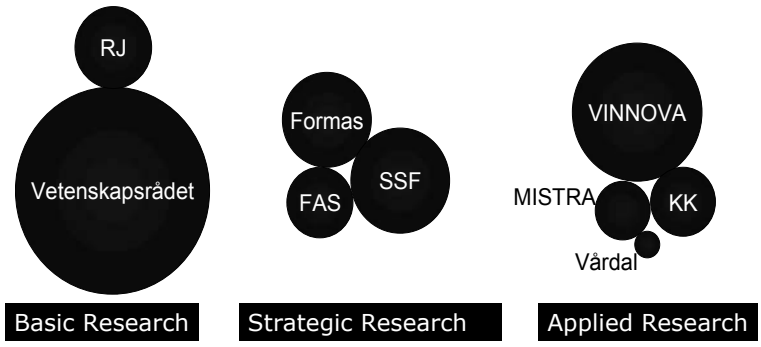


Figure 1. The Swedish Research Landscape (2003). Main financiers of research according to their general orientation. Size of each funding body is proportional to budget in 2003.

The third type of data that we will use in this paper is collected from the Thomson/ISI Internet Web of Science database. All papers from the selected 151 research leaders were downloaded using manual techniques for bibliometric identification. No automatic or matching techniques between CVs and ISI-data were used. Additionally, in order to build up reference values all journals in all subject categories where these researchers are publishing were downloaded. Data were then gathered in a SQL database and PHP programming techniques were applied. Using the ISI UT-number as identifier we managed, to connect the personal ID-code, name and department to author names in the downloaded ISI-data. This made it possible to get analytical results from the SQL-database in the form of exactly the author part of articles that were identified as belonging to the 151 university researchers. Included in these data is information about number of authors, number of addresses, author fraction of article, first/middle/last author. Gender of author follows with the personal ID-code.

We apply the bibliometric methods put forward by e.g. the Leiden and Leuven groups (van Raan 2004; Glänzel 1996). The methods for normalisation take all the relevant information that follows with the bibliometrics data into account. Each journal in the Thomson/ISI database is assigned to one or several subject codes (e.g. microbiology, biochemistry & molecular biology etc.). Our analysis focus on Citations per Paper (CPP) in relation to Journal Citation Score (JCS) and CPP in relation to Field Citation Score (FCS). In both cases the *mean* citation score for the journal set is the point of reference for the calculations (van Raan 2004). Self-citations are included in the data and we apply an open citation window. Nederhof & Visser (2005) shows why an open window gives an advantage. The bibliometrics is based on data for 1857 articles (full count).

Via the personal ID-code these results were coupled to financial data and an OLS regression analysis were performed. The analyses can e.g. answer the question whether it is good to have grants and contracts with many types of granting/contracting bodies (factor: broadness) or not. What impact follows from the share from different sources (factors; sharebas, sharestra, shareuse)?

In developing this paper we will add information on floor funding (ratio between floor funding and external funding for the department). Financial data on this aspect is not available per person as floor funding is related to the department. Additionally, we will include information on status of researcher (professor, associate professor, assistant professor), and also more detailed information on differences between departments. E.g. it would be interesting to be able to distinguish between fields of research that are prioritized by the university, or by the main funding bodies for research.

Findings

Descriptives are given in Table 1. Out of 151 researchers 123 are publishing in international journals. We have decided to include all persons in analyses as we are studying publishing behaviors and quality of publications. But, this is still an open question that we will have to dwell on in our future work.

As gender is one of the central dimensions for this paper we will start this section by some descriptive statistics on the results concerning differences between male and female researchers. The average sums of total grants (2000-2005) for females are about half the size of male researchers. Females are at larger departments, but in general they have the same ratio between full professors and other personnel. Female researchers are narrower than their male colleagues regarding the number of financing bodies, 1.9 for females and 2.1 for males. Shares from different types of funding sources are about the same, but although differences are not significant it is indicated that females have higher proportion of strategic research than males, likewise females are less dependent on basic science resources. Men do have a higher number of grants.

The average number of publication from this group is 12.3 publications during the period of eight years (1998-2005). Female researchers do have a lower productivity per person, 11.3 publications, while male researchers produce 17.2 publications in general. Evidently, productivity is lower, but counted as quality of publications (crown indicator CPP/FCSm) female researchers are doing better than their male colleagues. While females are 34 percent better than the world average, males are 27 per cent better only. The gender gap in productivity then is transformed to an efficiency gap in favor of female researchers that do have five per cent higher quality on their papers.

Table 1. Descriptives for 151 university researchers.

variable	observations	mean	stddev	min	max	percent	number
age	151	47.272	8.640	29	68		
male	124			0	1	62.9	78
total_grant	151	3.158	4.063	.5	26.98995		
broadness	151	1.947	0.807	1	4		
sharebas	151			0	1	9.5	14
sharestra	151			0	1	53.1	80
shareuse	151			0	1	35.5	53
no_grants	151	6.179	5.194	1	25		
personnel dept.	151	86.010	46.947	18	189		
ratio_prof/personnel	151	0.124	0.033	.0689655	.1764706		
no_P	124	14.976	14.997	1	76		
timescit	123	201.098	287.462	1	2022		
CPP	123	11.402	10.504	1	57.75		
CPP/JCSm	123	1.145	0.994	.1249116	10.54296		
JCS/FCS	124	1.271	0.403	.4174195	2.582656		
CPP/FCSm	123	1.317	0.798	.1153826	4.400838		

Results from the regression analyses are shown in Table 2 on the last page. Results indicate that for quantitative indicators (number of papers P and Times Cited) there are two significant variables: number of grants and total sum of grants. Qualitative variables like JCS/FCS and the crown indicator goes together with other variables. The impact factor score (JCS/FCS) seem to be negatively dependent on age and when we introduce personnel and ratio of professors these factors are significanct. CPP/JCSm is negatively related to size of department and ratio of professors. This indicates that smaller departments and fewer professors have a positive effect on journal citation scores. The so called Crown Indicator (CPP/FCSm) is also related to age, younger researchers do better quality, but also total sum of grants is of importance.

Table 2. OLS Regression analysis for 151 researchers

	(1)	(2)	(3)	(4)	(5)	(6)
LABELS	m1a	m2a	m3a	m4a	m5a	m6a
	no_p	timescit	cpp	cpp_jcs_m	jcs_fcs	cpp_fcs_m
age	0.0126	-1.906	-0.182	-0.00370	-0.0111*	-0.0194*
	(0.15)	(2.75)	(0.11)	(0.012)	(0.0043)	(0.0097)
male	1.701	38.12	0.360	-0.209	0.00688	-0.0952
	(2.41)	(43.7)	(1.76)	(0.20)	(0.068)	(0.15)
total_grant (million SEK)	0.920**	26.10**	0.342	0.0255	0.00992	0.0473*
	(0.32)	(5.84)	(0.24)	(0.027)	(0.0091)	(0.021)
broadness	3.313	7.974	1.299	0.00676	0.0143	0.0974
	(1.81)	(32.9)	(1.32)	(0.15)	(0.051)	(0.12)
sharebas	-6.512	955.3	51.54	3.438	-0.0626	3.276
	(66.6)	(1208)	(48.7)	(5.49)	(1.87)	(4.26)
sharestra	-4.318	617.0	31.07	3.950	-0.970	2.793
	(66.8)	(1213)	(48.9)	(5.52)	(1.88)	(4.28)
shareuse	-5.592	595.2	29.13	3.707	-0.901	2.852
	(66.8)	(1212)	(48.8)	(5.51)	(1.88)	(4.27)
no_grants	0.831**	11.69*	0.0436	-0.00652	0.00263	-0.0128
	(0.31)	(5.59)	(0.23)	(0.025)	(0.0087)	(0.020)
lpersonnel						
quotaprof_pers						
Constant	3.015	-559.0	-16.90	-2.407	2.549	-0.853
	(67.4)	(1222)	(49.3)	(5.56)	(1.90)	(4.31)
Observations	124	124	124	124	124	124
R-squared	0.38	0.44	0.32	0.04	0.32	0.11

	(7)	(8)	(9)	(10)	(11)	(12)
LABELS	m1b	m2b	m3b	m4b	m5b	m6b
	no_p	timescit	cpp	cpp_jcs_m	jcs_fcs	cpp_fcs_m
age	0.0575	-2.387	-0.200	0.00168	-0.0136**	-0.0192
	(0.15)	(2.82)	(0.11)	(0.013)	(0.0042)	(0.0099)
male	1.044	34.02	0.331	-0.271	0.0244	-0.120
	(2.43)	(44.5)	(1.80)	(0.20)	(0.067)	(0.16)
total_grant (million krona)	0.867**	25.71**	0.338	0.0206	0.0112	0.0451*
	(0.32)	(5.89)	(0.24)	(0.026)	(0.0089)	(0.021)
broadness	4.080*	11.90	1.310	0.0809	-0.00756	0.125
	(1.86)	(34.0)	(1.37)	(0.15)	(0.051)	(0.12)
sharebas	-9.704	1109	56.01	2.879	0.306	3.492
	(66.6)	(1218)	(49.2)	(5.44)	(1.84)	(4.30)
sharestra	-7.707	767.4	35.46	3.376	-0.600	2.997
	(66.8)	(1223)	(49.4)	(5.46)	(1.84)	(4.32)
shareuse	-8.282	728.2	32.98	3.232	-0.584	3.040
	(66.7)	(1220)	(49.3)	(5.45)	(1.84)	(4.31)
no_grants	0.785*	12.34*	0.0662	-0.0123	0.00538	-0.0128
	(0.31)	(5.64)	(0.23)	(0.025)	(0.0085)	(0.020)
lpersonnel	-0.0505	-0.144	0.00230	-0.00505*	0.00163*	-0.00159
	(0.029)	(0.52)	(0.021)	(0.0023)	(0.00079)	(0.0018)
quotaprof_pers	-59.25	652.5	24.47	-7.128*	3.251**	-0.269
	(42.6)	(780)	(31.5)	(3.48)	(1.18)	(2.75)
Constant	15.14	-755.7	-23.58	-0.855	1.780	-0.921
	(67.8)	(1241)	(50.2)	(5.54)	(1.87)	(4.39)
Observations	124	124	124	124	124	124
R-squared	0.39	0.45	0.33	0.08	0.36	0.12

Standard errors in parentheses
** p<0.01, * p<0.05

When we include personnel and ratio of professors to other personnel this effect disappears. The power of explanation is low for both of these quality variables which indicate that other factors should be included or that there are no significant patterns involved if we want to explain the quality of research.

Discussion

Our contribution indicates that gender of the researcher or project leader is of low importance for quality of research. Female researchers to some extent receive higher field normalized citations. In this investigation we test several variables towards quantitative and qualitative variables. Best fit is with Times Cited that goes together with number of grants and total grants. Productivity is a factor that we might be able to explain, but still, research quality is a tricky question.

This paper is research in progress. Other variables and new data will added in future work . In this paper we want to point out the possiblities and availability of the Swedish databases. Further work will include more of theoretical interpretation and surveys of the literature.

References

Bieber, Jeffery P.; Blackburn, Robert T. (1993) Faculty research productivity 1972-1988: Development and application of constant units of measure. *Res Higher Educ* 34(5):551-

Braxton, John M. & Alan E. Bayer. Perceptions of Research Misconduct and an Analysis of Their Correlates. *Journal of Higher Education*. Vol. 65, 1994, 351-372

Braxton, John M. and Bayer, Alan E., "Measures of Faculty Research Performance". In John W. Creswell [ed.], *Measuring Faculty Research Performance. New Directions for Institutional Research*. No. 50, San Francisco: Jossey-Bass, 1986.

Dietz & Bozeman (2005). Academic careers, patents, and roductivity … *Res Policy* 34:349-367

Fox, M (1992). Research, teaching and publication productivity. *Sociol Educ* 65: 293-305

Glänzel W. 1996. The need for standards in bibliometric research and technology. *Scientometrics* 35 (2): 167-176

Hattie J, Marsh HW (1996) The relationship between research and teaching: A meta-analysis. Review of Educ Res 66 (4): 507-542

Megdal, SB & Ransom MR (1985) Longitudinal changes in salary at a large public university: what respone to equal pay legislation. *Amer Econ Rev* 75 (2): 271-274

Moed, HF (2005). *Citation Analysis in research Evaluation*. Springer Verlag.

Nederhof & Visser (2004). Quantitative deconstruction of citation impact indicators. *J Doc* 60:658-672

Porter SR, Umbach PD 2001. Analyzing faculty workload data using multilevel modeling. *Res Higher Educ* 42 (2): 171-196

Sandström, U., Hällsten, M & Heyman, U. (2005) *Svensk forskningsfinansiering: inriktning och styrning* (reviderad version).[available at www.forskningspolitik.se/studier.asp]

Stephan and Levin (1992). *Striking the Mother Load in Science*. Oxford University Press, New York.

Stokes, D. (1997). *Pasteur's QuadrantBasic Science and Technological Innovation*. Brookings

Toutkoushian RK, Porter SR, Danielson C, et al. 2003. Using publications counts to measure an institution's research productivity. *Res Higher Educ* 44 (2): 121-148

Toutkoushian, R., Dundar, H., and Becker, W. 1998. The National Research Council graduate program ratings: What are they measuring? *Rev. Higher Educ*. 21: 427–443.

Toutkousinan et al (2003). Using publications counts to measure an institution's… *Res Higher Educ* 44:121-48

Van Raan, A.F.J. 2004. Measuring Science. [Eds.] Moed, Glänzel & Schmoch. Handbook of Quantitative Science and Technology Research. Dordrecht: Kluwer Acad Publ.

About the Identification of Technology Specific Keywords in Emerging Technologies: The Case of "Magnetoelectronic"

Edgar Schiebel and Marianne Hörlesberger

edgar.schiebel@arcs.ac.at, marianne.hoerlesberger@arcs.ac.at
Austrian Research Centers GmbH – ARC Tech Gate Vienna,
Donau-City-Straße 1, A-1220 Vienna (Austria)

Abstract
The objective of this contribution is the identification of relevant terms for emerging technologies. We take the case of *Magnetoelectronic*. The basis of our analysis is scientific literature in this field. We present three modifications of the well know indicator *term frequency inverse documents frequency* (tf-idf). At the beginning of the scientific / technology field very particular terms describe "exotic" research. These keywords occur only in a small spectrum of documents. Over the time physical laws, materials and methods establish and the corresponding terms come up in a lot of documents and technologies. We used different modifications of the tf-idf to identify these different terms.

Keywords
emerging technologies; magnetoelectronic; term frequency inverse document frequency; modified tf-idf; technology specific terms.

Introduction

This contribution deals with the detection of emerging technology terms out of scientific literature. The root of emerging technology has often been found in natural science. Therefore the analysis procedure starts with a specifically look at emerging technical fields with a close link to physics. The analysis presented here is done for a technology field called *Magnetoelectronic.*

The emphasis of this paper is the application of a modified tf-idf concept to generate a terminology framework for *Magnetoelectronic.* A good overview about the tf-idf is given by [Ferber, R. (2003)], [Robertson, S. (2004)], [van Rijsbergen, C.J. (1979)]. Karen Spärck Jones introduced the idf in 1972. The idf page tells us about the history and the problems round this weight and gives ideas for applications in different areas (see [Spärck, J. and Robertson (2006)]; and [http://www.soi.city.ac.uk/~ser/idf.html]).

Data

The data analysed here is a set of scientific literature of the field *Magnetoelectronic* from the time period 1994 and 2003. We got 281 different articles for *Magnetoelectronic*, and they were retrieved from the PASCAL database in INIST-CNRS[1].

The method introduced here requires reference data. The reference data are 39 further different scientific fields retrieved from PASCAL database. These fields deal with different technologies, biotechnologies and medical science. We instance only some of them to convey an idea: *Robot Control, Electronic Tubes, Semiconductors, Transistors, Optoelectronics, Molecular Electronics, Metallurgy, Surface Treatment of Metals, Road Traffic, Molecular Biophysics, Biophysics of Tissues, Organs and Organism, Imaging of Cardiovascular Systems, Nervous System and Sense Organs.* There are 74,981 different documents in all reference technologies.

1 The PASCAL database in INIST-CNRS (Nancy, France) is organised by means of classification codes, and using adequate classification codes, allocated to physics and technology, we recorded the investigated data.

Methodology

Modification 1
Using Ferber's definition of local weight we need some local text corpus representing one technology and a global corpus of reference documents.

We presume that a technology field is defined by a PASCAL-code; the terms in the descriptor fields have been classified according to the code. In this modification the local component of the weight (tf_{ij}) corresponds to the term frequency of the keywords that occur in all documents with the same PASCAL-code. The global component of the weight (idf_j) is measured on the basis of the keywords of all reference technology fields, what means that n is the number of all disposable documents and d_i is the number of those disposable documents where the term i occurs at least once as a keyword in a descriptor field.

The following formula accomplish more precision: *Modification 1*

$$w_{ij} = tf_{ij} \cdot idf_j \text{ where } tf_{ij} = \phi_{ij} \text{ and } idf_i = \ln(\frac{n}{\tau_i}) \quad (1)$$

a) i is a term, that occurs in a descriptor field of a document.
b) j is one technology field.
c) w_{ij} is the weight of a term i in the technology field j.
d) ϕ_{ij} is the frequency of the term i in the technology field j.
e) n is the number of documents of all reference documents.
f) τ_i is the number of all documents of the reference documents where i occurs at least once.

Modification 2
A dominating factor tf_{ij} can be corrected or attenuated. Ferber presents a rational function of the form

$$tf_{ij} = \frac{\phi_{ij}}{(1+\phi_{ij})} \quad (2)$$

Hence we get the formula: *Modification 2*

$$w_{ij} = tf_{ij} \cdot idf_j \text{ with } tf_{ij} = \frac{\phi_{ij}}{(1+\phi_{ij})}$$

Modification 3
Modification 1 stresses terms with high tf_{ij}, modification 2 privileges the value of idf_j. Therefore we introduce a third modification that upvalues terms with medium frequency. We propose a tf_{ij} with the shape of a S-curve that attenuates very small and very high frequencies. A simplified growth function is:

$$tf_{ij} = \frac{1}{(1+e^{(a-b_j \cdot \phi_{ij})})}. \quad (3)$$

This approach effects that small values of term frequency are attenuated as well as very high values.

$$a = 5 \quad (4)$$

$$b_j = \frac{5}{\frac{1}{m}\sum_{i=1,\ldots m}\phi_{ij}} \quad (5)$$

a and b are chosen as follows: 5 is the value for the inflexion point is in the middle of the codomain of the S-curve. The parameter b_j is defined by the mean of the frequency of all keywords in one technology and is to be calculated separately for each technology.[2]

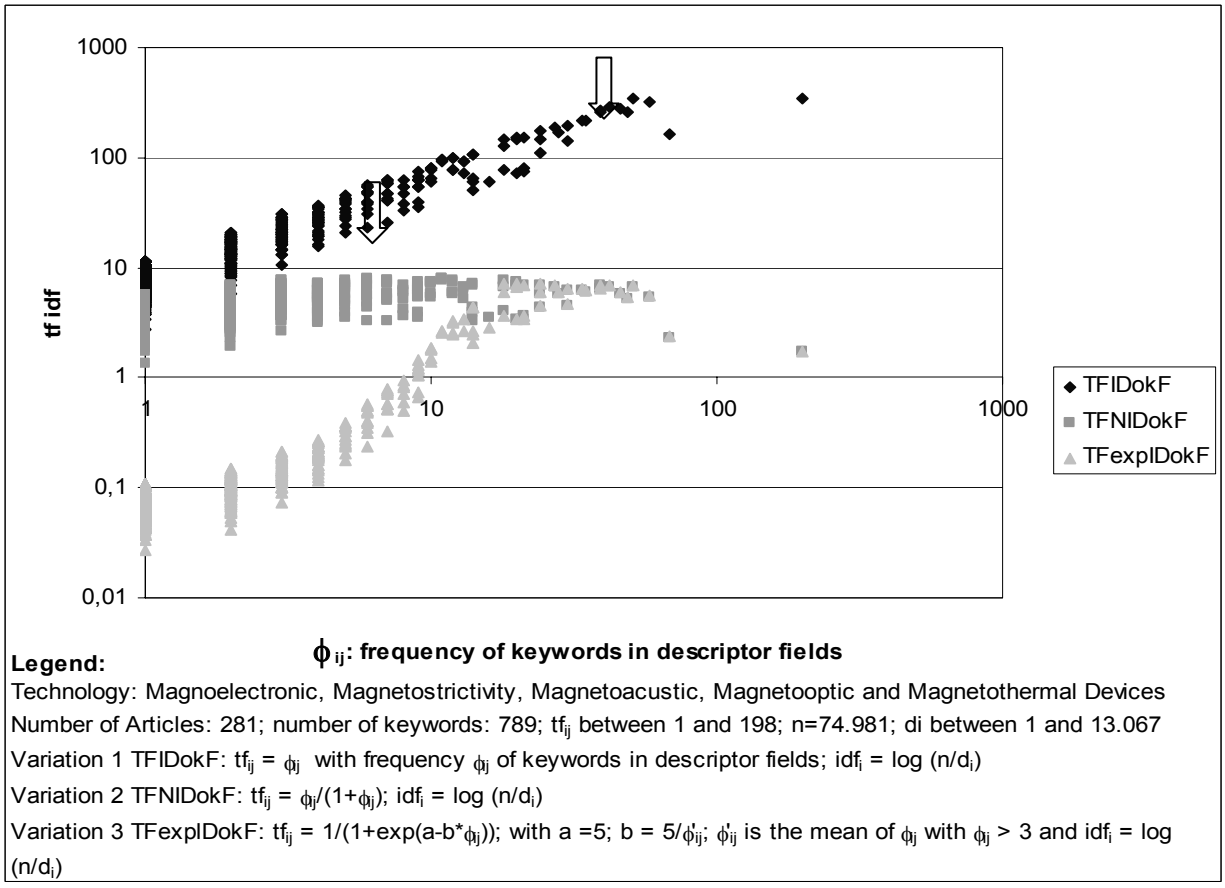
We get the formula: *Modification 3*

$$w_{ij} = tf_{ij} \cdot idf_j \text{ with } tf_{ij} = \frac{1}{(1+e^{(a-b_j \cdot \phi_{ij})})} \quad (6)$$

Results

The three modifications of the tf-idf were applied to the technology field *Magnetoelectronic.* The figure 1 shows the three different modifications.

Figure 1. Three Variations of tf-idf Weight for Keywords in Magnetoelectronic



Legend:

Technology: Magnoelectronic, Magnetostrictivity, Magnetoacustic, Magnetooptic and Magnetothermal Devices

Number of Articles: 281; number of keywords: 789; tf_{ij} between 1 and 198; n=74.981; di between 1 and 13.067

Variation 1 TFIDokF: $tf_{ij} = \phi_{ij}$ with frequency ϕ_j of keywords in descriptor fields; $idf_i = \log (n/d_i)$

Variation 2 TFNIDokF: $tf_{ij} = \phi_{ij}/(1+\phi_{ij})$; $idf_i = \log (n/d_i)$

Variation 3 TFexpIDokF: $tf_{ij} = 1/(1+\exp(a-b^*\phi_{ij}))$; with a =5; b = $5/\phi'_{ij}$; ϕ'_{ij} is the mean of ϕ_{ij} with $\phi_{ij} > 3$ and $idf_i = \log (n/d_i)$

Source: Calculated by Edgar Schiebel based on the described data above.

The x-axis represents the term frequency of the keywords and the y-axis shows the three modifications of tf-idf. The modification 1 is called TFIDokF. The weight grows with increasing frequency. Terms with a high frequencies show only a small differentiation caused by idf_j, because of the dominating tf_{ij}. The modification 2 is called TFNIDikF and shows that terms with a high tf_{ij} value (right arrow) are strongly brought down and terms with a very low term frequency (tf_{ij}) get a higher tf-idf weight. The effect of the S-curve (modification 3, named TFexpIDokF) is highlighted with the left arrow: terms

[2] Real frequency distributions of terms in documents tend to have a huge amount of terms with frequency 1,2,3,…what results in very low values for the mean and also for the median. Terms with low frequencies can be omitted, to get higher mean or medium values.

with a low frequency have a small idf_j weight and terms with high frequencies have a tf-idf weight that is dominated by the global weight idf_j.

Table 1. Terms with the different weights of the represented three modifications

	Modification 1		*Modification 2*		*Modification 3*
Terms	**TFIDokF**	**Term**	**TFNIDokF**	**Term**	**TFexpIDokF**
Electron spin polarization	347,07	Magneto-optical isolators	8,08	Interface magnetism	7,19
Experimental study	345,77	Magnetostatic wave devices	8,01	Magnetostrictive devices	7,08
Ferromagnetic materials	320,37	tunnelling magnetoresistance	7,87	Magneto-optical devices	7,06
Magnetic multilayers	290,08	magnetic tunnelling	7,81	Spin valve	7,00
Iron alloys	276,42	Magneto-optical devices	7,75	Magnetoresistive devices	6,96
Spin valve	273,18	Iridium alloys	7,74	Magnetic multilayers	6,91
Tunnel effect	261,29	Magnetic epitaxial layers	7,72	Magnetic sensors	6,88
Giant magnetoresistance	255,50	spin valves	7,69	Electron spin polarization	6,81
Magnetic thin films	218,85	Magneto-optical modulation	7,65	Exchange interactions (electron)	6,79
Magnetoresistance	215,21	Amorphous magnetic materials	7,62	Giant magnetoresistance	6,55
Cobalt alloys	194,12	Magnetoelectric device	7,59	Cobalt alloys	6,46
Magnetic sensors	186,84	Induced anisotropy (magnetic)	7,59	Magnetic thin films	6,43
Magnetorestrictive devices	172,77	Magnetic domain walls	7,59	Magnetoresistance	6,15
Cobalt	169,98	Magneto-optical sensors	7,55	Magnetostriction	6,05
Theoretical study	160,95	Magnetic thin film devices	7,55	Cobalt	6,05
Interface magnetism	154,55	Permalloy	7,51	Iron alloys	6,01
Magnetoresistive devices	153,55	magnetoelectronics	7,39	Nickel alloys	5,94
Magneto-optical devices	147,23	Interface magnetism	7,36	Ferromagnetic materials	5,52
Exchange interactions (electron)	145,85	spin polarised transport	7,35	Tunnel effect	5,33
Nickel alloys	144,88	Semimagnetic semiconductors	7,29	Magnetization	4,69
Magnetization	141,02	MIM structures	7,22	Aluminium compounds	4,47
Magnetostriction	126,35	Kerr magneto-optical effect	7,16	Magnetic switching	4,42
Aluminium compounds	109,10	Antiferromagnetic materials	7,14	Garnets	4,35
Magnetic switching	106,97	Magnetic switching	7,13	III-V semiconductors	3,70
Garnets	105,15	LEED	7,07	Gallium arsenides	3,67
spin valves	100,01	Dysprosium alloys	7,07	Ferromagnetism	3,42
Magnetic thin film devices	98,16	Spin waves	7,07	Sputtered coatings	3,39
Magnetostatic wave devices	96,13	Magnetostatic surface waves	7,06	Instrumentation	3,37
tunnelling magnetoresistance	94,44	Discontinuous metallic thin films	7,06	Measuring methods	3,35
Ferromagnetism	91,85	Magnetostatic waves	7,06	spin valves	3,26

Source: Calculated by Edgar Schiebel based on the data described above.

After looking at the statistics the ranking is represented in table 1. Considering the left column and its value “experimental study” is the keyword with the highest frequency in all documents and also in the field *Magnetoelectronic*. If “experimental study” is weighted by the basic tf-idf formula (1) (modification 1), it is ranked on the second position, although it has the smallest idf_j of all keywords. The reason is the extremely high frequency of 198 in comparison to the other keywords. A deeper view on the high ranked terms shows some categories:

1. *often used not meaningful terms:* experimental study, theoretical study.
2. *materials used with high frequency not only in magnet electronic but also in other technology fields:* ferromagnetic materials, iron alloys, cobalt alloys, cobalt, nickel alloys, aluminium compounds, garnets
3. *physical effects also relevant in other technologies and not only in the local technology:* electron spin polarisation, magnetic multilayers, giant magnetoresistance, magnetoresistance, tunnel effect, magnetization
4. *applications:* magnetic sensors, magneto optical device

As it could be expected modification 2 (TFNIDdok) pushes those terms to high ranks that occur in few reference documents. These terms occur mostly in articles of the theme *Magnetoelectronic*. These very specific terms show the following categories:

1. *terms that describe specific research fields:* magneto-optical isolators, magnetostatic wave devices.
2. *physical effects:* tunnelling magnetoresistance, magnetic tunnelling, spin polarised transport
3. *new materials research:* magnetic epitaxial layers, amorphous magnetic materials, perm alloy, antiferromagnetic materials.

The modification 3 (TfexpIDokF) causes a high ranking of term with medium ϕ_{ij} and low $d_{j.}$ These terms are technology specific and well established. They are also used in other reference technologies.

1. *applications:* magneto-optical devices, magnetoresistive devices, magnetic sensors, magnetostrictive devices.
2. *materials:* cobalt alloys, cobalt, iron alloys, nickel alloys, ferromagnetic materials.
3. *physical effects:* spin valve, electron spin polarization, exchange interactions (electron), giant magnetoresistance, magnetoresistance, magnetostriction, tunnel effect, interface magnetism.
4. *methods:* magnetic multilayers, magnetic thin films.

Conclusion

Each of the presented variations tells us different aspects about the technology field. Modification 1 provides only a rough overview about terms that are specific in the considered field and terms with cross-border function respectively. Modification 2 extracts really very specific terms for the considered technology field. Such terms indicate upcoming technology terms. Modification 3 helps to emphasize those keywords that have a medium frequency but are still established in considered the technology field or play also a role in other technologies. These three variations offer a basis for the construction of a term based description of a technology. However some expert knowledge is needed to classify the terms in dimensions of the technology.

The presented plan helps to identify specific technology terms in one technology field but is not appropriate to cross topics. How can the specific terms of such topics be detected? A further hypothesis is that at the beginning of the life cycle particular terms describe “exotic” research. Later in the life cycle the used terminology establishes and diffuses also into other technologies. The development of further indicators for detecting these characteristics will be published in a next paper.

References

Ferber, R. (2003*). Information Retrieval, Suchmodelle und Data-Mining-Verfahren für Textsammlungen und das Web,* Heidelberg: dpunkt.

Henkel, J. (2000). Magnetoelektronik aus Sicht der Innovationsforschung. Retrieved November 30, 2006 from: http://www.inno-tec.de/forschung/henkel/ME_IF_JH.pdf.

Iliopoulos, I.; Enright, A. J.; Ouzounis, C. A. (2001): Textquest: Document clustering of Medline abstracts for concept discovery in molecular biology. *Pac Symp Biocomput* 6 (pp. 384-395).

Robertson, S. (2004). Understanding Inverse Document Frequency: On theoretical arguments for IDF, *Journal of Documentation* 60 no. 5, 503-520.

Schmoch, U. (1997). Indicators and the Relation between Science and Technology. *Scientometrics*, Vol. 38, No. 1 (pp. 103-116).

Spärck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval: development and comparative experiments. *Information Processing and Management*. Vol. 36. (pp. 779-808, part 1 and pp. 809-840, part 2).

Spärck Jones, K. and Robertson, S. (2006). Inverse Document Frequency - The IDF page. Retrieved November 22, 2006 from: http://www.soi.city.ac.uk/~ser/idf.html.

van Rijsbergen, C.J. (1979), *Information Retrieval*, London: Butterworths.

Comparative Study between First and All-Author Co-Citation Analysis Based on Citation Indexes Generated from XML Data

Jesper Wiborg Schneider, Birger Larsen and Peter Ingwersen

jws@db.dk, blar@db.dk, pi@db.dk
Department of Information Studies, Royal School of Library and information Science
Birketinget 6, DK-2300 Copenhagen S (Denmark)

Abstract
The study presents a comparative analysis between first and all-author co-citation analyses, as well as comparison between two matrix generation approaches. We thus continue the latest research in author co-citation analysis (ACA), where the results of the traditional first-author analyses based on ISI citation indexes are challenged by incorporating all-authors from the cited references. Identifying all cited authors from references in source papers is an extremely cumbersome process if the Thomson ISI citation indexes are used as a basis. Due to the difficulty in obtaining all-author co-citation data few such studies exist. In order to study all-authors co-citation we use a citation index generated from documents in XML code. This allows us to carry out a comparative study between first and all-author co-citation analyses based on the hitherto largest set of references and the broadest domain of research.

Keywords
co-Citation; first-authors analyses; XML data

Introduction

Author co-citation analysis (ACA), introduced by White and Griffith (1981), is a technique for mapping the 'intellectual structure' of a research field, where the latter is defined as a coherent literature set. The intellectual structure is mapped from the oeuvres of the most cited and co-cited first authors in a particular literature set. Since its introduction, ACA has become a popular and much used technique. However, recently a debate concerning methodical procedures in ACA has emerged. Especially, the approach to ACA developed at Drexel University (e.g., White & Griffith, 1981; McCain, 1990) has been the focus of the current debate. Essentially, four methodical issues have been debated: 1) scalability (e.g., Chen, 1999), 2) units of analysis and their definition (e.g., Persson, 2001; Zhao, 2006; Rousseau & Zuccala, 2004), 3) the choice of proximity measures (e.g., Ahlgren, Jarneving, & Rousseau, 2003; Schneider & Borlund, 2007a; 2007b), and most recently 4) generation and transformation of matrices (Leydesdorff & Vaughan, 2006; Schneider & Borlund, 2007a). The present paper addresses the second and fourth issues in a comparative study of first and all-author co-citation analysis based on different matrix generation approaches in structured XML documents that allow for the construction of ad-hoc citation indexes.The paper is structured as follows. The following section discusses briefly previous research on all-author co-citation analyses and matrix generation. The proceeding section describes the research method of the study, i.e., data collection and data analysis. The next section presents and discusses the results, and the contribution ends with a conclusion.

Previous Work on All-author Co-citations and matrix generation

In several respects, the methodical approach to ACA developed at Drexel University has been shaped by specific technical features that have seemingly brought some constraints to the ACA methodology. Most important is the dependence upon the standardized cited reference strings in Thompson ISI's citation indexes, and the use of the SPSS statistical package as the tool for multivariate analyses. The most obvious example is that the cited reference strings only allows for first authors as units of analysis in ACA. As a result, ACA methodology only takes into account first authors in the definition of author co-citation counts. Two authors are considered to be co-cited when at least one document from each author's oeuvre occurs in the same reference list of a citing document, where an author's oeuvre is defined as all the works with the author as the first author (McCain, 1990). This definition has rarely been challenged. Persson (2001) is the first empirical study that compares the potential difference in intellectual structure between mappings done by first-author and all-author co-citation

analyses. The study is based on 7001 source documents from library and information science journals in the CD-ROM version of Social Science Citation Index 1986-1996. The study investigates how these source documents have been co-cited with each other within the dataset by use of multidimensional scaling (MDS). The co-citations for source documents amount to some 7% of the total number of references in the dataset; the remaining 93% go to non-source documents not indexed by the Thompson ISI citation indexes. The study demonstrates that first-author ACA leaves out several influential researchers compared to all-author ACA, although the subfield structure tends to be just about the same for both methods. The study is somewhat limited due to the dependence on a limited set of source documents, the sparse details provided concerning the definition and calculation of co-citations, and finally the informal evaluation procedures. Nevertheless, the results are indicative as they are somewhat confirmed in a smaller study done by Zhao (2006).

All-author vs. First-author Co-citation Analyses

Zhao (2006) is the hitherto must detailed theoretical and empirical investigation of all-author co-citation analysis, including a definition of co-citation counts reminiscent of the definitions given earlier by Rousseau and Zuccala (2004). The study defines three different counting methods: first-author co-citation; *inclusive* all-author co-citation; and *exclusive* all-author co-citation. Likewise, as a consequence of all-author co-citation analysis, the study redefines "…an author's oeuvre as all works with this author as one of the authors of each of the works." (Zhao, 2006, p. 1580). The distinction between inclusive and exclusive all-author co-citations refers to the immediate implication of the above definition of all-author co-citation counting of author's oeuvres, as two authors may also be considered as being co-cited when a paper that the two authors co-authored is cited. Thus, co-authorships when cited can also be counted into co-citations. This means that inclusive all-author co-citation analysis counts cited co-authorships, whereas exclusive all-author co-citation analysis does not. Typically author co-citations and co-authorships are treated as different units of analysis, where the former is used to map intellectual structures and the latter to investigate research collaboration. Rousseau and Zuccala (2004), in their definition, suggest that such an approach supports the view that authors, regardless of their overall authorship ranking, can contribute substantially to the development of a research area, and that it presents a more accurate portrayal of an individual author's contribution to a research area where high rates of co-authorship are prevalent.

Besides the novel definition of all-author co-citation counting, Zhao (2006) adheres to a traditional Drexel-approach to ACA (see below). The dataset was rather small: it consisted of 312 publications in PDF on the subject of XML identified using CiteSeer[1]. The 312 publications contained 4578 citations, which was used a basis for the co-citation analysis. The results of the study indicate that all-author co-citation counting creates more coherent groups of authors, which supposedly should be considerably clearer to identify and interpret. Nevertheless, due to the straightforward application of citation thresholds for including cited authors in the study, the results also show that all-author co-citation count can lead to identification of fewer specialties in a research field compared to first-author co-citation counting – that is when the same number of top-ranked authors is selected and analyzed (Zhao, 2006).

Zhao (2006) undoubtedly contributes considerably to our understanding of all-author co-citation analysis. However, for the time being, the results of the empirical study must be treated carefully until we have more substantial evidence that may or may not support its findings. The motivation for the present paper is therefore to continue the work of Zhao (2006) by further investigating inclusive all-author co-citation analysis in order to bring about deeper empirical understanding and evidence concerning this novel counting approach. The present study is the first in a series that addresses the research possibilities inherent in a citation index based on source documents formatted in XML. One such possibility is all-author co-citation analysis, and the present study is based on the hitherto largest set of citing documents applied in an all-author co-citation analysis.

[1] http://citeseer.ist.psu.edu/

Co-citation Matrix Generation
Most recently the role played by matrices in co-citation analyses has received attention. Leydesdorff and Vaughan (2006) demonstrate the fundamental difference between asymmetric data matrices ($n \times m$) and symmetric proximity matrices ($n \times n$), arguing that symmetric matrices of co-occurrence counts are *per se* proximity matrices and should be treated as such.

In the Drexel-approach to ACA, first author co-citation counts are obtained by online retrieval. Subsequently the co-citation counts are entered into a symmetric proximity matrix. However, the desire to apply factor analysis to ACA as a more detailed exploratory tool in order to identify latent structures and thus help interpret the mapping results, necessitates a symmetric proximity matrix of covariance or correlation coefficients. In traditional multivariate analyses such proximity matrices are derived from an asymmetric data matrix of variables by cases. However, such a matrix is not available in the Drexel-approach due to the paired online counting. As a result an unorthodox procedure is devised, where the proximity matrix of co-citation counts are transformed into an additional proximity matrix of derived correlation coefficients of first author co-citation profiles. Note that a linear transformation of a symmetric proximity matrix is not straightforward. A theoretical problem arises, as all relations in a symmetric matrix occur twice which evidently leads to a magnification. Further, the transformation also causes a fundamental problem in relation to the interpretation and treatment of diagonal values. In SPSS several possibilities for treating diagonal values are available, and the most commonly used in ACA is to treat the diagonals as missing data (e.g., McCain, 1990). Hence, rows are treated as cases and columns as variables, yet this procedure is only allowable when computing correlation matrices. The same practice is not applicable in SPSS if one wishes to transform the proximity matrix of co-citation counts into a similarity or distance measure. Missing data beyond doubt causes some loss of information in the matrix and therefore likely influence the ensuing ordination or clustering results. White (2003, p. 1250), nevertheless, asserts that the treatment of diagonal values is a minor problem. This may be true, depending on the data set at hand, but the generic problem arises due to the unorthodox matrix generation (transformation) approach and could be avoided if a more traditional procedure was applied. The question is whether the different approaches to matrix generation make a significant difference in the interpretation and mapping of ACA?

Consequently, the present study explores two of the recently debated methodical issues of ACA, first-author versus all-author co-citation analysis, and the influence of different matrix generation approaches upon mapping results. The major research questions explored in the study are:

- To what extend does a different data set support the previous findings of first-author versus all-author co-citation analysis?
- To what extend does different matrix generation techniques influence the interpretation and mapping of author co-citation data?

In order to answer these questions, we perform two first-author and two inclusive all-author co-citation analyses, one pair commencing from a data matrix and one pair commencing from a proximity matrix of co-citation counts.

Method

Data set
The citation index used in the current study was extracted from a corpus of full text XML documents. The corpus consists of 16,819 articles from the journals of the IEEE Computer Society from the years 1995 to 2004, and is part of the Information Retrieval test collection created by the Initiative for the Evaluation of XML Retrieval (INEX)[2]. While the Document Type Definitions (DTDs) used by publishers mainly aid in controlling the printing and publication process many of the so-called

[2] See Malik et al. (2005) and http://inex.is.informatik.uni-duisburg.de/2006/ for more information on INEX.

elements identified by XML tags have many other potential uses. In this paper we extract the following data from the XML version of the IEEE CS journals to form a citation index:

- *All* cited authors including their order in the cited document
- Cited titles of the cited articles or books
- Cited journal name of cited articles
- Cited year
- Cited volume and Cited issue of cited articles
- Cited page numbers (begin and end numbers)

For our purpose this data set is ideal: Compared to the Thomson ISI citation indexes we have direct access to all cited authors (100% coverage), and by working directly on the source files used in the production of the original citing articles we also have a range of high quality input data to generate the citation index. By relying on XML data we avoid some of the errors that other approaches have to deal with: So-called autonomous citation indexing (Giles, Bollacker & Lawrence, 1998) based on extraction of reference data from PDF files like those of Zhao (2006), e.g., CiteSeer.IST, Google Scholar[3] and Rexa[4], have to deal with many problems of segmentation and disambiguation of data from the raw PDF files. However, using the XML data we still have errors arising from the citing authors, that is, errors originating in sloppy reference practise. Some studies have reported quite large shares of references (some more than 50%) with errors originating from citing authors, see e.g. Lok, Chan & Martinson (2001) for examples from the medical domain. Our XML data makes it possible to investigate novel ways of automatically detecting such errors. This is, however, a subject for future research and not treated further here. For the present study we have chosen an approach proposed by Glänzel (1996) where each reference is represented by a cluster-key consisting of the two last digits of the cited year, the first four letters of the last name of the author, the volume number, the start page number and the first letter of the name of the cited journal. A key for Glänzel (1996) would then be: 96-GLAN-35-2-167-S. By reducing the references to this short form this key catches many of the variants arising from sloppy referencing. Errors in any of the key's elements are of course not corrected for, but the effect of these is hard to study on large datasets such as ours.

The 16,819 articles from the IEEE CS journals contained a total of 212,657 references (12.6 on average). After application of the cluster-key this was reduced to 132,311 unique references. For each of the cluster-keys that occurred more than once, the next step was to select one of the references to represent the whole cluster in order to be able to extract the cited authors needed in our analysis. For this purpose the cited titles were analysed: if one of the cited titles in a cluster had more occurrences than the rest this was chosen; if there two or more shared the top position the longest (and most specific) reference was chosen. Finally the data of the representative reference was used to represent all the references in the cluster. For the present study we have extracted two datasets: one with the ID of each citing article plus the cited first authors, and a parallel dataset with all cited authors extracted. The first-author dataset consists of 198,865 pairs of document IDs and cited first authors (13,792 references had no cited authors). The all-author dataset consists of 414,729 pairs.

Data analysis

We perform two first-author and two inclusive all-author co-citation analyses based on the two dataset and two different approaches to matrix generation. The commonly accepted Drexel-approach is applied, as well as an approach based on the conventional procedures for multivariate statistical analysis outlined in numerous textbooks. The one pair of first-author and inclusive all-author co-citation analyses commences from an $n \times n$ proximity matrix, which corresponds to the Drexel-approach. The other pair of first-author and inclusive all-author co-citation analyses commences from an $n \times m$ data matrix, which corresponds to conventional multivariate data analysis.

[3] http://scholar.google.com/
[4] http://rexa.info/

The basic components in the Drexel-approach are given above and are outlined in McCain (1990). An approach to ACA based on conventional procedures for multivariate data analysis could well include the same elements as the Drexel-approach; however, there is one tremendously important difference: Multivariate data analyses most often commence from an $n \times m$ data matrix (e.g., Lattin, Carroll, & Green, 2003). Factor analysis seeks a solution that focuses on the decomposition of the covariance or correlation matrix. This implies that the data matrix *must* be transformed into such a specific type of proximity matrix. Some multivariate techniques, such as MDS employ a proximity matrix as their input. Most commonly, such an input proximity matrix is *derived* from the traditional data matrix, by some suitable proximity measure (e.g., Lattin, Carroll, & Green, 2003). The transformation of the data matrix by use of a proximity measure results in an $n \times n$ matrix of inter-object proximities. Alternatively, as mentioned above, a proximity matrix can also be generated *directly*, for example from co-citation counts. Such data are perceived to be proximities, which can be added directly to a proximity matrix. Accordingly, we employ a conventional approach where we commence from a multivariate data matrix. This data matrix is the basis for factor analysis and a proximity matrix of correlations is derived from it. Subsequently, the derived correlation matrix is used as input for MDS.

The set of cited authors chosen for analysis is determined by citation frequency. Two sets of cited authors are needed. A set comprising first-authors of cited references only, and a second set comprising all-authors of the cited references. The latter is inclusive co-authorships, as described in above. Straight counting is applied in both instances. Note that we remove duplicate authors from individual references lists. For example, if cited author *X* appears 5 times in a specific reference list, he or she is only counted once, and the multivariate data matrix is thus binary. The main motivation for invoking duplicate removal is to reduce the likely effect of self-citations, especially in the case of all-author co-citation counting where multiple authorships can lead to an excessive number of self-citations. Contrary to Zhao (2006), who limited the cited authors to the first five, we include *all* authors of a cited reference in the counting.An arbitrary citation threshold of 75 cited authors are chosen for both cases of author co-citation. The overlap in authors between the two sets is 41.

The following matrices are generated: one first-author data matrix (75×2002), one all-author data matrix (75×3161), one proximity matrix of directly obtained co-citation counts between first-authors (75×75), and finally one proximity matrix of directly obtained co-citation counts between all-authors (75×75). The latter two proximity matrices are simply a multiplication of the respective data matrices with their transpose. Following the Drexel-approach, the two proximity matrices are transformed into matrices of correlation coefficients. Diagonal values are treated as missing data with the possible implications described in the introduction.

Non-metric MDS is applied to all four correlation matrices, i.e. the two derived from the data matrices and the two derived from the initial proximity matrices. We employ the PROXSCAL scaling routine in SPSS. Factors are extracted by principal components analysis with an oblique rotation (Oblimin in *XLSTAT*) that does not have the constraint of orthogonality as the factor rotation usually used in ACA. For want of a better solution, we apply the Kaiser-Guttman rule for determination of the number of factors extracted. The Kaiser-Guttman rule suggests that only those factors with associated eigenvalues that are strictly greater than 1 should be kept. MDS and factor analysis are thus applied as exploratory tools when investigating the grouping of authors and whether there is a difference between first and all-author co-citation analysis. The investigation into the likely difference of results when employing different matrix generation techniques is statistically validated by use of Procrustes analysis.

Procrustes analysis is a statistical technique for comparing two sets of data configurations for the same set of objects (Schönemann & Carrol; 1970; Gower, 1971). The technique is thoroughly introduced and demonstrated in Schneider and Borlund (2007b). Several approaches to Procrustes analysis exist; we employ the least squares optimization criterion and an orthogonal transformation matrix. The objective is to minimize the sum of the squared deviations, m^2, between points through translating, rotating and dilating one configuration to match the other target configuration. A typical Procrustes analysis simply provides a descriptive summary and graphical comparison of two configurations of

points. By employing a permutation approach to one of the data sets, we can determine whether the original m^2 is smaller than expected due to chance (see Schneider & Borlund, 2007a for details).

Results and discussion

As MDS configurations are the basis for our analyses we commence by presenting the four MDS solutions. Note that we in the present analysis do not map author names; the present evaluation is solely based on multivariate techniques, Procrustes analysis and Mantel test in combination with manual inspection. Figures 1 to 4 below illustrates the four MDS configurations. Figures 1 and 3 are based on the Drexel-approach, and Figures 2 and 4 on the conventional approach.

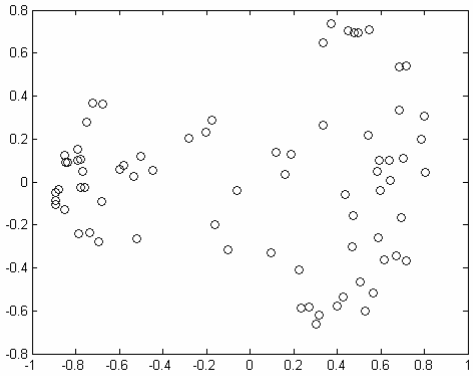


Figure 1. MDS configuration of first-author co-citation analysis – Drexel-approach.

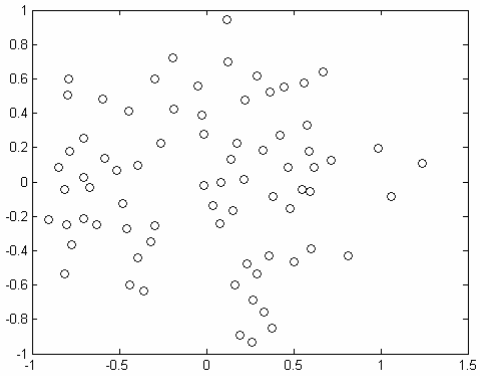


Figure 2. MDS configuration of first-author co-citation analysis – conventional approach.

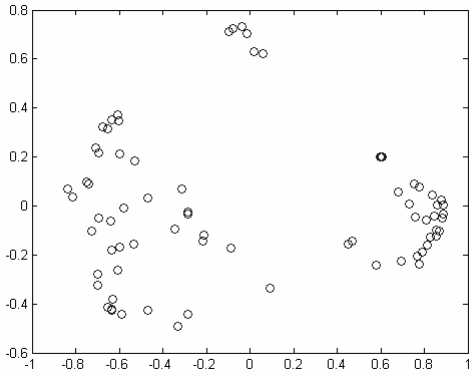


Figure 3. MDS configuration of all-author co-citation analysis – Drexel-approach.

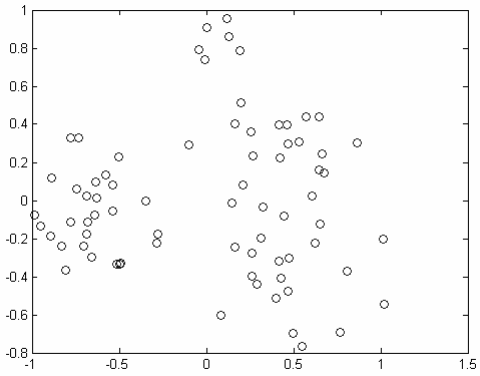


Figure 4. MDS configuration of all-author co-citation analysis – conventional approach.

A manual inspection of these configurations indicates that Figures 1 and 3 contain some noticeable structures. Likewise some structure is visible in Figure 4, while it is very difficult to identify any structure in Figure 2. Table 1 below gives the Stress-1 values for the 4 configurations.

Table 1. Stress-1 values for the four MDS-configurations illustrated in Figures 1-4.

Stress-1 values				
Configuration	**First-author/ conventional approach**	**First-author/ Drexel approach**	**All-author/ conventional approach**	**All-author/ Drexel approach**
2 dimensions	0.251	0.174	0.219	0.132
3 dimensions	0.186	0.102	0.181	0.096

None of these values are powerful. It can be inferred that the visibility of structure or rather lack of it in the MDS configurations above clearly is related to the meagre Stress-1 values. The most acceptable results for 2-dimensions are the two configurations based on the Drexel-approach. The consensus cut of level of 0.2 for Stress-1 values in non-metric MDS is only achieved for all configurations if we include a third dimension. Hence we have four configurations based on two different counting methods and two different matrix generation approaches. None of these configurations have remarkably low Stress-1 values; however, it is noticeable that two best and indeed acceptable configurations are based on the same matrix generation approach, i.e., the Drexel-approach. It is also noteworthy that the all-author configurations have lower Stress-1 values compared to first-author configurations in all cases. In particular, there is a noticeable reduction in Stress-1 in the 2D solution – indicating that the inclusion of all cited authors better fit the underlying data. This is of some importance as most MDS-based maps are presented in 2D. From the manual inspection, it also seems that the all-author ACA maps result in stronger concentrations of the co-cited authors into clusters, regardless of the approach to matrix generation.

Comparison of matrix generation techniques

One way of investigating whether the two matrix generation approaches provide different ordinations is to compare their solutions, based on the same proximity measure, for the same set of objects. Procrustes analysis compares and evaluates the resemblance between ordination solutions. Two Procrustes analyses are carried out, one for the two first-author configurations, and one for the two all-author configurations. Note Procrustes analysis is not possible between first and all-author configurations unless they contain identical objects. The target configuration in both cases is the MDS solution with the lowest Stress-1 value, i.e., the Drexel-approach in both cases. Consequently, the configurations based on the conventional approach are subjected to translation, rotation and dilation in order to match them with the target configurations. The remaining residuals for corresponding points in the two configurations after the least squares fitting provide the m^2 statistic. The m^2 statistic is subject to permutation procedure in order to test the significance of the comparison. Figures 5 and 6 below give the Procrustean superimposition plots for the two matching pairs.

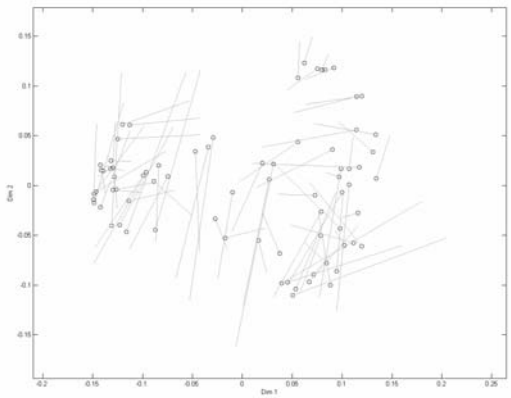


Figure 5. Procrustean superimposition plot of first-author co-citation analysis for the same set of objects based on a Drexel and a conventional matrix generation approach.

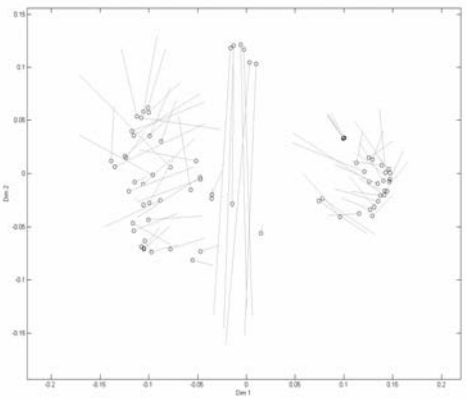


Figure 6. Procrustean superimposition plot of all-author co-citation analysis for the same set of objects based on a Drexel and a conventional approach.

At 1000 permutations, the Procrustes statistic, m^2, for the first-author comparison is 0.47 (p=0.0001) and m^2 for the all-author comparison is 0.50 (p=0.0001). The Procrustes statistic is essentially a dissimilarity measure of fit, thus the two results clearly indicate a significant difference in the ordination produced by the two matrix generation approaches. The circles in the Procrustean superimposition plots, Figures 5 and 6 are the authors plotted from the Drexel-approach configurations. The lines in the plots indicate residuals after matching the configurations based on the conventional

approach. A small vector residual indicates a close resemblance between the corresponding points and vice versa. It is evident from Figures 5 and 6 that some objects are represented very differently depending on the applied matrix generation approach. However, from the circles in the two plots, we can identify visible structures, and we can also observe that numerous residuals, of different length, move within these structures. Few, if any, shift between major groupings. So it seems that the two matrix generation approaches produces ordinations of some resemblance, however, in the present cases the Drexel-approach undoubtedly provides the best and most comprehensible configurations.

First-author versus all-author co-citation analysis

Factor analysis is traditionally applied in ACA to elaborate on the mapping results, i.e., to assist in finding latent structures among objects. Traditionally major factors in ACA are interpreted as 'research specialties' (cf. White & McCain, 1998). Hence in order to investigate whether the present data set support the previous findings of first-author versus all-author co-analysis we compare the four ACA on the basis of exploratory factor analyses. The most evident indictor for the prevailing latent structure in the data set is the major factors. Nevertheless, the determination and optimal extraction of major factors is a long standing dispute within the statistical community. A common practice is to use the Kaiser-Guttman rule.

The Kaiser-Guttman rule for extraction of factors in the present analysis results in a 29-factor model for the conventional approach to first-author co-citation analysis, explaining 59% of the variance; a 25-factor model for the conventional approach to all-author co-citation analysis, explaining 63% of the variance; a 16-factor model for the Drexel approach to first-author co-citation analysis, explaining 80% of the variance; and finally a 15-factor model for the Drexel-approach to all-author co-citation analysis, explaining 86% of the variance. Extraction of a smaller set of factors explaining a majority of the variance in the data set, no doubt implies that the latent structure in the data set should be more explicable and visible. In Zhao (2006) the inclusive all-author co-citation analysis provided a 5-factor model, explaining 97% of the variance, and the first-author co-citation analysis provided an 11-factor model, explaining 96% of the variance – indeed impressive results, which are interpreted so that first-author co-citation analysis provides more specialties compared to all-author co-citation analysis, the latter however provides more coherent groups.

Consider the two factor solutions based on the conventional approach first. As indicated by the Stress-1 values, the configurations based on the conventional approach on the whole have a poorer fit between its observations and derived proximities resulting in inferior mappings, compared to the Drexel-approach. Obviously, the two data matrices, upon which the conventional approach is based, are extremely sparse. Whereas the two proximity matrices, upon which the Drexel-approach is based, are relatively dense; the latter no doubt is profoundly influenced by the fact that the proximity matrices are treated as data matrices, doing this, means that the proximity values appear twice on each side of the diagonal. Treating a matrix this way is unorthodox to be sure; nevertheless, it seems to provide very clear results in ACA. Dense matrices contain few zero connections and consequently indicate a denser network of objects more suitable for measuring correlation coefficients. On the other hand, sparse matrices contain a considerable number of zero values, which makes computation of correlation coefficients more vulnerable (cf. Schneider & Borlund, 2007a). Remember that the basis for the Drexel-approach is a matrix of co-citations obtained by multiplying the data matrices by their transpose. Both types of matrices are transformed into matrices of correlation coefficients. As demonstrated above, the two approaches yield different configurations. If the data contour of the matrices is irrelevant to a proximity transformation, then the same set of objects represented by two different matrices, should obtain monotone rankings. If however zero values influence the computation of the correlation coefficient, then rankings will deviate. Two Mantel tests (Mantel, 1967; Schneider & Borlund, 2007b) between the two pairs of correlation matrices for the conventional and Drexel-approach respectively (i.e. the latter computed from the data matrices, and the former from the co-citation matrices) prove this. The Mantel statistic for the first-author correlation matrices gives *rho* = 0.683 (p=0.0001) and the Mantel statistic for the all-author correlation matrices gives *rho* = 0.793

(p=0.0001). This clearly indicates a decreasing monotonicity between the rankings of objects. This is illustrated in Figures 7 and 8.

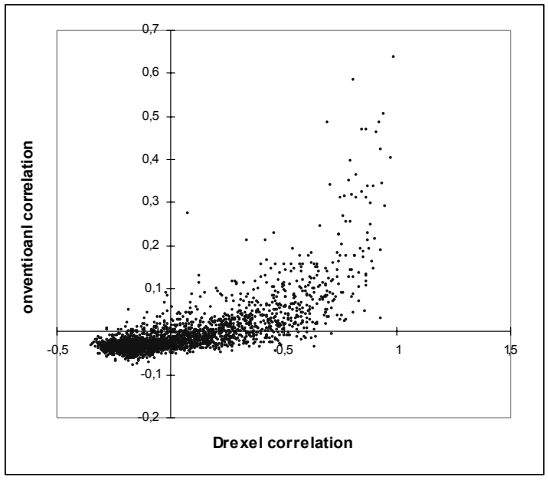


Figure 7. Mantel test for correlation matrices produced by the conventional and Drexel approach for the first-author co-citation analysis.

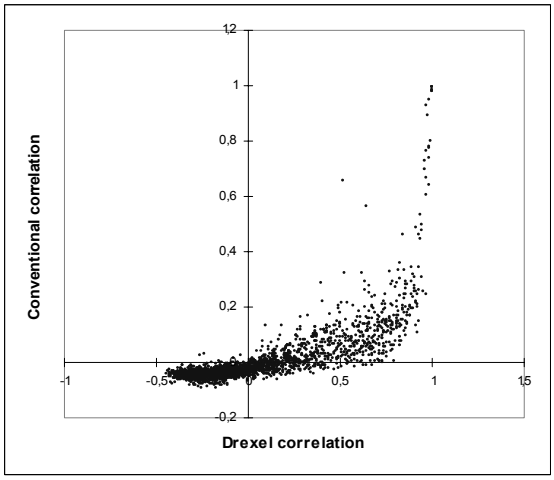


Figure 8. Mantel test for correlation matrices produced by the conventional and Drexel approach for the all-author co-citation analysis.

Note that the Mantel statistic is considerably higher for the all-author correlation matrices, a fact also indicated by the Stress-1 values for the respective pair of configurations. At first glance, this seems to contradict our claim as the dimensionality of the data matrix for the all-author co-citation analysis is considerably larger than its counterpart for the first-author analysis, 75×3161 and 75×2002, respectively. However, when computing the density of these matrices, it becomes clear that the larger all-author data matrix has a larger density (0.05) compared to the first-author data matrix (0.04). Consequently, even though the all-author data matrix is larger it also contains more non-zero values.
The latter results are tremendously important for a factor analysis as it decomposes a covariance or a correlation matrix. Applying factor analysis to a correlation matrix with only low inter-correlations, a likely consequence of sparse matrices, will require factor solutions with nearly as many factors (principal components) as there are original variables, thereby defeating the data reduction purposes of factor analysis. This is clearly the case in the present analysis for the matrices based on the conventional approach as indicated in Figures 10 and 12 below.

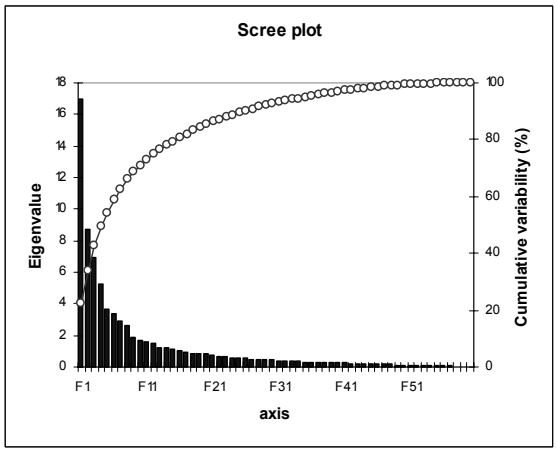


Figure 9. Factor analysis scree plot of first-author co-citation analysis – Drexel approach.

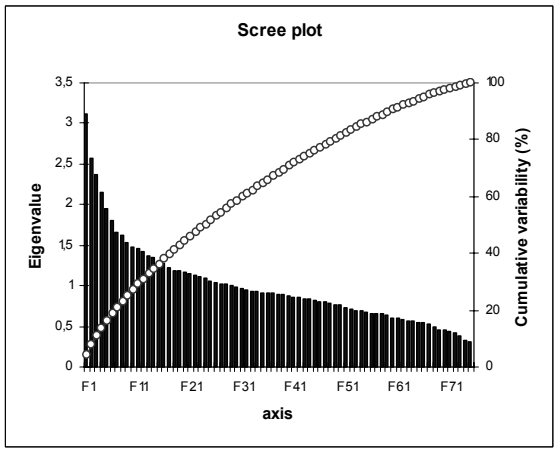


Figure 10. Factor analysis scree plot of first-author co-citation analysis – Conventional approach.

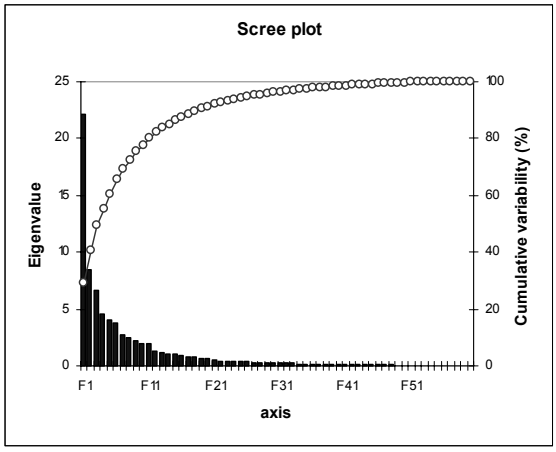


Figure 11. Factor analysis scree plot of all-author co-citation analysis – Drexel approach.

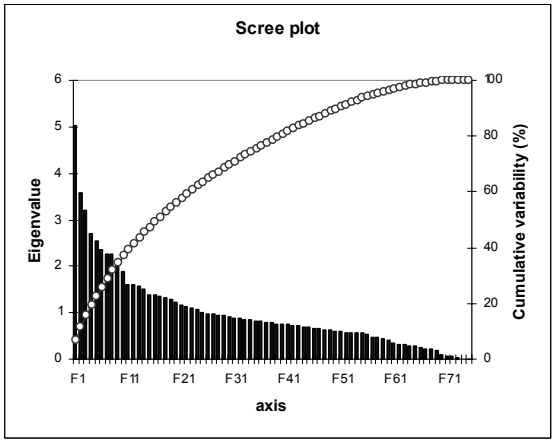


Figure 12. Factor analysis scree plot of all-author co-citation analysis – Conventional approach.

What is desirable when choosing major factors is clearly a shape of the histogram like the ones in Figures 9 and 11, preferably even steeper, and the cumulative variability curves like the one in Figure 11, where the total explained variance is spread among relatively few major factors. Consequently, for the present analysis only the factor analyses based on the Drexel-approach are reliable for identification of major factors or rather specialties within the IEEE data set. Consequently, we have a 16-factor model for the Drexel approach to first-author co-citation analysis, explaining 80% of the variance; and a 15-factor model for the Drexel-approach to all-author co-citation analysis, explaining 86% of the variance. A first glance at the eigenvalues for the extracted factors in Table 2 reveals a remarkable similarity.

Table 2. Eigenvalues for first and all-author co-citation analysis

	First-author co-citation analysis				All-author co-citation analysis		
	Eigenvalue	**Variability (%)**	**Cumulative %**		**Eigenvalue**	**Variability (%)**	**Cumulative %**
F1	16.944	22.233	22.233	F1	22.156	29,289	29.289
F2	8.689	11.402	33.634	F2	8.508	11.247	40.536
F3	6.926	9.088	42.722	F3	6.593	8.716	49.251
F4	5.248	6.886	49.609	F4	4.513	5,965	5.217
F5	3.650	4.790	54.398	F5	4.009	5.299	60,516
F6	3.376	4.429	58.828	F6	3.756	4.965	65.481
F7	2.934	3.850	62.678	F7	2.764	3.654	69.135
F8	2.645	3.470	66.148	F8	2.474	3.270	72.405
F9	1.841	2.415	68.564	F9	2.178	2.879	75,283
F10	1.706	2.238	70.802	F10	1.949	2.576	77.860
F11	1.564	2.052	72.854	F11	1.891	2.499	80.359
F12	1.484	1.947	74.801	F12	1.361	1.799	82.158
F13	1.241	1,628	76.429	F13	1.167	1,543	83.701
F14	1.180	1.548	77.977	F14	1.051	1.390	85.090
F15	1.103	1.447	79.424	F15	1.023	1.353	86.443
F16	1.013	1.330	80.754				

A close inspection of the factor patterns (not included in the paper), before and after oblique rotation, reveals an even more significant similarity between of the extracted factors. Remember that the two sets had an overlap of 41 authors. These authors represent in total 13 common factors in the first-author case, and 14 factors in the all-author case. The large majority of the remaining authors in both cases are spread among these common factors. The few left over authors in both cases represents the

remaining few factors, not unusually only one author are affiliated with such a factor. Figures 13 and 14 below demonstrate the similarity in structure, for factors 1 and 2, between the two author co-citation counting methods.

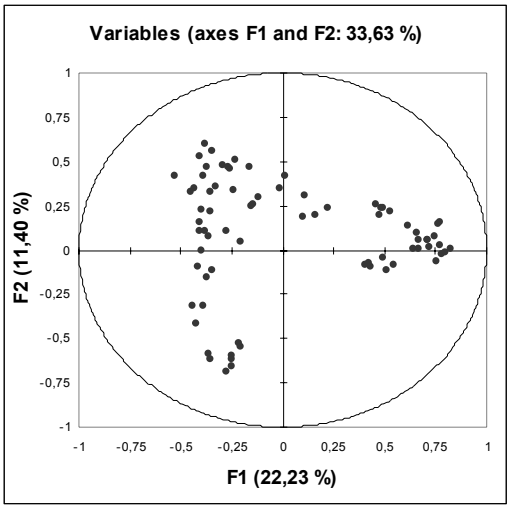


Figure 13. Mapping of two factors for the first-author co-citation analysis.

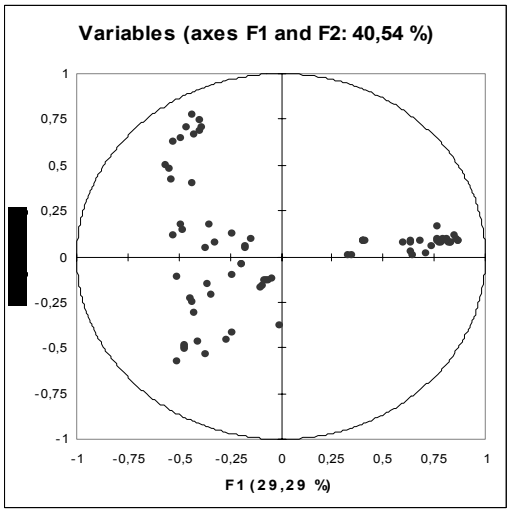


Figure 14. Mapping of two factors for the all-author co-citation analysis.

Perhaps the all-author solution gives a slightly more coherent grouping between the authors? At least its structure does support and elaborate the MDS solution presented in Figure 3. We cannot determine whether the small differences in extracted factors between the two analyses indicate whether first-author co-citation analysis is able to identify more specialties. Some evidence in the data set suggests that several of these authors perhaps should be treated as outliers. Consequently, from the present study we cannot confirm that all-author co-citation counts can lead to identification of fewer specialties, however, the study do confirm that all-author co-citation counting produce more coherent groupings amongst authors.

Conclusion

In this study we have presented a comparative analysis of first-author versus all-author co-citation analysis on the largest dataset so far used in an all-author ACA study. The results show that the inclusion of all cited authors can aid in producing 2D MDS visualisations that better fit the underlying data (i.e., have lower stress values), and that all-author ACA may lead to stronger concentration in the maps. In addition, we also study the differences between using two approaches for matrix generation: the popular Drexel-approach, as well as a conventional approach based on multivariate statistical analysis. Overall the two methods produce maps that are have some resemblances, but also many differences at the detail level. The Drexel-approach produces results that a have noticeable lower stress values and are more concentrated into groups.

The study also demonstrates the importance of sparse matrices and their potential problems in factor analysis. Sparse matrices are more suitable for principal components analysis. In the present study the sparse matrices deflated the extraction of factors. While, we can confirm that all-author co-citation analysis produce more coherent groupings, the present study cannot confirm previous findings that first-author co-citation analysis identifies more specialties. We need to investigate this further, as the removal of duplicates may have affected this result.

The dataset was drawn from full-text scholarly articles formatted in XML that allows precise extraction of a large range of features; most notably in relation to this study all cited authors. In future work we will addresses the research possibilities inherent in a citation index based on source documents formatted in XML.

References

Ahlgren, P., Jarneving, B., & Rousseau, R. (2003). Requirements for a cocitation similarity measure, with special reference to Pearson's correlation coefficient. *JASIST,* 54(6), 550-560.

Chen, C. (1999): Visualising semantic spaces and author co-citation networks in digital libraries. IP&M, 35(3), 401-420.

Giles, C.L., Bollacker, K. & Lawrence, S. 1998: CiteSeer: An Automatic Citation Indexing System. In: *Third ACM Conference on Digital Libraries*. ACM Press, New York, 89-98.

Glänzel, W. (1996): The Need for Standards in Bibliometric Research and Technology, *Scientometrics*, 35 (2), 1996, 167-176.

Gower, J. C. (1971). Statistical methods for comparing different multivariate analyses of the same data. In: Hodson, Kendall & Tautu (Eds.), *Mathematics in the Archaeological and Historical Sciences* (pp. 138-149). Edinburgh: Edinburgh University Press.

Lattin, J., Carroll, J.D., & Green, P.E. (2003). *Analyzing Multivariate Data*. Pacific Grove, CA: Brooks/Cole - Thompson Learning.

Leydesdorff, L., & Vaughan, L. (2006). Co-occurrence Matrices and their Application in Information Science: Extending ACA to the Web Environment. *JASIST,* 57(12), 1616-1628.

Lok, C.K.W, Chan, M.T.V & Martinson, I.M. (2001): Risk factors for citation errors in peer-reviewed nursing journals. *Journal of Advanced Nursing*, 32(2), 223-229.

Malik, S., Kazai, G., Lalmas, M. & Fuhr, N. (2006): Overview of INEX 2005. In: Fuhr et al.: *Proceedings of INEX 2005*. Berlin: Springer 2006, 1-15. (LNCS 3977)

Mantel, N. (1967): A technique of disease clustering and a generalized regression approach. *Cancer Research, 27*, 209-220.

McCain, K.W. (1990). Mapping authors in intellectual space: A technical overview. *JASIS,* 41(6), 433-443.

Persson, O. (2001): All author citations versus first author citations. *Scientometrics*, 50(2), 339-344.

Rousseau, R. & Zuccala, A. (2004): A classification of author co-citations: definitions and search strategies. *JASIST*, 55(6), 513-629.

Schneider, J.W. & Borlund, P. (2007a): Matrix comparison, Part 1: Motivation and important issues for measuring the resemblance between proximity measures or ordination results. *JASIST* (accepted for publication).

Schneider, J.W. & Borlund, P. (2007b): Matrix comparison, Part 2: Measuring the resemblance between proximity measures or ordination results by use of the Mantel and Procrustes statistics. *JASIST* (accepted for publication).

Schönemann, P.H., & Carroll, R.M. (1970). Fitting One Matrix to Another under Choice of a Central Dilation and a Rigid Motion. *Psychometrika,* 35(2), 245-256.

White, H. D. (2003). Author Cocitation Analysis and Pearson's r. *JASIST,* 54(31), 250–259.

White, H.D., & Griffith, B.C. (1981). Author co-citation: A literature measure of intellectual structure. *JASIS,* 32(3), 163-171.

Zhao, D. (2006): *Towards all-author co-citation analysis*. IP&M, 42(6), 1578-1591.

Do New SCI Journals Have a Different National Bias?[1]

R. D. Shelton[*], Patricia Foland[*], and Roman Gorelskyy[*]

[*]*shelton@wtec.org*
WTEC, 4800 Roland Avenue, Baltimore, MD 21210 (USA)

Abstract
National shares of worldwide publications in the Science Citation Index (SCI) have shifted recently. The long-term decline in U.S. share accelerated in the mid-1990s, and now the EU has joined this decline. Not coincidentally, the shares of some countries have increased sharply, particularly those of China, S. Korea, Taiwan, and Singapore. Since the SCI constantly adds new journals, one reason might be that newly added journals were more favorable to them. To test this, the database was partitioned into "old journals" (added before 1995) and "new journals," added afterward. The analysis was done for eight of the 20 fields of science defined by the National Science Indicator CD. In some fields, new journals were indeed much more favorable to the Asians. In some fields, however, new journals were actually more favorable to the U.S. In aggregate over the eight fields analyzed, the size of this effect was too small to account for much of the sharp changes in national shares. Furthermore tests between old and new journals find that differences in most fields are not statistically significant. The results provide evidence that the SCI can be used to accurately track national publication changes over time.

Keywords
Science Citation Index; bias; science leadership; publication share; american paradox.

Introduction

Since 1995 national shares of publications in the Thompson/ISI Science Citation Index (SCI) database have changed substantially. Figure 1 shows that the long-term decline of U.S. share accelerated in the mid-1990s. Also, in the late 1990s, the EU stopped gaining share and started losing it. The share of some Asian countries increased sharply from a insignificant amount in the 1980s to clearly cut into the western shares in recent years; the figure shows the share of the "Asian Tigers" -- China, S. Korea, Taiwan, and Singapore. The long-term decline of American share has resulted from slow growth in American paper production coupled with fairly rapid growth of the total of papers in the NSI annual CD--an average increase of about 16,000 per year. For example the U.S. share was 39.7% in 1981, declined to 37.6% in 1994, and then declined more rapidly to 33.3% in 2004. The rate of American decline has more than doubled in recent years: from 0.15 points per year in 1981-1994, to 0.39 points per year in 1995 - 2004. It is also important to note that the EU has slipped in share in the last four years: from 39.4% to 37.9%.

The big gainers are the Asian countries that are making very aggressive investments in R&D: China, Taiwan, South Korea and Singapore; together these four increased their share from only 0.6% in 1981 to 10.8% in 2004. As an example of this rapid rise, Jin and Rousseau (2005) have fitted an exponential curve to the Chinese publication count since 1991, calculating a doubling time of only 4.7 years. This shift in publication shares from the West to the East has already been the subject of several other analyses (Leydensdorff and Zhou 2005) and (Moed 2002). Most recently the Asian Tiger group took share from both the U.S. and EU, increasing almost 4 percentage points (7.0% to 10.8% share) in just the last four years. (These shares do not take into account the double counting of papers that have authors from more than one of the Asian Tiger countries; however based on analysis of such joint authorship in the online Science Citation Index database, this factor is estimated to not seriously affect the estimate of their share.)

[1] Sponsored by NSF grant ENG-0423742 and a sabbatical from Loyola College. (Bpaper2.28.7C.doc.)

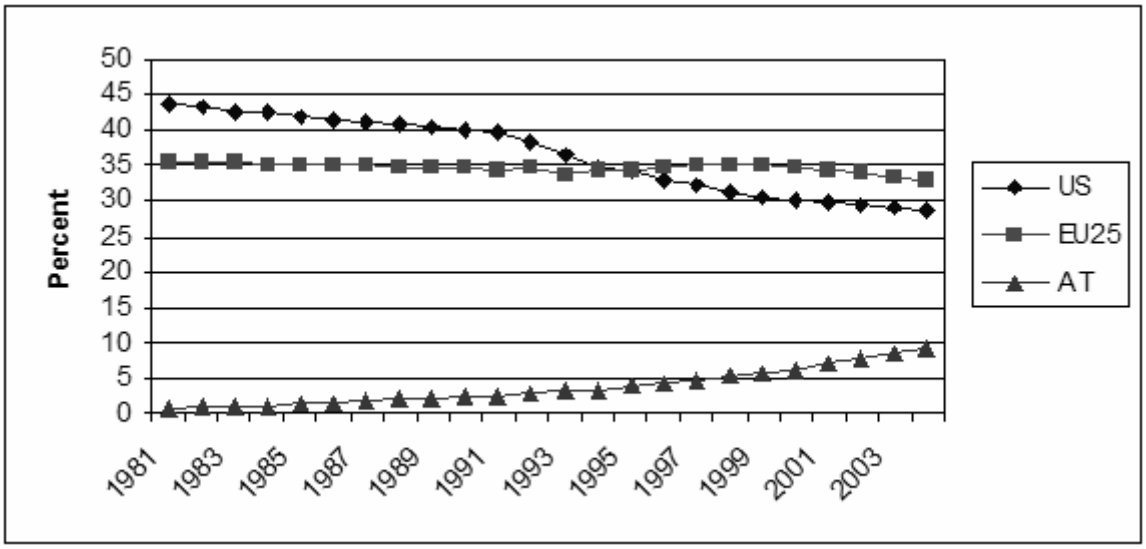


Figure. 1. Share of scientific publications in the 24 fields on the 2004 NSI CD. AT includes China, South Korea, Taiwan, and Singapore.

The Science Citation Index (SCI) database is widely used for measurement of output of scientific papers from nations, institutions, and individuals. Scientific prestige, tenure, and competition for scarce research funds can be at stake in measurements made from the database; not surprisingly, it has been much criticized for being biased. Since soon after its inception, it has been accused of being biased toward papers in the English language and those from the United States. A recent paper by Archambault and colleagues (2005) reviews this literature in connection with making the case that the SCI database is not well suited for rating the social sciences and humanities, even in its Social Science Citation Index (SSCI) variant.

Thompson/ISI adds new journals to its databases each year, which accounts for most of the annual increase in the number of total publications in the database (NSB 2006, Figure 5-34). While the process of adding new journals to the database is mostly objective, it does have a subjective component. International diversity is one of the criteria in making such selections (Testa 2004):

> *Geographic representation of a journal is another consideration. To meet the needs of its international subscriber base, ISI seeks to cover journals with international diversity among authors of both source articles and cited articles. To properly reflect the global context in which scientific research takes place, and to provide balanced coverage in each category, ISI seeks to cover the best regional journals as well.*

Zitt and colleagues (2003) have investigated national publication counts (and citation shares) as a function of the number of journals included in the ISI database, assuming that journals are included in order of their impact factors. In particular they found that for 1997 data, the publication share of the U.S. decreased as the number of journals in the database increases, i.e. as more lower-impact journals were added.

While there has been a gradual decline in U.S. share of papers in the SCI since 1981, after 1994 this decline accelerated sharply. Thus, there is the possibility that the selection criteria of newer journals are less favorable to the U.S. – perhaps partly in response to past criticism. This paper analyzes the journals added to the database after 1994 to determine whether the bias toward the U.S. has been reduced in these "new journals." The same techniques were also used to examine changes in the database as they affect the EU and AT regions.

NSF has done some similar analysis (NSB 2006, Note 39, p. 5-60), which shows that keeping the SCI database fixed with only the journals added before 1985 would even increase the calculated loss of American share.

First the methods and data used here will be described. Then the publication share results for each of the three regions will be summarized in turn. It will be seen that in some fields there were indeed substantial differences between the old and new journals. Sometimes, but not always, the new journals were more favorable to the Asian Tigers, and less so to the western countries. However, in every case analyzed the effect is so small on overall share that it cannot account for the substantial changes seen in recent years. Thus these shifts are real and not an artifact of the measurement methods used. Using this reassurance, Shelton (2006) has sought the underlying real reasons for these shifts. Using multiple linear regression analysis, he has provided data that suggest that since 1998 the Asian Tigers have taken *publication share* from western countries simply because they have sharply increased their world R&D *investment share.*

Databases and Methodology

The analysis mainly uses the online version of the SCI. Records from papers and other publications are constantly being added to this database, and not just for the current year. Thus measurements taken at different times yield slightly different results, even for dates several years ago. The CD version of the National Science Indicators (NSI) for 2004 was used for Figure 1 and to define the scope of the various fields analyzed through its lists of journals defining each of the 24 fields covered. In total the 2004 NSI listed 9129 journals in natural science fields plus 1723 others in social science fields: education, economics and business, law, and general social sciences. These social science fields are mostly indexed under the separate Social Science Citation Index (SSCI) online, although there is some overlap.

The procedure for testing the hypothesis that new journals were less favorable to the U.S. was as follows. Determination of when journals were added to the ISI database was by a brute force approach, although other approaches are possible. That is, for each journal in a particular field, the online SCI (and SSCI) database was queried to find all papers in that journal. Since the hit list is sorted chronologically, it was only necessary to go to the end of the file to find the publication date of the earliest paper. It was assumed that earliest date was the date that the journal was added to the database. Actually editors sometimes "back fill" data for earlier papers when their journals are first accepted by Thompson/ISI, so this procedure sometimes counts the entry date earlier than it should be. Another problem is that journals in the SCI simply change their name, making it difficult to determine from a search if they are truly new to the SCI database. In any case for this paper journals were individually classified as "old journals" if their earliest paper was before 1995, and as "new journals" if it was in 1995 or later.

Since this search of individual journals is tedious, it was not feasible to analyze the entire set of 10,900 journals. Thus the analysis focused on eight individual disciplines that include relatively few journals. Table A1(in the Appendix) shows the journal count, including the number of new journals identified for the disciplines analyzed.

After separating the journals for a particular discipline into these two classes, search filters were constructed to count papers for each year in the two classes. For example a search would be made for 1983, 1986, 1989, ... 2004 using:
SO = (New Journal #1 OR New Journal #2 OR ... OR New Journal #N)
Of course, searches prior to 1995 should yield no hits for new journals, which served as a partial check on the separation into classes. There is a limit of 50 logical terms in a SCI query, so that up to six filters had to be used for each discipline. These filters had to be further subdivided for the EU searches, since its 25 countries consume more than half of the allowable terms. Results from the separate filters were summed for each year to get the publications for the U.S., for the EU, for the AT, and for all countries for year in each class. The searches were for all types of publications and for all languages.

The first field analyzed was space science. It was found that new journals introduced into the SCI after 1994 were indeed not nearly so favorable to the U.S. as old journals introduced before 1995. In 2004

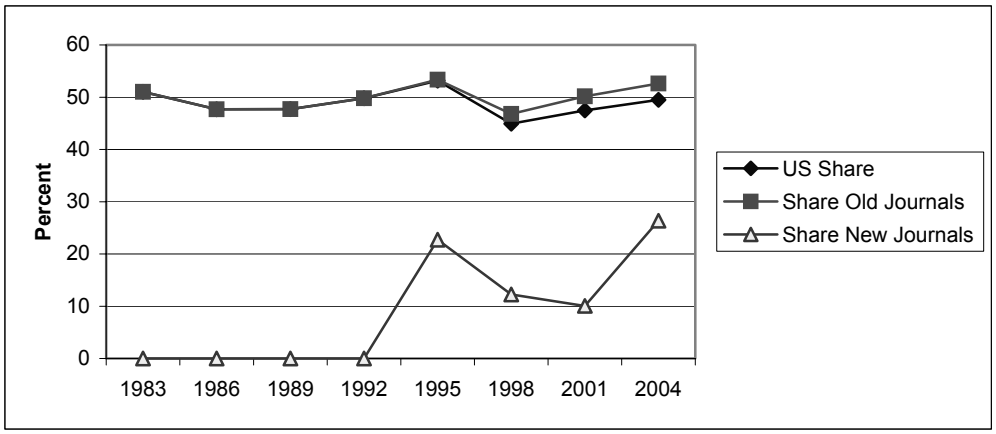


Figure. 2. U.S. share of papers in space science. There is a big difference between old and new journals, but keeping the database fixed with only old journals would add only 2% more in 2004.

52.6% of the papers in the 41 old journals had a U.S. author; only 26.4% of the papers in the 15 new journals had a U.S. author. Figure 2 shows some of these results.

While the difference between the two classes is large in this case, it is important to note that keeping the database fixed with only old journals would produce a U.S. share of about 52% in 2004, which is not substantially different from the 49.5% measured from the combined database of old and new journals. This is because there are relatively few new journals, and they are generally smaller; in 2004 only about 10% of the papers come from new journals.

Similar searches were done for seven other disciplines. The publication share results for the three regions will be discussed in turn. First Table A2 shows the worldwide totals of papers in the eight disciplines for 1989, 1995, and 2004. For 2004 counts are shown for papers in all journals, old journals only, and in new journals only.

National Shares in Old and New Journals

U.S. Shares

Table A3 lists the number of papers with one or more U.S. authors in eight fields of science for 1989, 1995, and 2004. U.S. paper counts for 2004 are tallied separately for all journals, old journals only, and new journals only. Table A4 shows the U.S. share in each field and year, which is just the ratio of corresponding entries in Tables A2 and A3. Figure 3 shows these percentage shares in columns. Comparing the last two columns in Table A4, it turns out that the first field analyzed, Space Science, had the greatest difference in share between old and new journals in 2004: 52.6% vs. 26.4% respectively. Other fields where new journals published a smaller share of U.S. papers include immunology, math and microbiology. However, in engineering, computer science, and pharmacology new journals were actually somewhat more favorable to the U.S. than old ones. Old and new journal counts were about the same for materials science. Furthermore, comparison of the "All Journals" and "Old Journals" columns in 2004 shows that the changes between old and new journals were not great enough to much affect the overall trends. The aggregate difference over the eight fields in 2004 was less than one percentage point. That is, there was a U.S. share of 28.9% for all journals and 29.2% for just the old ones. `Thus this effect was not likely to account for much of the more than 10 percentage point decline in U. S. share from 1989 to 2004 (for all 24 fields) shown in Figure 1.

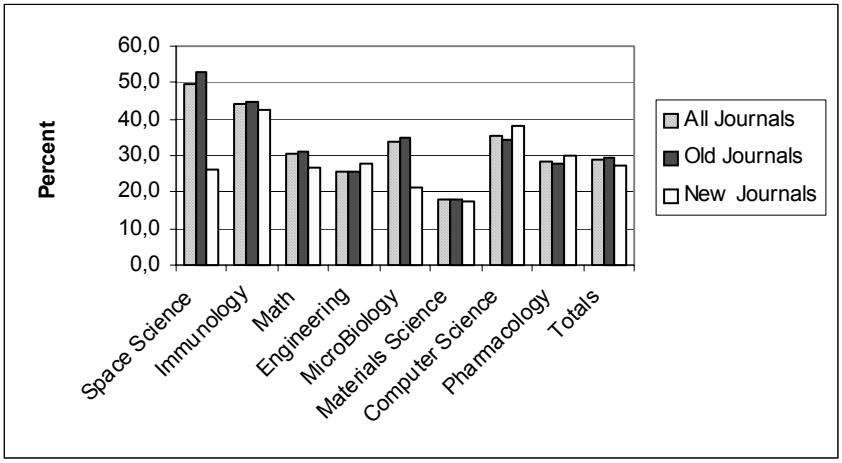


Figure 3. U.S. shares of old, new, and all journals in 2004.

EU Shares

Table A5 and Figure 4 show similar percentage share results for the 25 countries of the European Union. (The actual number of EU papers could be calculated by multiplying the corresponding entries of Tables A5 and A2.) Comparing the last two columns of again for Table A5, some fields show that new journals are less likely to publish EU papers than old ones: space science, immunology, math, and pharmacology, particularly. New journals in some fields are more favorable to the EU than old ones: engineering, microbiology, materials science, and computer science. Again, there is not much difference in any of the fields between the shares calculated between old journals and all journals, because new journals have relatively few papers. The aggregate over the eight fields is a difference between 33.4% for all journals and 33.7% for old journals alone. Thus the recent decline of EU share is not likely to be caused by shifts in the nature of the SCI database.

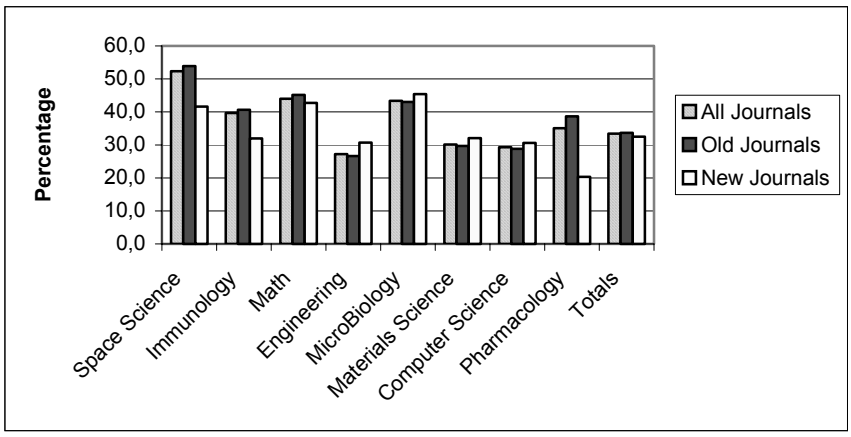


Figure 4. European Union publication shares in old, new, and all journals in 2004.

Asian Tiger Shares

Table A6 and Figure 5 show the percentage share results for the four Asian Tigers (AT): China, Taiwan, S. Korea, and Singapore. (The actual number of AT papers could be calculated by multiplying the corresponding entries of Tables A6 and 2.) Here the table shows the very rapid rise in share of this group, but what part of this can be attributed to addition of new journals to the SCI database that are more favorable to the Asians? Comparing the last two columns again, in only two fields are new journals much more likely to publish AT papers than old ones: space science and engineering. New journals in most fields are actually less favorable to the AT than old ones. In any case, there is again not much difference in any of the fields between the shares calculated between old

journals and all journals. The aggregate over the eight fields is a shows essentially no difference -- between 13.3% for all journals and 13.2% for old journals alone. Thus the recent rapid rise of AT share is not likely to be caused by shifts in the nature of the SCI database.

To make this point clearer, Figure 6 illustrates engineering, the field with the largest difference between old and new journals. It shows mainly the rapid rise of the Asian Tigers' publication share, and the new journal curve shows that they are indeed much more likely to publish papers from these countries than old journals. But because there are relatively few new journals, the miniscule difference between the old journal and all journal curves shows again that this effect is far too small to account for the rapid rise of Asian publication.

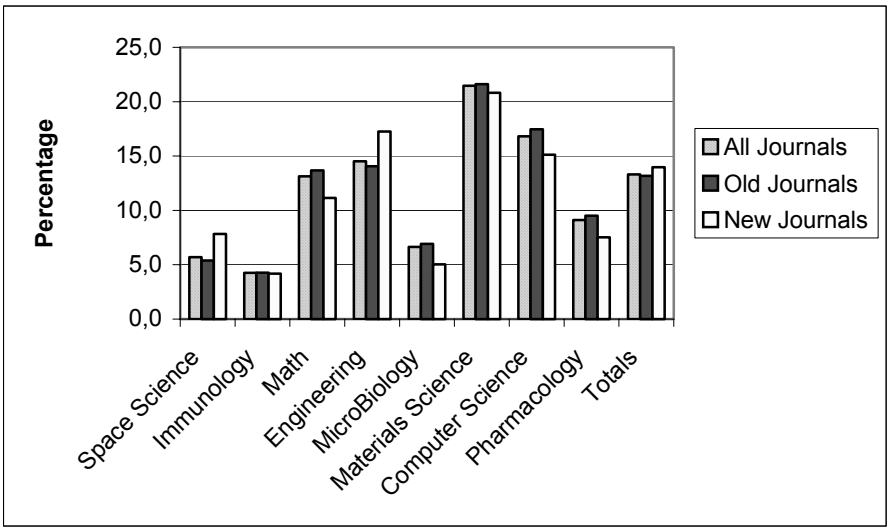


Figure 5. Asian Tigers share of papers in old, new, and all journals in 2004.

Tests of Statistical Significance of Difference Between Old and New Journals

While these differences between old and new journals are real for the data of 2004, they may well have occurred by chance, not due any systemic change in bias in 1995. To settle that issue, one needs to test the statistical significance of these differences assuming that the new journals were a sample from a larger population of possible new journals. This analysis, being even more tedious, is done for only six of the fields, and for the U.S. only. The results are summarized in Table A7.

Tests of significance required that the queries for share of U.S. publication be repeated at the individual journal level for the year 2004. This served as a check on the accuracy of the measurements described earlier. Two statistics were calculated for each journal.

The first statistic is simply the U.S. publication share of an individual journal in 2004. No weighting is applied for the size (number of papers) in the journal, so that the decision to add a small journal to SCI is treated as just as important as that for a large one. As shown in left half of Table A7, "Averages of Individual Shares," only two of the six fields had a statistically significant difference at the 0.05 level: space science and pharmacology.

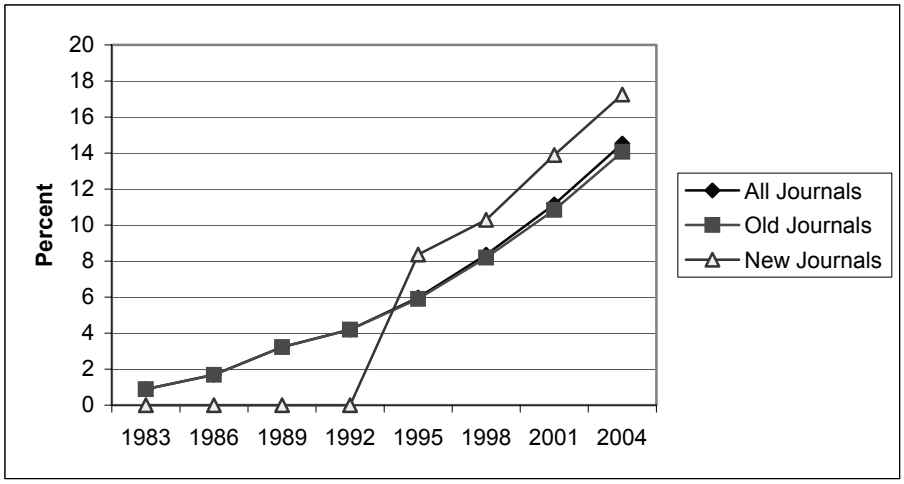


Figure 6. Asian Tiger share of engineering papers in old and new journals versus time.

The second statistic takes into account the size of the journal, weighting larger journals more heavily. Each journal's entry is its U.S. papers for 2004 normalized by the average number of U.S papers per journal in that set (old or new) for 2004. This statistic has the advantage that its mean is unbiased, i.e. it is the simple ratio of U.S. papers to total papers for that set (old or new), and is approximately the same as that calculated earlier. Despite what appear to be fairly large differences in the right half of Table A7 and fairly large samples, none of the six fields has a significant difference between old and new journals at the 0.05 level.

While other fields and the other regions were not analyzed, these results suggest that differences in those cases are probably not significant either.

Some recent work has analyzed the effect of journals that were in the SCI in 1994 changing their names after 1995. Such name changes caused these journals to be classified as new ones in this study when actually they should be classified as old ones. The effect of this correction on the previous findings was analyzed for four fields – it is even more tedious to tease out such name changes. Some of these results can be presented at the conference, and posted on the archival website. The important thing is that these corrections do not alter the key findings of this paper.

Since distortions in the SCI do not appear to account for the large shifts in shares between the U.S. and EU during the 1990s, the underlying reasons remain unclear. One known factor is that national investments in civilian research are more productive of scientific paper outputs than those in military research. Thus Europe's greater investment in the civilian component could help explain why it gained publication share from the U.S. during the early 1990s, and came to pass the U.S. in the mid-1990s. The even larger shifts in share more recently from both the U.S. and EU to the Asian Tigers are actually easier to explain: the Asian Tigers have been increasing their research investment so rapidly that they have taken both worldwide investment share and correlated publication share from the West (Shelton 2006).

Conclusions

The data presented here should reassure bibliometricians that their Science Citation Index has been an accurate indicator in measuring national shares of publications over recent years. Changes in the database are inevitable as important new journals emerge, old journals die or are reorganized under new names, and new nations become more important players in the game of scientific publication. However, addition of new journals to the database in the 1990s seems to have not distorted the national shares during an era when those shares were sharply changing.

Still, these findings depend somewhat on the discipline. American researchers in space science, or European researchers in pharmacology might like to know that new journals have been less likely to publish their papers than old journals. Chinese engineering researchers might like to know that new journals have been much more receptive to their papers than old ones.

References

Archambault, E, Vignola-Gagne E., Cote, G, Lariviere, V. & Gingras, Y. (2005), Welcome to the linguistic warp zone: Benchmarking scientific output in the social sciences and humanities. Proceedings of the ISSI 2005 Conference, Stockholm, July 24-28, 2005. pp. 149-158.

ISI (2004), National Science Indicators 1981-2003, Standard. Philadelphia: Thompson ISI. (CD)

National Science Board (NSB 2006) Science and Engineering Indicators 2006. Two volumes. Arlington, VA: National Science Foundation.

Jin, B., & Rousseau, R. (2005), China's quantitative expansion phase: Exponential growth, but low impact. Proceedings of the ISSI 2005 Conference, Stockholm, July 24-28, 2005. pp. 362-370.

Leydensdorff, L., Zhou, P. (2005), Are the contributions of China and Korea upsetting the world system of science? *Scientometrics*: 63: 617-630.

Moed, H. F. (2002), Measuring China's research performance using the Science Citation Index. *Scientometrics* 53: 281- 296.

Shelton, R. D. (2006), Relations between national research investment and publication output: Application to an American paradox. Ninth International Conference on Science and Technology Indicators, Leuven, Sept. 7-9, 2006. To appear in Scientometrics.

Testa, J. (2004), The Thomson Scientific Journal Selection Process, 2004. Retrieved from http://scientific.thomson.com/free/essays/selectionofmaterial/journalselection Feb. 28, 2007.

Zitt, M, Ramanana-Rahary, S., & Bassecoulard, E (2003), Correcting glasses help fair comparisons in international science landscape: Country indicators as a function of ISI database delineation. *Scientometrics*, 56: 2, 259-282.

Appendix: Detailed Tables

Table A1 Number of Journals Defining Disciplines on the 2004 NSI CD

Discipline	Total	Total	New
Agricultural Sciences	462		
Biology & Biochemistry	596		
Chemistry	649		
Clinical Medicine	1991		
Computer Science	216	211	78
Ecology/Environment	299		
Economics & Business	370		
Education	150		
Engineering	941	919	211
Geosciences	370		
Immunology	119	118	11
Law	123		
Materials Science	319	317	70
Mathematics	245	240	73
Microbiology	166	161	20
Molecular Biology & Genetics	211		
Multidisciplinary	166		
Neurosciences & Behavior	240		
Pharmacology	178	178	31
Physics	458		
Plant & Animal Science	857		
Psychology/Psychiatry	588		
Social Sciences, General	1080		
Space Science	58	56	15
Total	10852	2200	509

Table A2. Worldwide papers.

	1989	1995	2004	2004	2004
Journals	**All**	**All**	**All**	**Old**	**New**
Field					
Space Science	5244	6985	9846	8670	1176
Immunology	10843	14252	14191	13145	1552
Math	10720	13411	16109	12472	3582
Engineering	48046	68995	78101	66471	11630
MicroBiology	14222	15694	18316	16421	2704
Materials Science	15820	23448	33246	27039	6207
Computer Science	7124	10183	12213	8848	3365
Pharmacology	16189	17741	21878	17601	4277
Totals	128208	170709	203900	170288	33612

Table A3. U.S. Papers

	1989	1995	2004	2004	2004
Field	**All**	**All**	**All**	**Old**	**New**
Space Science	2503	3727	4874	4564	310
Immunology	4632	6202	6290	5633	657
Math	4403	4903	4881	3841	1032
Engineering	17128	22241	20217	17010	3207
MicroBiology	5082	5347	6175	5776	652
Materials Science	4275	6113	5914	4823	1091
Computer Science	2987	3815	4310	3031	1279
Pharmacology	5543	5949	6218	4938	1280
Totals	46553	58297	58879	49712	9167

Table A4. U.S. Share in Percent

	1989	1995	2004	2004	2004
Journals	**All**	**All**	**All**	**Old**	**New**
Field					
Space Science	47.7	53.4	49.5	52.6	26.4
Immunology	42.7	43.5	44.3	42.9	42.3
Math	41.1	36.6	30.3	30.8	28.8
Engineering	35.6	32.2	25.9	25.6	27.6
MicroBiology	35.7	34.1	33.7	35.2	24.1
Materials Science	27.0	26.1	17.8	17.8	17.6
Computer Science	41.9	37.5	35.3	34.3	38.0
Pharmacology	34.2	33.5	28.4	28.1	29.9
Totals	36.3	34.1	28.9	29.2	27.3

Table A5. EU Share in Percent

	1989	1995	2004	2004	2004
Journals	**All**	**All**	**All**	**Old**	**New**
Field					
Space Science	26.5	38.4	52.4	53.9	41.6
Immunology	32.5	37.5	39.7	40.6	32.0
Math	30.0	41.4	44.0	45.2	42.7
Engineering	16.6	24.2	27.2	26.6	30.7
MicroBiology	29.5	42.5	43.4	43.0	45.4
Materials Science	18.0	29.1	30.1	29.7	32.1
Computer Science	17.9	26.2	29.3	28.8	30.6
Pharmacology	31.7	40.5	35.1	38.7	20.3
Totals	22.8	31.3	33.4	33.7	32.5

Table A6. Asian Tigers Share in Percent

	1989	1995	2004	2004	2004
Journals	**All**	**All**	**All**	**Old**	**New**
Field					
Space Science	1.0	2.4	5.7	5.4	7.8
Immunology	0.5	1.0	4.3	4.3	4.2
Math	2.7	5.4	13.1	13.7	11.1
Engineering	3.2	6.0	14.5	14.1	17.2
MicroBiology	0.9	1.9	6.6	6.9	5.0
Materials Science	4.1	8.1	21.5	21.6	20.8
Computer Science	2.1	4.8	16.8	17.5	15.1
Pharmacology	1.9	3.1	9.1	9.5	7.5
Totals	2.5	5.0	13.3	13.2	14.0

Table A7. Summary of Individual Journal Shares for U.S. in 2004

	Averages of Individual Shares (Ignores number of papers in journal)				**Averages weighted by relative number of papers in journal (Ratios:US papers / All papers in set)**			
Field	**Old Mean**	**New Mean**	**t-test**	**Significant?**	**Old Ratio**	**New Ratio**	**t-test**	**Significant?**
pharmacology	29.2	41.3	0.025	yes	28.4	29.9	0.856	no
space science	42.4	23.7	0.023	yes	52.8	26.4	0.289	no
microbiology	25.9	27.8	0.669	no	35.1	24.1	0.258	no
math	33.5	30.1	0.210	no	30.8	28.8	0.684	no
immunology	39.5	42.2	0.648	no	42.9	42.3	0.973	no
computer science	35.6	36.2	0.863	no	33.8	38.8	0.551	no

Citation Structure of an Emerging Research Area: Organic Thin Film Transistors

Henry Small* and Phineas Upham**

*henry.small@Thomson.com
Thomson Scientific, 3501 Market St., Philadelphia, Pa. 19104 (USA)

**uphams@wharton.upenn.edu
The Wharton School, University of Pennsylvania, 3620 Locust Walk, Philadelphia, Pa. 19104 (USA)

Abstract
A case study of an emerging research area is presented dealing with the creation of organic thin film transistors, a subtopic within the general area called "plastic electronics." The purpose of this case study is to determine the structural properties of the citation network that may be characteristic of the emergence, early development and demise of a research area. Research on organic thin film transistors is highly interdisciplinary, involving journals and research groups from physics, chemistry, materials science, and engineering. There is a clear path to industrial applications if certain technical problems can be overcome. Despite the applied nature and potential for patentable inventions, scholarly publications from both academia and industry have continued at a rapid pace through 2006. The question is whether the bibliometric indicators point to a decline in this area due to imminent commercialization or to insurmountable technical problems with these materials.

Keywords
citation analysis; specialties; co-citation clustering; emerging areas; mapping

Introduction

Organic semiconductors have been known since the 1940s, and the first transistor based on an organic semiconductor was reported in 1986 (Collins). The 2000 Nobel Prize in Chemistry recognized the contributions to this area by Heeger, MacDiarmid, and Shirakawa for work done in the late 1970s. They were able to create semiconductors by doping the polymer polyacetylene. Conducting plastics have already been used in a number of applications, such as light emitting diodes. However, research in this field has recently picked up pace with new discoveries and the introduction of new materials.

Organic thin film transistors hold out the promise of the inexpensive manufacture of electronic and computing devices by printing on plastic. They have attracted much interest due to their potential applications in low-cost, light-weight, flexible, large-area applications such as smart cards, radio-frequency identification tags, flat panel displays, and computing in clothing. However, despite this progress and potential, several technical issues remain to be resolved before organic thin film transistors can be in wide scale use.

The field of organic thin film transistors came to our attention because it was the largest emerging cluster in an analysis of Web of Science data from the 1999-2004 period. This cluster first appeared in a research front analysis in August of 2004, when it consisted of only three highly cited papers. By year-end 2004 it had expanded to 26 highly cited papers.

Method

This area was identified using the method of co-citation clustering (Small 2006) during a routine analysis. An outline of this methodology is given in the following link: http://www.esi-topics.com/RFmethodology.html. To briefly review this process, the top 1% of highly cited papers is selected from each of 22 broad fields of science defined by journal sets for each year within a six year rolling file. The highly cited papers are subject to a single-link cluster analysis, adjusting a cosine normalized co-citation threshold to create a cluster within a prescribed size limit. This process creates approximately 5,000 clusters, and is repeated every two months with updated data for the six-year rolling file. Each set of clusters for one time period is matched against the clusters for the prior time

period to look for continuing areas as well as "new" clusters, that is, those containing highly cited papers that did not appear in any cluster in the prior period. These constitute the so-called "emerging research areas". Clusters are linked across time periods if they share continuing highly cited papers. Following these links across time allows us to create what are called cluster strings, the idea for which derives from Garfield's historiographs (2003). The methodology of co-citation combined with single linkage is particularly well suited to obtaining a rapid and broad overview of the research areas represented in a large set of highly cited documents. This provides a rough delineation and indicator of currently active specialties that can be refined using other more rigorous techniques, such as those described by Chen (2006) and Morris (2003). It is not claimed that the methodology used here provides the most accurate or refined picture possible of a research area, but rather an important initial view, and broad-based science monitoring strategy.

Additionally, each cluster can be displayed as a two-dimensional configuration generated using a force-directed placement algorithm (Small 2006). This procedure works by setting up attractive forces between co-cited papers and repulsive forces between all papers. The attractive force is directly proportional to the inter-paper distance multiplied by the cosine of the co-citation strength, and the repulsive force is inversely proportional to the square of the distance. The configuration is formed by successive addition of co-citation links, starting with only the strongest links that form a minimal spanning tree, and adding links successively in a series of iterations. In each iteration the positions of papers are adjusted to minimize the force on each paper. The amount of residual force per node at the end of the process is used as a goodness of fit measure. The limitations of this methodology are that weak co-citation links may not be taken into account in the positioning of nodes, that as in other ordination methods, such as multidimensional scaling, local minima may be reached, and that as in any attempt to represent complex networks in two dimensions only approximate fits may be achieved.

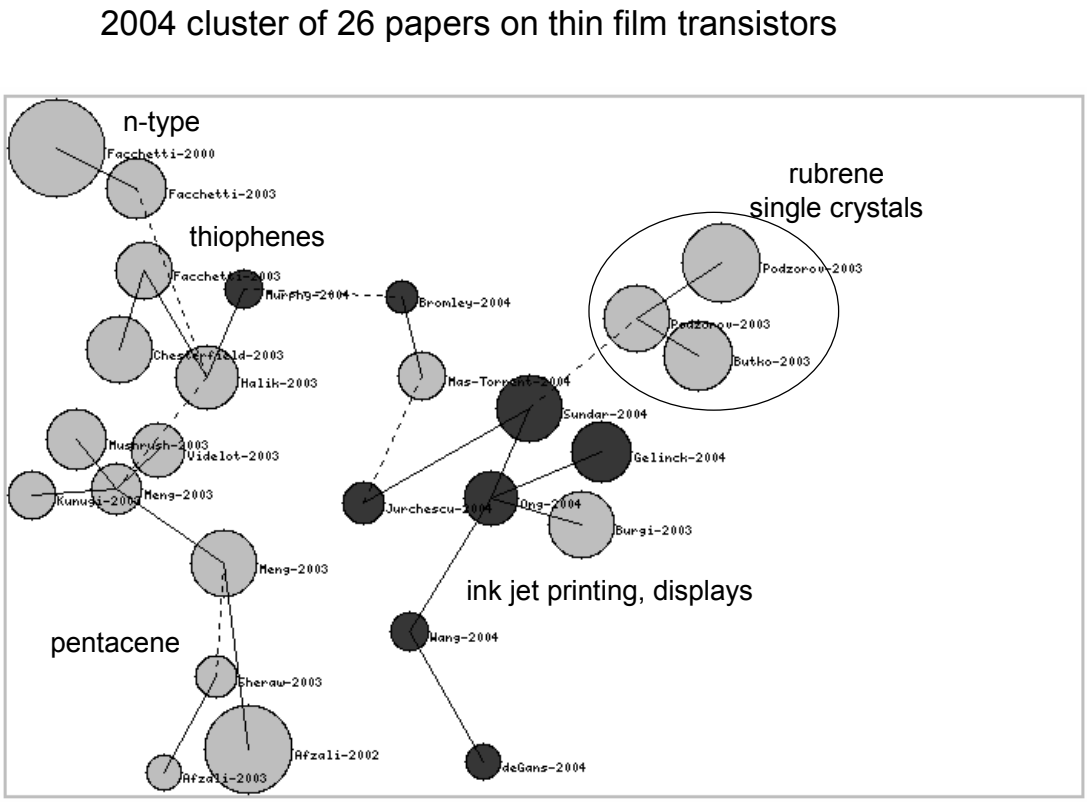


Figure 1. Map of the initial 2004 cluster on organic thin film transistors.

Emerging cluster on organic thin film transistors

The co-citation structure of the organic thin film transistor cluster is shown in Figure 1. The cluster consists of 26 highly cited papers. Each circle is a paper whose area is proportional to its citation count. Papers with the most recent publication year are darker in color. Papers are connected only by the strongest normalized co-citation links per paper (solid lines), plus additional weaker links sufficient to connect the remaining components (dashed lines). This forms a minimal spanning tree, with 25 links connecting the 26 papers. Actually there are 275 co-citation links among these 26 papers, that is, 85% of the theoretically possible connections. The residual force per node for this map is .064.

The cluster can be labelled by region according to the major themes of research. The oldest paper in the cluster from 2000 at the upper left is a demonstration of an n-type, or electron conducting transistor in an organic thin film. By contrast most organic transistors are p-type or hole conducting. Research tends to group around specific organic materials with transistor potential such as thiophenes, pentacene and rubrene. In addition, there is a grouping on the lower right dealing with fabrication methods such as printing on plastic, and applications such as flat panel displays. Many of these themes will continue in the next two years.

Co-citing contexts were examined for each pair of papers along the path of the spanning tree (Small 1986). The majority of links were what might be termed rhetorical modifications (paper A does X but paper B does Y). In other words an author would cite one paper as having somehow modified, qualified, or improved upon the results of the other paper. This finding illustrates the high degree of competition and disagreement in the field regarding the best technical approach to the problem. A smaller number of links might be characterized as applications (A has applied B), or redundancies (A is equivalent to B).

Citation and social characteristics

To examine some citation and social characteristics of the cluster we analyze the citing papers for the set of 26 highly cited papers. Earlier studies have shown that the authors within these clusters tend to know one another and form social groups (Crane). There are 228 citing papers, and 15 of these (6.6%) are in the set of 26 highly cited papers in the six-year period. This percent is a measure of the inbred nature of the cluster. The 228 citing items cite the highly cited papers 460 times. Seventeen percent of these citations were self-citations. The authors on the highly cited papers were also authors on 40% of the citing papers, accounting for about 50% of the citations. This means that highly cited authors remain involved in writing current papers and cite each other frequently. Thus, the leading authors are still active in the field, underlining the small-world nature of this group, where citing and cited populations are highly overlapped. The currency of the cluster is indicated by the percentage of cited items in the most recent year (2004), as well as the percentage of citing items in that year, as shown in Table 1. By way of comparison the table also shows similar measures for a sampling of other emerging clusters (Upham).

Table 1. Citation characteristics of emerging specialty clusters

	Thin-film transistors	**SARS**	**Ekpyrotic universe**	**Transgenic Cotton**	**ZnO nanowires**
% self-citations (all author)	16.9%	9.8%	13.6%	25.2%	13.2%
% citing papers by cited authors	39.9%	34.0%	37.1%	45.0%	27.0%
% cited items in most recent year	30.7%	92.0%	78.6%	85.7%	15.4%
% citing items in most recent year	78.9%	100%	100%	81.7%	63.2%
% citing items that are cited items	6.6%	5.0%	13.6%	16.7%	1.1%

The drive toward applications is evident if we examine the cited and citing papers. All but five of the 26 highly cited papers specifically mention possible applications in their first paragraphs.
Ten of the 26 papers also point to "significant recent progress" in their first paragraphs. This is echoed by the citing authors. The progress alluded to is often the increased charge carrier mobility, an important characteristic for the organic materials. The papers reveal a great deal of tinkering with systems in order to get better charge mobility. All investigators appear to be racing toward the same goal – finding a practical and economically viable device.

An industrial scientist interviewed for this study reported that the field was currently poised between basic research and applications, what the interviewee called the "development phase," with some significant obstacles to overcome before commercialization can begin. The focus on applications and economic payoff is reflected in the involvement of authors from companies such as Lucent, IBM, Infineon, Philips, and Xerox, often in collaboration with university partners. Table 2 shows the institutional mix of author addresses across sectors, both for the cited and citing populations.

Table 2. Sector of cited and citing authors for the initial cluster

Sector	Percent of cited addresses	Percent of citing addresses
Academic	60.7%	60.8%
Industry	33.6%	27.5%
Government	5.6%	11.6%

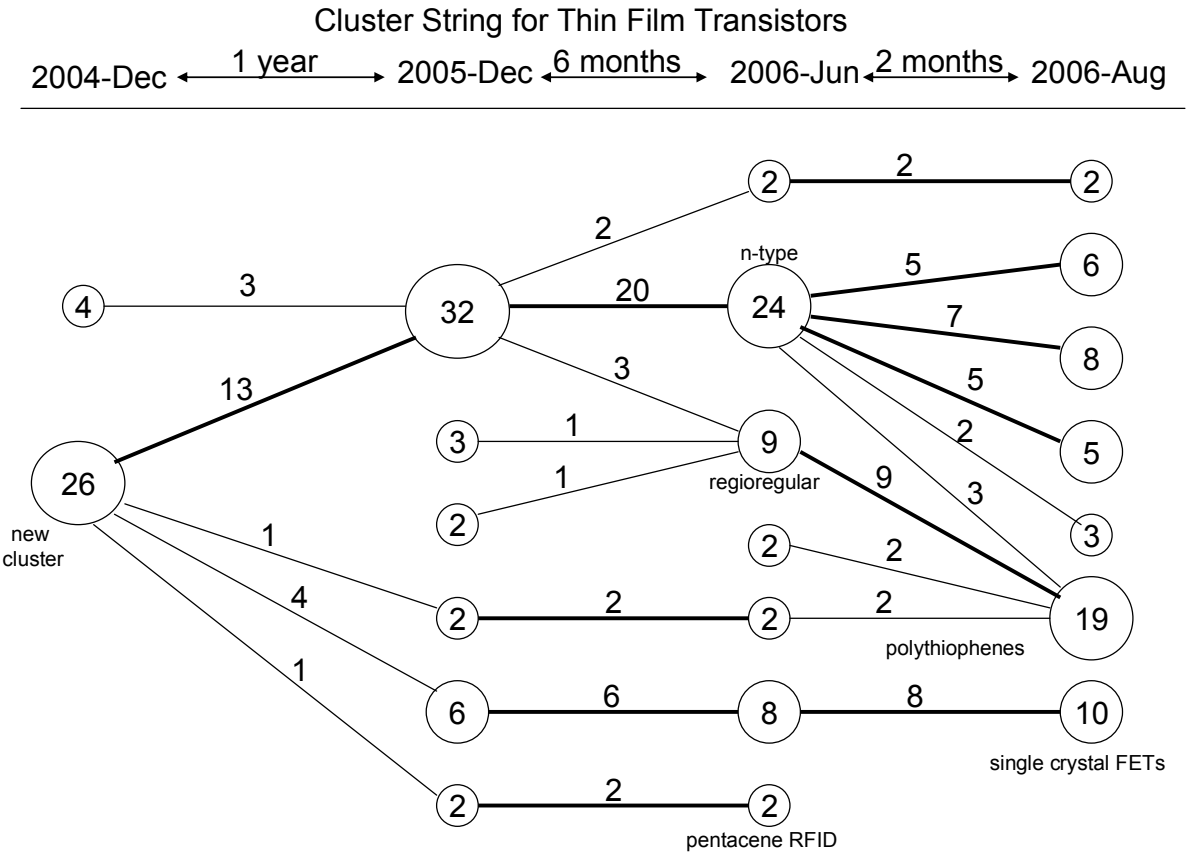


Figure 2. Cluster string representation of the evolution of the field of organic thin film transistors.

The respondent also pointed to the fact that large companies are entering the field and setting up expensive multidisciplinary research teams similar to those in academia, including materials scientists, physicists, and chemists. It was suggested that these new interdisciplinary teams may, in part, account for the upsurge in number of publications. However, the race for commercialization has also made companies and universities more secretive about their work, for example, making collegial interactions at conferences more difficult. In addition, often patents are applied for prior to the publication of results, which adds to the time lag for dissemination (Murray).

Thus, the emergence of this front in 2004 has to do with both the increase in activity sparked by new advances, as well as the increased investment of resources by academia and industry to overcome the remaining technical obstacles to commercialization. Ironically, if publication activity declines in the future, it might signal either the privatization of the field as patenting replaces publication, or alternatively the failure of the field to overcome the remaining technical barriers. Therefore, it is of interest to trace the subsequent history of this area up to the present to look for signs of commercialization, or intellectual decline.

Cluster evolution 2004 – 2006

A cluster string (Figure 2) is used to represent the evolution of the organic thin film transistors area (Small 2006). The string represents the complex splitting and merging pattern of cluster development from December 2004 through August of 2006. Each cluster is represented by a circle with the number of highly cited papers shown in the circle. For example, the initial cluster of 26 highly cited papers is shown on the left. The lines connect clusters in adjacent years that share continuing papers. The number of common papers is indicated on the line. Links are normalized using the cosine formula (number of common papers divided by the square root of the product of numbers of papers in the linked clusters). Only links with a cosine of 0.1 or higher are drawn. This breaks only one weak link that leads into organic chemistry. Note that summing the cluster sizes gives a growth in the specialty size of from 30 highly cited papers in 2004 to 53 in August 2006.

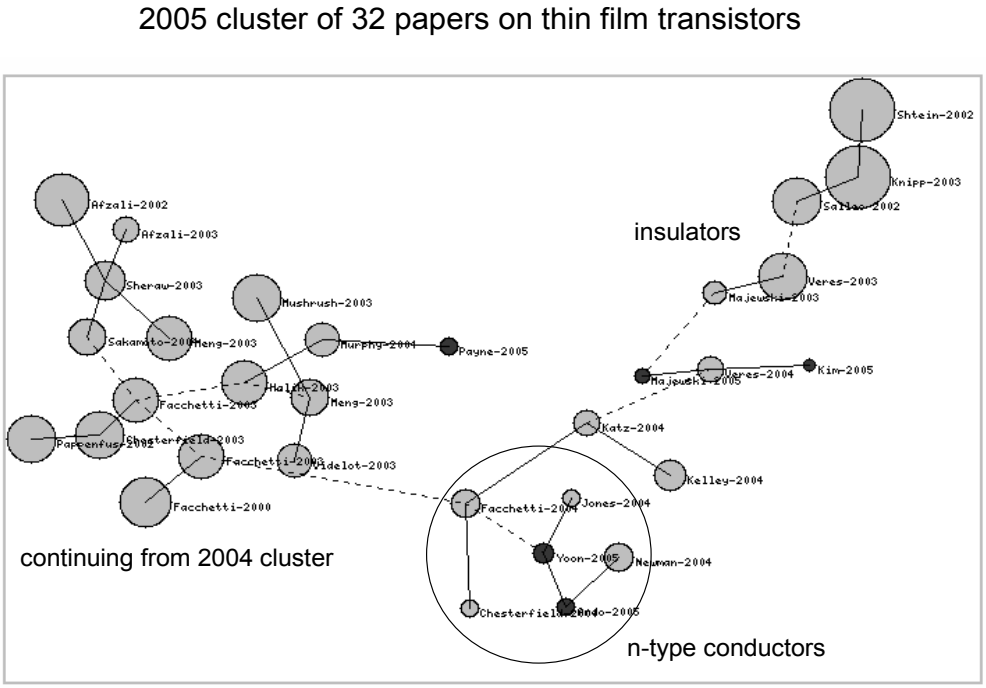


Figure 3. Map of the 2005 cluster on organic thin film transistors.

The initial cluster of 26 splits and leads to a large cluster of 32 papers in 2005. The map for the 2005 cluster (Figure 3) has three main regions. Almost all papers in the group on the left side of the map continue from the 2004 cluster. A new smaller group in the middle is on n-type conductivity in organic films, and an insulator and gate dielectrics region appears on the upper right. Three papers in this group at the upper right have merged from a small separate cluster in 2004 shown on the top left of the string (Figure 2).

The cluster in Figure 3 leads six months later to one in June 2006 consisting of 24 papers. This grouping (Figure 4) shows a much enlarged central area devoted to n-type devices, and a region on the lower left focused on thiophene containing remnants from the original 2004 cluster. In addition, there is a pentacene grouping at the upper right consisting of additional remnants, and a printing region at the upper left.

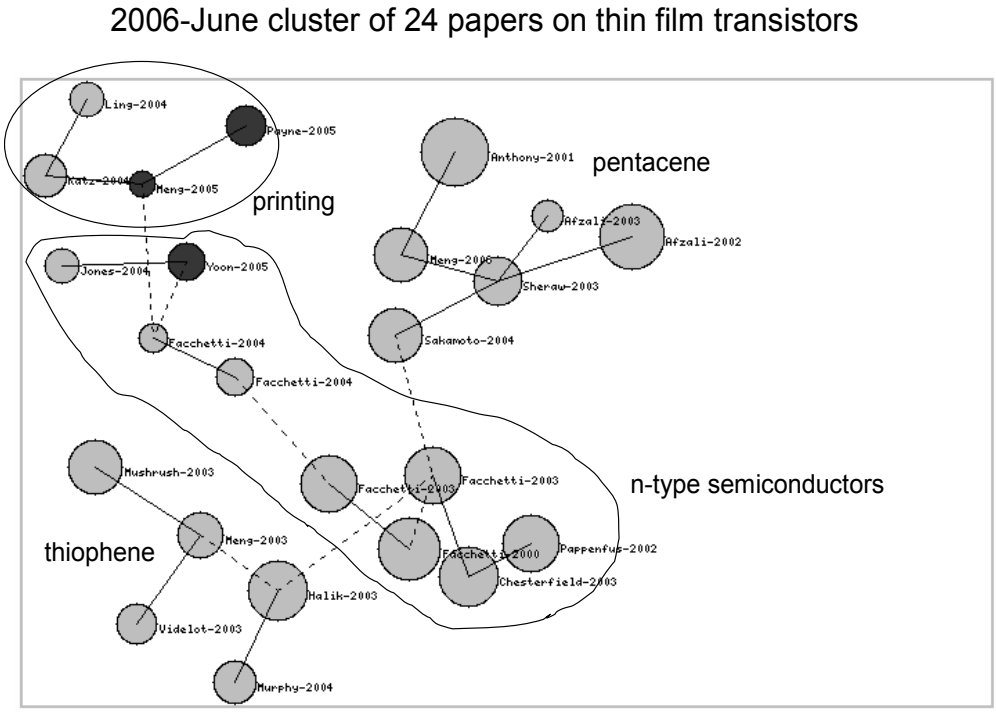


Figure 4. Map of the June 2006 cluster on organic thin film transistors

After only two months have elapsed, in August of 2006, this main cluster of 24 papers fragments (see Figure 2), one branch leading to what appears to be a new dominant area of 19 papers which has formed by the merging of four independent lines of research previously separate from the main path of development. Thus, in August 2006 there appears to be a reorganization and coalescence of research around a new approach, and a fragmentation and diminution of the previously dominant group. The largest of the resulting fragments from the break-up of the prior dominant group is a cluster of eight papers dealing almost exclusively with n-type semiconductors, with papers originating primarily from two research labs. Such short-term disruptions have been observed in prior co-citation studies (Sullivan et al; Small 1977).

The new dominant area of 19 papers (Figure 5) receives its main influx from a cluster of 9 papers in June 2006 focused on what is called regioregular organic structures and also on the role of molecular weight of the organic materials for improving charge mobility. All 9 papers from this prior group are incorporated into the cluster of 19 papers and comprise nearly all of the upper regions on the map, specifically, the groups labelled “regioregular thiophenes” and to its right the group labelled

"insulators and gate dielectrics". From the previous dominant group of 24 papers, the new group of 19 papers receives three papers concerned with printing methods that are circled on the lower right of Figure 5. Two smaller groups that merge into the cluster of 19 may hint at the growing sense that practical applications are not far off. One of these small groups deals with "device physics" and "new applications". The other group, circled at the lower left labelled "high performance thiophenes", suggests that there is a new emphasis on devices with high performance with progress accelerating.

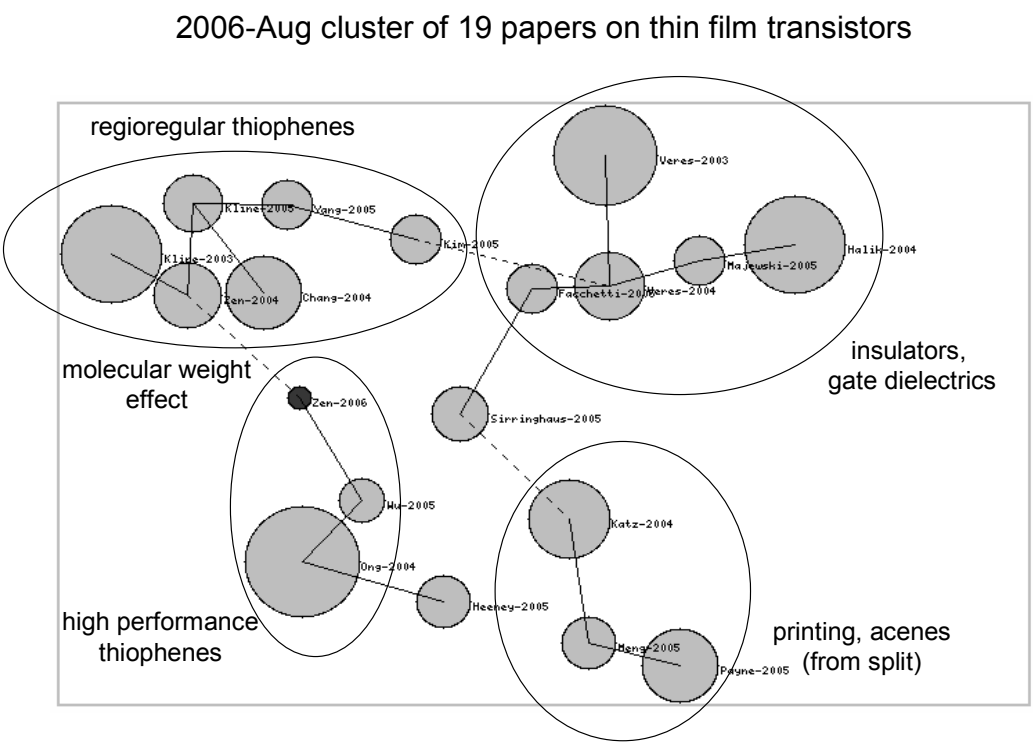


Figure 5. Map of the August 2006 cluster on organic thin film transistors

One way of thinking about cluster strings is to focus on the proportion of new or continuing papers with respect to some base year. Selecting Dec-2004 as the base, it is easy to discern that the regioregular region is the most dynamic and rapidly growing of the larger groups. This is further confirmed if we track the string to the end of 2006 (data not shown), when we find that 90% of the papers in that sub-grouping are new with respect to 2004. In addition, we find that four out of five small clusters (containing fewer than five papers) are 100% new with respect to 2004. Such new micro-areas show that the field is still highly innovative, and may be moving in new directions.

Conclusions

While it is too soon to say whether any of these papers represents a significant breakthrough, or a clear path toward applications, the recent reorganization of the research front and switch to a new dominant cluster may signal important progress. The overall pattern of development in the field of organic thin film transistors has been one of twigging or the spawning of new lines of research. This is to be expected in an area where knowledge is expanding rapidly. Only one area seems to have ceased, namely pentacene RFIDs.

The opposite of twigging, namely coalescence, is also evident at specific points in time and may be more critical for the development of the specialty. For example, there is the merging in June 2006 leading to the formation of the cluster of nine papers, and in August 2006 with the formation of the cluster of 19 papers. In these cases the coalescence or merging, which takes place within an overall pattern of twigging, signals the formation of a new dominant line of research. A number of specific concepts seem to be connected with this coalescence, including, "regioregularity" in the polymer

structure, the use of molecular weight to improve performance, and the applications of insulators and gate dielectrics.

Throughout both the splitting and merging processes of specialty evolution it is clear that certain structures of strongly co-cited documents are conserved, rearranged, and exchanged from one cluster to the next. These units of exchange may consist of only a few papers and they are highly topic specific, perhaps behaving as concept markers for the specialty. These conserved groups are reminiscent of the memes of Dawkins (1976), units of cultural transmission. Examples are paper sets on insulator/gate dielectrics, regioregular structure, n-type conducting, and printing. How this phenomenon relates to Chen's "pivot points" remains to be investigated (2006). We also see that new small clusters can merge with larger groupings later on and be incorporated into their structures, representing perhaps a delayed recognition of their relevance. Other patterns are more stable and represent relatively isolated groups that continue without much change from year to year, for example, the single crystal field effect transistor group which splits off from the original cluster and continues as a separate area. Whether in forecasting future developments it is best to track the new micro-clusters rather than the larger established groupings remains a topic for future research.

If we apply the rule that a cessation of publication would signal that the field has moved into an applications phase, where patenting and industrial secrecy would hold sway, we do not see as yet any evidence of this happening.

References

Chen, C. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology*, 57(3), 359-377.

Collins, G. P. (2004). Next stretch for plastic electronics. *Scientific American*, 291: 74-81.

Crane, D. (1972). Invisible Colleges Diffusion of Knowledge in Scientific Communities. Chicago, IL, University of Chicago Press.

Dawkins, R. (1976). The Selfish Gene. Oxford, UK : Oxford University Press.

Garfield, E., A. I. Pudovkin, et al. (2003). Why do we need algorithmic historiography? *Journal Of The American Society For Information Science And Technology* **54**(5): 400-412.

Morris, S.A., Yen, G., Wu, Z., & Asnake, B. (2003). Timeline visualization of research fronts. *Journal of the American Society for Information Science and Technology*, **55**(5): 413-422.

Murray, F. S. (2005). Do intellectual property rights hinder the free flow of scientific knowledge? NBER Academic Science and Entrepreneurship Conference.

Small, H. (1977). A co-citation model of a scientific specialty: a longitudinal study of collagen research. *Social Studies of Science,* **7**: 139 – 166.

Small, H. (1986). The synthesis of specialty narratives from co-citation clusters. *Journal of the American Society for Information Science,* **37** (3): 97 – 110.

Small, H. (2006). Tracking and predicting growth areas in science. *Scientometrics*, 68 (3): 595 – 610.

Sullivan, D., White, D.H., et al. (1977). Co-citation analyses of science: an evaluation. *Social Studies of Science* **7**(2): 223-240.

Upham, P., Small, H. (2007). Emerging research fronts in science and technology: Patterns of new knowledge development. To be published.

Issues in "Blogmetrics" - Case Studies Using BlogPulse to Observe Trends in Weblogs

Alastair Smith

AlastairSmith@vuw.ac.nz
School of Information Management, Victoria University of Wellington, PO Box 600, Wellington (New Zealand)

Abstract
Weblogs ("blogs") have emerged as a significant communication medium. The BlogPulse Trend Search tool was used for a number of case studies, examining trends in the use of terms in blogs. The paper discusses methodology issues in searching blogs, concluding that blogs can indicate new trends in popular culture, language, and science. However interpretation of trends requires careful examination of the postings, to ensure that an apparent trend is not due to problems in terminology, spam, trivial references to the concepts, etc. In particular, there seems to be no easy way to link postings to particular countries.

Keywords
weblogs; search engines; longitudinal studies; methodology; case studies.

Introduction

Weblogs ("blogs") have emerged as a distinct form of communication where non-technical individuals can easily comment on events and issues, and provide links to other blogs. Search engines have emerged that enable searches of blogs to be carried out, raising the possibility of bibliometric research into blogs - "blogmetrics".
This study uses the BlogPulse Trend Search tool (http://www.blogpulse.com) to show trends in the use of terms in blogs over time. This paper uses a case study approach to examine the methodology issues that arise in using commercial search engines to investigate blogs. Some questions that are addressed:
To what extent do terms used in searches reflect the content of the blogs found by the search?
To what extent are research oriented topics reflected in blog traffic?
Can national differences be tracked using blog searches?

Literature review

The growth of blogs has lead to a number of research papers on the topic. Reynes-Goldie (Raynes-Goldie, 2004) examined the blog server LiveJournal as a new way of managing information and creating knowledge. Dickie (Dickey, 2004) investigated the use of blogs in distance education to overcome the feelings of isolation experienced by student teachers. The use of blogs by politicians has been studied by Auty (Auty, 2005) and Jackson (Jackson, 2006). Clyde (Clyde, 2004) examined 55 blogs maintained by libraries, and discussed the use of blogs by libraries to communicate with their clients. Bar-Ilan (Bar-Ilan, 2005) monitored a set of blogs over two months, and demonstrated their use as "information hubs" summarising and linking to more substantial information. Herring et al (Herring, Scheidt, Wright, & Bonus, 2005) examined a random sample of blogs, and argued that blogs tended to be individual and intimate, rather than externally focussed. Adar and Adamic (Adar & Adamic, 2005) examined the propagation of information through the blogosphere, using an "epidemic" model. Cohen and Krishnamurthy (Cohen & Krishnamurthy, 2006) carried out a crawl of blogs, identifying the overall size of the blogosphere and its link structure. Delwiche (Delwiche, 2005) compared news stories covered by blogs with those covered by the conventional news media, finding that blogs had moved from technology to broader political issues.

The research indicates that blogs are a growing method of communication, and are worthy of bibliometric investigation.

Methodology

The BlogPulse search engine is described by Glance, Hurst and Tomokiyo (Glance, Hurst, & Tomokiyo, 2004). In the current study, searches on a number of terms that might be expected to show

trends in blog discussion were carried out on the BlogPulse Trend Search tool. These produced a graph of the frequency of hits over time. Up to three searches at a time can be compared. It is possible in the BlogPulse Trend Search tool to examine the specific blog postings that gave rise to the hits on a particular date. For this study, a sample of hits were examined to see if the postings were truly reflective of the content expected. Searches were carried out in October/November 2006. Graphs are presented courtesy of Nielsen Buzzmetrics, http://www.blogpulse.com who allow use for research purposes.

Results

This section will examine and interpret a number of case study searches. Some representative graphs are included.

Case study 1: blogging a trend

A simple example of the BlogPulse Trend Search tool is to see when a new trend is noticed in the Blogosphere. For example in mid-October 2006 the use of an actress as a search intermediary, Ms Dewey (http://www.msdewey.com/), elicited discussion on many blogs, with the peak shown in Figure 1.

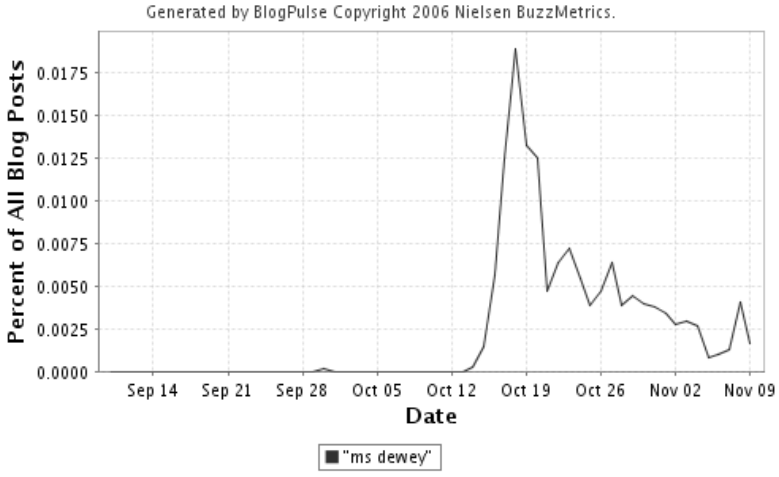


Figure 1. Ms Dewey.

Case study 2: sports terminology

"Football" means gridiron in North America, rugby in Australasia, and soccer in the rest of the world.

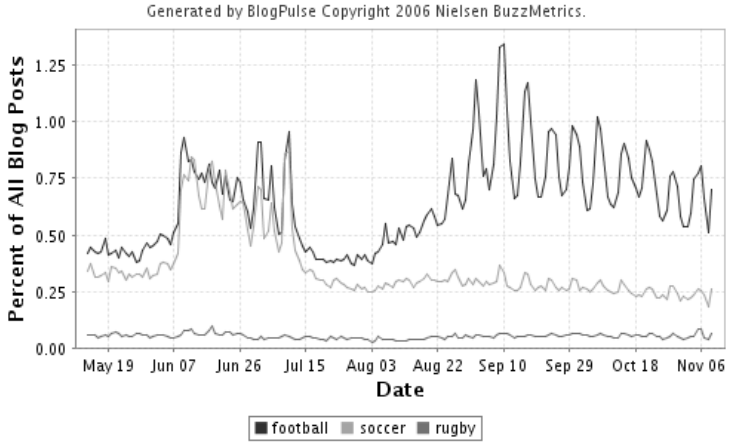


Figure 2. Football

This is illustrated in Figure 2. The June/July peak corresponds to the 2006 FIFA World, so the terms "football" and "soccer" are both used frequently. After this, "soccer" falls off, since it is a minority interest in the countries where the term is used. "Football" continues strongly, although European bloggers are discussing soccer, the North Americans gridiron. In fact many of the references are to football as a background to daily life: "my family sits at home watching football", so the occurrence of a term is not necessarily an indication that it is the subject of the blogging.

Case study 3: Searching for a country's postings
Bloggers tend to use blogging sites which are generally in the .com domain, so the URL gives no indication of the blogger's country. To test if a search on a country name identifies blogs from that country, a phrase search was carried out on three country names: New Zealand, Britain, and United States. A random sample of the first 10 entries were checked.

Table 1. relationship of blog posting to country

Country	Blog originating in country	Tourist account of country	Blogger is expatriate of country	Incidental mention of country	Country is topic
Britain	2	-	-	7	1
New Zealand	2	1	1	7	1
United States	9	-	-	-	1

For Britain and New Zealand, most postings were to incidental mentions for example "Things I'd like to do: Travel: where? Japan, Scotland, New Zealand". The differing United States result reflects the fact that most bloggers are based in the US. The use of a country name in a blog posting is not an indication that the posting is related to that country.
Case study 4: universities' profile in the blogosphere
The names of three NZ universities were searched in Blogpulse (Figure 3). VUW has more mentions than Auckland or Canterbury Universities, although Auckland and Canterbury are larger universities. However this is misleading: the 4 September spike has many postings related to the death of Steve Irwin, born in Victoria, Australia. Also, a number of "victoria university" postings relate to other institutions, for example in Victoria Canada.

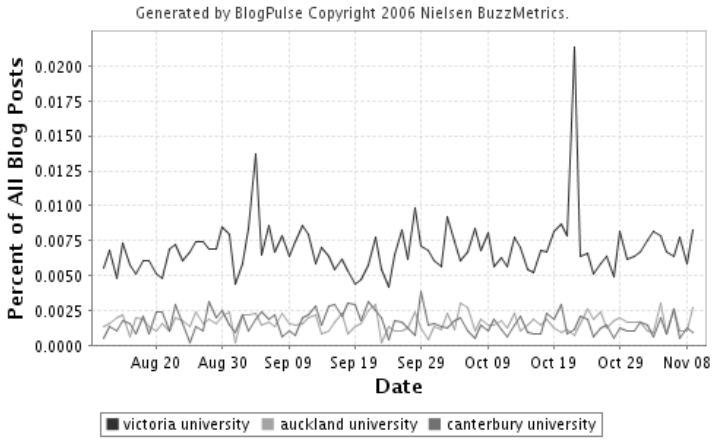


Figure 3. NZ Universities

Case study5 : triple helix in the blogosphere?
The triple helix of University, Government, and Industry has been discussed in the bibliometric literature (e.g. Leydesdorff & Curran, 2000). A search (Figure 4) appears to show a spike at 15 October in discussion of university/government and university/industry relations. However they are spam - spurious blogs set up with random phrases that on this day happen to include "government" and

"university". A few days later the spike had disappeared both from Blogger (the host) and from BlogPulse.

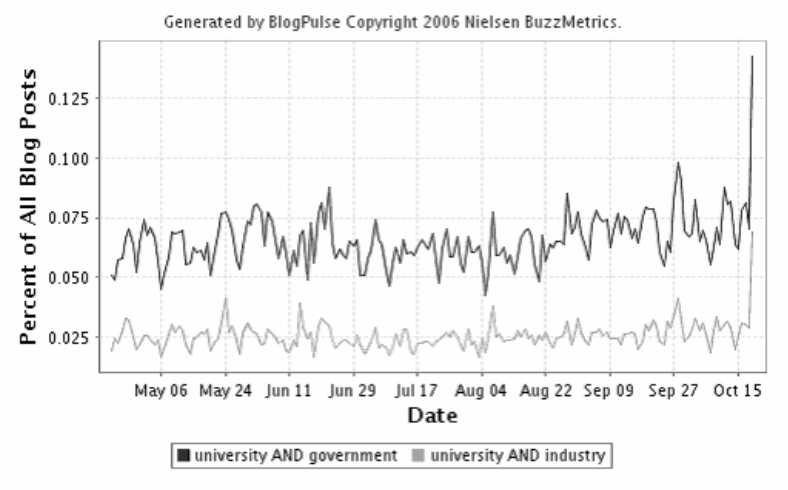


Figure 4. Triple helix

Case study 6: science in the blogosphere - stem cell research

Figure 5 shows very clear peaks related to news of embryonic stem cell research (banning of US federal funding, an actor's appeal for treatment, controversy in the US elections).

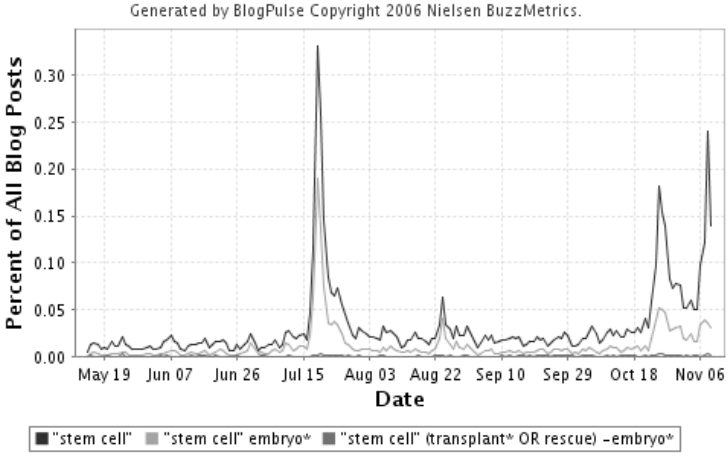


Figure 5. Stem Cell Research

However there is a non-controversial use of stem cells, from donors or from patients themselves, in transplant/rescue treatments of cancer. While the peaks relate to controversial embryonic stem cell use, a steady but small level of posting relates to, for example, patients blogging during their treatment.

Conclusion

The blogosphere may not have achieved the importance peer reviewed literature, but it is an area deserving of bibliometric research. Commercial blog search tools offer opportunities for this, but conclusions need to be drawn with care.

- Search tools can show trends in popular culture, particular if these are described by unique terms (e.g. Case Study 1)
- Interpretation requires awareness that words are being searched, rather than concepts, and national differences in language need to be taken into account (e.g. Case Study 2)
- Reference to a term (e.g. football) may be background to daily life rather than the subject of the blogging. This is a difference from searching for keywords in more focussed research material. (e.g. Case Study 2)
- There appears to be no easy way to track blogs relating to specific countries. (Case Study 3)

- References to institutions can be "polluted" by other uses of the words in the institution name (e.g. Case Study 4)
- It is important to sample blog postings to ensure that they are discussing topics being investigated. The prevalence of spam blogs in particular means that sudden recent spikes need to be treated with some scepticism. (e.g. Case Study 5).
- Scientific developments and controversies are reflected in blog discussions and this could be a useful way of evaluating the impact of science on society (e.g. Case Study 6)

BlogPulse is not designed for bibliometric purposes, so it is possible that dedicated search engines might reveal more about the blogosphere in future. Research blogs are appearing (e.g. Wikimetrics Research Blog, http://wm.sieheauch.de/) in which researchers offer commentary on their research area. In future a specialised search engine ("BlogScholar", perhaps, by analogy with GoogleScholar) would cover research blogs, providing early warning of research trends.

References

Adar, E., & Adamic, L. A. (2005, Sept. 19-22, 2005). Tracking Information Epidemics in Blogspace. Paper presented at the *Web Intelligence 2005 Conference, Compiegne, France*. Retrieved November 23, 2006 from http://www.hpl.hp.com/research/idl/papers/blogs2/

Auty, C. (2005). UK elected representatives and their weblogs: first impressions. *Aslib Proceedings*, 57(4), 338 - 355.

Bar-Ilan, J. (2005). Information hub blogs. *Journal of Information Science*, 31(4), 297-307.

Clyde, L. A. (2004). Library weblogs. *Library Management*, 25(4/5), 183 - 189.

Cohen, E., & Krishnamurthy, B. (2006). A short walk in the Blogistan. *Computer Networks*, 50(5), 615-630.

Delwiche, A. (2005). Agenda-setting, opinion leadership, and the world of Web logs. *First Monday*, 10(12). Retrieved November 23, 2006 from http://www.firstmonday.org/issues/issue10_12/delwiche/

Dickey, M. D. (2004). The impact of web-logs (blogs) on student perceptions of isolation and alienation in a web-based distance-learning environment. *Open Learning: The Journal of Open and Distance Learning*, 19(3), 279-291.

Glance, N. S., Hurst, M., & Tomokiyo, T. (2004). BlogPulse: Automated Trend Discovery for Weblogs. Paper presented at the *WWW2004, May 17-22, 2004, New York, NY* Retrieved November 23, 2006 from http://www.blogpulse.com/papers/www2004glance.pdf

Herring, S. C., Scheidt, L. A., Wright, E., & Bonus, S. (2005). Weblogs as a bridging genre. Information *Technology & People*, 18 (2), 142-171.

Jackson, N. (2006). Dipping their big toe into the blogosphere: The use of weblogs by the political parties in the 2005 general election. *Aslib Proceedings*, 58(3), 292-303.

Leydesdorff, L., & Curran, M. (2000). Mapping University-Industry-Government Relations on the Internet: The Construction of Indicators for a Knowledge-Based Economy. *Cybermetrics*, 4(1), paper 2. Retrieved November 23, 2006 from http://www.cindoc.csic.es/cybermetrics/articles/v4i1p2.html

Raynes-Goldie, K. (2004). Pulling sense out of today's informational chaos: LiveJournal as a site of knowledge creation and sharing. *First Monday*, 9(12). Retrieved November 23, 2006 from http://www.firstmonday.org/issues/issue9_12/raynes

University-Industry-Government Relationships Manifested Through MSN Reciprocal Links

David Stuart and Mike Thelwall

dp_stuart@hotmail.com, M.Thelwall@wlv.ac.uk
University of Wolverhampton, School of Computing and IT, Wulfruna St, Wolverhampton, WV1 1SB (UK)

Abstract
The web provides a free and timely source of information which is not limited to any single sector of society, thus providing a vital source for the investigation of inter-sector collaboration. This paper describes a pilot study investigating the potential of reciprocal links to identify collaborations between the university, industry, and government sectors. Using the Windows Live search-engine's recently introduced operator *linkfromdomain:* in conjunction with the established *linkdomain:* operator, it finds that a high proportion of such 'MSN reciprocal links' reflect a collaboration between the real world organisations they represent. Whilst more research is necessary it clearly shows the potential of reciprocal links for investigating organisational collaboration.

Keywords
webometrics; collaboration; reciprocal links.

Introduction

The potential of the World Wide Web to provide information about the relationships between different organisations has been the focus of a number webometric of investigations (e.g., Vasileiadou & van den Besselaar, 2006). The web is a free and timely source of information, and unlike the more traditional surrogates of bibliometric investigation, i.e. patents and scholarly articles, the web is not limited to any single sector of society but rather can be used for investigation of the relationships across the different sectors, a factor that is seen as increasingly important to innovation process (e.g. Lundvall, 1992; Etzkowitz & Leydesdorff, 2000; Gibbons et al., 1994).

Webometric investigations are restricted, however, by the data collection tools available. As data collection tools improve it is necessary to re-examine previous investigations to determine whether alternative methodologies are available. The new Windows Live (formerly MSN) search engine operator, *linkfromdomain:*, introduced in October, 2006 (MSDN, 2006), enables the manipulation of web data in a way that was previously only possible with a web crawler. It also allows the investigation of reciprocal links: where the first web site links to a second web site, and the second web site links back to the first web site. Previously such investigations were restricted to a small group of web sites that could be crawled by the investigator with a web crawler, or the links between specific web sites could be investigated with a search engine.

This study investigates what "MSN reciprocal links" (defined below) can reveal about the relationships between organisations by re-examining the case of the UK's West Midland's automobile industry, the focus of a recently published study (Stuart & Thelwall, 2006). The original study was carried out before the new operator was made available. It took a set of web sites involved in the automobile industry in the West Midlands and, using Google (via its web API - application programming interface), determined whether there were URL citations of any of the web sites on any of the other web sites within the set. This paper (a) changes the data type of the previous paper from URL citations to reciprocal links and (b) extends the scope of the connections to include sites outside of the core set investigated, using the following questions.

- What kind of university-industry-government collaborations, if any, are reflected by MSN reciprocal links?
- Do MSN reciprocal links identify more collaborative relationships than previous methodologies?

Research methodology

Data collection
As with the previous investigation (Stuart & Thelwall, 2006) the web sites used for this investigation are the 13 higher education institutes in the West Midlands region (West Midlands Higher Education Association, 2006), the 14 web sites identified by the automotive industry unit of the Department and Trade and Industry as having a significant role in the region (DTI, 2006), and the 38 West Midlands local councils identified on the government web site (DirectGov, 2006).

A program was written to determine those web sites that each of the 'seed' URLs had reciprocal links with. For each of the seed URLs a query was sent to the MSN API as many times as necessary to retrieve either: all the reciprocal web sites identified (if less than 1,000), or the MSN imposed maximum of 1,000 where more than 1,000 reciprocal websites were identified:

e.g. linkdomain:wlv.ac.uk linkfromdomain:wlv.ac.uk –site:wlv.ac.uk

Where more than one URL was identified for an organisation both URLs were used.

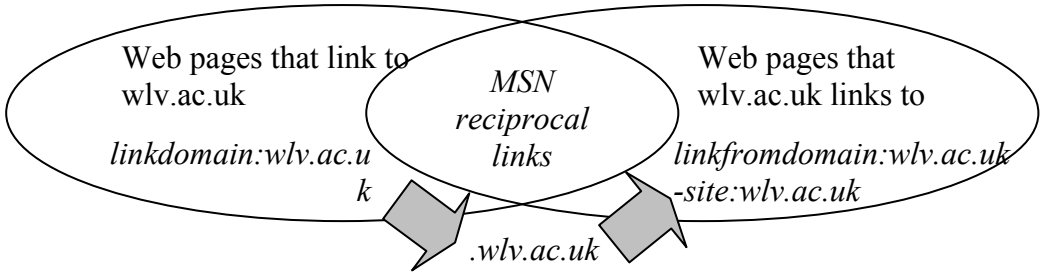


Figure 1. MSN reciprocal links

Since it is the first use of the *linkfromdomain:* operator it is necessary to briefly discuss the results that are provided when the operator is used on its own and in conjunction with the *linkdomain:* operator. The *linkfromdomain:* operator provides a list of pages that are linked to from the domain name entered. Whilst it has been reported that it doesn't provide a list of outlinks for sub-domains (SearchEngineWatch, 2006), further investigations indicated that whilst it does not currently allow a search to be restricted to sub-domains, those pages within the sub-domains are indexed and included in the search at the domain name level, e.g., *linkfromdomain:wlv.ac.uk* will return those pages linked to by pages in the sub-domain *linkanalysis.wlv.ac.*uk even though the query *linkfromdomain:linkanalysis.wlv.ac.uk* would produce no results.

Combining the *linkfromdomain:* operator with the *linkdomain:* operator for the same web site will produce a list of pages which are both linked to from the domain, and link to the domain. It should be noted however that MSN reciprocal links are not necessarily reciprocated; there is still a perspective that needs to be taken into consideration.

Figure 2 provides an example of non-reciprocated reciprocal links. Whilst the query "*linkdomain:websiteA.com linkfromdomain:websiteA.com"* would return one of web site B's pages amongst the hits, "*linkdomain:websiteB.com linkfromdomain:websiteB.com"* would not return either of web site A's pages, as there are no pages on web site A that are both linked to by web site B and link to web site B.

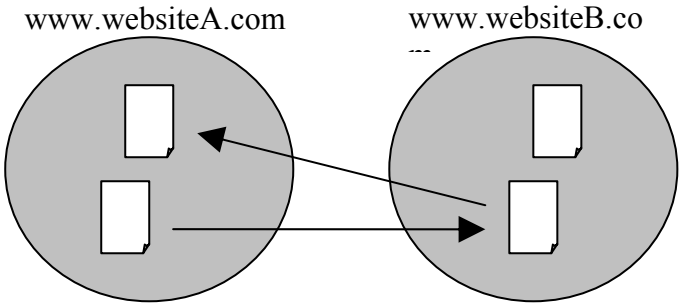


Figure 2 . Example of non-reciprocated reciprocal links

A program was also written to determine the number of different web sites that had MSN reciprocal links from the original seed web sites. For the purposes of this investigation web sites are operationalised on a lexical basis. The retrieved URLs were 'cleaned' by removing anything following the top-level domain name and then deleting duplicates. Due to the large number of web sites with shared web hosting arrangements reflected in their domain names, web sites with different sub-domain names were treated as different web sites.

Classification of relationships between the organisations
For each of the three sectors (i.e., academia, industry, and government) 50 MSN reciprocally linked web sites were investigated, where 50 reciprocally linked organisations were identified, irrespective of the reciprocally linked web site's domain. To prevent bias towards the larger organisations the 50 reciprocally linked organisations were taken from a subset based on up to 10 reciprocal links for each of the organisations in the study.

The classification scheme addresses the question of whether a collaborative relationship can be determined between the seed organisation and the organisation or individual whose web site it has a reciprocal link with.

Table 1. Classification scheme of relationship between organisations

Type of relationship
1. Collaborative relationship (formal or informal)
a. Subsidiary web site
b. Subsidiary web site – jointly owned with partners
c. Non-subsidiary web site
2. Non-collaborative relationship

Results
As can be seen from Table 2 a wide variation was found amongst the number of sites with which each of the seed web sites had reciprocal links. With 1,607 different web sites being identified as having MSN reciprocated links from the initial seed web sites.

Table 2. Number of reciprocally linked web sites

	Minimum	**Maximum**	**Mean**	**Median**
University	2	508	124	41
Industry	0	10	2	1
Government	0	67	11	8

Table 3. Shows that a high proportion of reciprocally linked organisations reflect collaborations between the organisations.

Table 3. Types of relationship between MSN reciprocally linked organisation

	Type of 'Seed' Organisation		
	University	**Industry**	**Government**
1. Collaborative Relationship			
a. subsidiary web site	13/50	23/30	18/50
b. subsidiary web site – with partners	2/50	1/30	2/50
c. non-subsidiary web site	22/50	4/30	24/50
2. Non-collaborative relationship	13/50	2/30	6/50
Percentage of collaborative relationships	74%	93%	88%

Discussion

Previous investigations into the ability of web links to provide indicators of collaboration between organisations suggested the low proportion of links reflecting collaboration would necessitate too much human intervention to make it a viable course of study (Stuart, Thelwall, & Harries, forthcoming). However, this pilot investigation of MSN reciprocal links, enabled through the recent introduction of the *linkfromdomain:* operator, suggests that further investigation is necessary before the notion of links as indicators of collaboration is cast aside.

The study found MSN reciprocal links to a large number of web sites outside the original seed web sites, even though MSN reciprocal links may be considered a fairly restrictive type of relationship in comparison to the existence of a single inlink or outlink. This coupled with the high proportion of the MSN reciprocal links that were found to reflect collaboration between organisations, far higher than in previous investigations of outlinks (Stuart, Thelwall, & Harries, forthcoming), suggests that there is a role for web links in the investigation of relationships between organisations.

Whilst web sites from all three sectors have MSN reciprocal links, there are clear differences in both the type of collaborative relationship that is reflected, and the percentage of MSN reciprocal links that reflect collaboration; most noticeably the high proportion of MSN reciprocal links in the industry sector that target subsidiary web sites. The lack of outlinks from the industry sector has been noted in previous studies (Shaw, 2001; Stuart & Thelwall, 2006), and it is found again in this study. It seems likely that the different linking practices of the different sectors needs to be reflected in different data collection methodologies. Whereas an analysis of outlinks for university and government web sites is likely to produce too many links to be of any use when investigating collaboration, it may be a more appropriate data collection technique for industrial organisations.

This study also highlights the large number of subsidiary web sites. Part of the reason for this is the large number of individuals that were using the facilities offered by blogging web sites, and other web publishing sites, for official organisational work. Such use is unsurprising: little skill is necessary in setting up such web sites, and it bypasses the official bureaucracy that may impede publication on the primary organisational web site. As well as having implications for the calculation of impact factors for web sites, they also cause dead-ends when investigating collaborative networks as they generally use a sub-domain of a blog site, and as has already been mentioned, the *linkfromdomain:* operator does not currently work for sub-domains.

It seems likely that the differences in the proportion of MSN reciprocal links that reflect collaboration is a sign of the size and scope of the web sites based in the different sectors. The idea that reciprocal links are more likely to connect collaborative organisations than is a single outlink, is based on the supposition that when sites are linking to one another it is less likely to be for non-collaborative reasons. The larger the web site and the more diverse the content held on it, the more likely the returning link is for a non-collaborative reason.

A potential weakness in using reciprocal links to find collaborations between different organisations is the large number of intra-organisational links which are showing up as reciprocal links between

different web sites; how can intra-organisational relationships be distinguished from inter-organisational relationships? Whilst it seems likely that satellite web sites have a different linking structure with their main organisational web site than with other organisational web sites, this is an area which needs further investigation.

Conclusion

The purpose of this pilot study was to establish the feasibility of utilising MSN reciprocal links to investigate collaboration between organisations. It has shown that whilst there are limitations in the use of search engines for webometric investigations, and specifically limitations with MSN's *linkfromdomain:* operator, a lot of potentially useful information is now available through MSN's search engine which can tell us a lot about the collaboration between different organisations. There is still a lot more research that is necessary before MSN reciprocal links can be used as indicators of collaboration between organisations:

- How much of the collaboration that is within the web link structure is being lost by focusing on reciprocal links rather than outlinks?
- How can intra-organisational collaboration be distinguished from inter-organisational collaboration?
- How great a hindrance is the operator's current inability to deal with sub-domains to mapping the collaboration network beyond the first generation of URLs created by the seed URLs?
- The feasibility of measuring the strengths of collaborations identified by reciprocal links.
- Alternative methods of data collection for certain commercial organisations.

Hopefully, however, as more search engines increase the accessibility of their databases, webometrics will be able to provide more indicators.

References

DirectGov (2006). Local councils in the West Midlands. Retrieved November 8, 2006 from:http://www.direct.gov.uk/Dl1/Directories/LocalCouncils/LocalGovernmentByRegion/fs/en?CONTENT_ID=4003650&chk=J1Wt8n.

DTI (2006). The automotive industry – West Midlands. Retrieved November 8, 2006 from: http://www.autoindustry.co.uk /regions/westmidlands/2?s=-1b1kskgcar80wj6d.

Etzkowitz, H. & Leydesdorff, L. (2000). The dynamics of innovation: From National Systems and "Mode 2" to a Triple Helix of university-industry-government relations. *Research Policy*, 29(2), 109-123.

Gibbons, M., Limoges C., Nowotny, H., Schwartzman, S., Scott, P. & Trow. M. (1994). *The new production of knowledge: The dynamics of science and research in contemporary societies.* London: Sage Publications.

Lundvall, B-A. (ED.) (1992). National Systems of Innovation: Towards a system of innovation and interactive learning. London :Pinter.

MSDN. (2006). Search Macros: LinkfromDomain. Retrieved November 20, 2006, from: http://blogs.msdn.com/livesearch/archive/2006/10/16/search-macros-linkfromdomain.aspx.

Shaw, D. (2001). Playing the links: interactivity and stickiness in .com and "not.com" websites. *First Monday,* 6(3). Retrieved November 8, 2006 from: http://firstmonday.org/issues/issue6_3/shaw/index.html.

Stuart, D. & Thelwall, M. (2006). Investigating triple helix relationships using URL citations: a case study of the UK West Midlands automobile industry. *Research Evaluation*, 15(2), 81-152.

Stuart, D., Thelwall, M. & Harries, G. (forthcoming). UK Academic web links and collaboration – an exploratory study. *Journal of Information Science.*

SearchEngineWatch (2006). Windows Live Search Adds linkfromdomain Command. Retrieved November 15, 2006 from: http://blog.searchenginewatch.com/blog/061017-094100.

Thelwall, M. (2002). Conceptualizing documentation on the Web: an evaluation of different heuristic-based models for counting links between university web sites, Journal of the American Society for Information Science and Technology, 53(12), 995-1005.

Thelwall, M. & Stuart, D. (2006). Web crawling ethics revisited: Cost, privacy, and denial of service. *Journal of the American Society for Information Science and Technology*, 57(13), 1771-1779.

Vasileiadou, E. & van den Besselaar, P. (2006). Linking shallow, linking deep. How scientific intermediaries use the Web for their network of collaborators. *Cybermetrics*, 10(1), paper 5. Retrieved September 3, 2006, from: www.cindoc.csic.es/cybermetrics/articles/v10i1p5.html.

West Midlands Higher Education Association. (2006). Retrieved November 8, 2006 from: http://www.wmea.ac.uk/universities.asp.

A Structural Analysis of Benchmarks on Different Bibliometrical Indicators for European Research Institutes Based on their Research Profile

Bart Thijs[*] and Wolfgang Glänzel[*,**]

[*]*Bart.Thijs@econ.kuleuven.be, Wolfgang.Glanzel@econ.kuleuven.be,*
Katholieke Universiteit Leuven, Steunpunt O&O Indicatoren, Dekenstraat 2, B-3000 Leuven (Belgium)

[**]*glanzw@iif.hu*
Hungarian Academy of Sciences, ISPR, Nádor u. 18, H-1051 Budapest (Hungary)

Abstract
The present study is part of an ongoing project on clustering European research institutions according to their publication profiles. Using stopping rules suggested by Duda and Hart (1973), eight clusters have been found the optimum solution for the classification. Aim of the present study is a structural analysis of a set of benchmarks for the evaluation of research performance of specialised and multidisciplinary institutions. A breakdown by subject fields is used to characterise field-specific peculiarities of individual clusters by bibliometric indicators and to allow comparison within the same and among different clusters. Finally, these benchmarks can then also be used to study national research performance on basis of the institutional classification.

Keywords
research institutes; benchmarking; classification

Introduction
Benchmarking studies often face the problem of incomparability of research institutes on bibliometric indicators (even after normalisation). Publication and citation indicators have proven to be strongly field dependent (see, for instance, Glänzel, 2001). To overcome these problems, we develop a method to cluster research institutes in order to obtain groups of 'likewise' institutes in terms of their publication profile. Benchmarking can thus take place within the same group. Using stopping rules suggested by Duda and Hart (1973) and supported by Milligan and Cooper (1985) we found 8 eight groups that can be characterised by their main focus 1. Biology, 2. Agricultural Sciences, 3. Multidisciplinary, 4. Geo & Space Sciences, 5. Technical and natural Sciences, 6. Chemistry, 7. General and Research Medicine, 8. Specialised Medicine. Discriminant analysis is then used to create a classification model.

In this paper we want to investigate whether this classification of research institutes into groups of likewise institutes active in the same areas really can create a better set of comparable institutes. Therefore we'll look at several commonly used publication and citation indicators, investigate whether there are substantive differences between groups and apply suitable statistical tests.

Data sources and data processing
Data were extracted from the yearly updates of the Web of Science database of Thomson-ISI (Philadelphia, PA, USA). Only papers of the document type article, letter and review indexed in the 2001 to 2005 volumes have been selected. This data has undergone a detailed cleaning and processed to bibliometric indicators. Publications were assigned to countries and main institutions according to the address in the by-line of the paper. A 3-step assignment procedure was developed. For each country under study, a list of distinct names of institutes as occurring in the extracted addresses is compiled. This list contains thus all possible synonyms and spelling variance/errors of research institutes. Secondly, each entry in this list with a number of publications above a certain threshold was assigned, if possible, to a unique, known institute. Finally, this thesaurus is matched with all addresses in a paper's by-line. For this study we selected 15 European countries (EU15 without Greece but including Switzerland). This resulted in a set of 2775 institutes. More than 80% of all addresses of these countries could be assigned to one of the institutes.

Subject classification of the publications was based on the field assignment of journals according to sixteen major fields of science developed in Leuven and Budapest (Glänzel & Schubert, 2003). These fields are Agriculture & Environment, Biology (Organismic & Supraorganismic Level), Biosciences (General, Cellular & Subcellular Biology, Genetics), Biomedical Research, Clinical and Experimental Medicine I (General & Internal Medicine), Clinical and Experimental Medicine II (Non-Internal Medicine Specialties), Neuroscience & Behaviour, Chemistry, Physics, Geosciences & Space Sciences, Engineering, Mathematics, Social Sciences I (General, Regional & Community Issues), Social Sciences II (Economical & Political Issues) and Arts & Humanities. Journals can also be classified as Multidisciplinary.

Citations received by these papers have been determined for three-year citation window beginning with the publication year on the basis of an item-by-item procedure using special identification-keys (so-called cluster-keys) made up of bibliographic data elements. Citation data up to 2005 has been extracted from the WoS.

All 2775 institutes were assigned to one out of eight different groups based on their research profile using a classification model developed Thijs & Glänzel (2007). Each group can be characterized by an average research profile. There are 7 specialised groups with a main focus on one or two of the 16 fields mentioned above. One group is characterized by its multidisciplinary profile. Table 1 lists the possible groups together with their share in the total set of institutes.

Table 1. Eight institutional clusters and their share in the total set of institutes

Group	**Code**	**Share**
Group 1 (Biology)	BIO	7.7%
Group 2 (Agriculture)	AGR	5.4%
Group 3 (Multidisciplinary)	MDS	26.1%
Group 4 (Geo- & Space Science)	GSS	3.2%
Group 5 (Technical & Natural)	TNS	14.1%
Group 6 (Chemistry)	CHE	6.1%
Group 7 (General & Internal Med.)	GRM	12.3%
Group 8 (Non-internal Med. Spec.)	SPM	25.1%

Methods and Results

Several commonly used bibliometric indicators are used to get more insight into the differences between the 8 groups.

- *Mean Observed Citation Rate* (MOCR) is defined as the ratio of citation count to publication count. MOCR is here based on 3-year citation windows.
- *Share of cited papers* (Cited%) is the share of papers receiving at east oner citation in a 3-year citation window beginning with the year of publication.
- *Share of author self-citations* (SCit%) is the percentage of author self-citations in all citations received in the same period.
- *Share of international co-authorship* (Coop%) is the percentage of internationally co-authored papers in all publications.

For each of these 2775 institutes all indicators have been calculated. Shares and mean values for each of the 8 groups are listed in Table 2. Indicator values in Table 2 are in line with our expectations. The citation impact and share of cited papers in the multidisciplinary and the life-science clusters is distinctly higher than that in the natural, applied and technical sciences. The question arises of whether the same field characteristics can be found in general and specialised clusters. The second question arises in the context of national peculiarities of research performance. Therefore, a breakdown by subject fields as well as by countries is used to compare indicators commonly used for benchmarking.

Publication activity and citation rates have discrete distributions which, however, belong to the domain of attraction of the normal distribution. Although none of these indicators are normally

distributed, neither have they equal variances over the different groups, based on the central limit theorem we can safely apply Gaussian statistics and standard ANOVA. Non-parametric tests are used as well. The results of these statistical tests are shown in Table 3. All presented statistics in this table are significant at 0.01 level. This means that these indicators have indeed different values over the distinct groups.

Table 2. Basic indicators by profile clusters

Indicator	Bio	AGR	MDS	GSS	TNS	CHE	GRM	SPM
MOCR	4.38	2.86	5.16	4.64	2.93	2.72	6.43	4.62
Cited%	79.7%	77.5%	77.4%	76.7%	68.5%	70.4%	81.4%	72.3%
SCit%	32.3%	34.2%	28.1%	36.9%	36.7%	33.8%	19.9%	19.0%
Coop%	43.8%	33.9%	39.2%	60.7%	44.3%	41.0%	32.3%	26.1%

Table 3. Statistical tests on basic indicators

Indicator	F (ANOVA)	X^2 (Kruskal-Wallis)	X^2 (Median Test)
MOCR	32.16	376.8	314.6
Cited%	26.82	205.9	189.1
SCit%	80.57	705.4	738.8
Coop%	40.76	350.7	367.8

In order to get more insight in the underlying differences we have broken down the profiles of the individual clusters by subject fields. In what follows, we present group 6 (institutes specialised in chemistry) as an example. 77% of the publication output of cluster CHE is devoted to chemistry. The share of this field is still significant in clusters MDS and TNS, less significant in BIO and AGR and almost marginal in the two medical groups and in GSS. Nonetheless we are interested in the performance of chemistry research in the other clusters too. Moreover, we would like to see a finer grained profile of chemistry research in the different institutional groups. Table 4 presents the citation-based indicators in this field organised by the eight clusters.

Table 4. Citation-based indicators in chemistry by clusters

Indicator	Bio	AGR	MDS	GSS	TNS	CHE	GRM	SPM
n	2160	850	97653	235	38980	5698	296	1108
MOCR	4.10	2.96	4.32	3.96	3.75	3.40	4.28	4.03
Cited%	84.0%	73.6%	80.1%	71.1%	75.2%	70.0%	84.5%	78.1%
SCit%	33.6%	33.4%	36.2%	34.3%	37.3%	31.7%	30.3%	30.1%

Chemistry papers published by multidisciplinary and by medical institutes have a higher citation impact than those by chemistry institutes themselves. The corresponding expected values (based on journal impact measures) reflect the same tendencies. Multidisciplinary and medical institutes publish on average in higher-impact journals than scientists at chemistry institutes. A closer look at subfield profile reveals important reasons for these deviations. The two medical clusters show a clear similarity, particularly, the predominance of subfields C1 and C3, while the cluster CHE (and similarly TNS) are characterised by a more applied profile. Only the multidisciplinary cluster is characterised by a more balanced subfield representation. The deviations of the citation indicators from the CHE standard, on one hand, and from the European total, on the other hand, can largely be explained by these profiles. The multidisciplinary cluster deserve its name at this level, too, since it profile largely coincides with that of the European total in the field. The two medical clusters do research especially in those disciplines which are relevant for their profile, namely, analytical, inorganic & nuclear chemistry and organic & medicinal chemistry

Table 5. Subfield profile of chemistry research in different clusters

Subfield	MDS	TNS	CHE	GRM	SPM	Total
C1	25.0%	18.3%	**12.5%**	40.2%	40.5%	21.8%
C2	10.4%	12.1%	**17.3%**	5.4%	8.0%	12.0%
C3	18.2%	6.1%	**9.2%**	42.6%	24.2%	14.4%
C4	22.5%	23.9%	**22.1%**	4.1%	7.6%	21.4%
C5	5.9%	8.0%	**13.1%**	3.0%	4.2%	6.6%
C6	23.0%	39.5%	**36.0%**	2.4%	10.0%	27.2%

(C1: Analytical, Inorganic & Nuclear Chemistry, C2: Applied Chemistry & Chemical Engineering, C3: Organic & Medicinal Chemistry, C4: Physical Chemistry, C5: Polymer Science, C6: Materials Science)

Comparing the percentages in the TNS and CHE cluster results in a X^2-value of 4.08 with 5 degrees of freedom. This means that the two distributions are not significantly different. There's also no difference between the distribution of the MDS group and the total set of chemistry publications (X^2=1.18). However, comparing other pairs results in significant X^2-values. The pair GRM and SPM has a X^2-value of 11.43 which is significant at 0.05 level.

Conclusion

The analysis of the bibliometric characteristics of the profile clusters of European universities reveal interesting details of institutional research profiles. The breakdown of clusters by subjects reflects the fine structures of their publication profiles. The analysis extended to all subject fields will provide the groundwork for sound intra- and inter-cluster comparisons of research performance at the meso level. The national breakdown and a dynamical look at the evolution of the fine-profile of the eight clusters will be tasks of future research.

References

Aksnes, D.W. (2003). A macro-study of self-citations. *Scientometrics*, 56 (2), 235–246.

Duda, R.O., Hart, P.E. (1973), *Pattern Classification and Scene Analysis*. New York: Wiley.

Glänzel, W., Schubert, A.(2003), A new classification scheme of science fields and subfields designed for scientometric evaluation purposes, *Scientometrics*, 56 (3), 357–367.

Milligan, G.W., Cooper, M.C. (1985), An examination of procedures for dermining the number of clusters in a data set. *Psychometrika*, 50(2), 159-179.

Snyder, H., Bonzi, S. (1998), Patterns of self-citation across disciplines. *Journal of Information Science*, 24 : 431.435.

Thijs, B., Glänzel, W. (2006), The influence of author self-citations on bibliometric meso-indicators. The case of European universities. *Scientometrics*, 66 (1), 71-80.

Thijs, B., Glänzel, W. (2007), A structural analysis of publication profiles for the classification of European research institutes. *Scientometrics*, to be published.

Research Cooperation within Europe: Bibliometric Views of Geographical Trends and Integration Processes [1]

Robert J.W. Tijssen and Thed. N. van Leeuwen

tijssen@cwts.leidenuniv.nl
Centre for Science and Technology Studies (CWTS) Leiden University (The Netherlands)

Abstract
This paper presents the first results of an on-going research project devoted to the geography of European science, more specifically to identifying and analyzing structural determinants of international scientific cooperation within Europe. The empirical data are extracted from co-authored research publications by scientists and scholars located in different European countries and regions. The preliminary results disclose underlying patterns and developments within international scientific cooperation during the years 2000-2005 at the level of three country blocks (i.e. EU-15 member states, New EU-25 member states, EU acceding countries and candidate countries), individual countries, and the level of NUTS2 regions (provinces). We find that intra-European cooperation has increased very significantly, largely following the general trends towards further internationalization of science. The European Union is also characterized by a complex and highly dynamic system of research cooperation between its various member states and candidate member states. Data at the regional level provide new insights into underlying patterns and trends, such as the pronounced role of the United Kingdom and Paris (France) as drivers of research partnering within the European landscape.

Keywords
scientific cooperation; Europe; NUTS Regions; co-authored research publications; european research area

Introduction

At the Lisbon European Council in 2000, the European Heads of State made an undertaking to develop a 'Europe of knowledge'. However Europe's universities, research organizations and centers of research excellence are scattered across the continent and all too often their efforts fail to create regional clusters in the absence of adequate networking and a critical mass of institutional cooperation. A related European Community initiative called the *European Research Area* (ERA) is dedicated to improving the coordination of research activities at both the regional, national and EU level. This paper deals with these three levels of the scientific integration within Europe, more specifically the countries within the European Union and those that wish to join. There is a great need to monitor this integration process by means of indicators that facilitate trend analyses and in-depth comparisons of science connectivity trends and 'knowledge spillover' effects, but systematic empirical data is scarce, especially at the regional level (e.g. Greunz, 2005).

Methodology and information sources

In line with the traditional bibliometric approach of measuring the strength of scientific collaborative links (e.g. Moed et al., 1992; Luukkonen et al., 1993; EC, 2003; Zitt & Bassecoulard, 2004; Luwel, 2006), the empirical data for this study were derived from co-authored research articles, more precisely those that were published in 2000-2005 in peer-reviewed science journals indexed by the *Web of Science* (WoS), an international bibliographical database produced by Thomson Scientific. The WoS, indexing some 9,000 journals, is considered to be the most comprehensive and reliable source of information on basic research activity across all countries and fields of science. It provides a good to excellence representation of internationally accepted high-quality 'mainstream' research, especially within the natural sciences, life sciences and medical sciences. The bibliographical records within this database usually include author affiliate addresses with the city names and/or postal codes. The European regions used in this analysis comprise NUTS 2 regions (provinces) in the case of larger

[1] We are very grateful to our co-workers Erik van Wijk and Suze van de Luijt for research assistance, and we would also like to thank Marge Fauvelle (EC, DG Research) and Torbioern Carlquist (EUROSTAT) for kindly supplying us information on the NUTS definitions of European regions.

European nations, accompanied by the smaller European nations (defined as NUTS1 or NUTS2 regions): Cyprus, Denmark, Estonia, Iceland, Latvia, Liechtenstein, Lithuania, Luxembourg, Macedonia (FYCOM), and Malta. However, most of the regional publication output data was collected at the NUTS3 level (city regions and towns) and subsequently aggregated to the NUTS2 level. Their publication output was gathered by matching elements from the author affiliate addresses in the WoS records to the official EUROSTAT definition of NUTS3 regions. These NUTS3 city names and postal codes were matched to the (cleaned and standardized) WoS-based city names and/or postal codes. In cases of a conclusive match, the corresponding NUTS 3 codes were linked to the research publication.[2] The publications were attributed in full to all pairs of countries or regions that could be identified in the author affiliate addresses.

Results

The first stage of the analyses dealt with the overall co-publication trends within Europe at the pan-national level comprising of three country blocks according to their current membership status within the European Union:[3]

(a) EU-15 member states (Austria, Belgium, Denmark, Finland, France, Germany, Greece, Ireland, Italy, Luxembourg, The Netherlands, Portugal, Spain, Sweden, and United Kingdom);

(b) New EU-25 member states (Cyprus, Czech Republic, Estonia, Hungary, Latvia, Lithuania, Malta, Poland, Slovakia, Slovenia);

(c) EU acceding countries and candidate countries (Bulgaria, Croatia, Macedonia/FYROM, Romania and Turkey).

Figure 1 depicts the growth rates of co-publications in these country blocks broken down into six categories of pair-wise relationships. The global trend in international co-publications ('International cooperation worldwide') was added for general reference. Using the co-publication output levels in 2000 as a baseline, we find an intra-EU15 growth rate of about 25% in 2005 which comparable to the worldwide trend. However, since Europe's publication output in the WoS increased by only 7% during those years, European research cooperation is clearly becoming an even more pronounced within ERA.

Remarkably, very little growth occurred between the EU15 member states and the 'EU25' member states. According to these data, the newcomers prefer to cooperate with partners in neighboring central European states, no doubt partly due to their common historical and cultural backgrounds and their post-WWII political heritage. In contrast, the collaboration linkages between the EU15 and the group of acceding/candidate countries, including some former communist states, follow Europe's average growth rate. These varying preferences and trends suggest a diversity of economic, scientific, cultural and geographical factors that are driving these research cooperation patterns which warrant further investigation. The remainder of this particular 'research in progress' paper however focuses exclusively on geographical factors.

Detailed analysis of co-publication data at level of individual countries throws more light on where significant growth occurred. Table 1 lists the pairs of European countries broken down by the top 10 'fast growers' and the top 10 of EU15's largest partnerships. None of the largest 'power house' pairs however exceed the 'intra EU15' average growth rate of 33% listed in Figure 1. Hence, the largest EU15 growth rates are generated among the cases with relatively low numbers of co-publications, a clear sign of catching up and convergence of research collaboration propensities is taking place. Nonetheless, it is quite revealing that even in the case of these 'power house' pairs, where large numbers of co-publications are generated often based on long-standing relationships, we still find

[2] Such a match proved impossible in a significant fraction of the cases because the author address information was incomplete or missing, or the NUTS3 definitions provided insufficient distinctive information. The success rate of matching author addresses to NUTS3 regions exceeded 98% of the publications for most countries.

[3] The EU has grown in size with successive waves of accessions. The European Union welcomed ten new countries in 2004: Cyprus, the Czech Republic, Estonia, Hungary, Latvia, Lithuania, Malta, Poland, Slovakia and Slovenia. Bulgaria and Romania expect to follow soon, and Croatia and Turkey have entered membership negotiations. Note that the 2000-2005 publication data refer to research done during the years 1998-2003, prior to the entry of these countries and a time-period in which intra-EU15 scientific cooperation was one of the dominating features of the European scientific landscape.

significant growth rates during the past few years. The statistics on the 'fast growers' indicate that the largest growth rates reflect significant upward trends from low baseline levels: the average annual volumes of co-publications in these cases is less than 20 per year. Some cases involve partnerships between new EU25 member states and the main contributors of the large growth rates between these countries as indicated in Figure 1.

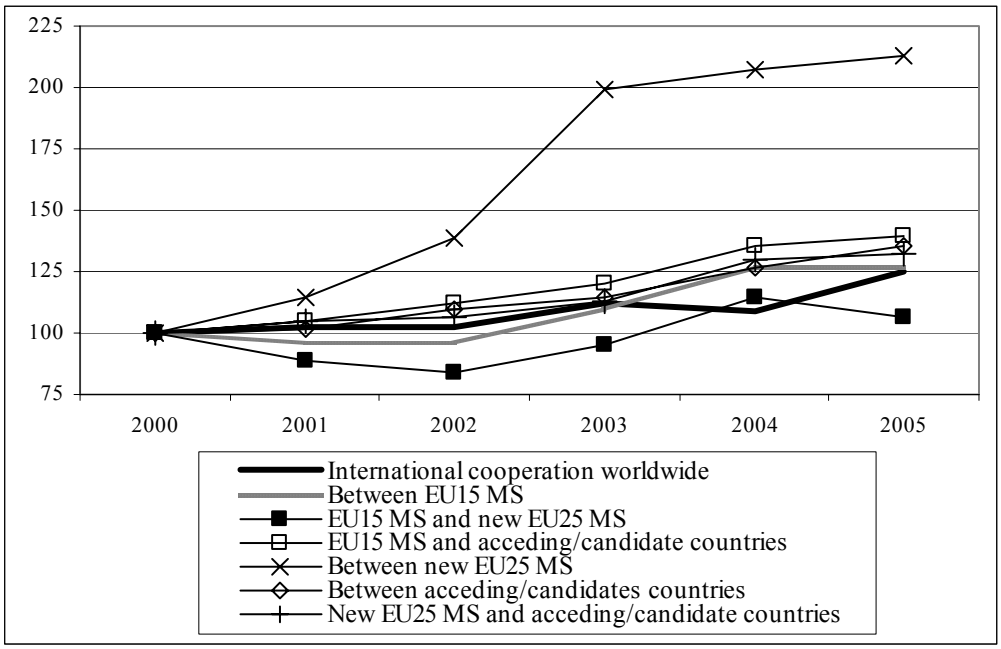


Figure 1. Growth rates of international scientific cooperation Trends in co-publication output frequencies (% growth; with frequencies in 2000 as a baseline)

Expanding the above findings to the regional level of Europe's provinces (NUTS2 regions), we address the following questions: which pairs of regions are the main contributors in size to international interregional collaboration? And which pairs exhibit the largest growth rates in recent years? Table 2 displays the top 10 largest pairs and the top 10 fastest growing pairs of transnational collaborating regions. The top 10 largest all relate to the Paris region in France, where major sections of French science are concentrated. Owing to the magnitude and regional distribution of French publication output the 'Paris effect' produces large numbers of co-publications with regions in other leading European countries. All these large-scale and intensifying collaborative linkages between regions relate to the EU15. In contrast, all fast growing partnerships involve pairs of EU15 regions and non-EU15 'catching up' regions. In most cases the co-publication output increased from levels as low as 1 or 2 cases in 2000 up to about 100 six years later. Interestingly, all but one of these cases includes a UK region. The UK therefore seems to be one of Europe's major actors in forging new scientific partnerships that succeed in producing results for research articles in (English language) science journals.

What's next?

The co-publication data reveal cooperation patterns and integration processes driven the regional concentration and distribution of science, language, and historical developments and traditional preferences. Obviously, we need to expand this analytical approach to unravel the mix of interrelated determinants that are driving scientific cooperation within Europe, notably the distribution of economic development and scientific performance across Europe's NUTS2 regions. Special attention will be devoted to examining their scientific strengths and weaknesses (both in terms of size, research performance, and scientific specialization) as determinants for joint research activities and partnering. This next stage will also include in-depth statistics on collaborative linkages at the NUTS3 level of cities and towns.

Table 1. Scientific cooperation within Europe: co-publication connectivity between nations

*Top 10 largest partnerships between member states in terms of co-publication output.
**Top 10 fastest growing partnerships in terms of growth in co-publication output (baseline: year 2000).Threshold value for countries: minimum of 60 co-publications in 2000-2005.

		Co-publication output	Growth (%)
Co-publication pairs of countries		**2000-2005**	**2000-2005**
Top 10 largest*			
France	*United Kingdom*	16 908	26
Great Britain	*Netherlands*	10 717	27
Germany	*Italy*	10 588	30
United Kingdom	*Italy*	8 957	25
Belgium	*France*	7 194	30
Belgium	*United Kingdom*	4 258	30
Denmark	*Germany*	3 742	17
Netherlands	*Sweden*	3 590	24
Denmark	*United Kingdom*	3 341	9
Denmark	*Sweden*	3 225	24
Top 10 fastest growing*			
Croatia	*Netherlands*	95	867
Bulgaria	*Ireland*	68	850
Greece	*Turkey*	84	725
Czech Republic	*Ireland*	79	600
Ireland	*Romania*	62	500
Croatia	*Spain*	81	440
Hungary	*Turkey*	113	380
Denmark	*Turkey*	93	340
Croatia	*Poland*	78	320
Bulgaria	*Turkey*	72	220

Table 2. Scientific cooperation within Europe: co-publication connectivity between provinces

*Top 10 largest partnerships between regions of different countries in terms of co-publication output.
**Top 10 fastest growing partnerships between regions of different countries in terms of growth rate in co-publication output (baseline: year 2000). Threshold values for regions: a minimum of 60 co-publications in 2000-2005.

		Co-publication output	Growth (%)
Co-publication pairs of NUTS2 regions		**2000-2005**	**2000-**
Top 10 largest*			
Île de France (FR)	*Inner London (UK)*	2 048	36
Île de France (FR)	*Oberbayern (GER)*	1 867	31
Île de France (FR)	*Lombardia (IT)*	1 649	38
Île de France (FR)	*Cataluña (ES)*	1 546	88
Île de France (FR)	*Berkshire,Buckinghamshire,*	1 505	6
Île de France (FR)	*Berlin (GER)*	1 377	52
Île de France (FR)	*Zuid-Holland (NL)*	1 329	87
Île de France (FR)	*Noord-Holland (NL)*	1 285	78
Île de France (FR)	*Berlin (GER)*	1 272	33
Île de France (FR)	*Karlsruhe (GER)*	1 268	33
Top 10 fastest growing*			
South Western Scotland (UK)	*Cantabria (ES)*	118	3700
Outer London (UK)	*Arnsberg (GER)*	102	3600
Dresden (GER)	*Campania (IT)*	113	3500
Greater Manchester (UK)	*Puglia (IT)*	98	3400
Greater Manchester (UK)	*Mecklenburg-Vorpommern (GER)*	102	3300
Greater Manchester (UK)	*Umbria (IT)*	82	3300
West Midlands (UK)	*Liguria (IT)*	93	3300
Surrey, East and West Sussex (UK)	*Umbria (IT)*	83	3300
Surrey, East and West Sussex (UK)	*Mecklenburg-Vorpommern (GER)*	94	3100
Surrey, East and West Sussex (UK)	*Liguria (IT)*	92	3100

References

EC (2003), *Third European Report on Science and Technology Indicators 2003: Towards a Knowledge-based Economy*, Brussels, European Commission, Report EUR 20025.

Greunz, L. (2005). Intra- and inter-regional knowledge spillovers: Evidence from European regions, *European Planning Studies*, 13, 449-473.

Luwel, M. (2006). Benchmarking the EU research area's integration, Presentation at the 9th International Conference on Science & Technology Indicators, 7-9 September 2006, Leuven, Belgium.

Moed, H.F., De Bruin, R.E., Nederhof, A.J. & Tijssen R.J.W. (1991). International scientific co-operation and awareness within the European Community: problems and perspectives, *Scientometrics*, 21, 291-311.

Luukkonen, T., Tijssen, R.J.W., Persson, O. & Sivertsen, G. (1993). The measurement of international scientific collaboration, *Scientometrics*, 28, 15-36.

Zitt M. & Bassecoulard E. (2004), Internationalisation in science in the prism of bibliometric indicators: Journals, collaboration and geographical distribution. In: Moed, H., Glänzel, W. & U. Schmoch (Eds.), *Handbook of Quantitative Science and Technology Research*, Dordrecht: Kluwer Academic Publishers.

Content Assisted Web Co-Link Analysis For Competitive Intelligence[1]

Liwen Vaughan* and Justin You**

*lvaughan@uwo.ca
Faculty of Information and Media Studies,University of Western Ontario, London, Ontario, N6A 5B7, (Canada)

**justin.you@apacbridge.com
ApacBridge Consulting, 8 Northgate Street, Ottawa, Ontario, K2G 6C7 (Canada)

Abstract
Building on a previous study that succeeded in mapping business competition positions at an industry level using Web co-link analysis, the current study attempted to improve Web co-link analysis by adding Web page content to obtain the mapping at a particular market segment level. This method combines the ideas of Web content mining (keyword) with Web structure mining (hyperlink). The method was tested in the WiMAX sector of the telecommunication industry. WiMAX is defined as Worldwide Interoperability for Microwave **Acc**ess by the WiMAX Forum, the exclusive organization dedicated to certifying the interoperability of broadband wireless access products (Wikipedia, 2006). Specifically, the keyword WiMAX was incorporated into queries that searched for co-links to pairs of company Websites. Two sets of data were collected: one with the proposed method and one with co-link search alone. The resulting two data matrices were analyzed using multidimensional scaling (MDS) to generate maps of business competition. The comparison between the two maps shows that the proposed method produced a map focusing on the WiMAX sector. The study also proposed the measure of *reduction of co-link count* that can be used to gauge the effectiveness of focusing the analysis on a particular sector. The *reduction of co-link count* could also be an easy and pragmatic measure for an analysis of a company's competitiveness in a particular market segment.

Keywords
competitive intelligence; web co-link analysis; web structure mining; web content mining

Background of the Study

Web hyperlink analysis has become an established area of Webometrics as the result of extensive research in recent years (Björneborn & Ingwersen, 2004; Thelwall, 2004). Link analysis has been used specifically for competitive intelligence (Reid, 2003; Thuraisingham, 2003; Wormell, 2001). An early study (Vaughan & You, 2005) successfully used co-link data to map competitive positions of a group companies in the telecommunication industry. If page X and Y are both linked to by page Z (i.e. page X and Y both have inlinks from page Z), then X and Y are co-inlinked. The concept of co-inlinks (called co-links later in the paper) is analogous to the concept of co-citation (Small, 1973) in bibliometrics. The current study attempts to improve the early result from mapping at an industry level to a sector level. The sector level mapping provides a more focused and detailed picture of the competitive landscape of a sector within an industry.

We propose a method that combines page content with co-link data to achieve this sector level mapping. Specifically, we choose a keyword that represents the sector and add this keyword to the co-link search queries. Co-linking pages (pages that initiated the co-links) that do not have this keyword in its page content (i.e. page texts) will be filtered out from the co-link search results. The effect of this filtering is to remove pages that co-linked the two companies for reasons other than the companies' activities in this sector. From a methodological point of view, the proposed method combines Web content mining with Web structure mining for competitive intelligence. This will become clearer if we review Web data mining as follows.

[1] This study is part of a larger project funded by the Initiative on the New Economy (INE) Research Grants program of the Social Sciences and Humanities Research Council of Canada (SSHRC). Research assistant Karl Fast helped with the programming work.

Web data mining can be classified into the following three sub-areas based on the type of data used (Madria, Bhowmick, Ng, & Lim, 1999): Web content mining, Web structure mining, and Web usage mining. Web content mining uses Web page content. Web structure mining tries to discover the model underlying the Web hyperlink structure. Web usage mining tries to discover patterns from Web usage data (Lu, Yao & Zhong, 2003). Mena (2001) suggests specific ways of Web usage mining for business purposes. Web page content is often represented by keywords on Web pages. This idea has been used in information retrieval and in other areas of information science such as tracing language evolution and the spread of ideas (Thelwall & Price, 2006). Keywords on Web pages are important objects of Web content mining for competitive intelligence (Madnick and Siegel, 2002). For example, Liu, Ma, & Yu (2003; 2001) compared patterns of keywords on the users' site with its competitor's site to discover unexpected information for business competition.

Methodology

The proposed method was tested in a group of companies in WiMAX sector of the telecommunication industry. WiMAX is defined as Worldwide **I**nteroperability for Microwave Access by the WiMAX Forum, the exclusive organization dedicated to certifying the interoperability of broadband wireless access products (Wikipedia, 2006). WiMAX is an emerging broadband wireless access technology. The WiMAX sector was chosen for two reasons. First, the acronym WiMAX, instead of the complete spelling of the term, is usually used in the industry and on Websites. This unique acronym is ideal to test the proposed method as incorporating this keyword into co-link data collection is likely filter out Websites that co-linked two companies for reasons other than WiMAX. Second, the variety of WiMAX companies makes them good candidates to test the method. Some large companies such as Ericsson and Samsung have broad product portfolios while other smaller companies specialize in WiMAX products or WiMAX chips. These two types of companies are likely to have different co-link patterns: the specialized companies will have co-links related to the WiMAX market segment while the companies with broad product lines will have co-links associated with a variety of market segments. We hypothesized that the proposed keyword method will distinguish the two.

The companies in the study were selected through the following process. First, all 43 companies included in a reputable market research and analysis report on WiMAX industry (Maravedis, 2006) were considered candidates of the study. These companies are major WiMAX producers worldwide. The Website of each company was then searched in major search engines such as Google and Yahoo! and their URLs recorded. We could not find the Website for one company, probably because the company no longer existed. Another company's Website showed that it was a service company rather than a product company so it was excluded from the study. Two other companies were acquired by larger companies so they were not considered as independent companies in the study. As a result, 39 companies remained in the study. See Appendix 1 for the list of the companies, their URLs, and the groups that they belong (e.g. WiMAX Infrastructure , WiMAX chip). This classification of companies into groups was done by that original report (Maravedis, 2006).

We searched for co-links to each pair of these companies using search engine MSN. Data collection result was a matrix of 39 by 39 symmetrical by the diagonal. Each cell of the matrix shows the number of Web pages that co-linked the pair of companies in question. Two sets of data were collected. One set was the co-link count alone and the other set added the keyword WiMAX in search queries (details below). The results from the two data sets could then be compared to determine if adding the keyword in co-link searches does achieve the goal of providing a competitive intelligence analysis at a sector level.

MSN was used for data collection as two other two major search engines in the market, Google and Yahoo!, could not serve the purpose of the study. Google can only search for total inlinks, i.e. it cannot filter out internal links in the search result. Google (2006) documentation on back link search query states "No other query terms can be specified when using this special query term". The "link" query needs to be combined with the "site" query in order to filter out internal links. Yahoo! was the search engine used in inlink data collection (e.g. Ortega et al, 2006; Vaughan & You, 2005) in recent years after it acquired AltaVista and AllTheWeb around the March of 2004. However, at the time of

data collection for this study (early Nov. 2006), we found that Yahoo! did not support co-link search although inlink search still seemed to work. A co-link query such as (link:www.uwo.ca –site:uwo.ca) AND (link:www.ubc.ca –site:ubc.ca) was interpreted by Yahoo! as a keyword search because the word "link" and the URL in the query were bolded in the search result screen as they would be in a keyword search. Many retrieved pages did not have the actual links that we were looking for but had the word "link" on them. So Yahoo! could not be used for this study.

The query syntax used to collect the two sets of data is shown in Table 1 in a hypothetical scenario of searching for co-links to www.abc.com and www.xyz.com. Note that MSN, like other major search engines, adds Boolean operator AND by default in between query terms so the AND operator is omitted. In other words, the query syntax for the data without the keyword is effectively "(link:www.abc.com –site:abc.com) AND (link:www.xyz.com –site:xyz.com)".

Table 1. MSN Query Syntax

Type of data collected	MSN query
Without the keyword	(link:www.abc.com –site:abc.com) (link:www.xyz.com –site:xyz.com)
With the keyword WiMAX	((link:www.abc.com –site:abc.com) (link:www.xyz.com –site:xyz.com)) WiMAX

The "link" command of MSN finds Web pages that link to a particular URL (in this study, links to a company homepage rather than all pages of the company Website). The "linkdomain" command of MSN will search for Web pages that link to all pages of a site. We decided to use the link command as an earlier study (Vaughan & You, 2005) showed that data collected using this command generated better mapping result than that using the "linkdomain" command. A content analysis study (Vaughan, Kipp, & Gao, submitted) that examined reasons of co-linking found that links to homepage were more likely to be business related than links to non-homepage. This confirms that "link" command is better than "linkdomain" command for business Websites.

The co-link data collected need to be normalized to obtain a relative measure of the number of co-links because a co-link count of 5 is very high if the number of links pointing to each Website is 6 while it will be low if the number of links pointing to each Website is 100. The normalization is done through Jaccard Index as follows:

$$NormalizedColinkCount = n(A \cap B) / n(A \cup B)$$

Where A is the set of Web pages which links to Website X

B is the set of Web pages which links to Website Y

$n(A \cap B)$ is the number of pages which link to both Website X and Website Y, i.e. the raw co-link count

$n(A \cup B)$ is the number of pages which link to either Website X or Website Y.

Multidimensional Scaling (MDS) is then applied to the normalized co-link matrices using SPSS. The MDS output is a map that positioned each company according to their similarity as measured by the number of co-links (The more co-links a pair of companies have, the more likely that they are similar or related companies). The higher the co-link count, the closer the two companies will be placed in the MDS map. Essentially the map will cluster competing companies together as competing companies are similar or related companies.

Findings

Mapping Result Comparison

Figure 1 is the MDS mapping result for the data set that were collected using co-link queries alone (i.e. the first search query in Table 1) while Figure 2 is the result for the data set that added the keyword

WiMAX into the co-link search queries (i.e. the second search query in Table 1). The stress values of both MDS analysis are below 0.05 which show that there is a very good fit between the input data and the output maps (Meulman & Heiser, 2001, 201).

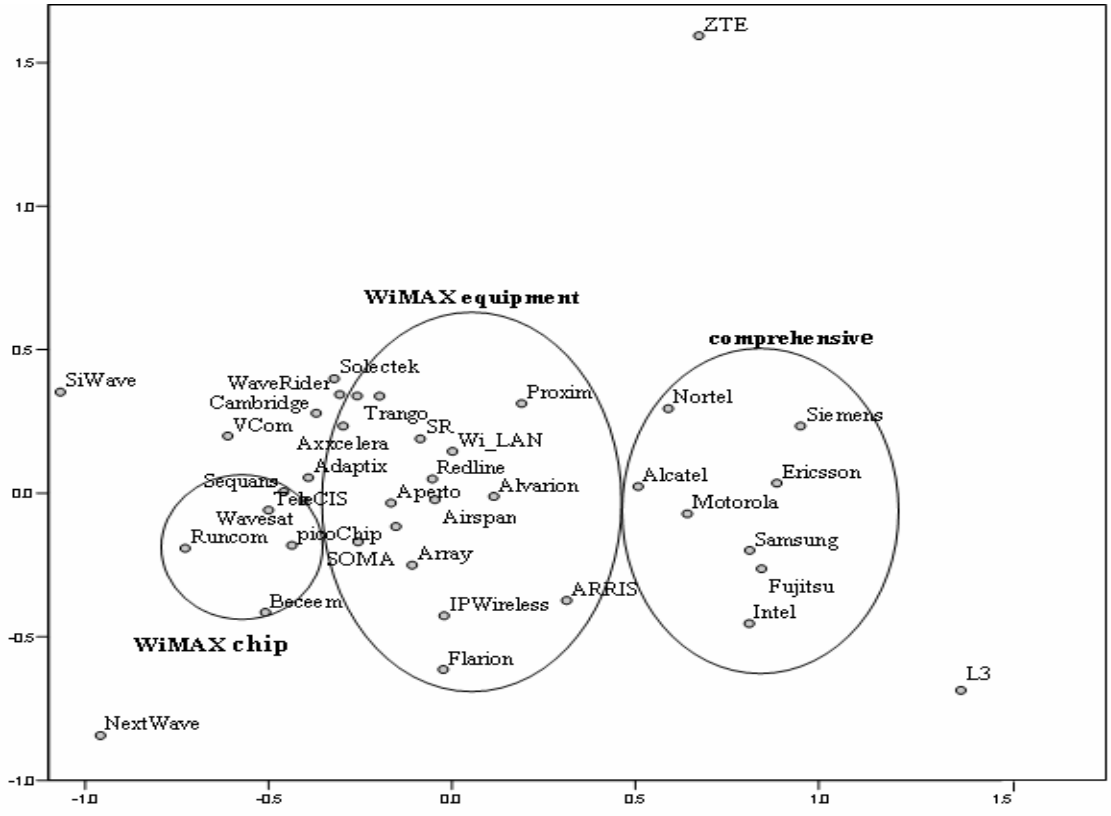


Figure 1. MDS map without the keyword

In Figure 1, company positions reflect their overall strength in the telecommunication marketplace, which includes WiMAX market and non WiMAX markets. There are three major groups of companies in this map:

1. Large tier 1 companies are grouped in the "comprehensive" category. These are big telecommunication equipment companies such as Ericsson, Alcatel and Nortel which have a variety of telecommunication products including optical, switches, routers and wireless products. The only exception in this group is semi-conductor company Intel. Intel is associated with these tier 1 companies duo to the fact that Intel provides telecommunication chip solutions for these tier 1 players. The competition among these big companies is more intensive than their competition with other small companies because they compete for major market shares rather than niche markets. ZTE, a major Chinese telecommunication equipment company, does not have many co-links with the rest of tier 1 players and is thus shown as an outlier in Figure 1. This reflects the fact that ZTE focuses on the Chinese market and has little competition with other tier 1 players in the major global market.
2. "WiMAX chip" group. This group mainly includes specialized WiMAX chip companies such as Beceem, picoChip, and Runcom. The Web co-link map clearly indicates the nature of their competitive relationship because they are clustered together inside the circle while other WiMAX chip companies such as SiWave and Next Wave are placed on the outskirt of Figure 1.
3. "WiMAX Equipment" companies. This is a group of specialized WiMAX equipment companies including Alvarion, Airspan, and Proxim. These companies are much smaller than tier 1 telecommunication companies and focus on WiMAX equipment niche market as their main business.

Figure 1 has proved that co-link analysis is able to cluster companies in terms of their overall competitive positions in the telecommunication industry. The result matches well with the classification of these companies listed in Appendix 1.

Figure 2 is the result for the data set that added the keyword WiMAX into the co-link search queries (i.e. the second search query in Table 1). Companies are clustered into five groups and show a sector level competition landscape:

1. "WiMAX 1" equipment companies. This group includes the market leaders of WiMAX equipment companies such as Alvarion, Proxim, Motorola, and Redline. These companies are grouped together because they are the leaders in the WiMAX market (Alvarion, 2006). WiMAX infrastructure companies such as Alcatel, Siemens and chip company Fujitsu have closer tie with this group than with "WiMAX infrastructure" group. This is a change compared with their positions in Figure 1 as the result of introducing the keyword.
2. "WiMAX 2" equipment companies. This group also includes major WiMax players like Flarion, Arraycomm, IP Wireless and SOMA. These companies compete with each other more than they do with the companies in the first group (Alvarion, 2006)
3. "WiMAX 3" group. This group mainly includes smaller WiMAX companies. They have much less market shares than the previous two groups. These companies can be perceived as second tier WiMAX companies.
4. "WiMAX chip" group. These companies mainly focus on WiMAX chip business and compete with each other in chip business.
5. "WiMAX infrastructure" players which provide infrastructure equipment for WiMAX networks. This group contains Ericsson, Nortel and Samsung and the largest WiMAX chip vendor Intel. However, their co-link relationship or competitive relationship with the companies in the "WiMAX 1" group is not very strong. This suggests that they are not key competitors for the market leaders of WiMAX 1 companies yet.

In summary, the results from our proposed method of keyword assisted co-link analysis match very well with results obtained by separate market research (Alvarion, 2006) and the company classification shown in Appendix 1. This has demonstrated a practical application of using content assisted co-link Web mining analysis for competitive analysis and business intelligence research.

Co-link Count Comparison

The above comparison of the mapping results provided a visual contrast of the two data sets. We then calculated the *reduction of co-link count* to gain a numerical contrast of the comparison.

Reduction of co-link count = (A-B)/A

Where A =co-link count without keyword

B =co-link count with keyword

This calculation was applied to each cell of the co-link matrix. The overall pattern is that tier 1 comprehensive companies have much larger *reduction of co-link count* than specialized WiMAX companies. For example, the co-link counts among these comprehensive companies (Alcatel, Nortel, Siemens, Ericsson, Motorola, and Siemens) dropped between 91% to 96% due to their less strong positions in WiMAX market than their positions in the overall telecommunication market. On the other hand, the drops of the co-link counts among specialized WiMAX companies (in groups 1, 2 and 3 of Figure 2) were much smaller, ranging from 9% to 40%. Link count changes also clearly demonstrated the effectiveness of using the keyword for a sector level analysis. Without the keyword, the top 10 companies of inlink counts are Intel and all the tier 1 comprehensive companies followed by Alvarion and Proxim at 9^{th} and 10^{th}. With the keyword WiMAX, Alvarion moves to the 1^{st} position, Proxim, Redline, Arraycomm, and Wi-LAN move to 7^{th} – 10^{th}. Nortel, Samsung and Fujitsu dropped out from the top 10 list.

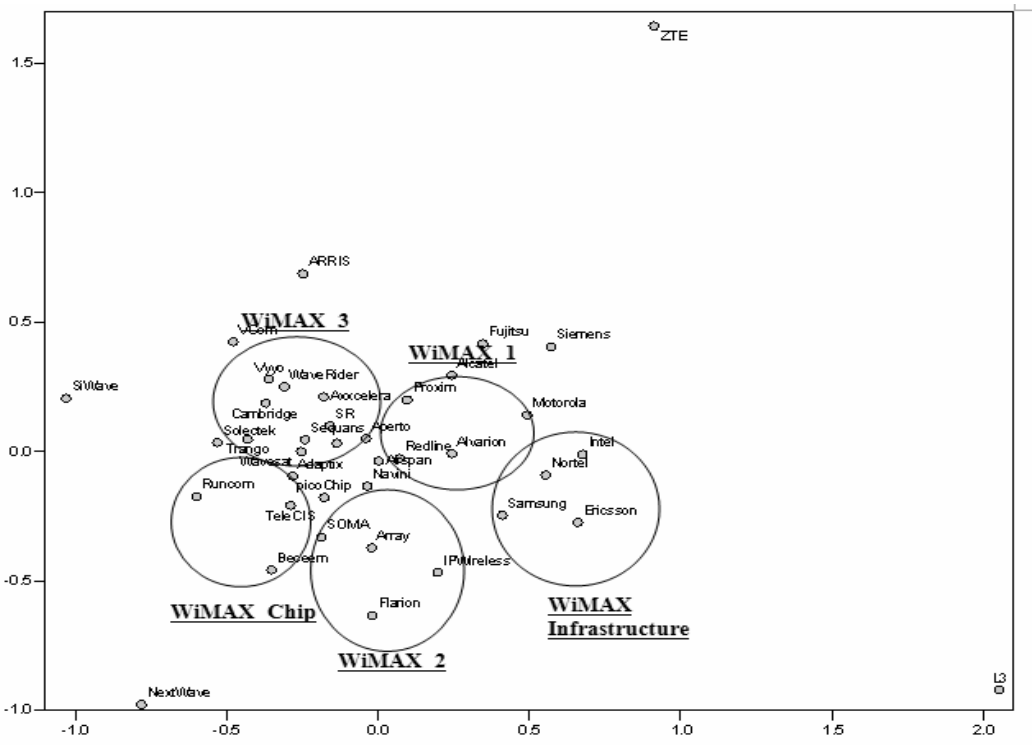


Figure 2. MDS map with Keyword WiMAX

This suggests that the proposed method is effective in focusing the co-link counts and inlink counts in the target sector because a larger proportion of the links that the broad, comprehensive companies receive are for their products other than WiMAX equipments. Once the keyword WiMAX was added into the co-link queries, those linking pages were filtered out.

Conclusions and Discussion

Based on a previous study that succeeded in mapping business competition positions at an industry level using Web co-link analysis, the current study attempted to augment or boost Web co-link analysis by adding Web page content to obtain the mapping at a sector level. This method combines the ideas of Web content mining (keyword) with Web structure mining (hyperlink). The method was tested in the WiMAX sector of the telecommunication industry. The keyword WiMAX was added into queries that searched for co-links to pairs of company Websites. The two sets of data (one without the keyword and one with the keyword) were analysed with MDS. The comparison between the two MDS maps shows that the proposed method produced a map more focused on the WiMAX sector. However, the contrast between the two maps is not as strong as what was achieved in an earlier study (Vaughan & You, 2006) that applied essentially the same keyword method to the DSLAM sector. The *reduction of co-link count* measure was not used in that study. We are investigating the differences between the two sectors to gain a better understanding of the underlying mechanism.

The study also proposed the *reduction of co-link count* measure and tested it in the data collected. The comparison between bigger comprehensive companies and smaller specialized WiMAX companies in the *reduction of co-link count* showed that adding the keyword into co-link queried is useful in focusing the analysis into the target sector. The analysis showed that *reduction of co-link count* could be a measure a company's relative strength in a particular sector. It allows the link and co-link analysis to be zoomed into a particular sector. It could be a very easy and pragmatic measure for competitive intelligence analysis and for daily business practice.

As an exploratory study, the project showed that the method is promising. However, more research is needed to improve and refine the method. As telecommunications is perhaps one of the industries that make the most extensive use of the Web, it is not known how well the method will work in other

industries. More testing in other sectors or industries is needed to determine the general applicability of the method. The method proposed is novel but relatively primitive in that only one keyword was used in co-link search queries. This simple keyword method cannot be applied directly to other sectors that do not have an unambiguous terms such as WiMAX. Future studies will try to find out how to make more sophisticated use of keywords so that the method will be more generally applicable. The current study showed quantitatively that the keyword method is effective. We intend to carry out a content analysis that compares two sets of linking pages (one retrieved without the keyword in co-link queries and one with the keyword) to examine qualitatively if the keyword method does filter out pages that co-linked companies for their activities outside the target sector.

References

Alvarion (2006). Corporate fact sheet. Retrieved Nov. 27, 2006 from http://www.alvarion.com/upload/images/mediakit/factsheet/Corporate_FS_revD_LR.pdf.

Björneborn, L. & Ingwersen, P. (2004). Toward a basic framework for webometrics. *Journal of the American Society for Information Science and Technology*, 55(14), 1216-1227.

Google (2006). Google SOAP Search API Reference. Retrieved Aug. 18, 2006 from http://www.google.com/apis/reference.html#2_2.

Liu, B, Ma, Y., & Yu, P. S. (2003). Discovering business intelligence information by comparing company Web sites. In Zhong, N., Liu, J., & Yao, Y. (Eds) *Web Intelligence*. Berlin: Springer, 105-127.

Liu, B., Ma, Y., & Yu, P. S. (2001). Discovering Unexpected Information from Your Competitors' Web Sites, In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug. 26-29, 2001, San Francisco, U.S.A.

Lu, Z., Yao, Y. & Zhong, N. (2003). Web Log Mining. In Zhong, N., Liu, J., & Yao, Y. (Eds) *Web Intelligence*. Berlin: Springer, 173-194.

Madnick, S. and Siegel, M. (2002). Seizing the Opportunity: Exploiting Web Aggregation, *MIS Quarterly Executive*, 1(1), 35-46.

Madria, S., Bhowmick, S.S., Ng, W.K. & Lim, E.P. (1999). Research Issues in Web Data Mining. In *Proceedings of Data Warehousing and Knowledge Discovery, First International Conference*, Florence, Italy, Aug. 30 – Sept. 1, 1999. Lecture Notes in Computer Science, Vol. 1676, 303-312

Maravedis (2006). WiMAX *and Broadband Wireless (Sub-11Ghz) Worldwide Market Analysis & Trends 2006-2012*. Retrieved Nov. 16, 2006 from http://www.wimax-industry.com/mk/mrv/maravedis/mkmv1b.htm.

Mena, J. (2001). WebMining for Profit: E-Business Optimization. Boston: Digital press.

Meulman, J. & Heiser, W. (2001). *SPSS Categories® 11.0*. Chicago, USA: SPSS Inc.

Reid, E. (2003). Using Web link analysis to detect and analyze hidden Web communities. In: D. Vriens, (ed). *Information and Communications Technology for Competitive Intelligence* (pp. 57-84). Hilliard, Ohio: Ideal Group Inc.

Ortega J. L, Aguillo, I., Cothey, V., & Scharnhorst, A. (2006). Maps of the academic web in the European Higher Education Area - an exploration of visual web indicators. In *Book of Abstracts, 9th International Conference on Science & Technology Indicators*, pp. 107-110, Leuven, Belgium, Sept. 7-9, 2006.

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, July-August, 1973, 265-269.

Thelwall, M. & Price, E. (2006). Language evolution and the spread of ideas: A procedure for identifying emergent hybrid word family members. *Journal of the American Society for Information Science and Technology*, 57(10), 1326-1337.

Thelwall, M. (2004). *Link analysis: An information science approach*. San Diego: Academic Press.

Thuraisingham, B. (2003). *Web Data Mining and Applications in Business Intelligence and Counter-Terrorism*. Boca Raton, Florida: CRC Press.

Vaughan, L. & You, J. (2006). Keyword enhanced Web structure mining for business intelligence. In *Proceedings of the 2006 IEEE/ACM International Conference on Signal-Image Technology and Internet-Based Systems*, to be held in Hammamet, Tunisia, Dec. 17–21, 2006.

Vaughan, L. & You, J. (2005). Mapping business competitive positions using Web co-link analysis. In *Proceedings of ISSI 2005 – the 10th International Conference of the International Society for Scientometrics and Informetrics,* (P. Ingwersen & B. Larsen eds.), pp. 534–543, Stockholm, Sweden, July 24–28, 2005.

Vaughan, L., Kipp, M.E.I., Gao, Y (submitted). Are co-linked business Websites really related? A qualitative study. Paper under review in *Online Information Review*.

Wikipedia (2006). WiMAX. retrieved Nov. 16, 2006 from http://en.wikipedia.org/wiki/WiMAX.

Wormell, I.(2001). Informetrics and Webometrics for measuring impact, visibility, and connectivity in science, politics, and business. *Competitive Intelligence Review*, 12(1), 12-23.

Appendix 1 – Companies in the Study

Company Name	URL
WiMAX equipment	
Adaptix	http://www.adaptix.com
Airspan	http://www.airspan.com
Alvarion	http://www.alvarion.com
Aperto	http://www.apertonet.com
ArrayComm	http://www.arraycomm.com
ARRIS	http://www.arrisi.com
Axxcelera	http://www.axxcelera.com
Cambridge Broadband	http://www.cambridgebroadband.com
QUALCOMM Flarion	http://www.flarion.com
IPWireless	http://www.ipwireless.com
L3 Communications	http://www.l-3com.com
Navini Networks	http://www.navini.com
Proxim Wireless	http://www.proxim.com
Redline Communications	http://www.redlinecommunications.com
Solectek	http://www.solectek.com
SOMA Networks	http://www.somanetworks.com
SR Telecom	http://www.srtelecom.com
Trango Broadband Wireless	http://www.trangobroadband.com
VCom	http://www.vcom.com
Vyyo	http://www.vyyo.com
WaveRider	http://www.waverider.com
Wi-LAN	http://www.wi-lan.com
WiMAX Infrastructure	
Alcatel	http://www.alcatel.com
Ericsson	http://www.ericsson.com
Motorola	http://www.motorola.com
Nortel	http://www.nortelnetworks.com
Samsung	http://www.samsung.com
Siemens	http://www.siemens.com
ZTE	http://www.zte.com
WiMAX chip	
Beceem	http://www.beceem.com
NextWave	http://www.nextwave.com
Fujitsu	http://www.fujitsu.com
Intel	http://www.intel.com
picoChip	http://www.picochip.com
Runcom	http://www.runcom.com
SEQUANS	http://www.sequans.com
SiWave	http://www.si-wave.com
TeleCIS Wireless	http://www.telecis.com
Wavesat	http://www.wavesat.com

Introducing the Contemporary Contribution Index for Characterizing the Recent, Relevant Impact of Journals

Peter Vinkler

pvinkler@chemres.hu
Chemical Research Center, Hungarian Academy of Sciences, Pusztaszeri út 59-67, H-1025 Budapest (Hungary)

Abstract
The Garfield (Impact) Factor characterizes the measure of the up to date specific contribution of scientific journals to the total impact of the journals in a special field. A new indicator (Contemporary Contribution Index, CCI) was introduced in order to characterize the relative contribution of journals to recent, relevant knowledge of a corresponding field. The CC Index relates the number of citations received by a journal in a given year to the total number of citations obtained by all journals of the corresponding field in that year. Mean Garfield Factors and mean Contemporary Contribution Indexes were calculated for some fields and several journals. No significant correlation was found between the Garfield Factor (GF) and Contemporary Contribution Index (CCI) of journals. The ratios of the GF to CCI referring to the corresponding top 10, 20 or 50 per cent of the journals ranked by decreasing GF and CCI, strongly differ by field.

Keywords
eminence of journals; highly cited papers; impact factor; reference standard

Introduction

Scientometric assessment methods preferably apply indicators for assessing publication performance as follows: *Journal Paper Productivity* (JPP: number of journal papers per researcher per year), *Publication Strategy* (PS: mean weighted impact factor of journals where the papers assessed were published), or *Relative Publication Strategy* (RPS: PS related to the mean impact factor (GF_m) of the journals dedicated to the field or subfield where the authors assessed are active), *Relative Subfield Citedness* (RW: number of citations obtained related to the total number of papers assessed multiplied by GF_m) or *Relative Citation Rate* (RCR: mean number of citations by paper related to the PS index), and *Number of Highly Cited Papers* related to the number of researchers or to the total number of papers assessed (Vinkler, 2006). For calculating the PS, RPS, RW, and RCR indicators Garfield (Impact) Factors representing the eminence of scientific journals are widely applied. For calculating HCPs we should determine a citedness threshold for journal papers. For obtaining appropriate standards bibliometric factors, which have different impact on the scientometric indicators of papers or journals of different fields, may cause great difficulties.

Majority of the scientometric distributions, e.g. the publication productivity or number of citations obtained by individuals or teams, the number of citations received by papers or journals may be (strongly) skewed (Seglen, 1992, 1994).

There are many researchers who can publish in journals with high Garfield (Impact) Factor, but they obtain only a relatively low number of citations. It is often forgotten that the publication in eminent journals offers only a possibility for obtaining high numbers of citations. According to the Gospel: "For many are called, but few are chosen". (The Gospel according to St. Matthew, Chap. 22/14; The Holy Bible, British and Foreign Bible Society, London, 1611). This phenomenon corresponds in scientometrics to the well-known fact that many papers published in journals of relatively high GF will be cited less frequently than the average and only few papers obtain high number of citations. This regularity may be termed as the "Invitation Paradox".

The GF, relating the number of citations obtained to the number of citable items, is a *specific* index. It was suggested for comparing citedness of journals of different size. This advantageous feature brings forth, however, some drawbacks. Namely, journals with low or extremely low number of papers may have very high GF. According to the Weber-Fechner law of physiology (see, Egghe & Rousseau, 1990)

the impact is a logarithmic function of the effect. Consequently, significantly greater efforts should be made by journals to arrive at the same level of citedness by publishing many (> 500) papers than publishing relatively low number of papers (< 100).

In his seminal paper Garfield (1976) used two indicators for characterizing journal significance: the total number of citations and the impact factor. The first index refers to the total number of citations obtained in a special year by papers published at any time.

The Garfield (Impact) Factor (GF), being a short term citedness indicator, may represent the *up to date specific contribution* of a journal to the total up to date impact of the journals devoted to a respective field or subfield (Vinkler, 2004). In contrast to this, the total number of citations obtained in a special year to all papers published at any time in a journal, may characterize the *total impact* of that journal on recent publications of the corresponding field.

Only part of the total information published may be regarded as *information required for generating new information.* This part can be termed *as recent, relevant knowledge* (RRK). Consequently, calculating with cumulative number of papers for presenting the relevant knowledge of a field whilst neglecting the aging of information − is misleading. The content of RRK dynamically changes due to permanent information production, evaluation, modification, and aging processes (Vinkler, 2002). The amount of RRK in a given year in a given field may be approximated by the *number of references* in journals devoted to that field. Because of multiplication (i.e. the same paper may be referenced several times), the total number of references may be regarded as the weighted amount of information applied.

The distribution of the age of references in journal papers published in a given year reveals that only 5-10 per cent of the total refers to papers published earlier than 15-20 years (Glänzel & Schoepflin, 1999; Glänzel, Schlemmer & Thijs, 2003). Consequently, information in papers published within the period of 0-15 years prior to recent papers and referenced recently, may be regarded as the main body of recent relevant knowledge (RRK). The amount of contribution to RRK may be calculated for journals or for sets of papers, as well.

Based on the above consideration, we may define a Contemporary Contribution Index (CCI) for journals according to Eq.1. The CC Index may represent the relative contribution of journals to the total recent relevant knowledge of a field or subfield.

$$CCI_{j,y} = \frac{C_{j,y}}{\sum_{j=1}^{J} C_{j,y}} \qquad (1)$$

where $CCI_{j,y}$ is the CC Index for the j-th journal in year *y*, $C_{j,y}$ is the total number of citations obtained by the j-th journal in year *y* to papers published in any year, and *J* is the number of journals of a field or subfield. Relating the CC Index to the number of papers published, we can calculate a *specific* contribution index independent of the size of journals.

Finding appropriate *absolute reference standards* is one of the central issues of evaluative scientometrics. There are, however, no absolute norms available for information productivity or impact to be attained, at present. Standards should be calculated with great care because the bibliometric factors influencing most scientometric indicators differ by field. It seems obvious to apply generally the mean values of the top 1, 10 or 20 per cent of papers ranked, e.g., by decreasing number of the citation rate index, as normalization factors.
The aim of the present work was to calculate mean Garfield (Impact) Factors (GF) and Contemporary Contribution Indexes (CCI) for different subsets of journals in different fields and to find appropriate reference standards for relative indicators.

Results

The means of GFs and CCIs were calculated for the Top 1, 10, 20, and 50 per cent of journals in some fields. The data (Table 1) show that the GF of journals in Polymer Science is lower (1.376) than that in Physical Chemistry (2.349) and Neuroscience (3.079), on average. The median value is significantly lower than the corresponding mean for each field (0.926, 1.557, 2.363, for polymer, physical chemical and neuroscience journals, respectively). The mean GF of all subsets (e.g., Top 10%; Top 20%) is highest for Neuroscience journals, followed by journals in Physical Chemistry and Polymer Science. In contrast to the GF data, the 10^2 CCI values are highest for polymer journals for the total set as well as for all subsets.

Table 1. Mean Contemporary Contribution Index (m 10^2 CCI) and mean Garfield (Impact) Factor (mGF) calculated for the total sets and some subsets of journals by field.

Set	Number of journals			mGF (SD)			$m10^2$ CCI (SD)		
	PhCh	Polym	Neuro	PhCh	Polym	Neuro	PhCh	Polym	Neuro
Total	77	111	200	2.349 (2.696)	1.376 (1.933)	3.079 (3.173)	0.901 (1.604)	1.299 (3.514)	0.500 (0.998)
Top 50 %	38	55	100	3.774 (3.248)	2.230 (2.428)	4.868 (3.648)	1.840 (2.066)	2.537 (4.717)	0.935 (1.273)
Top 20 %	16	22	40	6.125 (4.146)	3.654 (3.514)	7.467 (4.663)	3.462 (2.449)	5.201 (6.447)	1.824 (1.654)
Top 10 %	8	11	20	8.467 (4.890)	4.982 (4.504)	10.421 (5.116)	5.178 (2.468)	8.636 (7.882)	2.792 (1.900)
Top 1 %	1	1	2	17.857 (0.000)	16.045 (0.000)	22.567 (2.286)	9.269 (0.000)	26.865 (0.000)	7.431 (2.842)

PhCh: Physical Chemistry
Polym: Polymer Science
Neuro: Neuroscience
SD: standard deviation

GF: publication time-period, t_p = 2003-2004; citation year = 2005
CCI: citation year = 2005; publication time-period, t_p = any year prior to 2005

Top 50, 20, 10 and 1 %: number and data of journals of the corresponding set. The journals are ranked by decreasing value of the Garfield (Impact) Factor (GF) and Contemporary Contribution Index (CCI), resp.
Source of data: Thomson Scientific ISI, Journal Citation Reports, 2005.

The number of journals in the Top 10% set is relatively low (11, 8, 20, resp.), consequently the application of indicators referring to the Top 20% set could be recommended as standard. However, the hierarchical level of the standards should be selected not only according to methodological viewpoints but considering the aspects of science policy, as well. (It should be noted that the results referring to the Top 10% sets and those for sets containing the square roots of the total number of journals do not differ significantly.)

The GF or CCI ratios of the same subsets (Table 2, e.g., m(Top 10%)/m(Top 20%)) yield similar values in the different fields (e.g., 1.38, 1.37 and 1.39 for Physical Chemistry, Polymer Science and Neuroscience, resp.). The CCI ratios are higher than the GF ratios. Among the CCI ratios there are two extremes (m(Top 10%)/m(Total) = 6.65 and Mean/Median = 6.50).

The (GF/CCI) ratios referring to the same subset (e.g. Top 10%/Top 10%) are significantly different, e.g., 1.63, 0.58 and 3.73 for Physical Chemistry, Polymer Science and Neuroscience, resp. The data in Table 3 reveal that the GFs are lower for polymer journals (the ratios are lower than unity except for the total set). The ratios, however, are higher than unity for the other fields.

The CC Index may depend on several factors, which are partly different from those influencing the GF. Among these factors the number of papers published during the whole life time of the journal and citations obtained recently, may play a determining role. The number of citations to be obtained depends on several factors, e.g., development rate, aging of information, mean number of references, etc. Therefore, a cross-field comparison aimed at setting an "absolute" rank of significance or impact for the journals is hardly possible. In contrast to this, the short-term impact of journals may be appropriately represented by the Garfield (Impact) Factor, whereas the long-term impact by the CCI values within well selected fields.

Table 2. Relation of the mean of Garfield (Impact) Factors (GF) and Contemporary Contribution Indexes (CCI) of different subsets of journals by field.

	$\frac{m(Top\,10\%)}{m(Top\,20\%)}$	$\frac{m(Top\,10\%)}{m(Top\,50\%)}$	$\frac{m(Top\,10\%)}{m(Total)}$	$\frac{m(Top\,20\%)}{m(Total)}$	$\frac{Mean}{Median}$
GF ratios					
Physical Chemistry	1.38	2.24	3.60	2.61	1.51
Polymer Science	1.37	2.23	3.62	2.66	1.49
Neuroscience	1.39	2.14	3.38	2.42	1.30
CCI ratios					
Physical Chemistry	1.50	2.81	5.75	3.84	2.83
Polymer Science	1.66	3.40	6.65	4.01	6.50
Neuroscience	1.53	2.99	5.58	3.65	2.87

m: mean of GF of journals of the corresponding set
Source of data: Thomson Scientific ISI, Journal Citation Reports, 2005.

The most striking feature of the relationship between the number of papers published by the journals studied and their Garfield (Impact) Factors (GF) and Contemporary Contribution Indexes (CCI) is that the GF and CCI values show no significant correlation (Table 4). This finding supports the idea that the two indicators are independent measures characterizing *different aspects* of the citedness of journals. The difference is due to the characteristics of the indicators, i.e. the GF is a *specific* index, whilst the CCI is a *distribution* (contribution) index (Vinkler, 2001). The lack of significant correlation between the number of papers (P) published and the impact of journals (i.e. GF) indicates that the quantity of information published may hardly play any role in determining the general quality (impact) of the journals. (It should be noted that the number of papers published in the period of 1990-2005 and that published in 2003-2004 show significant correlation.) The journals with a very low number of papers and very high GF may be regarded as exceptions within the total set of journals. In contrast to the above, the Contemporary Contribution Index (CCI) shows highly significant correlation with the number of papers published by the journals in each subfield. (It should be noted that identical correlation can be calculated with the total number of citations obtained). It is obvious, that the measure of contribution to the total impact of journals in a field made by a given journal shall be in direct correlation with the corresponding number of citations, if the impact is measured by the total number or share of citations received.

Table 3. Relation of the mean Garfield (Impact) Factor (GF) to the mean Contemporary Contribution Index (CCI) of identical subsets by field.

(GF/CCI) ratios				
	$\frac{m(Top10\%)}{m(Top10\%)}$	$\frac{m(Top20\%}{m(Top20\%)}$	$\frac{m(Top50\%)}{m(Top50\%)}$	$\frac{m(Total)}{m(Total)}$
Physical Chemistry	1.63	1.77	2.05	2.61
Polymer Science	0.58	0.70	0.88	1.06
Neuroscience	3.73	4.09	5.21	6.16

In order to study the GF and CCI indicators on a lower level, we analyzed the first six journals ranked by GF and 10^2 CCI of the three fields (Tables 5 and 6). Only three of 18 journals are recorded in both lists (Macromolecules, Neuron, J. Polym. Sci., Pol. Chem.). This feature may indicate that GF and CCI values do not run parallel. The journals relatively highly cited, publishing a low number of papers are preferably placed in the GF list whilst those with a great number of papers are recorded in the CCI list. The mean number of papers in journals in the GF list is as follows: 107, 808 and 241 and that in the CCI list: 2901, 1899 and 1446, for Physical Chemistry, Polymer Science and Neuroscience, respectively. The cross-field rank numbers clearly show the different aspects of the GF and CC Index. The sums of the cross-field rank numbers of journals on physical chemistry are similar according to the GF and CC Index. In contrast to this, the sum of ranks is very different for polymer journals (84 *vs* 41) and for neuroscience journals (34 *vs* 77). It should be noted that the higher rank numbers correspond to lower GF or CCI indexes. According to the long-term impact (CCI) polymer journals are preferred, whilst the short-term impact of journals exceeds the long-term effect in Neuroscience.

Table 4. Pearson's correlation coefficients between the number of papers published in 2003-2004 (P) and Garfield (Impact) Factor (GF, 2005) and Contemporary Contribution Indexes (CCI, 2005) of journals by field.

		P	CCI
	Neuroscience	0.868*	
CCI	Polymer Science	0.833*	
	Physical Chemistry	0.917*	
	Neuroscience	0.110	0.353
GF	Polymer Science	0.128	0.259
	Physical Chemistry	0.007	0.137

*significant at the $p < 0.01$ level

For obtaining sets of journals as absolute reference standards for calculating relative impact indicators selecting the top 0.1, 1.0, 10.0, 20.0 or 50.0 per cent of the total of a field may be appropriate. The data in Table 2 reveal that the different bibliometric factors working in different fields hardly have any influence on the citedness ratios of the different top sets of *journals*. The question may be raised whether the ratios of the citedness values of different elite sets of *papers* would show similar trends.

The top 0.01, 0.1, 1.0 and 10.0 per cent of papers were selected in several fields, and the citedness threshold values were determined (Table 7). The citedness threshold (CT) value corresponds to the lowest citations per paper index of a paper belonging to the respective subset (i.e. top 10 per cent, where the papers are ranked by decreasing number of citations). The normalized CT values show (Table 7) that the ratios between the Citation Threshold (CT) values are similar for the subsets of top 10.0 to 1.0 per cent (e.g., nCT = 4.94 and 4.14 for papers in the top 10 per cent for Mathematics and Neuroscience and Behavior, respectively). This observation would indicate the application of the corresponding citation rate values of the subsets (top 1.0 or 10.0 per cent of papers) as *reference standards*. The differences between the normalized Citedness Threshold (nCT) values are significantly

higher for the subsets containing the top 0.10 or 0.01 per cent of papers (e.g., for Mathematics nCT = 253.91 and only 90.01 for Neuroscience and Behavior, for the top 0.01 per cent of papers). The relatively great difference may be attributed to the relatively low number of papers in the subset. Table 7 shows also the data referring to the total ISI data base. This data base also contains citedness data of papers of social sciences and refers to the period of 1981-1997 in contrast to the 1995-2005 data of papers by field, i.e., Chemistry, Clinical Medicine etc. Nevertheless, the corresponding nCT values of the ISI data base are commensurable with those of the fields.

Table 5. The first six journals ranked by the Garfield (Impact) Factor (GF) in some fields.

Journal	P (2003-2004)	GF (2005)	Cross-field rank	10^2 CCI (2005)	Ratio (GF/10^2 CCI)
Physical Chemistry					
Surf. Sci. Rep.	21	17.857	3	0.413	43.23
Nat. Mater.	256	15.941	5	0.844	18.89
Annu. Rev. Phys.	37	13.405	9	0.681	19.68
Chem.					
J. Photoch.	24	8.167	11	0.072	113.43
Photobio. C					
Adv. Funct. Mater.	278	6.677	12	0.528	12.65
Prog. Nucl. Mag.	26	6.462	13	0.216	29.92
Res. Sp.					
Sum of rank numbers		53(31%)			
Polymer Science					
Prog. Polym. Sci.	67	16.045	4	1.618	9.92
Adv. Polym. Sci.	47	4.319	14	1.063	4.06
Macromolecules	2738	4.024	15	26.865	0.15
Biomacromolecules	566	3.618	16	1.608	2.25
Macromol. Rapid.	422	3.126	17	2.250	1.39
Comm.					
J. Polym. Sci. Pol. Chem.	1008	3.027	18	5.593	0.54
Sum of rank numbers		84(49%)			
Neuroscience					
Annu. Rev. Neurosci.	143	24.184	1	0.836	28.94
Nat. Rev. Neurosci.	49	20.951	2	0.824	25.41
Nat. Neurosci.	333	15.456	6	1.840	8.40
Trends Neurosci.	166	14.325	7	1.419	10.10
Neuron	665	14.304	8	4.433	3.23
Prog. Neurobiol.	90	11.789	10	0.767	15.36
Sum of rank numbers			34(20%)		

P: number of papers
Source of data: Thomson Scientific ISI, Journal Citation Reports, 2005.

The data in Table 2 and 7 refer to different sets of items. Nevertheless, m(Top 10%)/m(Total) citedness ratios referring to journals may be compared with the normalized Citedness Threshold values (Top 10%/Total) of papers. The corresponding ratio for the journals of Neuroscience and Physical Chemistry is 3.38 and 3.60, resp. (Table 2), whilst the ratio for papers in Neuroscience and Behavior and Chemistry is 4.14 and 4.19, resp. (Table 7). The values may be regarded as commensurable.

Table 6. The first six journals ranked by the Contemporary Contribution Index (10^2 CCI) in some fields.

Journal	P (2003-2004)	10^2 CCI (2005)	Cross-field rank	GF (2005)	Ratio (GF/10^2 CCI)
Physical Chemistry					
J. Phys. Chem. B.	4433	9.269	4	4.033	0.44
Langmuir	3327	8.525	6	3.705	0.43
Chem. Phys. Lett.	3400	7.476	7	2.438	0.33
Chem. Mater.	1551	5.213	10	4.818	0.92
J. Phys. Chem. A.	2845	4.971	11	2.898	0.58
Surf. Sci.	1848	4.001	15	1.780	0.44
Sum of rank numbers			53(31%)		
Polymer Science					
Macromolecules	2738	26.865	1	4.024	0.15
Polymer	1875	11.317	2	2.849	0.25
J. Appl. Polym. Sci.	3428	9.062	5	1.072	0.12
J. Polym. Sci. Pol.	1008	5.593	8	3.027	0.54
Chem.					
J. Membrane Sci.	761	4.851	12	2.654	0.55
Synth. Met.	1586	4.784	13	1.320	0.28
Sum of rank numbers			41(24%)		
Neuroscience					
J. Neurosci.	2521	9.440	3	7.506	0.80
Brain Res.	2145	5.421	9	2.296	0.42
Neuron	665	4.433	14	14.304	3.23
J. Neurophysiol.	1238	3.476	16	3.853	1.11
Comp. Neurol.	892	3.319	17	3.855	1.16
J. Neurochem.	1216	3.175	18	4.604	1.45
Sum of rank numbers			77(45%)		

P: number of papers
Source of data: Thomson Scientific ISI, Journal Citation Reports, 2005.

Conclusion

A new indicator was introduced for characterizing the contribution of journals to the recent relevant knowledge of a corresponding field or subfield. The Contemporary Contribution Index (CCI) offers an alternative for the evaluation of the impact of scientific journals. It depends on the aim of the given study whether the use of Garfield (Impact) Factors or CC Indexes can be recommended. For analyzing the recent impact of journal papers published recently, the GF indicator, whereas studying the recent impact of the body of information of all relevant papers, the Contemporary Contribution Index may be applied.

There are several methods for obtaining eminent sets of journals or papers by different scientometric indicators. Selecting sets of papers belonging to the top 1-50 per cent of the total items, seems to be appropriate, depending on the size of the set studied and purpose of the assessment.

The different bibliometric factors working in different subject fields may cause great differences in the value and ratios of citedness indicators (e.g. Garfield (Impact) Factor, GF or Contemporary Contribution Index, CCI). Accordingly, mean GFs and mean CC Indexes were found to be different for the different eminent sets of papers. However, the ratio of the indexes (e.g. GF(mTop 10%)/GF(m(Top 20%)) shows no field effect. In contrast to this, the (GF/CCI) ratios referring to the same subsets (e.g., m(Top 10%)/m(Top 10%)) show significant field effect (Table 3).

Table 7. Citedness Threshold (CT) and Normalized Citedness Threshold (nCT) values for sets of most frequently (Top 0.01; 0.1; 1.00 and 10 per cent, resp.) cited papers in some subject fields.

Set	Total set*		Chemistry		Clinical Medicine	
	CT	nCT	CT	nCT	CT	nCT
Top 0.01 %	951	110.97	1056	142.90	1478	151.90
Top 0.10 %	316	36.87	311	42.08	517	53.13
Top 1.00 %	106	12.37	102	13.80	167	17.16
Top 10.0 %	23	2.68	31	4.19	42	4.32
Total(LtC)	8.57	1.00	7.39	1.00	9.73	1.00

Set	Engineering		Mathematics		Physics		Neuroscience and	
	CT	nCT	CT	nCT	CT	nCT	CT	nCT
Top 0.01 %	326	113.19	617	253.91	1022	153.68	1370	90.01
Top 0.10 %	119	41.32	180	74.01	341	51.28	581	38.17
Top 1.00 %	46	15.97	42	17.28	107	16.09	223	14.65
Top 10.0 %	14	4.86	12	4.94	28	4.21	63	4.14
Total(LtC)	2.88	1.00	2.43	1.00	6.65	1.00	15.22	1.00

nCT: normalized values to the mean citation rate of the total within the corresponding field
Papers published in 1995 and cited in 1995-2005
Source of data: ISI, ESI Web of Knowledge, Thomson Scientific, Version 1.0
Note: The Citedness Threshold value is the lowest number of citations obtained by a paper in the corresponding subset.

*Source of data: H. Small & D. Pendlebury (http://physics.bu.edu/~redner/projects/citation/isi.html). Papers published in 1981 and cited during the period January 1981-June 1987. Total number of papers in ISI studied: 783329
Total (LtC): Long-term Citedness values refer to the mean citations per paper index within the total set.

References

Egghe L., & Rousseau R. (1990). Introduction to Informetrics. *Quantitative Methods in Library, Documentation and Information Science*. Amsterdam: Elsevier Science Publishers.

Garfield, E. (1976). Significant journals of science. *Nature*, 264, 609-615.

Glänzel, W., Schlemmer, B. & Thijs, B. (2003). Better late than never? On the chance to become highly cited only beyond the standard bibliometric time horizon. *Scientometrics*, 58, 571-586.

Glänzel, W., & Schoepflin, U. (1999). A bibliometric study of reference literature in the sciences and social sciences. *Information Processing & Management*, 35, 31-44.

Seglen, P. O. (1992). The skewness of science. *Journal of the American Society for Information Science*, 43, 628-638.

Seglen, P. O. (1994). Causal relationship between article citedness and journal impact. *Journal of the American Society for Information Science*, 45, 1-11.

Vinkler, P. (2001). An attempt for defining some basic categories of scientometrics and classifying the indicators of evaluative scientometrics. *Scientometrics*, 50, 539-544.

Vinkler, P. (2002). The institutionalization of scientific information. The Scientometric Model (ISI-S MODEL). *Library Trends*, 50, 553-569.

Vinkler, P. (2004). Characterization of the impact of sets of scientific papers: The Garfield (Impact) Factor. *Journal of the American Society for Information Science and Technology,* 55, 431-435.

Vinkler, P. (2006). Composite scientometric indicators for evaluating publications of research institutes. *Scientometrics*, 68, 629-642.

Beyond Rankings: the Role of Large Research Universities in the Global Scientific Communication System

Martijn S. Visser, Clara M. Calero Medina and Henk F.Moed

{visser,clara,moed}@cwts.leidenuniv.nl
Centre for Science and Technology Studies (CWTS), Universiteit Leiden, P.O. Box 9555, 2300 RB Leiden (The Netherlands)

Abstract
This paper presents preliminary findings of a study that aims to assess the role of large research universities in the global scientific communication system. For this study a special database was created containing the publication output in Web of Science journals of the largest research universities. Its primary aim is to illustrate how a structural analysis of the global academic system may provide policy relevant insights that are crucial for a proper interpretation of bibliometric rankings. It presents two analyses. The first focuses on differences between the US and the European academic system. It is proposed to further analyse and interpret these differences in terms of a distinction between a concentrated and a distributed model of research quality among universities in national or supra-national academic systems. The second analysis is based on a network approach, revealing how the research activities carried out in one university depend upon those conducted in another academic institution. This dependence is measured by citation links. It addresses the question how one can distinguish between 'core', 'intermediate' and 'peripheral' research universities in the network.

Keywords
rankings; world universities; citation impact; citation networks; concentrated vs. distributed models; national academic systems.

Introduction

As internationalisation and globalisation in academic research and teaching proceeds, more and more attempts are made to identify top research universities from a global perspective. Universities are increasingly competing for research funds, research students and researchers in the global research area. Their reputation as research universities is a crucial factor in such a competitive system. Therefore, members of the international scientific community, officials responsible for institutional, national and supra-national science policies, and the wider public need 'objective', 'reliable' information about the research performance of universities.

Comparative analyses of the performance of universities at a national level, focusing on particular research fields or disciplines, have been carried out for many years. For instance, in 1995 the US National Research Council (NRC), the working arm of the National Academy of Science and the National Academy of Engineering, published a report presenting a quality rating of PhD programs at 274 US institutions in 41 fields, based on surveys sent to faculty (Goldberger et al., 1995).

A recent phenomenon is the compilation of rankings of universities from a supra-national or global perspective. For instance, the European Commission published in the recent European Science Indicators Reports listings of European universities presenting their bibliometric scores. Global rankings of universities were published by the Jiao Tong University in Shanghai (SJTU 2005) and by Times Higher Education Supplement (THES 2005). For a thorough review of these two rankings, the reader is referred to Van Raan (2005).

In rankings of universities differences between disciplines tend to become submerged, and the focus is primarily on straightforward counts of publications, citation impact, external funding, awards and so on. Little attention is paid to the particular role institutions play within the national and global network of knowledge diffusion and knowledge use. Rankings are in a sense one-dimensional: entities are ordered by descending score on one particular statistic. Of course one can design some compound statistic, taking into account scores on several aspects rather than one. But then the problem is: how

should one weigh the importance of the various aspects? Moreover, rankings disregard how the performance of one entity depends upon that of others (e.g., Calero Medina and Moed, 2006).

This paper presents bibliometric characteristics of the 386 most frequently publishing world universities. Its primary aim is to illustrate how a structural analysis of the global academic system may provide policy relevant insights that are crucial for a proper interpretation of bibliometric rankings. It presents two analyses. The first focuses on differences in citation impact between the US and the European academic system among disciplines. It is proposed to further analyse and interpret these differences in terms of a distinction between a concentrated and a distributed model of research quality among universities in national or supra-national academic systems. A second analysis examines these contrasting models in more detail by means of a network approach. It aims to characterise differentiated academic systems by showing how research activities carried out in one university depend upon those conducted in another academic institution. This dependence is measured by citation links, i.e., essentially based on the number of times papers published by one university are cited by those published by another institution.

A key question is which of the contrasting models is the most appropriate in the various countries, especially which model provides the most optimal conditions for 'top' research. The intent of this paper is not to resolve this complex question but to differentiate between the research performance of academic systems in a more precise and accurate way than rankings schemes do.

Data and Methodology

Assignment of articles to universities; accuracy
The data collection process aimed at defining the article output of European universities publishing at least some 500 papers during the time period 1997-2004. For non-European universities the process identified the articles of the 200 most frequently publishing universities. Articles were assigned to universities on the basis of the information on the institutional affiliations of authors, included in the corporate address field. Two rounds were carried out.

In a first round, papers were selected with the name of a university (and its major departments) mentioned explicitly in the address. Name variations were taken into account. For instance, Ruprecht Karls University is a name variant of the University of Heidelberg and Université Paris 06 of Université Pierre et Marie Curie. For European universities, this round took into account all variations occurring 5 or more times. For non-European universities this threshold was set to 25.

A specific problem in the assignment of papers to universities concerns the different organisational relationships across academic systems that exist between faculties of medicine and their teaching hospital(s). The degree of affiliation may range from tight integration to occasional collaboration. As a consequence the assignment of papers to universities may not be consistent across academic systems if based solely on the working addresses mentioned in research papers. For that reason additional papers were selected in a second round from affiliated, teaching hospitals on the basis of an author analysis. This round added to a particular university's article output selected in the first round papers from affiliated hospitals, published by authors who did not explicitly mention this university's name in their institutional affiliation, but who showed strong collaboration links with that university, as its name appeared in the address lists of at least half of their papers. In this way, for instance, a part of the papers containing the address Addenbrookes Hospital was assigned to University Cambridge, and a part of the papers with the address Hospital La Pitié Salpetrière to University of Paris VI, and another part to University of Paris V.

Since the de-duplication and counting process of European universities took into account only name variants occurring 5 or more times, an overall accuracy rate for this group of universities is estimated to be about 95 per cent. It is somewhat higher for universities with a large number of published articles than it is for universities with smaller publication volumes. For non-European universities it is around

90 per cent. It is important to note that the data were not verified by representatives of institutions, except in the case of Flemish universities.

Universities analysed in this paper
This paper analyses the set of universities that published more than 5,000 articles in Web of Science journals during 1997–2004, or on average more than 625 papers per year during this time period. It contained 386 universities, and is denoted as the global or world set, containing world universities. In view of the collaboration among institutions, resulting in co-publications by scientists from two or more institutions, it would be more precise to state that the universities contributed at least one author to more than 625 papers per year. Technically, this number is denoted as an integer count.

Citation impact indicators calculated
Citation analysis is an appropriate tool to measure intellectual influence or prominence of research groups and institutions in their fields (e.g., Garfield, 1979). The *Normalised Citation Impact* calculated in this paper is defined as the average number of citations per article published from a university, relative to the world citation average in the subfields in which it is active. It is also denoted below as 'citation impact' or 'impact per paper'. Details can be found in Moed (2005). A value of 1.0 indicates a citation impact that is equal to the world citation average.

Research disciplines and their coverage by the Web of Science
In this paper, a classification of research articles into 15 disciplines is used, based upon a grouping of journal categories available in the *Web of Science*: *biological sciences related to humans; biological sciences related to animals and plants*; *molecular biology and biochemistry; clinical medicine physics; chemistry; applied physics and chemistry; engineering; geosciences; mathematics; psychology; social sciences related to medicine and health; other social science; humanities*. For more details the reader is referred to Moed (2005, p. 187–189). *WoS* coverage is excellent or good in most disciplines, except in parts of social sciences and humanities. Therefore, the indicators calculated in this paper may not properly reflect a university's position in the latter domains of scholarship, especially when it is located in a non-English speaking country.

Differences between US and European academic system
In the set of 386 world universities, 172 are located in Europe, and 122 in the USA. Table 1 gives for each geographical region the mean and quartiles of the distribution of normalised citation impact among universities. The table shows that US universities tend to have a higher normalised citation impact than European academic institutions: 1.55 versus 1.11. The 75th percentile of the distribution for Europe is lower than the 25th percentile for the USA. The third column shows that the 172 European universities account for around 72 per cent of the total European university output. The 122 US universities published about 83 per cent of the total US university output. This percentage is higher than the 72 per cent obtained for Europe, and indicates that there is a stronger concentration of published articles among US universities than there is among European institutions, in agreement with earlier analyses published by Matia et al. (2005).

Table 1. Citation impact distribution for European and US universities

Region	***No Universities***	***% Papers from univs***	***Normalised citation impact distribution***			
			Mean	***P25***	***P50***	***P75***
Europe	172	72	1.11	0.99	1.10	1.22
USA	122	83	1.55	1.32	1.54	1.72

Legend to Table 1. Mean, P25, P50, P75: The mean, 25th, 50th (i.e. the median) and 75th percentile of the distribution. % Papers from univs: A rough estimate of the percentage of the total university article output from a country/region published by the universities in the set of 386 world universities. Both percentages are rough estimates, as the number of articles published by the *total* collection of universities in Europe or the USA is not precisely known in this study.

In order to further characterize differences among European and US universities, an institution's citation impact was analysed *per discipline*, using a classification of research articles into 15 disciplines presented in Section 2. For each institution the number of disciplines was determined in which it was 'world leader', i.e., ranked among the top 10 or top 25 per cent according to normalized citation impact in the set of 386 world universities described in Section 2 above. For each region, the number and percentage of universities was determined that was world leader in at least one discipline, and for these institutions the average number of such 'top' disciplines per university was computed. These indicators were calculated for all universities in the set, and also for the 'very best' universities in their region, i.e., being among top 25 percent in their region on the basis of their overall normalised citation impact.

Table 2. Analysis disciplines in which universities are 'world leaders'

Indicator	***All universities***		***Very best 25 % universities***	
	Europe	***USA***	***Europe***	***USA***
Number of universities	172	122	43	31
Among the world top 10 % universities in a discipline				
No (%) universities with at least one 'top' discipline	44 (26 %)	99 (81 %)	29 (67 %)	31 (100 %)
Average number of 'top' disciplines per univ	1.8	5.1	2.1	9.3
Among the world top 25 % universities in a discipline				
No (%) universities with at least one 'top' discipline	112 (65 %)	119 (98 %)	43 (100 %)	31 (100%)
Average number of 'top' disciplines per university	3.2	8.4	5.4	12.3

The results are presented in Table 2. The upper half of this table present the outcomes when the concept of 'world leader' in a discipline is defined as being among the *top 10 per cent* among all 386 world universities in that discipline. In the lower half, the criterion for being world leader is somewhat relaxed, and defined as belonging to the *top 25 per cent* in a discipline. Focusing on the upper half, the second and third column show that 26 per cent of the 172 European universities analysed in this section are world leaders (among the top 10 per cent) in one or more disciplines. For the 122 US universities this percentage is 81. The last two columns focus on the 'very best' universities in a region, and show percentages of 67 for European and 100 for US institutions. The average number of disciplines in which European universities are world leader is substantially lower than that for their US counterparts.

Analysing the appearance among the top 25 per cent rather than the top 10 percent in a discipline (lower half of Table 2), it can be seen that in the set of all 172 European universities 65 per cent is world leader in at least one discipline. In the set of the 43 very best European universities this percentage amounts to 100, equal to that for the US. However, the number of disciplines in which a European university is world leader is on average substantially lower than that calculated for US universities: 5.4 versus 12.3. It needs emphasizing that this average of 2 top disciplines for European universities is the mean of a rather skewed distribution. Several universities located in the UK are world leader in a number of disciplines similar to that for US universities.

Discussion of preliminary findings

In a recent paper by Lambert and Butler (2006), the authors claim that European universities – with some exceptions, mainly in the UK – are 'mediocre'. The first results from our analysis indeed revealed that US universities are highly overrepresented in the top of the world ranking based on published article output, and particularly in that based on citation impact. However, one may question whether the term 'mediocre' is appropriate to qualify the position of European universities in the world rankings. But even if one adopts this qualification from Lambert and Butler, it needs emphasising that 'mediocrity' of a university does not necessarily imply that it is mediocre in *all*

disciplines. A key finding from this paper is that an European 'top' university tends to be among the best 25 per cent in the world in at least one discipline, although the number of disciplines in which it is world leader is on average substantially lower than that calculated for a US top university.

A distinction can be made between two models for distributing 'top' research among universities: a concentration model in which a limited number of big research universities carries out research at a top level in a wide range of disciplines, and a distributed model, in which top research is more evenly distributed among universities, and a strong link between teaching and research is maintained.

Remaining work: Citation relationships among universities

A more detailed analysis will be made of the structure of European, US and other (supra-) national academic systems and the extent to which these systems are structured according to one of the two contrasting models described above. The analysis will apply the concepts and methodology that have been developed in graph theory and social network analysis on the knowledge flows between the 386 universities. Who is using whose research results, and how reciprocal is the relationship between knowledge users and knowledge producers? At this moment the research is not yet finalised.

The following specific questions will be addressed for separate disciplines:

1) How large is the combined contribution of the selected research universities to the total of citation flows?
2) Can we detect a core/periphery structure among research universities and is a dichotomous or a more continuous model more appropriate to describe this structure?
3) Which trends are visible in the recent past?

Matrices of the citation relationships between citing and cited universities are analysed to mark the different positions of research universities in the global scientific communication system. The division of distinct roles among academic organisations will be used to characterise academic systems and differentiate between them. Structural differences that can be detected between academic systems will be combined with research performance indicators as measured by normalised citation impact to identify possible relations and patterns.

References

Calero Medina, C.M. and Moed, H.F. (2006). *Depicting the landscape of research universities*. Paper presented at the Ninth International Conference on Science and Technology Indicators, Leuven (Belgium), 7-9 September 2006.

Garfield, E. (1979). Citation *Indexing. Its theory and application in science, technology and humanities*. New York: Wiley.

Goldberger, M.L, Maher, B.A, and Flatteau, P.E. (1995). *Research-doctorate programs in the United States: Continuity and change*. National Academy Press.

Lambert, L. and Butler, N. (2006). *The Future of European Universities: Renaissance or Decay?*. London: Centre for European Reform.

Matia, K., Amaral, L.A.N., Luwel, M., Moed, H.F., Stanley, H.E. (2005). Scaling phenomena in the growth dynamics of scientific output. *Journal of the American Society for Information Science and Technology,* 56, 893-902.

Moed, H.F. (2005). *Citation Analysis in Research Evaluation.* Dordrecht: Springer.

Moed, H.F. Bibliometric Rankings of World Universities. CWTS Report 2006-01. Retrieved January 8, 2007 from: http://www.cwts.nl/hm/bibl_rnk_wrld_univ_full.html.

SJTU (2005). Academic Ranking of World Universities -2005. Shanghai Jiao Tong University, Institute of Higher Education. Retrieved January 8, 2007 from: http://ed.sjtu.edu.cn/rank/2005/ARWU2005Main.htm.

THES (2005). World University Rankings. Who is Number One? The Times Higher Education Supplement, February 24, 2005; April 29, 2005; May 2, 2005. Retrieved January 8, 2007 from: http://www.thes.co.uk/worldrankings/.

Van Raan, A.F.J. (2005). Challenges in Ranking of Universities. Retrieved January 8, 2007 from: http://www.cwts.nl/cwts/AvR-ShanghaiConf.pdf.

Mapping Institutions and Their Weak Ties in a Research Specialty: A Case Study of Cystic Fibrosis Body Composition Research[1,2]

Liying Yang*, Steven A. Morris** and Elizabeth M. Barden***

*yangly@mail.las.ac.cn
National Science Library of Chinese Academy of Science 33 Beisihuan Xilu, Zhongguancun,Beijing 100080, (P.R.China)

**steven.morris@bakerhughes.com
Baker-Hughes Inc. 2001 Rankin Road, Houston, Texas 77073（USA）

***embarden@comcast.net
Barden Consulting, 22 Federation Rd., Bedford, NH, 03110 (USA)

Abstract
Using a collection of papers gathered from the Web of Science on the topic of cystic fibrosis body composition research, we demonstrate analysis and visualization methods that show the collaboration structure of institutions in the specialty and the researchers that function as weak ties among them. Institution names were extracted from the collection of papers and disambiguated using the Derwent Analytics (v1.2) software product. Institutions were clustered into collaboration groups based on their co-occurrence in papers. A crossmap of clustered institutions against research fronts, which were derived using bibliographic coupling analysis, shows the research fronts that specific institutions participate in, their collaborator institutions and the research fronts in which those collaborations occurred. A crossmap of institutions to author teams, derived from co-authorship analysis, reveals research teams in the specialty and their general institutional affiliation, and further identifies the researchers that function as weak ties and the institutions that they link. This case study reveals that the techniques introduced in this paper can be used to extract a large amount of useful information about institutions participating in a research specialty.

Keywords
institutional collaboration; mapping; visualization; cystic fibrosis body composition

Introduction

A research specialty is a self-organized social organization, which can be described by the different features associated with it: 1) a research paradigm, 2) knowledge structure, 3) personnel, 4) institutions, 5) specialized vocabulary, 6) collaboration structure, 7) research output, 8) and domain journals. Institutions provide the nurturing environment and infrastructure for researchers and play an important role in the evolution of a specialty. The mapping of institutional activity in a specialty will be helpful for decision making by policy makers and research funding agencies.

In previous examples of mapping institutions, research teams, derived from co-authorship patterns, are used to reflect the collaboration pattern among institutions (Nagpaul, 2002; Kretschmer *et al*, 2005; Havemann, *et al*, 2006). Katz and Martin (1997) described the distinction between collaboration and co-authorship, and pointed out that the two are not necessarily equivalent. Due to many problems experienced when extracting institutions, few researchers have attempted to map institutions from journal literature. As a notable exception, Borner, *et al* (2006) described techniques for extracting and normalizing institution names based on co-citation links, and analysed the diffusion of scholarly knowledge among major U.S. research institutions.

In recent years, social network theory was introduced in bibliometric study (Leydesdorff, 1998; Debackere & Clarysse,1998; Kretschmer & Aguillo, 2004). Network theory focuses on organizations'

[1] This work was partially supported by the Ministry of Science and Technology of China (Grant 2004CCC00400).
[2] The author thanks Professor Jin Bihui of the Library of Chinese Academy of Sciences for her valuable help and encouragement.

and authors' knowledge interaction and communication, and views social relationships in terms of nodes and ties (Wasserman, *et al*, 1994). Nodes are the individual actors within the networks, and ties are the relationships between the actors. Granovetter (1973) discriminated between 'strong' ties that occur among close co-workers, and 'weak' ties that occur between individuals that interact infrequently. He noticed that by using weak ties, individuals could communicate beyond their well defined social boundaries.

The concept of weak ties in social networks was extended to the study of scientific specialties by Chubin (1976), who conjectured that an important class of weak tie researchers consists of marginal workers with no incentive to pursue research within existing paradigms. In addition, Bjorneborn and Ingwersen (2001) noticed the weak tie phenomenon in webometrics perspectives. Wagner (2005) argued that weak ties, evidenced by geographically remote collaboration, can promote new knowledge creation.

Detecting and visualizing researchers that function as weak ties between institutions will be helpful for demonstrating communication of knowledge, and may further identify the structure of invisible colleges. Using the coauthorship network of a specialty, it is possible to use graph layout software such as Pajek (Wouter, *et al*, 2005) or UCINET[3] to visualize author teams and weak ties between teams. On such a map, teams appear as closely co-located clumps of authors with dense intra-team link networks, while weak ties appear as authors at the ends of inter-team links. This visualization technique is difficult to use in practice because the denseness of author labels interferes with easy interpretation of the plots.

In this article, we will describe techniques to extract and clean up institutions in a collection of papers covering a specialty. The extracted institutions were used to build a paper to institution matrix for analysis and mapping of institutional collaboration. We apply timeline and crossmapping techniques to visualize the collaboration structure of institutions, show institutional links to individual researchers and research teams, and most importantly, show new techniques for visualizing weak ties in a specialty. We demonstrate these techniques through a case study of cystic fibrosis body composition, a medical specialty. A subject matter expert (SME) was used in our study to guide data collection and to validate analysis and visualization results.

Experimental Technique

Collection of data

The source data for this study was gathered on August 7, 2006 from ISI/Thompson's Web of Science (WOS) product, using the following topic query: *"cystic fibrosis" AND ("nutritional status" OR "body composition" OR "fat mass" OR "fat free mass" OR "lean body mass" OR "body cell mass" OR "bone mineral content" OR "total body water" OR "bone mineral density" OR "body density" OR "dual energy x-ray absorptiometry" OR "bioelectrical impedance" OR "air displacement plethysmography" OR "anthropometry")*. The query was applied against publication years from 1900 to 2006. This query was developed by our SME to retrieve papers on the specialty from the PUBMED archive for an earlier contract study.

Description of data

The downloaded source data consists of Web of Science (WOS) records in ISI tagged file format, containing records for 622 papers, with 22257 citations to 12411 references, and 3164 authorships by 2032 authors. Publication dates of the papers range from 1978 to August, 2006.

Extraction of institutions

The task of extracting institutions from the WOS source file is critical for successful analysis of participation of institutions in a research specialty. To achieve the desired results, researchers face two main obstacles to successful extraction of institutions: 1) initial extraction of institution names, and 2)

[3] Analytic Technologies, P.O. Box 920089, Needham, MA 02492, USA

disambiguation of institution names. Note that in WOS files the authors are not positively associated with their institutions (Katz & Martin, 1997), so every author and his or her corresponding institution cannot be directly matched.

To extract and disambiguate institution names, we employed *Derwent Analytics*[4]. The software can input collections of papers in WOS tagged file format and transform that data into a database configuration file. Fields, such as author field, affiliation field, and journal field are extracted from the raw file using pattern matching, rule-based extraction, and natural language processing techniques. Our institution extraction process consisted of the following steps:

1. Import of raw source file and extraction of institution name. In the raw WOS record file, the institution name is obtained from the author affiliation address, which is tagged "C1" in the WOS format. In each record each author may have more than one affiliation, and furthermore, if two or more authors in the paper have the same affiliation, that affiliation is only listed once in the paper's affiliation field. Derwent extracts one institution per C1 line as an 'unnormalized' institution. In our study, 558 unnormalized institutions were extracted from the source file.
2. Disambiguation of institution names. In many cases, different institutions might have the same name, or alternately, the same institution might have several different names. For accuracy, it is important to separate or combine such ambiguous institution names. Derwent Analytics relies on dictionary and matching rules to perform the combine function, but it cannot separate the different institutions that have the same name. In our case study, manual editing was required to correct 5 names (such as "Children's Hosp", "Hosp Sick Children" and so on) that Derwent Analytics could not correctly separate. This manual step required about 20 man-hours of labor.
3. Producing an institution-paper matrix. Based on the normalized institution list, Derwent Analytics was used to produce a paper to institution matrix. Institutions associated with five or more papers were retained for analysis, resulting in a matrix of 50 institutions to 622 papers.

Co-occurrence and similarity computation

Data was loaded into three matrices for analysis: 1) a paper to reference matrix, 2) a paper to paper author matrix, and 3) a paper to institution matrix. Cooccurrence counts were computed and converted to similarities using the Dice formula (Salton, 1989). Distances for clustering were computed by subtracting similarities from unity. All clustering analysis was done using hierarchical agglomerative clustering with Ward's method linkage. Three types of co-occurrence clustering were applied to the data:

1. Clustering of papers by common references, that is, bibliogrphic coupling clustering (Morris, *et al*, 2003). This type of analysis produces groups of papers that tend to use the same base knowledge, as represented by the common references that they cite. We shall refer to these groups of papers as 'research fronts' (Morris *et al*, 2003), although a longer and more appropriate term would be 'bibliometric research fronts'. For this analysis, those papers not having 5 or more bibliographic coupling counts to at least one other paper were discarded, leaving 461 papers for clustering. The resulting clustering hierarchy was truncated to produce 20 research fronts.
2. Clustering of paper authors by common paper, that is, co-authorship clustering. This type of analysis tends to produce groups of authors corresponding to research teams (Melin & Persson, 1996). In this analysis, authors having less than 5 papers were discarded, leaving 83 highly productive authors for analysis.
3. Clustering of institutions by common paper. This type of analysis tends to produce groups of institutions that collaborate on common research projects, or that have researchers with affiliations in both institutions, as will be explained below. In this analysis, institutions having less than 5 papers were discarded, leaving 50 highly productive institutions for analysis.

[4] Thomson Scientific, 3501 Market St., Philadelphia, PA., 19104, USA

Visualization

Visualization of research front structure, institution collaboration structure and author collaboration structure was conducted using timeline techniques (Morris, *et al*, 2003) and crossmapping techniques (Morris & Yen, 2004). These are matrix-based techniques that display the matrix of occurrence counts between entities of two different entity-types in the collection.

Figure 1. shows a timeline of research fronts in the collection of papers. Papers are plotted on the y-axis in horizontal tracks by research front, and on the x axis by publication date. The research front clustering dendrogram and research front numbers are shown on the left side of the plot, while the research front labels appear on the right side. Papers are plotted as circles whose size is proportional to the number of times the paper has been cited, and whose shading is proportional to the number of times the paper was cited in the final year of the collection. This allows easy identification of important papers, and currently 'hot' topics in the collection.

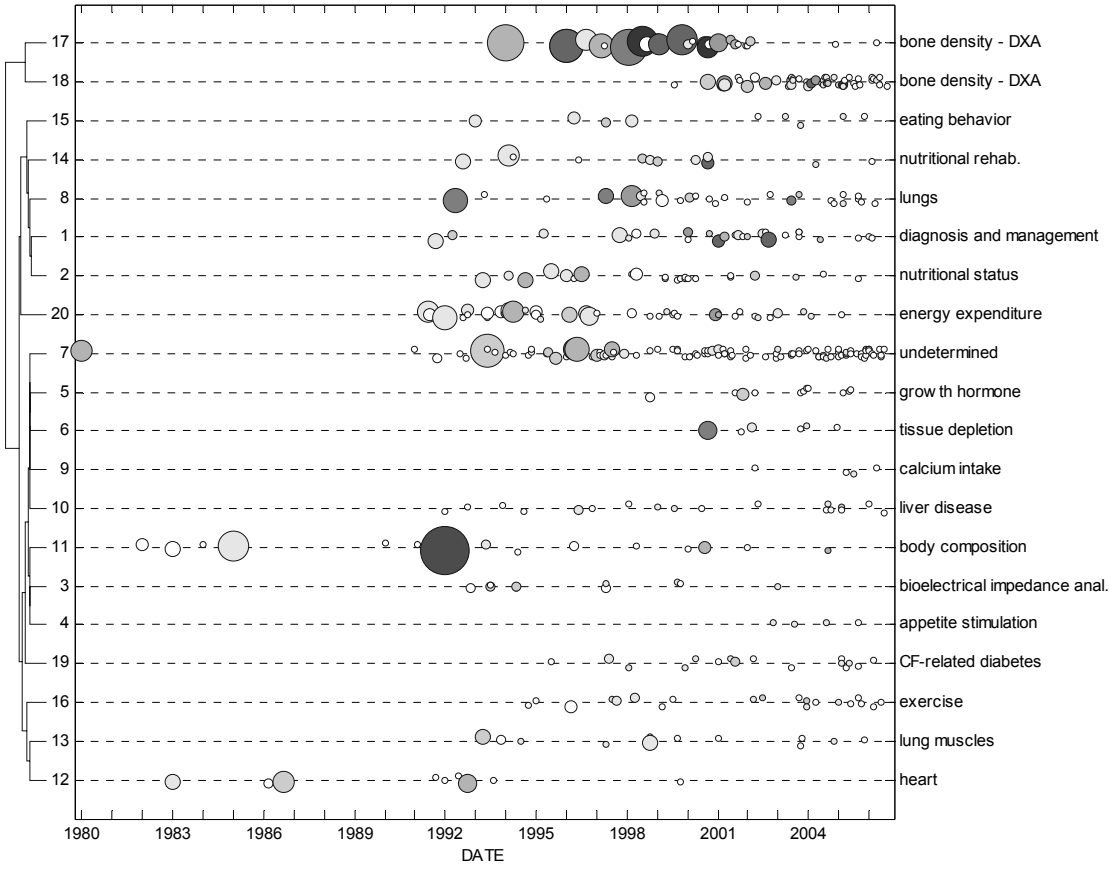


Figure 1. Timeline of research fronts in the CF body composition data set.

Figure 2 shows a crossmap of research fronts to institutions in the collection. The research fronts, on the y-axis, have the positions, research front numbers, labels, and clustering dendrogram taken from the timeline of Figure 1. The clustering dendrogram and institution numbers, which are arbitrarily assigned during clustering, are shown at the top of the plot. Corresponding institution labels are shown along the bottom of the plot. Given institution i on the x-axis, and research front j on the y-axis, the size of the circle at position (i,j) is proportional to the number of papers in research front j that are associated with institution i through author affiliation. This plot was constructed to visualize the relation of institutions to research fronts, and by inference the sub-topics, in the specialty. The plot also gives indications of each institution's research in terms of the number of research fronts it participates in, the productivity of each institution, and the collaboration structure of the institutions.

Figure 3 shows a crossmap of institutions against paper authors. The institution numbers, labels, and clustering dendrogram are the same as those from the y-axis of the research front to institution

crossmap of Figure 2. The paper author numbers, arbitrarily assigned during clustering, and the clustering dendrogram appear above the plot. The corresponding author names appear below the plot. Given author *i* on the x-axis, and institution *j* on the y-axis, the size of the circle at (*i*, *j*) is proportional to the number of papers to which author *i* and institution *j* were both associated. This plot was constructed to give visual indications of the productivity of individual authors, their relation to specific institutions and the breadth of author participation in terms of the number of institution that author is associated with. The plot also indicates the basic collaboration structure among the authors, which shows clearly research teams in the specialty. We also use this plot to visualize weak ties among institutions in the specialty.

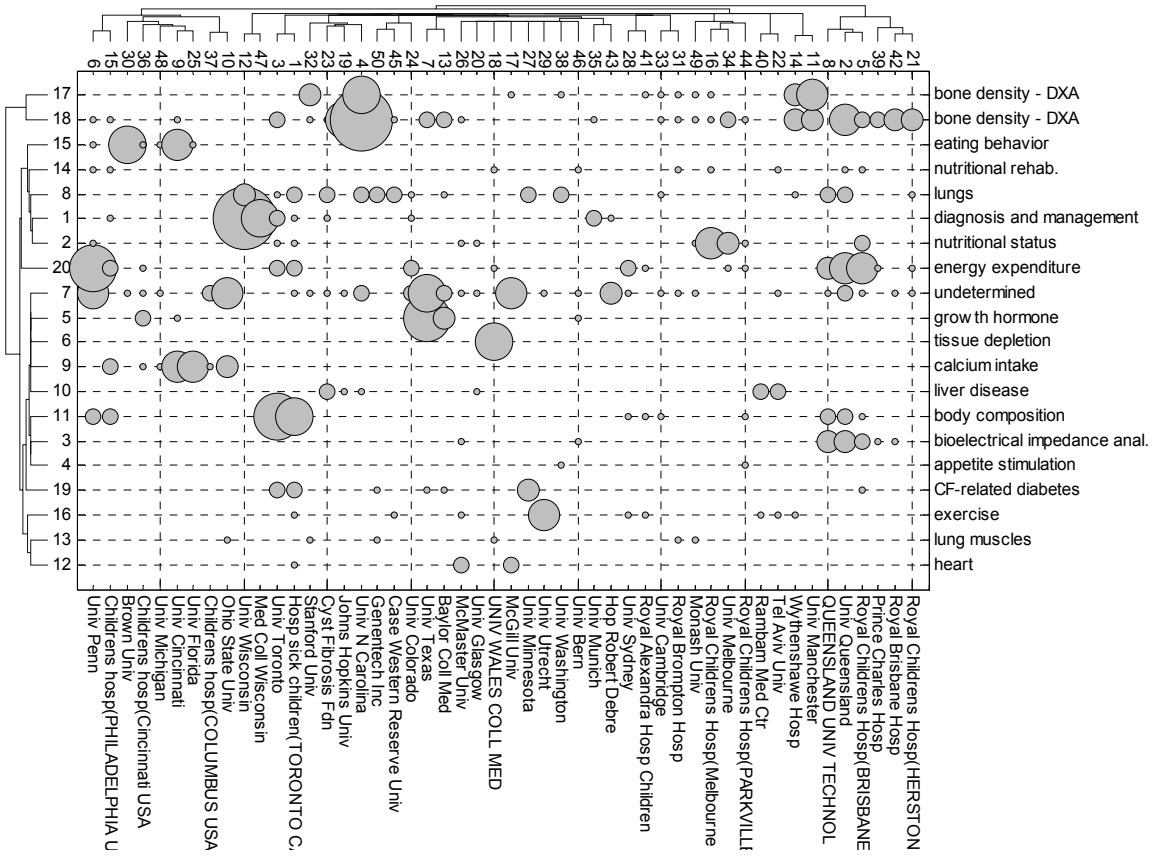


Figure 2. Crossmap of research fronts against institutions in the CF body composition data set.

Case Study: Cystic Fibrosis Body Composition

Validation with SME

Our subject matter expert, the third author of this paper, is a medical researcher with recent experience in the specialty, who has recently completed a review of the specialty's literature under contract from a university hospital. The SME is familiar with the specialty's research topics, key papers, leading researchers and leading institutions.

The visualization and analysis results of this study were presented to the SME for assessment of conclusions against expert knowledge. The SME was provided with the three visualizations of this paper (Figures 1, 2 and 3) and an automated report which contained summary data for each research front:

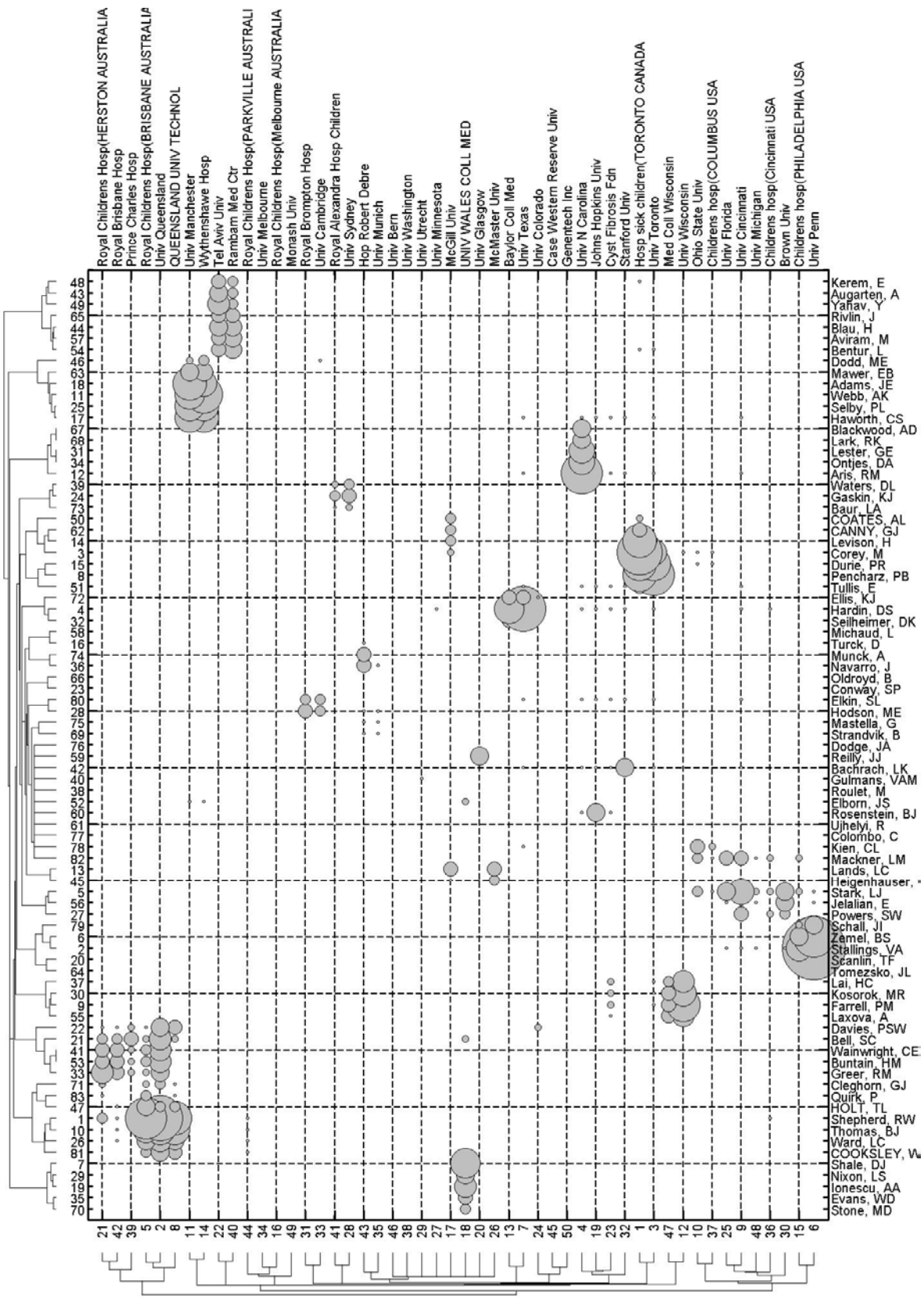


Figure 3. Crossmap of institutions against authors in the CF body composition dataset.

1. A list of papers in the research front, including authors, title, journal, published year, volume, and page.
2. A list of references cited, ranked by number of citations received.
3. A list of paper authors, ranked by number of papers authored.

The SME was asked to provide background on the specialty and to assess the analysis and visualization results for accuracy.

Background
Cystic fibrosis (CF) is a commonly occurring inherited disorder characterized by chronic pulmonary inflammation and gastroenterological abnormalities. Many children with CF experience growth retardation, malnutrition, and premature mortality due to respiratory insufficiency. Severe growth delay and poor nutritional status in terms of reduced height for age, low body weight for age, and pronounced deficits in lean body mass are predictors of increased morbidity and mortality. Nutrition intervention studies have suggested that improvements in nutritional status may mitigate declines in pulmonary status, yet understanding of the etiology of malnutrition and its role in disease progression in CF remains incomplete.

Classic studies in the 1980s demonstrated that diet therapies which resulted in an increase in body weight were associated with improvements in growth, clinical status, and survival. These findings, coupled with subsequent research into the nature of energy requirements and negative energy balance as well as improvements in pharmacological options, have contributed to considerable advances in the nutritional and clinical management of patients with CF (Pencharz & Durie 2000), so that life expectancy has increased into the third decade of life.

Research topics regarding body composition in CF tend to be divided into subcategories:

- Many researchers restrict their investigations according to age (pediatric versus adult).
- Some investigators focus on a given technique and its utility in characterizing body composition in healthy individuals and in those with various diseases.
- Some investigators focus on a given population (such as subjects with CF) and examine body composition in that context using multiple techniques.
- Research may be further restricted to the study of certain body composition characteristics, such as bone mineralization and density, or abnormalities in the distribution of extracellular fluid, and their relationship with morbidity or outcome.
-

Description of research fronts
Because of limited space in this paper, we will only cover notable highlights in the results. Research fronts are displayed in the timeline of Figure 1. Using the abbreviation RF for ‘research front’, note that RF 17 and RF 18 are papers concerned with bone mineral density. This group shows a stairstep pattern of papers with many highly cited early papers (large circles), and currently highly cited papers (red circles). When using timeline techniques, this stairstep characteristic is typical for ‘hot topic’ research fronts that develop after some discovery (Morris & Boyack, 2005). In this case, a new method of measuring bone density, dual x-ray absorptiometry (DXA), was developed and became available in the early 1990’s. Papers in RF 17 exploited this technique as well as earlier techniques that characterized bone health. Papers in RF 18 heavily cite papers in RF 17 and also heavily cite a 1999 study by Haworth which established by DXA that low bone density was common in a large group of patients with varying degrees of clinical severity in their lung disease. This finding, in conjunction with the accessibility of DXA and a growing population of adult patients with CF, has resulted in prolific research into bone health in patients with CF that is ongoing today.

Analysis of institutional participation in research fronts
Examining the research front to institution crossmap of Figure 2, the following interpretation is possible:

- Clustering and dendrogram seriation used in the crossmapping technique have placed the institutions into three general groups. The left half of the map, left to right from Institution 6 to Institution 13 corresponds generally to U. S. and Canadian institutions. The right half of the map, left to right from Institution 35 to Institution 21 corresponds generally to Australian and European institutions. The center, left to right from Institution 26 to Institution 46 corresponds to institutions that do not show evidence of collaborating with other institutions, that is, the institutions that have little similarity to other institutions based on co-authorship in papers. In our experience, such central placement of items with low similarity is a characteristic of the crossmapping technique.
- There is a distinct pattern of affiliation of universities with local hospitals, for example, the University of Pennsylvania, on the extreme left, is matched with co-located Children's Hospital of Philadelphia, which is located adjacent to U. of Pennsylvania on the crossmap. Several of these university-hospital close pairings are evident, particularly on the Australian/European side of the map. Our SME conjectures that these pairings reflect single researchers that are affiliated with both a hospital and a university medical school. An examination of the affiliation lists in selected papers generally verifies this conjecture.
- The following research fronts have dominant institutions:
    - RF 1, 'diagnosis and management' – U. of Wisconsin and associated hospital.
    - RF 15, 'eating behavior' – Brown U. and U. of Cincinnati
    - RF 2, 'nutritional status' – U. of Melbourne and associated hospital.
    - RF 5, 'growth hormone' – U. of Texas
    - RF 6, 'tissue depletion', U. of Wales
    - RF 9, 'calcium intake', U. of Cinncinnati, U. of Florida, and Ohio State U.
    - RF 16, 'exercise', U. Ultrecht
    - RF 20, 'energy expenditure', U. of Pennsylvania, U. of Queensland and associated hospital
    -

Analysis of team participation in institutions and weak ties

The institution to author crossmap of Figure 3 is used to investigate participation of teams in institutions and to further indicate which authors function as weak ties among the institutions. In this figure the institutions appear on the y-axis using the same dendrogram and serial order as they appeared on the research front to institution crossmap of Figure 2. This allows authors that correspond to particular institutions in the author-institution crossmap to be related back to research topics through the research front-institution crossmap. The following indications can be derived from the institution-author crossmap:

- In the upper left, a group of Australian authors appear as two collaborating teams: left to right authors 81 to 47 form a team drawn from U. of Queensland, Queensland U. of Technology, and Brisbane Royal Childrens Hospital. The authors from Brisbane Royal Childrens Hospital also participate with other Australian hospitals such as Prince Charles Hospital, but do not collaborate with either of the Australian Universities when doing so.
- A distributed group of researchers appears at the bottom left center of the map, consisting of authors drawn from the Eastern U.S. These include researchers at U. of Pennsylvania, the Children's Hospital of Philadelphia, Brown University (Rhode Island), U. of Michigan, U. of Cincinnati, U. of Florida, Children's Hospital – Columbus (Ohio), and Ohio State U. (Columbus). Our SME conjectures that this may indicate a group of authors that have participated in a multi-center trial, where patients from several institutions are pooled to produce a study with a sufficiently large number of subjects. Results from such a study could be authored by researchers from all participating institutions.

Note in the institution-author crossmap that authors, which correspond to columns on the map, typically have a large circle that presumably corresponds to their home institution, and several smaller circles that represent the institutions that the author has collaborated with. For example, Author 3, M. Corey, has a large circle corresponding to Toronto Hospital for Sick Children and also U. of Toronto, presumably the author's home institution. M. Corey also has collaboration links to McGill U., U. of

Wisconsin, Ohio State U. (in Columbus), and Columbus Children's Hospital. This is a good indication that Corey functions as a weak tie between her institutions and the other institutions listed. Thus, from this characteristic of the institution-author crossmap, it can be used to visualize and identify weak tie authors and the institutions that they connect.

Note in the institution-author crossmap that authors 42, 4, 51, 12, and 17 form a series of links to several institutions across the lower right of the map: U. of Texas, U. of North Carolina, Cystic Fibrosis Foundation, Stanford U., and U. of Cincinnati. This system of links corresponds to coauthorship of these authors and institutions of "Consensus statement: Guide to bone health and disease in cystic fibrosis" published in *Journal of Clinical Endocrinology And Metabolism* in 2005. This indicates that these researchers function as weak ties among their institutions. Our SME conjectures that a system of professional ties probably existed among these researchers prior to writing the consensus paper.

Conclusion

We have shown that Derwent Analytics software can be used to extract institutions from ISI files. Using extracted institutions, timeline and crossmapping techniques were applied to visualize the institutions' participation in the specialty. Using the author to institution crossmap, we investigated the detection of weak ties and their institutional connections.

In our case study, the SME validated that we have found the main institutions in cystic fibrosis body composition specialty; and that the timeline technique correctly illustrated that bone mineral density research exploiting DXA is a 'hot topic.' The detailed information about the collaboration structure among these institutions and the research fronts they participated in was described. In addition, the weak tie authors and their institutions were observed directly. The resulting visualizations were examined and interpreted by our SME, demonstrating their readability and practical value for researchers, policy makers and funding agencies.

Extracting institutions is a critical step in our study, though Derwent Analysis can do this work easily, it can not normalize institution names as well as we anticipated, particularly, it cannot distinguish the different institutions that have the same name and different addresses. In our study, we had to expend 20 hours of manual work to separate them, a process prone to manual errors.

The study presented here illustrates the need for 'smarter' institution extraction software. While the use of the crossmapping techniques for finding weak ties appears to be theoretically well-founded, the great abundance of researchers with multiple affiliations appears to have caused interpretational ambiguities, a fact that was immediately apparent to our SME when evaluating the visualizations. What is necessary is the development of institution extraction software that can reliably connect authors to their affiliations. This software will be difficult to develop, but will allow significant advances in bibliometric analysis of institutional collaboration in specialties.

References

Bjorneborn, L., & Ingwersen, P. (2001). Perspectives of webometrics. *Scientometrics,* 50(1), 65-82.

Borner, K., Penumarthy, S., Meiss, M., & Ke, W. M. (2006). Mapping the diffusion of scholarly knowledge among major us research institutions. *Scientometrics,* 68(3), 415.

Chubin, D. E. (1976). Conceptualization of scientific specialties. *Sociological Quarterly,* 17(4), 448-476.

Debackere, K., & Clarysse, B. (1998). Advanced bibliometric methods to model the relationship between entry behavior and networking in emerging technological communities. *Journal of The American Society for Information Science,* 49(1), 49.

Granovetter, M. S. (1973). The strength of weak ties. *American Journal of Sociology,* 778(6), 1360-1380.

Havemann, F., Heinz, M., & Kretschmer, H. (2006). Collaboration and distances between German immunological institutes. *Journal of Biomedical Discovery and Collaboration,* 1(1), 6-6.

Katz, J. S., & Martin, B. R. (1997). What is research collaboration? *Research Policy,* 26(1997), 1-18.

Kretschmer, H., & Aguillo, I. F. (2004). Visibility of collaboration on the web. *Scientometrics,* 61(3), 405.

Kretschmer, H., Kretschmer, U., & Kretschmer, T. (2005). Visibility of collaboration between immunology institutions on the web including aspects of gender studies. *Paper presented at the 10th International*

Conference of the International Society for Scientometrics and Informetrics (pp.750-760). Stockholm, Sweden : Karolinska University Press.

Leydesdorff, L. (1998). Theories of citation? *Scientometrics,* 43(1), 5-25.

Melin, G., & Persson, O. (1996). Studying research collaboration using co-authorships. *Scientometrics,* 36(3), 363-377.

Morris, S. A., & Boyack, K. W. (2005). Visualizing 60 years of anthrax research, *10th International Conference of the International Society for Scientometrics and Informetrics* (pp. 45-55). Stockholm, Sweden: Karolinska University Press.

Morris, S. A., & Yen, G. (2004). Crossmaps: Visualization of overlapping relationships in collections of journal papers. *Proceedings of the National Academy of Sciences of the United States,* 101(suppl. 1), 5291-5296.

Morris, S. A., Yen, G., Wu, Z., & Asnake, B. (2003). Time line visualization of research fronts. *Journal of the American Society for Information Science and Technology,* 54(5), 413-422.

Nagpaul, P. S. (2002). Visualizing cooperation networks of elite institutions in India. *Scientometrics*, 54(2), 213.

Pencharz PB, Durie PR (2000). Pathogenesis of malnutrition in cystic fibrosis, and its treatment. *Clinical Nutrition,* 19(6), 387-394.

Salton, G. (1989), Automatic *Text Processing: the Transformation, Analysis and Retrieval of Information by Compute.,* Reading, MA:Addison-Wesley, 1989

Wagner, C. S. (2005). Six case studies of international collaboration in science. *Scientometrics,* 62(1), 3-26.

Wasserman, S., Faust, K., & Iacobucci, D. (1994). *Social Network Analysis: Methods and Applications. Cambridge*: Cambridge University Press.

Wouter, D.N., Andrej, M. & Batagelj V. (2005) *Exploratory Social Network Analysis with Pajek*. New York,Cambridge University Press

Profiles of Technological Capabilities of the Chinese Academy of Sciences (CAS) –A Comparison of Patenting Activities of the CAS with other National Level Institutions[1]

Yang Zhiping, Fang Shu, Cheng Yunwei, Wang Chun,Wen Yi, Hu Zhengyi and Zheng Yi

{Yangzp, fangsh, chenyw, clas.ac.cn, wangc, weny, huzy, zhengy}@clas.ac.cn
The Chengdu Library of the Chinese Academy of Sciences, No.16 South Sec.2 Yihuan Road Chengdu 610041(P.R.China)

Abstract
The purpose of this study is to compare patent activities within the Chinese Academy of Sciences(CAS) with France National Center for Scientific Research(CNR) and Max Planck Society(MPS) which have the similar national role outside China based on patent statistic. The results suggest that CAS tended to file more patents in its home country, while CNRS and MPS showed the most interest in patenting aboard. The tendence of technology international markets for three institutions all was to Australia, United States, Europe and Japan. For technology priority areas, three institutions had the same preference in the pharmaceutical field. Comparable with other two institutions, CAS had less technology inventive and collaborative activities than other two institutions. Three institutions have presented patterns of cross-sectoral collaboration with industry and university. And CNRS showed the strongest technology cooperation capability in its home country.

Keywords
patent quantitative analyse; patent family; technology field; technology cooperation; national institution

Introduction

As a national research team, Chinese Academy of Sciences (CAS) has been being developed a sound scientific and technological innovation system and built itself into a scientific research base at advanced international level, and a base for promoting the development of China's high and new technology industries. In order to be able to reflect what the international level it is in the technology capabilities, there is a need to assess technology innovation capabilities for CAS.

Patent statistics have been used to analyse technology capability over the last 20 years.(Lionel & Pari 2004, Ashton W.B 1983). Current many studies on patent activities in China were focused on national patent analysis of technology strength and trend evaluation.(Yun Liu 2006, Bhattacharya S 2004). These were mostly focused on macro-level data. A few studies were addressed on China industrial fields. In additional, there has been few publications comparing the patent activities with those within other international institutions outside China. In this paper, we developed a quantitative analysis on patent activities of CAS, and compared patenting activities of the CAS with those of France National Center for Scientific Research(CNR) and Max Planck Society(MPS), which have the similar national roles in their countries with CAS. The aim of this study was to address the following questions:

- How does CAS compare to the other institutions having the similar national duty function in terms of patent activities?
- What are the patenting strategies of three institutions distributed in the world?
- What are the inventive features among three institutions, especially for CAS?
- What have been the technology accumulation and invention fields in the last fifteen years?
- What were patterns of cross-sectoral collaboration in technology innovation?

Date source and method

Date source and search strategy

The paper was based on patent datum collected in the Derwent Innovations Index (DII) which is available in ISI web of Knowledge produced by Thomson Company in USA. With regard to the retrieval strategy, search was carried out not only by using Derwent patent assignee codes, but also by

[1] * This paper was based on the part report of CAS annual patent analysis. Authors would like to thank Pari Patel for his insights and help of this study. And also thanks for the conference reviewer's comments.

adopting complex search queries. We selected all possible English names, synonyms as well as abbreviations of the CAS including its affiliated institutes and companies as retrieval terms. For CNRS and MPS, we also used their own assignee codes, CNRS and PLAC, and possible English name spellings in our search queries.

Data time and others

The period of Derwent patent basic year was researched from 1990 to 2004 in the collection data. Considering the patent data collection lagged by Derwent, we limited the database years from 1990 to 2005. In this article, such concepts as basic patent and basic year are applied for data statistics and analysis. According to the introduction of DII database, basic patent means the first member of a patent family, while this member is not necessarily published before all other members, normally a basic patent is the first patent record cited into Derwent World Patents Index(Derwent Aanlytics VantagePoint User Guide 2003). In this study, the Derwent Analytics software and other self-developed analysis software were used to analyse the patent data.

Comparison of patent families of three institutions

By comparing patent families, we can see the global picture of where three institutions have chosen to seek protection for an invention. Not only is it important from a legal point-of-view, but also can tell us a great deal about the owner's plans for the eventual marketing of the product（Patent Family Searching Using Derwent World Patents Index）

Fig 1 shows the annual changes in the number of patent families and of patent family members filed by three institutions from the basic patent year of 1990 to 2004. The ration of the number of patent family in three institutions was 13:3:1(CAS, CNRS, MPS). Not only had CAS the highest increasing rate among three institutions, but had it the largest number of patent family members. The difference for CAS between the number of patent families and its member was the lowest among three institutions.

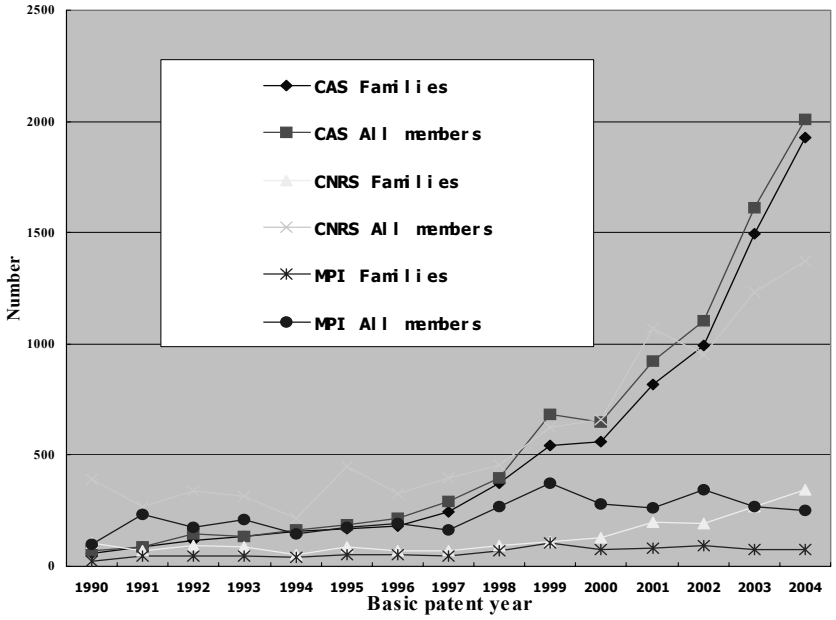


Figure.1 The evolution of patent family numbers of three institutions

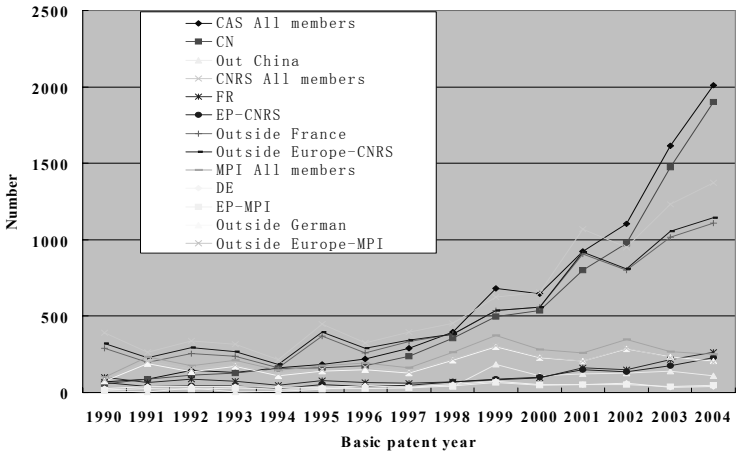


Figure 2. The Evolution of the number of patents in different applying countries or organizations: three institutions

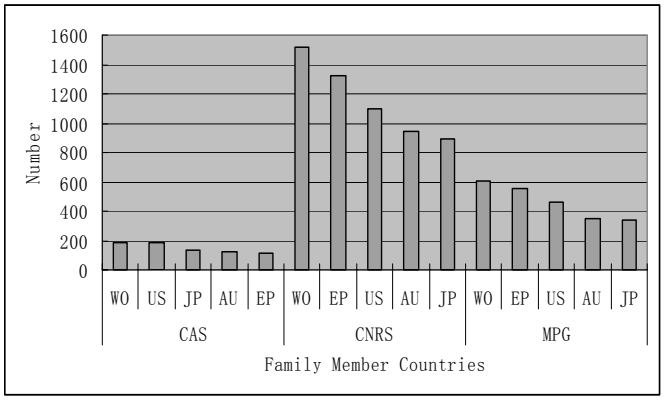


Figure 3. Distribution of Foreign filing patents of three institutions

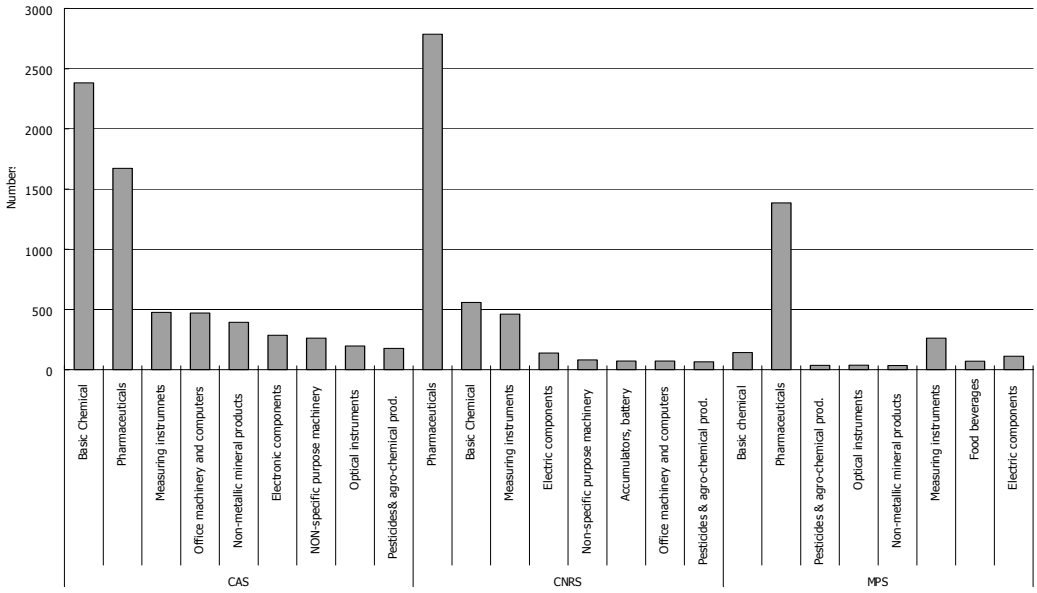


Figure 4. Distribution of technology fields according to Top 20 IPC with the largest patent number

Considering the 'home advantage' effect, we discussed the patent filing outside its own country. Fig 2 shows evolution of the number of patents filed in the world during 15 years. Fig 3 shows distribution of top rank countries (regional, organizations) of patent filed by three institutions, which exclude their home countries. The Top 5 filed countries or organizations in three institutions all were World Intellectual Property Organization, Europe, United States, Australia and Japan, with slight differences in the rank-ordering for three institutions. From them, we know that the number of filing foreign countries of CAS was relatively less than those by CNRS and MPS, although the total patent number by CAS was the largest. The phenomenon reflects that the patent strategic operation of CAS was more tend to file patent in its home country than those of CNRS and MPS. Moreover, The top rank order of patent filed CNRS was the same as MPS, which are were World Intellectual Property Organization (WIPO), European Patent Office (EPO), United States, Australia, Japan.

Analysing of distribution of technology activities of three institutions

Not only does patent classification analysis provide a better understanding of the current situation of a country's core technologies, but also helps to evaluate the potential of a specific or rising industry. (H. Y. LIN etc, 2003). Patent classification indexing, such as including International Patent Classification Codes (IPC) and Derwent Manual Codes (DMC),etc., have been widely studied in assessing the technology described in a large set of patent records(Technology Assessment Using Patent Literature------STN Europe). However, the IPC and DMC in the patent data don't directly correlate with the specific technology field. In this article, in additional to IPC and DMC studies, we also used the results of the project of "Linking technology areas to Industrial Sectors" studied by ISI OST and SPRU to correlate 43 technical fields to IPC(Ulrich, Francoise & Pari,2003).

Comparison of distribution of technology priorities

The IPC divides patentable technology into 8 key areas (http://www.bl.uk/ collections/patents/class.html). The rank according to the patent numbers of IPC can reflect technology priorities of one institution. Fig 4 shows distribution of technology fields of three institutions, according to TOP 20 patent numbers of IPC. The technology fields were named by the research results done by ISI OST and SPRU. In the Top 20, kinds of technology fields in patent filed by CAS were less than those by CNRS and MPS. Its technology priorities were mainly focused on the technology innovation of pharmaceuticals and basic chemical. Although CNRS and MPS also had the same activities, technology preference fields of two institutions were more diverse. Compared with CAS, both of CNRS and MPS tended to focus on the technology innovation of the measuring instrument and electronic components. In addition to, CAS had the same patent activities in the fields of non-metallic mineral products and food beverages as MPS.

Comparison of Share of patent countries in technology fields, according to TOP 10 IPC

Country or regional distribution of technology fields can reflect an institution intention on technology markets(Michael,2005). Considering the research result above, we here addressed how distribution of three institutions in pharmaceutical fields was. Table 1 shows distribution of selected foreign countries or organizations in pharmaceutical field for three institutions, according to TOP 20 patent numbers of IPC in specific a country (organization). In pharmaceutical, the most applying countries or organizations of CAS were EPO, Japan, Australia, World Intellectual Property Organization, United States, Japan, Korean and Brazile.

In ASIA, CNRS and MPS focused on patenting in pharmaceutical field, followed by measuring instruments, electrical components and basic Chemicals. CNRS and MPS respectively applied 81% and 84% of Chinese patents in Pharmaceuticals, according to TOP 20.

There is a strange phenomenon that numbers of France application patent by MPS was very low in our data collection. So we specially searched DII on Nov. 26. The result from the search year from 1963 to 1997 was 116 items. Here we could not give any conclusion on it.

Table.1 Distribution of pharmaceutical field according to Top 20 IPC with largest patent numbers

Patent codes	WO	EP	US	JP	CN	AU	DE	FR
CAS	79%	85%	66%	84%		80%	56%	19%
CNRS	82%	82%	82%	78%	81%	84%	73%	
MPS	79%	79%	77%	79%	84%	80%		

Analyzing the inventive features

DMC is applied to the inventive/significant features of the invention using the Documentation Abstract as the source document. The Codes are assigned by a team of Derwent analysts who have been specifically trained in the application of these codes(Derwent Analytics VantagePoint User Guide 2003). Here we use the codes to study the technology inventive features of three institutions. From the rank of TOP 20 technology inventive areas according to the number of DMC, the conclusion can be seen below:

- The similar inventive activities for three institutions were D05-H09(TESTING AND DETECTION)， D05-H12E(VECTORS (1994-)， D05-H12A(WILD-TYPE CODING SEQUENCES)，B04-E08(VECTORS, PLASMIDS, COSMIDS, TRANSPOSONS)，B14-H01(ANTICANCER GENERAL AND OTHER)， D05-H08(CELL OR TISSUE CULTURE)，S03-E14H(INVESTIGATION METHODS FOR BIOLOGICAL MATERIAL).
- Comparable with CNRS and MPS, CAS focused on 13 separate inventive activities. CNRS and MPS have more common inventive activities than CAS.

Table 2. TOP20 manual codes of CAS,CNRS and MPS by basic patent year, from 1990 to 2004

	CAS		CNRS		MPS	
	Manual Codes	**Total**	**Manual Codes**	**Total**	**Manual Codes**	**Total**
TOP1	J04-E04	280	D05-H09	397	D05-H09	253
TOP2	D05-H09	152	D05-H12E	249	D05-H12E	133
TOP3	B14-H01	131	D05-H12A	218	D05-H12A	131
TOP4	D05-H12A	128	B04-E08	213	B04-E08	91
TOP5	B11-B	115	*B12-K04E*	168	*D05-H11*	89
TOP6	H04-F02E	111	B14-H01	157	*B12-K04E*	81
TOP7	D05-H08	110	*D05-H11*	140	*B04-G01*	76
TOP8	D05-H12E	96	*B04-F0100E*	126	*B04-E01*	75
TOP9	E11-Q01	85	*B04-E01*	122	S03-E14H	73
TOP10	E11-Q02	81	*D05-H14*	116	B14-H01	73
TOP11	N06-E01	70	D05-H08	112	*B12-K04*	71
TOP12	A10-D	68	*B12-K04F*	111	*D05-H14*	71
TOP13	D05-H04	67	*S03-E14H*	108	*B04-N04*	71
TOP14	B04-A10	66	*B04-G01*	108	*B12-K04A*	69
TOP15	B04-E08	66	S03-E14H	106	*B04-F0100E*	62
TOP16	S03-E14H	62	*B04-N04*	102	D05-H12	58
TOP17	E31-N03	61	*B12-K04A*	95	*S03-E14H4*	54
TOP18	S03-E09C	60	B04-C01	94	D05-H16B	51
TOP19	D05-H13	58	D05-H07	91	D05-H08	50
TOP20	J04-E01	55	*B12-K04*	*89*	*B12-K04F*	49

Analysis of Patent collaboration of three institutions

Patent Co-assigneeship reveals the changing social structure of the conduct of technology innovation. It can reveal industry-academic institutions collaboration and the emphasis of inventive features.

Comparision of the Number and Change of Patent Co-assignees

Table 3 shows the change number of co-assignees of patents owned by three institutions. The increase number of patent co-assignees of every institution from 1990 to 2004 is more than ten times. The number of the co-assignees of CAS patents was smaller than those of CNRS and MPS patents. The number of the co-assignees of CNRS patents is as 4 times as that of CAS patents, and as 3 times as

that of MPS patents. So CNRS showed the strongest technology cooperation capability. The number of the co-assignees of MPS patents in each basic patent year was higher than that of CAS patents.

Table 3. The Number of the Co-assignees of patents owned by CAS, MPS and CNRS

Basic Patent Year	90	91	92	93	94	95	96	97	98	99	00	01	02	03	04	Total
CAS	4	3	7	6	1	15	38	31	18	54	44	61	139	81	52	464
MPS	6	22	15	21	22	27	25	18	55	51	47	101	168	117	76	535
CNRS	34	32	46	40	32	48	38	83	86	98	157	356	455	514	456	1604

(Note: Patent cooperation between institutes of CAS is not accounted, but cooperation with individuals is within. That is, all co-assignee organizations of CAS patents are NON-CAS organizations.)

Analysis of Top 10 Patent Co-assignees

- The ration of the number of top 10 Patent Co-assignees in three institutions was 1:13:7.7(CAS, CNRS, MPS). Besides, not only had CNRS the largest in total co-assignee patent numbers, but also it had the largest numbers in every co-assignee. The co-assignee number of CAS has the least among others. The categories of the top 10 co-assignees of three institutions include universities, corporations as well as research institutes. It presented that all have corporation with main technology innovation sectors.
- In top 10 patent co-assignees of every institution, common patenting inventive features are D05 (FERMENTATION INDUSTRY), B04 (NATURAL PRODUCTS (OR GENETICALLY ENGINEERED), POLYMERS), B14 (PHARMACEUTICAL ACTIVITIES). These also are the most inventive features of patent co-assignees of CAS. Compared with inventive features of CNRS and MPS patents with top 10 co-assignees, CAS showed the limited inventive features.
- In top 10 co-assignees of CAS, there are seven international organizations among CAS's top co-assignees which are DOKURITSU GYOSEI HOJIN RIKAGAKU KENKYUSH, JAPAN SCI & TECHNOLOGY CORP, SHIMADZU CORP, YAMANOUCHI PHARM CO LTD, TORAY IND INC, KENT RIDGE DIGITAL LABS, and UNIV INST SYSTEMS SCI NAT SINGAPOR. We may conclude CAS are more than happy to cooperate with foreign companies(especially Asia companies) in technology fields, such as fermentation industry, natural products and pharmaceutical.
- Top 10 co--assignees of CNRS are COMMISSARIAT ENERGIE ATOMIQUE,INSERM INST NAT SANTE & RECH MEDICALE, INST PASTEUR,UNIV PASTEUR LOUIS,ECOLE SUPERIEURE CACHAN LESIR,INRA INST NAT RECH AGRONOMIQUE,UNIV MONTPELLIER II,RHONE-POULENC RORER SA (AVENTIS PHARMA SA), INST ROUSSY GUSTAVE. It showed that home country sector was a key collaborator of inventive features with CNRS. Besides, CNRS cooperated with almost every assignee in natural products fields，pharmaceutical and fermentation industry. And it showed most activated in cooperation among others.
- Top 10 co-assignees of MPS are BASF AG, BAYER AG, DEUT KREBSFORSCHUNGSZENTRUM, PLANTTEC BIOTECHNOLOGIE GMBH, FRAUNHOFER GES FOERDERUNG ANGEWANDTEN, SUGEN INC, SONY INT EURO GMBH, YISSUM RES & DEV CO, UNIV GEORG AUGUST GOETTINGEN, REGENERON PHARM INC. There were six companies (institutions or universities) belong to home country. Others are USA, Japan and Israel. MPS has cooperated with almost every top10 co-assignees in fermentation, except with BASD AG.

Conclusion

The numbers of patent application filed by CAS have increased significantly during last fifteen years. Although CAS had the largest numbers of patent application, a large portion of them was Chinese patent. Its foreign patent member ratio to all members (number) was lowest among three institutions. CNRS and MPS had strategic target in foreign technology market. Besides, CAS showed international

industry-university corporation activity. But the number of the co-assignees were small. Thus, CAS needs to take some measures on patent policy making in future to focus on filing foreign patens in the best priority application countries. Also CAS needs to improve their technology innovative capabilities to get the international level in technology innovation, which can be achieved by technology cooperation and its technology innovation field expanding and priority.

References

Lionel Nesta and Pari Patel.(2004): National patterns of technology accumulation:use of patent statistics,Handbook of quantitative science and technology research:the use of publication and patent statistics in studies on R&D systems

Ashton, W.B., Campbell, R.S. etc, Patent analysis as a technology forecasting tool, *Economics and Policy Analysis Occasional Paper No.60,* Richland.

Yun Liu, Guo-Ping Cheng, Yu Yang; (2006): Patent applications of the top 500 foreign investment corportations in China. *Scientometric*, 68(1), 167-177

Bhattacharya S.(2004).Mapping inventive activity and technological change through patent analysis: A case study of India and China; *Scientometrics*, 61(3),361-381

Thomson Derwent(2003) Derwent Analytics Vantagepoint User Guid.

The Dialog Corporation. Patent Family Searching Using Derwent World Patents Index. Retrieved Feb.24,2007 from: http://support.dialog.com/searchaids/dialog/pdf/dwpi_fam.pdf

H. Y. LIN, H. W. LIN(2002), From imitation to innovation: the study of patents from Taiwan. *Taiwan Economic,Research Monthly*, 25 (5), 23-35.

Fiz Karlsruhe-STN Europe, Technology Assessment Using Patent Literature, Retrived Nov. 16,2006 from http://www.cas.org/patents/techpatlit.pdf

Ulrich Schmoch, Francoise Laville, Pari Patel, Rainer Frietsch;(2003). Linking technology areas to industrial sectors-Final Report to the European Commission, DG Research

http://www.bl.uk/collections/patents/class.html

Michael O'keeffe.(2005).Cross comparison of US, EU,JP and Korean Companies patenting activity in Japan and in the Peoples Republic of China; W*orld patent information :27:125-134*

Aspden, H.(1983): Patent statistics as a measure of technology, *World Patent Information. 5 (3):170-173*

Astrid Bachmann.(1998).Profiles of corporate technological capabilities--a comparison of large british and German pharmaceutical companies, *Technovation,18 (10):593-604*

Choung, J-Y.(1995). 'Technological Capabilities of Korea and Taiwan: An Analysis Using US Patent Statistics', *STEEP Discussion Paper* No 26, November, SPRU, Brighton

Reinhard Haupt.(2005). Patent analysis of a company's technology strength, *Cactus:229-249*

National Science Foundation(2004). Science & Engineering Indicators,5-37~60

An Analysis of the Connection between Researchers' Productivity and their Co-authors' Past Attributions, Including the Importance in Collaboration Networks[1]

Fuyuki Yoshikane[*], Takayuki Nozawa[*], Susumu Shibui[*] and Takafumi Suzuki[**]

[*]*fuyuki@niad.ac.jp, nozawa@niad.ac.jp, shibui@niad.ac.jp*
Faculty of University Evaluation and Research, NIAD-UE (National Institution for Academic Degrees and University Evaluation), 1-29-1 Gakuen-nishimachi, Kodaira, Tokyo 187-8587 (Japan)

[**]*qq16116@iii.u-tokyo.ac.jp*
Graduate School of Interdisciplinary Information Studies, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033 (Japan)

Abstract
Many studies have analyzed the "synchronic" correlation of properties, i.e., the correlation of properties in the same period, between authors and their co-authors. However, the "diachronic" correlation of properties, i.e., the correlation between their subsequent and precedent activity, has not yet been sufficiently studied using quantitative methods. This study pays attention to not only productivity but also the importance in the collaboration network as a measure of the researcher's activity, and clarifies whether there is any connection between (i) the researcher's activity subsequent to a collaboration and (ii) the collaborator's precedent activity, aiming at deriving knowledge about the diachronic effect of collaborators. The targets analyzed in this study are the newcomers who had first appeared in the domain of computer science in 1998 and their co-authors. The results show very little correlation between the subsequent activity of newcomers and the preceding activity of their co-authors. However, there are significant differences in the co-author's precedent activity between the newcomers who have published no papers after their first publication (those who have not become visible researchers) and the newcomers who have published at least one paper after then (those who continue to produce papers).

Keywords
co-authorship network; research collaboration; research productivity; educational effect

Introduction

There have been a large number of studies that analyze the correlation of properties between authors and their co-authors, some focusing on the productivity of papers and others on affiliation, age, and sex, etc (e.g., Kretschmer, 1994; 1997; Kretschmer & Gupta, 1998; Kundra & Kretschmer, 1999; Bahr & Zemon, 2000). These studies mainly examined the synchronic correlation between these properties (both are of the same period). However, the diachronic correlation of properties, that is, the correlation between their subsequent and precedent activity, has not yet been sufficiently studied using quantitative methods. This study pays attention to not only productivity but also the importance in the collaboration (co-authorship) network as a measure of the researcher's activity, and clarifies whether there is any connection between the researcher's activity subsequent to a collaboration and the collaborator's (co-author's) precedent activity, aiming at deriving knowledge about the diachronic effect of collaborators.

In recent years, certain factors, such as the specialization of researchers and the growth of interdisciplinary fields, have caused researchers to collaborate (Cason, 1992; Andersson & Persson, 1993; Bird, 1997; Bordons & Gómez, 2000). Networking and communication through collaborating and publishing co-authored papers seem to have had very large effects on the research activity under such situations. In particular, for new researchers who have not yet established their research styles and methods due to their having just emerged into their academic field, collaborators (co-authors) are considered to have significant effects on their future careers as researchers. From the viewpoint of productivity, for instance, Bindon (1981), Long & McGinnis (1985), and Fonseca, Velloso, Wofchuk

[1]. This work was partially supported by Grant-in-Aid for Young Scientist (B) 18700244 (2006) form the Ministry of Education, Culture, Sports, Science and Technology, Japan, and we would like to show our gratitude to the support.

& Meis (1998) have described the influence of supervisors or mentors on their students. Long & McGinnis (1985) noted that the performance of mentors collaborating with their students increases the students' later publications and citations. Bindon (1981), on the other hand, noted that academic association with the staff of an industrial research institute during graduate research has a positive impact on student careers in industry.

For example, it serves following purposes to clarify the correlation between the researcher's activity "subsequent" to a collaboration and the collaborator's "precedent" activity; (1) to give information used in choosing one's collaborators, and (2) to predict the effect of education in training students as researchers (in domains where graduate students generally publish co-authored papers with their supervisors in the process of education). We examine the co-author's activity from the viewpoint of not only productivity but also the importance in the co-authorship network.

This paper is organized as follows. First we describe the target domain and the source data, and then, after explaining our methodology, we show and discuss the results of our experiments examining the diachronic correlation of properties between authors and their co-authors.

Data

We selected computer science as the target domain. Because, in this domain, research collaboration is very active, one should take into account collaboration networks in the analysis of researchers' performance. As we mentioned above, the effect of collaborators is considered to be presented especially to researchers in the early stages of their careers. So, we targeted those who had published a paper for the first time in a core journal of this domain as the first author, that is, newcomers who had just entered this domain as full-fledged researchers, and examined the correlation between the activity of the newcomers subsequent to their emergence and the activity of their collaborators precedent to the emergence of the collaborating newcomers.

In many cases of co-authorship, at least in computer science, the last author (the author whose name is listed last in co-authorship credits) plays the role of supervisor or mentor considered to have a large impact on the collaborators. For instance, Liang et al. (2001) and Liang, Liu & Rousseau (2004) reported typical name order patterns found in co-authored papers in which the student's (or younger researcher's) name preceded that of the supervisor (or elder researcher). Yoshikane, Nozawa & Tsuji (2006) also observed that the last author often has a special function in the collaboration network of computer science. Thus, this study focused on the last author as the most important collaborator.

The data used for measuring the productivity of the researcher and for observing the co-authorship network were extracted from the *SCI* (*Science Citation Index*) CD/DVD-ROM version, provided by Thomson Scientific. The *SCI* evaluates the quality of journals with qualitative criteria based on peer review and citation analysis (Testa, 2004) and selects only the core journals that satisfy the criteria. We regarded the journals selected by the *SCI* and classified in the category of "computer science, theory & methods" in the list of source publications arranged by subject category, which is given on the Thomson Scientific web site: http://www.thomsonscientific.com/cgi-bin/jrnlst/jlsubcatg.cgi?PC=K, as the core journals in the domain of computer science.[2] This category includes 37 journals (e.g., *Journal of Algorithms*).

The researchers to be analyzed are the newcomers who appeared in the domain of computer science (i.e., published their first papers in this domain) in 1998, and the last authors listed on those newcomers' first papers. This study defines a newcomer in 1998 as one who published a paper (the "document type" of which is "article" in the *SCI*) as the first author in one of the core journals in 1998 and had not published a paper in the same domain for the preceding 7 years from 1991 to 1997. Moreover, we selected the newcomers who published only one paper (co-authored paper) in 1998 in order to obtain homogenous samples. In order to observe those newcomers' activity subsequent to

[2]. The area of application is also active in computer science. However, as the core of computer science is considered to be the theoretical area, we chose the category of "theory & methods."

1998 and their co-authors' activity precedent to 1998, we extracted the bibliographic data of all of the papers published in the 37 journals between 1991 and 2005 (i.e., the 7 years preceding and 7 years subsequent to 1998) from the database.

It is necessary to identify authors' names, that is, to distinguish the same name for physically different persons and to integrate different names for the same person, when collecting data from bibliographic databases. For identifying authors' names, there are methods such as using information which serves to distinguish authors (e.g., affiliation) and integrating a pair of names which differ only in the existence of a middle name(s) (e.g., Hayashi & Tomizawa, 2006). However, in the *SCI*, authors do not necessarily correspond one-to-one with affiliations in credits. Besides, in some cases an author has two (or more) affiliations, and in others there is no description about affiliations. Unfortunately, descriptions in the *SCI* are not always sufficient to strictly identify authors. As for the most productive authors, we carefully checked the authors' affiliations, recognizing the above limitations. Furthermore, we integrated pairs of names which are distinguished only by the existence of a middle name(s) or by differences in the use of upper and lower case letters.[3]

Table 1. Basic quantities of the data.

NP	TA	DA	A_{av}	P_{av}	DA_{new}
29820	69240	34374	2.32	2.01	641

Table 1 shows the basic quantities of the data: the number of papers, *NP*; the total number of author tokens, *TA*; the number of different authors, *DA*; the average number of authors per paper, A_{av} (= *TA*/*NP*); the average number of papers per author, P_{av} (= *TA*/*DA*); and the number of newcomers in 1998, DA_{new}. The number of newcomers fulfilling the above conditions is 641.

Methodology

Indices

The number of papers published for 7 years subsequent to 1998 (from 1999 to 2005) is used as an indicator to observe the productivity of newcomers subsequent to their emergence into the domain. For counting the number of papers of newcomers, we adopt the complete count, the adjusted count, and the straight count. *CMP* (the complete count), *ADJ* (the adjusted count), and *FST* (the straight count) represent the number of papers *of newcomers* measured by each method, respectively. *ADJ*, which is normalized by the number of co-authors, is referred to examine the productivity considering the proportion of contribution in research collaboration, while *FST*, which is defined as the number of papers published as the first author, is referred to examine the productivity as the leader of research.[4]

On the other hand, as for the past activity of the newcomers' collaborators (the last authors in the newcomers' first papers), this study measures both the productivity and the importance in the co-authorship network for 7 years preceding to 1998 (from 1991 to 1997). For counting the number of papers of the collaborators, we adopt the complete count, the adjusted count, the straight count, and furthermore the last author count which we define as the number of papers published as the last author. *SCMP* (the complete count), *SADJ* (the adjusted count), *SFST* (the straight count), and *SLST* (the last

[3]. We processed 2,179 variants of names. Although it is possible that authors with a middle name and without it are different persons, we assume that the same first and family names, without regard to presence of a middle name(s), represent the same person in almost all cases. If more than one author with a different middle name(s) (e.g., W. B. Mao and W. H. Mao) matched an author without a middle name (W. Mao), we regarded each of them (W. B. Mao, W. H. Mao, and W. Mao) as different persons.

[4]. This is based on the assumption that the first author designs the whole research as the leader and plays the special role, which is different from other co-authors' roles. We can consider this assumption to be valid, at least to some extent, because there are guidelines endorsing it in computer science (Zobel, 1999).

author count) represent the number of papers *of the collaborators* measured by each method, respectively. *SLST* is referred to examine the productivity as the supervisor or coordinator.

The global structure of the network as well as direct relations between authors should be considered when observing the importance in the co-authorship network because information, knowledge, or connections spread beyond direct ties. We assume the following network model, for setting the indices to measure the importance of the newcomers' collaborators.

- Assuming directed graphs where the ties are oriented from co-authors to the first author for each paper.
- Assuming weighted graphs where the strength of co-authorship relations is taken into account.

As indices of direct relations, we use indegree SD_{in} and outdegree SD_{out} in this model, that is, the number of researchers who have published co-authored papers in which the target researcher is the first author and the number of researchers who have published co-authored papers as the first author with the target researcher.

As for indices of the importance in the global structure including indirect ties, we use SC_l and SC_f by which the importance as the leader and that as the co-operator (supporter) are measured, respectively. Yoshikane, Nozawa & Tsuji (2006) proposed SC_l and SC_f, modifying the HITS (Hyperlink-Induced Topic Search) algorithm (Kleinberg, 1998). The two indices are obtained by applying the following substitutions (followed by normalization).

$$SC_l(n_i) = \sum_{j=1}^{g} a_{ij} SC_f(n_j) \qquad (1)$$

$$SC_f(n_i) = \sum_{j=1}^{g} a_{ji} SC_l(n_j) \qquad (2)$$

where g represents the number of nodes in the network, that is, the number of researchers in the domain. a_{ij} represents the value in cell (i, j) of the adjacency matrix A of the co-authorship network, and is equal to the relationship strength of the tie oriented from n_j to n_i, that is, the number of co-authored papers where n_i is the first author and n_j is his co-author. The value of diagonal cells a_{ii} is 0.

Here, we assume the mutual dependency that "a researcher who assists important leaders plays an important role as the co-operator, and a researcher who organizes important co-operators plays an important role as the leader". In the formulas (1) and (2), by repeating recursive substitution, the global structure of the co-authorship network is reflected in the importance of each researcher.

This study refined the calculation of $SC_l(n_i)$ and $SC_f(n_i)$ by normalizing a_{ij} (the relationship strength) according to the number of authors of each co-authored paper, considering that the relationship among authors of a paper become less intimate as the number of co-authors increases. After assigning 1 to $SC_l(n_i)$ and $SC_f(n_i)$ as the initial value, we calculate $SC_l(n_i)$ and $SC_f(n_i)$ by recursively repeating substitution and normalization of vectors. (Substitution and normalization are repeated 10 times.)

Among the four indices considering the co-authorship network, SD_{in} and SC_l correspond to the importance as the leader (the first author), while SD_{out} and SC_f correspond to that as the co-operator (other co-authors). Also, SD_{in} and SD_{out} take into account only direct relations, while and SC_l and SC_f take into account the global structure of the network.

Analysis of Correlation

We examine the correlation between (i) the productivity of newcomers subsequent to their emergence into the domain and (ii) the activity of their co-authors precedent to it, with following two viewpoints.

6. Is there a positive correlation between the productivity of newcomers and the past activity of their co-authors?
7. Is there any difference in the co-author's past activity between the newcomers who have published no papers after their first publication (those who have not become visible researchers) and the newcomers who have published at least one paper after then (those who continue to produce papers)?

It is meaningful to examine with respect to not only the first viewpoint but also the second one when this analysis is applied, for example, to the prediction of educational effects. Even if a researcher has published only a few papers, he might be good in research quality. Therefore, in some cases, one makes much of the ability to continue to produce papers as a researcher regardless of the quantity of papers.

Regarding the first viewpoint, we calculate the correlation coefficients between the family of indices to evaluate the productivity of newcomers subsequent to their emergence into the domain {*CMP*, *ADJ*, *FST*} and that to evaluate the activity of their co-authors precedent to it {*SCMP*, *SADJ*, *SFST*, *SLST*, SD_{in}, SD_{out}, SC_l, SC_f}. As these indices are expected to follow not the normal distribution but the power law distribution,[5] we adopt the Spearman's rank-order correlation coefficient.

Regarding the second viewpoint, we compare the newcomers whose value of *CMP* equals 0 (*CMP* = 0) and those whose value of *CMP* is greater than 0 (*CMP* > 0), with regard to the mean value for each of indices to evaluate the activity of their co-authors {*SCMP*, *SADJ*, *SFST*, *SLST*, SD_{in}, SD_{out}, SC_l, SC_f}, and test the significance of the difference between the two groups by the Wilcoxon rank sum test.

Results and Discussion

Correlation for All

The results of calculating the correlation coefficients between the productivity of the 641 newcomers subsequent to their emergence into the domain {*CMP, ADJ, FST*} and the activity of their co-authors precedent to it {*SCMP*, *SADJ*, *SFST*, *SLST*, SD_{in}, SD_{out}, SC_l, SC_f} are shown in Table 2. All of the values of the correlation coefficients are between 0.02 and 0.09, that is to say, very little positive correlation is observed for all the combinations of the former and latter families of indices.

The number of papers as the first author for newcomers (*FST*) has "relatively" high correlation with the same type of index for their co-authors (*SFST*). The newcomer's productivity as the leader might be slightly affected by the same ability of his co-author. As for the number of papers, including all papers published as a co-author, of newcomers (*CMP* and *ADJ*), it is the importance as the co-operator (supporter) in the global structure of the collaboration network (SC_f) that has the highest correlation among the indices to evaluate the past activity of their co-authors.

However, the values of the correlation coefficients between them are not "absolutely" high.[6] Also, no distinct correlation is found in the scatter chart for each combination of the indices. The scatter charts (the bubble charts where the size of the bubbles expresses the frequency) of the number of papers of newcomers (the complete count), i.e., *CMP*, against the eight indices for their co-authors, i.e., *SCMP*, *SADJ*, *SFST*, *SLST*, SD_{in}, SD_{out}, SC_l, and SC_f, are shown as examples in Figure 1.

[5]This is expected from Lotka's law (Lotka, 1926), for example.

[6]The "synchronic" correlation after the emergence of newcomers, that is, the correlation between the newcomers' productivity from 1999 to 2005 and the co-authors' activity during the same period, is somewhat high (around 0.3). It seems to be due to the fact that some of newcomers have been continuing to collaborate and to publish co-authored papers with the same co-authors after then.

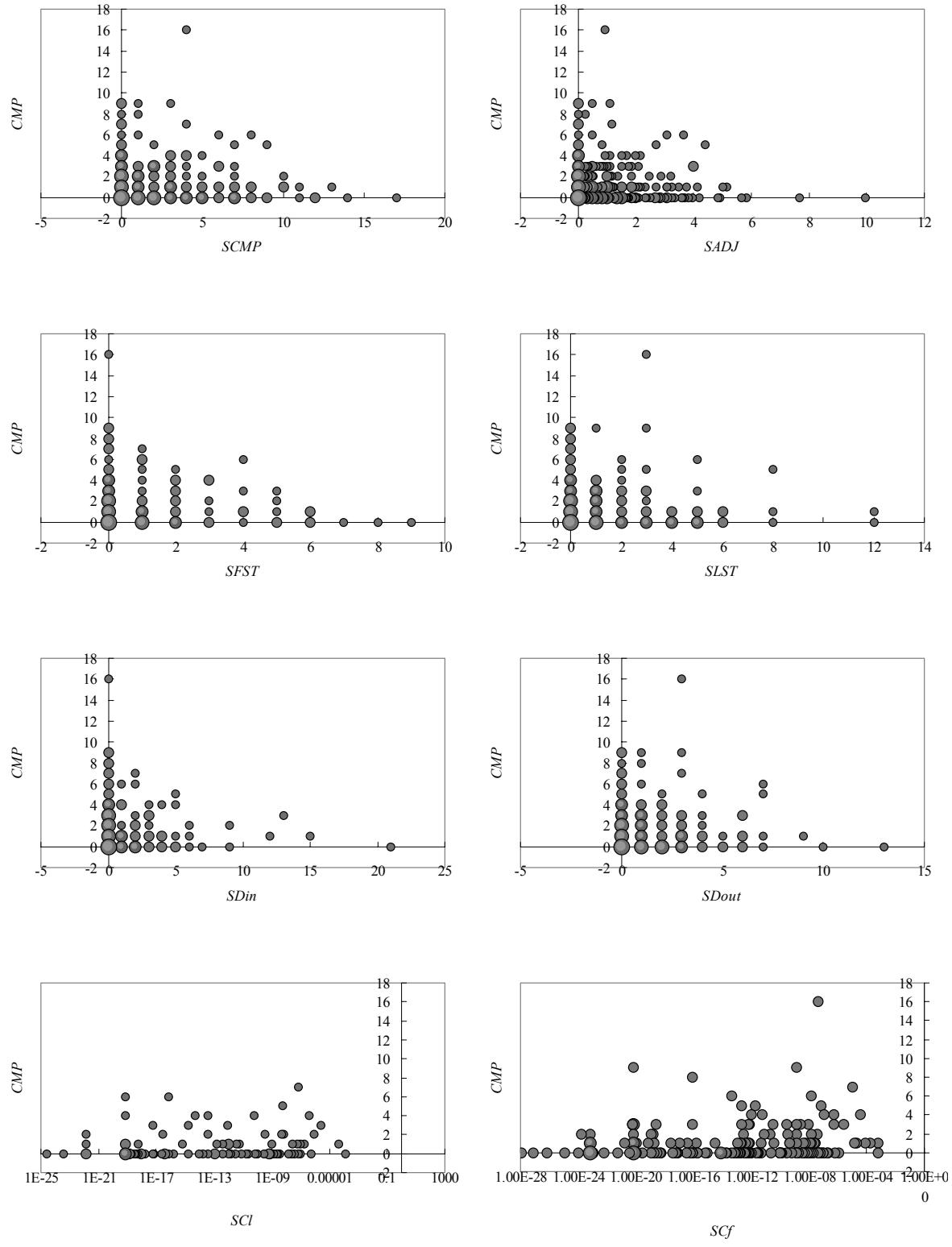


Figure 1. Bubble charts of the number of papers of newcomers against the indices for their co-authors.

Table 2. Correlation between the subsequent productivity of newcomers and the precedent activity of their co-authors.

	CMP	**ADJ**	**FST**
SCMP	0.085	0.083	0.079
SADJ	0.083	0.082	0.079
SFST	0.046	0.049	0.084
SLST	0.046	0.041	0.020
SD_{in}	0.020	0.020	0.050
SD_{out}	0.072	0.067	0.050
SC_l	0.021	0.020	0.051
SC_f	0.090	0.084	0.073

As regards other domains, for instance, Long & McGinnis (1985) reported a positive correlation between the productivity, i.e., the number of papers, of mentors and that of their collaborating students in biochemistry, while some studies reported that the productivity of mentors is not directly related to their teaching effectiveness (e.g., Rushton, Murray & Paunonen, 1983). The results of this study show that, in computer science, neither the researchers' productivity nor their importance in the collaboration network is necessarily related with the subsequent productivity of their collaborators. Computer science is a relatively new domain. It may be assumed that the effect of collaborators on newcomers is weaker in this domain than in more traditional domains, such as biochemistry.

Differences between the Two Groups

Among the 641 newcomers, 232 newcomers have produced at least one paper subsequent to their emergence into the domain ($CMP > 0$) while the remaining 409 have produced no papers subsequent to then ($CMP = 0$). Table 3 shows the correlation coefficients between the subsequent productivity of the 232 newcomers ($CMP > 0$) and the precedent activity of their co-authors. The correlation coefficients, with the exception of that between *FST* and SC_l, become lower if excluding the 409 newcomers who have produced no papers after their first publication ($CMP = 0$) in the calculation (compare Table 3 with Table 2). It can be supposed that the past activity of co-authors is related to whether newcomers do/can not produce a paper after their first publication, rather than to how many papers they produce.

Table 3. Correlation between the subsequent productivity of newcomers ($CPM > 0$) and the precedent activity of their co-authors.

	CMP	**ADJ**	**FST**
SCMP	0.020	0.007	0.051
SADJ	0.005	0.002	0.046
SFST	0.011	0.031	0.106
SLST	0.040	0.008	-0.016
SD_{in}	0.016	0.015	0.094
SD_{out}	0.047	0.011	0.023
SC_l	0.015	0.013	0.094
SC_f	0.077	0.034	0.061

Next, we compared the two groups of newcomers, i.e., the 409 newcomers who did not continue to produce a paper subsequent to their emergence into the domain ($CMP = 0$), and the 232 newcomers who did produce at least one additional paper subsequent to that point in time ($CMP > 0$), in terms of their co-authors' past activity. The means of the indices for each group are shown in Table 4. The co-

authors of the latter group show higher values for all indices except for SC_l. The value of the latter group is about 2.8 times as high as that of the former group for SC_f, and about 1.3 times for the other indices. In particular, for *SCMP* and *SADJ,* which represent the productivity where roles in co-authored papers (the first author, the last author, or other co-authors) are not taken into account, and for SC_f, which represents the importance as co-operators in the global structure of the networks, the co-authors of the latter group show significantly higher values than those of the former group ($p < .05$).

There are three indices which represent the level of activity and importance as the co-operator, that is, *SLST*, SC_{in}, and SC_f. Among them, only SC_f shows a significant difference between the two groups of newcomers. SC_f is the index that also reflects the importance of researchers assisted by the target researcher, and it represents "the achievements of the target researcher in assisting researchers who are actively performing research as the leader". It is implied that such achievements (or experiences) of researchers affects the productivity of their collaborating newcomers.

Table 4. Comparison of the two groups of newcomers in terms of their co-authors' past activity.

	CMP > 0	**CMP = 0**
SCMP	*1.51	*1.17
SADJ	*0.63	*0.52
SFST	0.49	0.38
SLST	0.70	0.59
SD_{in}	0.59	0.42
SD_{out}	0.88	0.65
SC_l ($\times 10^7$)	2.04	3.50
SC_f ($\times 10^6$)	*3.99	*1.43

*significant ($p < .05$)

Conclusions

This study analyzed the correlation between the productivity of newcomers subsequent to their emergence into a new domain and the precedent activity of their co-authors with the aim to derive knowledge about the effect of collaborators on their collaborating partners. Very little correlation was observed between the number of papers of newcomers and the past activity of their co-authors. On the other hand, whether the newcomers continued or ceased producing papers was related to the past activity of their co-authors. It is possible that, newcomers are helped to successfully start their careers as researchers by having the opportunity of touching the information, knowledge, and connections which their collaborators possess, though whether the newcomers can perform actively is mainly determined by other factors, such as their own effort and motivation (Frieze, Knoble & Mitroff, 1980), rather than the effect of their collaborators.

Some of the newcomers in this study may have been researchers who had temporarily moved from a related domain to the computer science domain searching for an opportunity for publication, and therefore they may not have intended to continue producing papers in the same domain. If we had excluded them from our data, then we could have possibly observed more clearly the characteristics of the newcomers who really intended to go on to careers in the computer science domain.

We focused on the productivity of papers in this study. The correlation from other viewpoints, such as the topics dealt with in papers, should be also examined in further studies in order to clarify the effect of collaborators in further detail.

References

Andersson, A. E. & Persson, O. (1993). Networking scientists. *Annals of Regional Science*, 27(1), 11-21.

Bahr, A. H. & Zemon, M. (2000). Collaborative authorship in the journal literature: Perspectives for academic librarians who wish to publish. *College & Research Libraries*, 61(5), 410-419.

Bindon, G. (1981). Output measures of cooperative research: The case of the Pulp and Paper Research Institute of Canada. *Scientometrics*, 3(2), 85-106.

Bird, J. E. (1997). Authorship patterns in marine mammal science, 1985-1993. *Scientometrics*, 39(1), 99-105.

Bordons, M. & Gómez, I. (2000). Collaboration networks in science. In B. Cronin & H. B. Atkins (Eds.), *Web of Knowledge: A Festschrift in Honor of Eugene Garfield* (pp. 197-213). Medford: Information Today.

Cason, J. A. (1992). Authorship trends in poultry science, 1981 through 1990. *Poultry Science*, 71(8), 1283-1291.

Fonseca, L., Velloso, S., Wofchuk, S. & de Meis, L. (1998). The relationship between advisors and students. *Scientometrics*, 41(3), 299-312.

Frieze, I. H., Knoble, J. M. & Mitroff, I. I. (1981). American university students' beliefs about success in science: A case study. *Scientometrics*, 3(2), 115-126.

Hayashi, T. & Tomizawa, H. (2006). Restructuring the Japanese national research system and its effect on performance. *Scientometrics*, 68(2), 241-264.

Kleinberg, J. M. (1998). Authoritative sources in a hyperlinked environment. In *Proceedings of 9th Annual ACM-SIAM Symposium on Discrete Algorithms* (pp. 668-677). New York: Association for Computing Machinery.

Kretschmer, H. (1994). Coauthorship networks of invisible-colleges and institutionalized communities. *Scientometrics*, 30(1), 363-369.

Kretschmer, H. (1997). Patterns of behaviour in coauthorship networks of invisible colleges. *Scientometrics*, 40(3), 579-591.

Kretschmer, H. & Gupta. B. M. (1998). Collaboration patterns in theoretical population genetics. *Scientometrics*, 43(3), 455-462.

Kundra, R. & Kretschmer, H. (1999). A new model of scientific collaboration Part 2: Collaboration patterns in Indian medicine. *Scientometrics*, 46(3), 519-528.

Liang, L., Kretschmer, H., Guo, Y. & Beaver, D. (2001). Age Structures of scientific collaboration in Chinese computer science. *Scientometrics*, 52(3), 471-486.

Liang, L., Liu, J. & Rousseau, R. (2004). Name order patterns of graduate candidates and supervisors in Chinese publications: A case study of three major Chinese universities. *Scientometrics*, 61(1), 3-18.

Long, J. S. & McGinnis, R. (1985). The effects of the mentor on the academic career. *Scientometrics*, 7(3), 255-280.

Lotka, A. J. (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences*, 16(12), 317-323.

Rushton, J. P., Murray, H. G. & Paunonen, S. V. (1983). Personality, research creativity, and teaching effectiveness in university professors. *Scientometrics*, 5(2), 93-116.

Testa, J. (2004). The Thomson Scientific journal selection process. Retrieved November 16, 2006 from: http://scientific.thomson.com/free/essays/selectionofmaterial/journalselection/.

Yoshikane, F., Nozawa, T. & Tsuji, K. (2006). Comparative analysis of co-authorship networks considering authors' roles in collaboration: Differences between the theoretical and application areas. *Scientometrics,* 68(3), 643-655.

Zobel, J. (1999). Guideline on Research Practice in Computer Science. Retrieved November 20, 2006 from: http://goanna.cs.rmit.edu.au/~jz/conduct.html.

A Bibliometric Analysis of siRNA Fundamental Research and Patent Activities

Weiping Yue*, Yuntao Pan**, Zheng Ma** and Xishan Wu**

*weiping.yue@thomson.com
Thomson Scientific, Beijing 100080, (P.R.China)

**panyt@istic.ac.cn, mazheng@istic.ac.cn, wuyishan@istic.ac.cn,
Institute of Scientific & Technical Information of China, Beijing 100038 (P.R. China)

Abstract
In recent years, siRNA technology and application have attracted much attention in life sciences. This paper intends to explore the dynamics of the interaction of siRNA fundamental research and its technology application. Publication and patent data on siRNA were collected from *ISI Web of Science* and *Derwent Innovation Index*. *Thomson Data Analyzer* was used to do the bibliometric analysis. The analytical result shows a steady growth in outputs in the form of patents and publications since 2001. Top organizations and clusters on main siRNA research and application are identified. Existing links between fundamental research and patenting activities are also found from the comparison study.

Keywords
siRNA; Thomson Data Analyzer; patent analysis; factor map

Introduction

The Nobel Prize in Physiology or Medicine for 2006 was jointly awarded to Dr. Fire and Dr. Mello for their discovery of RNA interference (RNAi) mechanism (Nobel Foundation, 2006). RNAi has enormous potential and has already found applications in determining gene function, pathway analysis, identifying and validating targets, functional screening, and gene redundancy screenings. As key to RNAi technology, short interfering RNA (siRNA) is a relatively new technique, but in recent years its applications have attracted much attention in life sciences (Dorsett & Tuschl, 2004).

This paper aims to explore the dynamics of the interaction of siRNA basic research and its technology application. Publication and patent data of siRNA were collected from *ISI Web of Science (WoS)* and *Derwent Innovation Index (DII)*. *Thomson Data Analyzer (TDA)* was used to do the bibliometric analysis and to further answer the When, What, Who, Where questions.

Methods

Literature records on siRNA were obtained by searching the following keywords in the field of *WoS* Abstract, Title, Keyword and Keyword Plus: "small interfer* RNAs" OR "small interfer* RNA" OR siRNAs OR siRNA OR "short interfer* RNA" OR "short interfer* RNAs". Patent data were gathered from *DII*. The same keywords used in *WoS* were also applied to the patent search in *DII*. Moreover, in order to improve the level of precision and accuracy, Derwent manual code B04-E07C that represents the concept of siRNA was also used in searching *DII*.

TDA, a powerful text mining tool to analyze data from a variety of rich information sources, was applied to analyze the literature data and patent data on siRNA. In addition to its flexible importing and analyzing abilities, *TDA* also provides advanced data cleaning functions to minimize the effects of differences in spelling words, technology terms, organizations, and even countries. In this study, the field of patent assignee, IPC code, Derwent class, publication country, and author keywords have been cleaned before doing further analysis. Analytical functions such as List, co-occurrence, factor map, technology reporting were used to find the trends and technological landscape in the field of siRNA.

Results

There are in total 7,559 literature records were retrieved from *WoS* and 2,192 patent records found in *DII*. The patent/literature ratio 29% shows relatively strong patent activities in the applied field. The siRNA literature started in 2001, peaked in 2006 at 2550 papers, demonstrating a strong growth model.

Likewise, a similar trend in patent literature growth can be found, implying that the fundamental research has been quickly transferred to applied research as measured by patent protection (Figure 1). The siRNA patent family year started from 2002 and also peaked in 2006 at 1273 records, illustrating a steady increase continuing to the present day. The comparison data in Figure 1 confirms that the fundamental research is feeding through to applied research and innovation (Tansey & Stembridge, 2005).

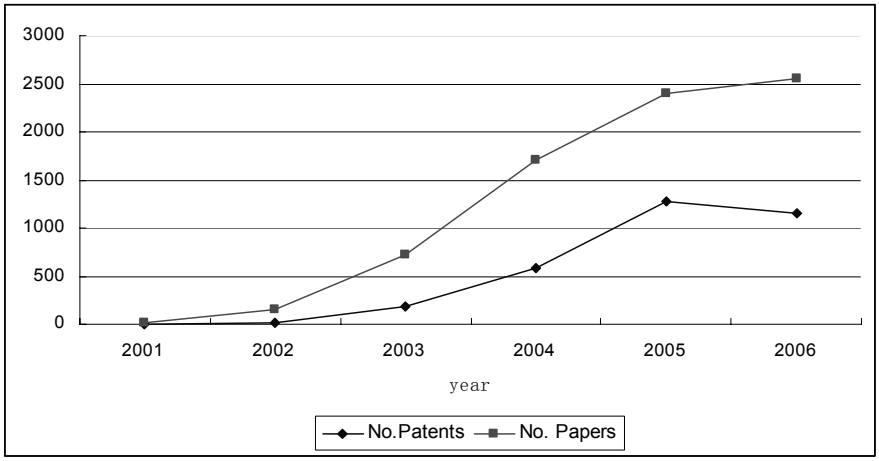


Figure 1. Comparison of the number of papers vs. patents on siRNA.

Where is research/patent activity being conducted?
The publication countries of 7559 papers included 62 different countries, which represents siRNA research has been expanded in many countries. However, much of the fundamental research as measured by scientific publications is being done in USA (3789 out of 7559 papers). Japan (868) and Germany (607) ranked second and third. China ranked the fifth with a total number of siRNA publication 372. Likewise, the siRNA patenting activity is predominantly in the USA (1565 out of 2192 patents) as well. Japan with 205 patents also ranked second. The above data demonstrate that the steady acceleration both in basic research and patenting activity are all driven by US.

Who is conducting research/patenting?
As we may expect, much of the fundamental research in siRNA is being conducted by the major US universities. It is found that 86% of USA publications are from the top three ranked universities: Univ Texas (247), Harvard Univ (232), and Univ Penn (110). Univ Tokyo with 186 publications also shows a strong research profile on siRNA. For patent activities, USA companies and universities also dominate patenting of siRNA technology. The top four patent assignees are all from USA. SIRNA THERAPEUTICS INC with 113 basic patents ranked first, followed by ISIS PHARM INC, UNIV CALIFORNIA and UNIV MASSACHUSETTS. Univ Tokyo with 36 patents ranked the fifth.

A close look shows that there are 88 organizations that have both patents and scientific publications in the field of siRNA. Table 1 shows the detailed fundamental research and patent activities of the top 10 organizations in terms of the number of patents. Most of them are universities and research institutes.

siRNA is a rapidly growing field. In the last three years (2004-2006), more than 2000 institutions stepped in the field and during the same period about 160 organizations no longer published on siRNA. Similar trend is also found in the applied technology area. However, compared to the fundamental research area, the number of new entrants (approximately 550 organizations) in patent activity is much lower than the one in basic research. We may project that in the future there will more organizations transfer their basic research outcomes into patents.

Table 1. Organizations having both fundamental research and patent activities in siRNA.

Author Affiliation/Patent Assignees	No. Patents	No. Papers	Patent/literature ratio
Univ Tokyo	36	186	19.4%
Univ Massachusetts	36	93	38.7%
Univ Michigan	7	54	13.0%
Scripps Res Inst	5	53	9.4%
Cold Spring Harbor Lab	7	49	14.3%
Univ Chicago	6	48	12.5%
Univ Pittsburgh	4	42	9.5%
Univ Rochester	10	34	29.4%
Univ Washington	9	34	26.5%
Univ Minnesota	1	33	3.0%

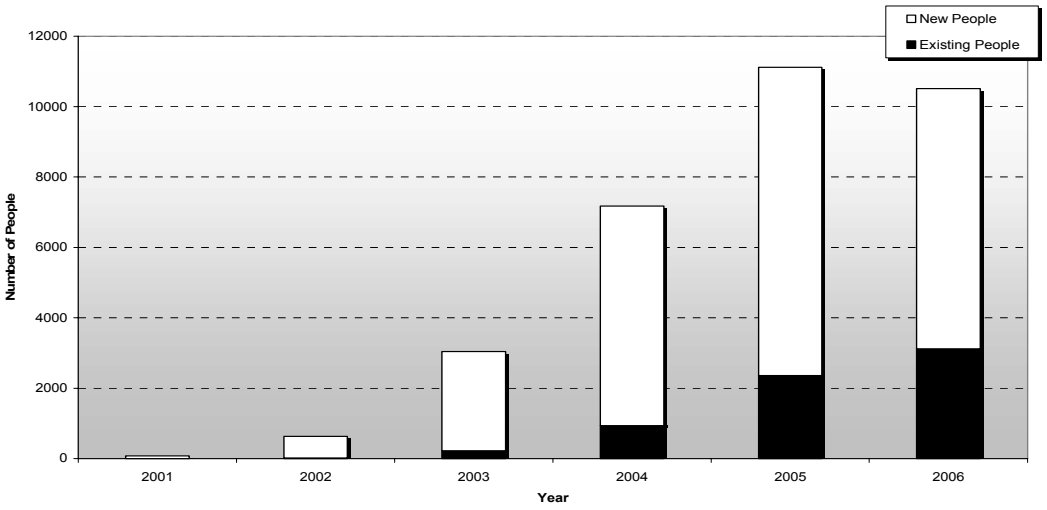


Figure 2. Number of people active in siRNA fundamental research by year.

This study doesn't aim to investigate the detailed researchers/inventors portfolio. Instead it intends to show the development in the area of siRNA. Figure 2 illustrates the change of people active in siRNA fundamental research by year. It clearly shows that there are more and more new researchers publishing in this relatively new field. Meanwhile, it also can be found that the figure of the existing researchers becomes stable in 2005 and 2006. In the technology activities, it follows the similar pattern. However, the number of inventors is far less than the active researchers, which indicates that siRNA patent activities will become more active in the near future.

What is the emerging research or technology in siRNA?
Technology report in TDA provides information on the terms first used in the last three years. In the scientific literature, the following terms appeared between 2004 and 2006: microarray (31), mitochondria [31], reactive oxygen species [30], STAT3 [24], cell proliferation [23], oligonucleotide [23], nitric oxide [21], endothelium [21], cisplatin [20], cytokines [19], calcium [19], lung cancer [18]. In the patent records, there are over 400 manual codes appeared in the last three years; the top three manual codes are: B14-H01E1 (Colon cancers), B14-S21 (Cell therapy), B04-F11B0E (RNA virus general genetically engineered).

What 'clusters' of fundamental research exist?
A Factor Map in TDA is a graphical representation of the results of a Principal Components Analysis (PCA). The PCA finds the list items that frequently occur together in the dataset. Figure 3 shows a

factor map on author keywords. Each node in the map represents a cluster of terms. The lines between nodes represent a measure of similarity between the two clusters of terms. The thickness of the line indicates the degree of similarity - a number between 0 and 1. Figure 3 shows part of a factor map on author keywords in the scientific literature. We can tell that RNAi is the center of the map, linking to other clusters, such as HBV; implying that RNAi technology can be applied to HBV and hepatitis therapy.

Figure 3. Factor map on author keywords.

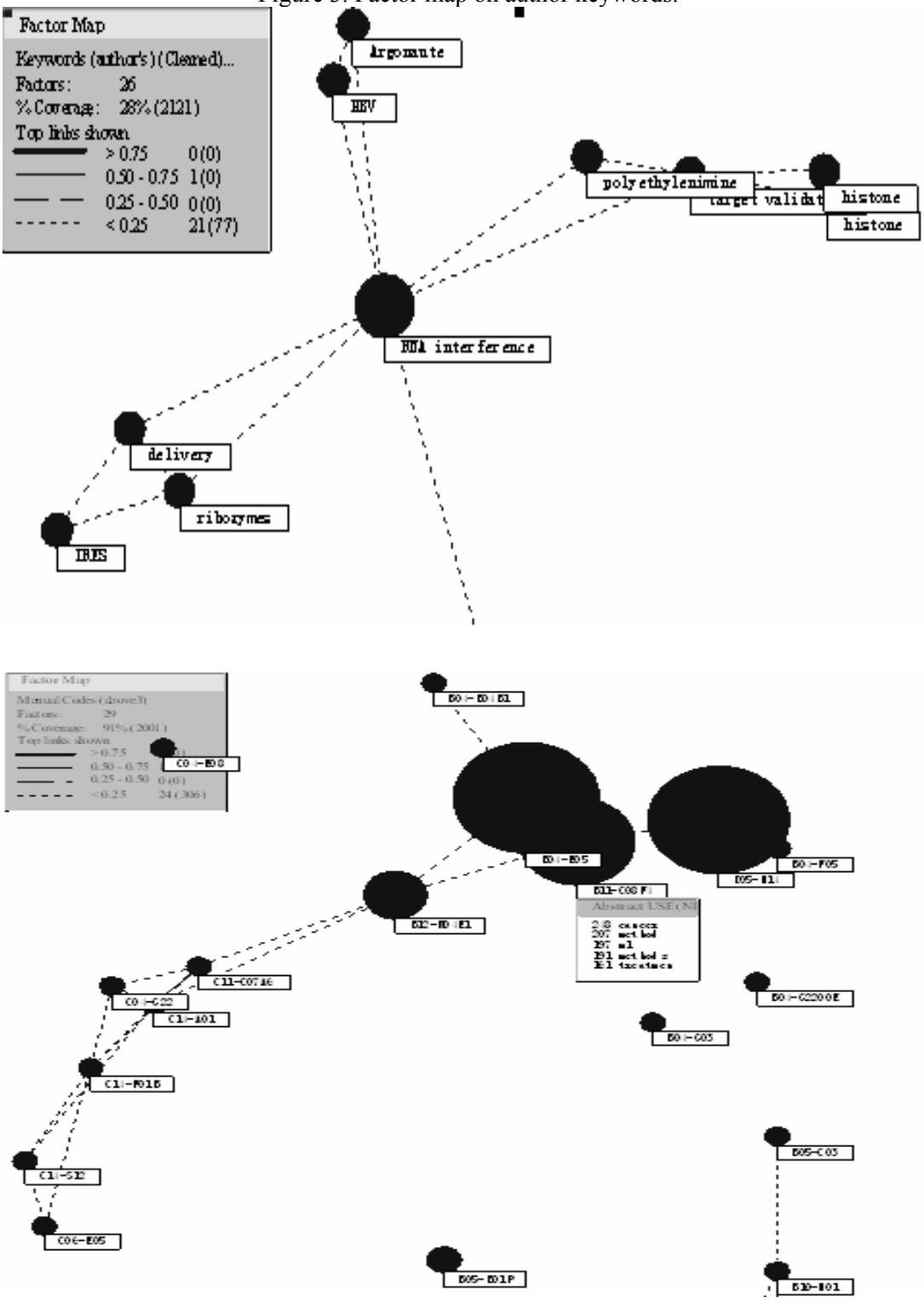


Figure 4. Factor map on Derwent manual codes.

Figure 4 illustrates a factor map in siRNA patent activity. Three major clusters are found on the top of this map. By examining the field of Abstract Use, we found that these clusters are linked to the cancer treatment. In other words, siRNA technique has been mainly applied in oncology to treat different type of cancers.

Conclusion

The analytical result shows a steady growth in outputs in the form of patents and publications since 2001. USA demonstrates a leading position in both the fundamental research and patent activity of siRNA. The predominance of the US is perhaps driven at least to some extent by the breakthrough scientific research on RNAi mechanism. Top organizations and clusters on main siRNA research and application are identified. Existing links between science and technology are also found from the comparison study.

References

Dorsett, Y., & Tuschl, T. (2004). siRNAs: Applications in functional genomics and potential as therapeutics. Nature Reviews Drug Discovery, 3, 318-329.

Nobel Foundation (2006). The Nobel Prize in Physiology or Medicine 2006. Retrieved December 12, 2006 from: http://nobelprize.org/nobel_prizes/medicine/laureates/2006/.

Tansey, M., & Stembridge, B. (2005). The challenge of sustaining the research and innovation process. World Patent Information, 27, 212–226.

Vantage Point (2006). Thomson Data Analyzer.

Revisiting Country and Institution Indicators from Citation Distributions: Profile Performance Measures

Michel Zitt[*,**], Elise Bassecoulard[*], Ghislaine Filliatreau[**], Suzy Ramanana-Rahary[**]

**zitt@nantes.inra.fr, bassecou@nantes.inra.fr*
**Lereco, INRA, Nantes (France)*

***filliatreau@obs-ost.fr, ramanana@obs-ost.fr*
Observatoire des Sciences et des Techniques (OST), Paris (France)

Abstract
This communication presents some aspects of an on-going research on a class of indicators of scientific visibility. Relying on a new usage of classic citation distribution studies, the "activity index in citation classes" (AIC) describes the "relative citation profile" or distribution of actors' articles among quantile classes of citations defined by the world reference. AIC exists in multiple variants reproducing various types of field-normalization, as well as direct or cumulative (AICC) presentation. The profiles of AIC can be summarized by empirical indexes, or parameters of simple laws, which are good candidates for indicators of overall citation performance. As profile performance indicators derived from AICC and classical "relative impact" (RI) indicators basically reflect the same information, their comparison is appealing. Empirical profiles obtained for national or institutional actors also raise the question of homogeneity of scientific literature: the high tail (excellence) and especially the low tail of the profile can show large deviations to the medium range of profile reflecting the general performance.

Keywords
science indicators; excellence measures; **d**atabase coverage; citation distributions; relative impact; power laws

Introduction

Bibliometric indicators are increasingly used in evaluation and positioning of scientific players in various contexts: assessment of individuals, labs, institutions, regions, nations. Pitfalls of crude bibliometric-based assessment and funding have also been put forth (*Butler, 2003; Weingart, 2005*), as well as the limits of "hit-parades" of science (*Van Raan, 2005*). Bibliometric measures are data-dependent and field-dependent, as shown in numerous classic studies (*Schubert & Braun, 1996, Murugesan & Moravcsik, 1978*) and one of the lasting issues is the delineation of reliable sets for establishing reference values (*Zitt et al., 2003, 2005*). We study here the over or under representation of an actor in the quantiles of citation of world literature. We will focus here on the macro level (countries). The application to evaluation of universities and research organizations, easily derived, will be reported in the full report. In the first section we present the methods, namely the "relative citation profiles" measures. In the second section we mainly address the macro-level (countries) indicators.

Methods

Relative profiles on citation rankings

The family of indicators presented describes the relative citation profile of actors, namely series of Activity Indexes in Citation Classes (AIC; in cumulative form: AICC) which requires, apart from usual prerequisites of citation indicators (counting option, citation window), that all articles in the reference (say the world literature, with many options of ranking, see below §1.2) be ordered by citation score, giving quantiles and relative ranks. The rank (r) is represented in abscissas, with conventionally, r=1 for the top-level, with the excellence class conventionally on the left side. A variant is based on actual scores of citations instead of ranks (on related questions, however with different aims, see *Glänzel & Schubert, 1988*; *Glänzel, 2007*).

- the subset of articles of a particular actor is distributed among citation classes defined for a reference, here the world. The relative distribution of the particular actor's articles reflects its involvement in the high or low-visibility articles in each discipline. The AIC is defined as the ratio of the actor's presence to the expectation, in principle the nominal value of the class (for example 5% for the top-cited half-decile of "excellence"), but sometimes a different value because of ties, quite frequent in the lower classes. If the actor places 7.5% of its production in this top 5% class with nominal size, the ratio AIC is 1.5 for this class.

- we address both “direct profiles” and “cumulative profiles”. The latter (AICC) is based on cumulative classes (quantiles or ranks) from the top one, and is appropriate for deriving robust indicators.

A flexible representation
By choosing various ranking rationale for articles (abscissas), the representation is flexible and exhibits both

a) a "cross-scale" capability: reflecting on option several level of normalization

As shown by *Zitt et al.* (2005), field-normalized indicators such as relative impacts are quite unstable when the level of normalization varies. Citation profiles embody normalization choices by changing the ranking option reflected on the x axis arrangement and scale. Most common options are, from fine to coarse grain: within the journal (all the top cited articles in any journal, regardless of the journal level, will receive top ranks); the "specialty" (say Thomson subject category); the sub-discipline; the large academic discipline; and all science, i.e. no field-normalization.

b) an ability to address "structural" indicators

Field-normalized indicators have inspired a large literature, reviewed for example by *Schubert & Braun* (1996). The relation between indicators at different levels of normalization can be expressed by "structural" indicators. For example a global citation indicator may be seen as the product of the actor's "impact factor" (solely based on journals' IF of his/her published articles; this is an example of structural indicator) and of a "relative citation ratio" (RCR), a type of journal-level normalized indicators (*Schubert & Braun, 1986*). AIC scale on the x axis can be arranged to build profiles on structural indicators as well as normalized indicators.

Derived indicators and relation to impact measures
The relative profile of publication is a particular presentation of the citation distribution normalized to the world reference. Synthetic "meta-indicators" can render the general trend of the profile over r. The typical meta-indicator is the slope of the profile with axes scales reflecting simple linearized models (e.g. double-log: power law) when applicable. Although power laws are pervasive in bibliometrics, as shown by an abundant literature since *Price (Price, 1976, Rousseau, 1990, Bookstein, 1990)*, their adequacy however is often discussed *(Burrell, 2001, Van Raan, 2001)*. Other simple models perform well, including in our particular context.

Various empirical approximations can be used to estimate the parameter(s), from sophisticated fitting to crude estimations based on a few points, for example the standard slope estimation of linear models. The measures for competitive literature must exclude the low tail. The inclusion of the high tail as such -- it is included by construction of cumulative classes, but not necessarily specified as particular points in the series -- sometimes irregular and subject to fluctuations, remains to discuss, depending on the target.

Also, a profile performance indicator (PPI) is quite efficiently built in a moving average rationale, from the difference of smoothed ordinates for distant classes i and j:

$PPI_{t,u}(i^*,j^{**})=[AICC_t(i)^*-AICC_t(j)^{**}] / [(i_u)^* -(j_u)^{**}]$

t and u denote transformations of variables y and x, e.g. log; (*) and (**) moving average filters in the neighborhood of i and j. An example of simple empirical index for the range of competitive literature is the slope profile indicator, where t and u are log transformations and (*),(**) a smoothed value on three points:

PPI(i*,j**) = PPI{5%,10%,20% -- 40%,60%,80%} = [log AICC(i)* - log AICC(j)*] / [log(i)* - log (j)**]

where for example: log AICC(i)* = (1/3) (log AICC(0.05) + log AICC(0.10) + log AICC(0.20)) and so on.

The log form in the numerator is consistent with the definition of AIC in terms of ratios, and robust towards a complete removal of the last class and a subsequent recalculation of AIC on the remaining classes: AIC and AICC are not altered in this process.

The relation of the AICC on cumulative classes and RI (relative impact) on the same range is interesting, since both indicators aim at describing the actor's relative position with respect to the world reference, and embody the same basic information respectively on world's and actor's citation distributions. It can be demonstrated that in a particular theoretical distribution (Pareto-I) applying both to world's and actors' citation score distributions, RI would be a most synthetic indicator, constant over r for an actor, its level fixing the exponent of AICC distribution. These conditions are not met in reality and this advantage vanishes. On countries series RI is in fact subject to a trend in r (in a different context see *Zitt et al., 2003*) so that, in practice, the knowledge of both level (e.g. intercept) and trend of RI is necessary, whilst a single parameter of slope can usually approximate the AICC profile, featuring a quite synthetic indicator of citation performance.

Three literatures?
We will show an evident irregularity in the low tail, and a more disputable one in the high-tail in actors' profiles. One may ask about the ability of models based on the "competitive range" indicators to predict the tails, if only in weak terms (rank correlation of actors). This will be examined both at the macro and micro-level.

Measures of actors' performance

Competitive literature and other literature
The collapse of global citations distribution (world level) in the lower tail is well known and mixes several phenomena in addition to possible low performance of scientists: role of the nationally-oriented journals (*Zitt et al., 2003)*, role of "transfer" literature where prominent scientists also occasionally publish. The demarcation line between competitive and other literature is not conspicuous, but the country-level observation is helpful. Fig. 1 shows a collection of actors ranging from top (Switzerland, US) thr. BRICK countries, with a linear scale on the two axes. Most profiles show a striking discontinuity around the point 80 (80% of literature), sometimes started earlier. In the last class, the 20% less cited, mainstream countries are strongly under-represented, and BRICK over-represented. The discontinuity is resilient towards changes in axes scale reflecting simple models. The less cited class is prone to particular phenomena, and must be excluded from fitting. If we restrict to linearized models with one parameter (slope), the best performing model differ amongst group of countries: by and large, power law performs slightly better for emerging countries profiles, log model slightly better for best performing countries. The straight linear model performs well in the limit case of average level countries (almost flat profiles), but cannot be generalized. The practical differences are usually low in terms of R-square. In comparable ranges, correlations on the 23 countries ranking are quite high, almost always Spearman > +0.99 thr. +0.999. Besides, the moving average approach

gives a very strong agreement with the fitted slopes, for example Spearman > +0.999 for the 3 points average measure above. The ranking of countries on the latter measure is displayed in Table 1.

Table 1 Countries ranking on impact performance: slope indicator = moving averages 3 points - Citation window: 4 years - smoothing 3 years 2000-2002 - competitive literature excl 0-2% top, 80-100% bottom

rank on 23 countries	country	PPI	rank on 23 countries	country	PPI
1	CHE	-0.162	13	ISR	-0.016
2	USA	-0.137	14	FRA	-0.015
3	NLD	-0.097	15	AUS	-0.012
4	DNK	-0.093	16	ITA	0.014
5	ISL	-0.077	17	ESP	0.077
6	GBR	-0.076	18	JPN	0.095
7	BEL	-0.054	19	KOR	0.127
8	DEU	-0.053	20	CHN	0.176
9	CAN	-0.051	21	RUS	0.191
10	AUT	-0.038	22	BRA	0.255
11	SWE	-0.033	23	IND	0.306
12	FIN	-0.030	(provisional results)		

The truncation of the low-tail allows a more reliable performance measure for Sweden, Japan, Spain whose avoidance of the low tail artificially enhances performances measured on the whole range (without truncation), and conversely for Russia or China, whose high activity in the low-tail artificially leads to underestimate the overall performance.

Ultra-competitive literature: "excellence classes"

Excellence indicators attract an increasing attention in research evaluation (e.g. *Tijssen, 2002*). AIC profiles visually express the relative position of an actor in citation classes, including excellence - with respect to the statistical sensitivity due, by construction, to the size of these classes. For several countries, the top-class is clearly deviating from the trend in any sensible model. In the data considered here up deviations of the first 1% class are noted for Israel, Finland, Island, UK, Spain. In other cases, the excellence class will be considered deviant or not depending on the model; log and power-law models tend to underestimate the activity of emerging countries in excellence, and overestimate it for mainstream countries.

Micro-scale application: universities and PROs

AIC/AICC indicators stated above were first developed for the institutional level analysis (*op.cit. 2002*) because of their descriptive ability for various visibility classes, allowing the competitive literature and other literature to be embraced and distinguished. The synthetic indicators summarizing the slopes of the profiles (PPI) remain relevant at this level as well, with respect to the possible fluctuations for small actors. Applications will be shown in the full report, let us only mention that the issues stated above at the country level are also met at the university level: risk of misleading measures when keeping the low tail, which favors some actors avoiding this class but quite average in competitive literature, question of alignment of excellence measures on the global trend of the profile, comparison with relative impact.

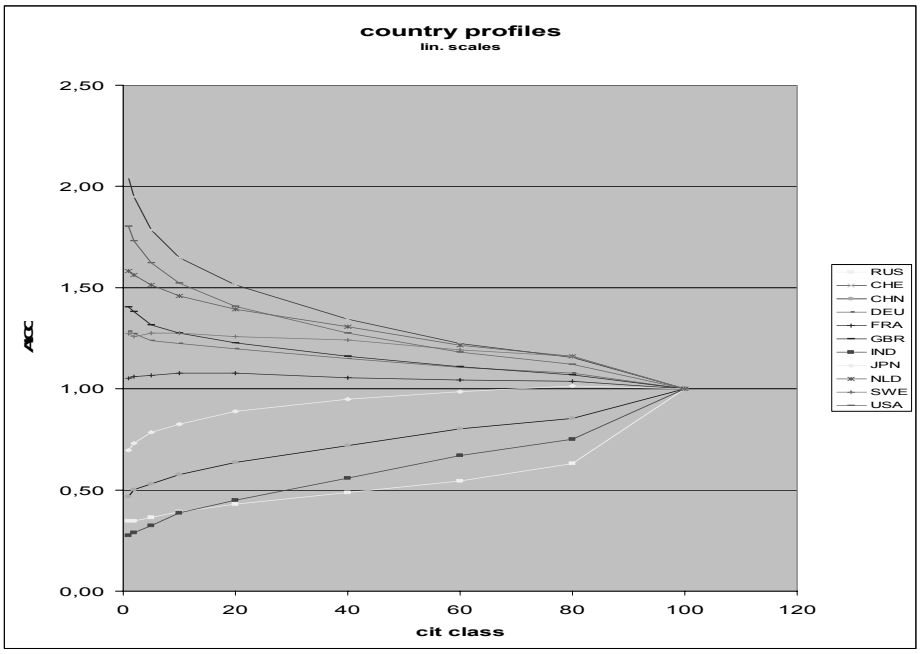


Figure 1. Citation profiles for some countries

Order of series from top to bottom (value at 1% class): Switzerland, USA, NLD, UK, DEU, SWE, FRA, JPN, CHN, RUS, IND -- NB the ranking on slopes does not reflect this order.

Conclusion

We propose here new or renewed performance indicators based on the deviation of actor's activity to the world reference in the various classes of visibility (Activity Index in Citation Classes, direct: AIC or cumulated: AICC). Corresponding meta-indicators such as the Profile Performance Index are efficient summaries of actors' citation performance. The study entails a comparison, both in the case of theoretic models and empirical description of actors, with classic "relative impact" indicators which contains the same information. It also addresses possible divergences and irregularities of the upper tail and the lower tail (the three literatures hypothesis). Both the macro and the micro-level are considered: the macro-level is more appropriate for testing general models, providing a background for the micro-level (universities) that will be detailed later.

Eventually, indicators design should take into account the trade-offs between the reluctance of decision-makers to handle too many figures (the decisional Occam's razor), and the crucial necessity of varying the points of view.

- in the particular range of competitive research, a "slope indicator" such as PPI described above is probably the most concise and straightforward measure of average performance throughout the range of literature, more economic than the equivalent modeling based on RI, whose interpretation should take into account both the trend and the level. On other grounds (convergence issues, see *Ramanana-Rahary et al, 2007* for RI), both indicators have (almost) the same limits.

- the parsimony principle however should not be extended much farther. A variety of representations should be kept by taking several vantage points, either on the type of literature considered amongst the three forms considered here, or on the setting of the "zoom" (normalization levels).

References

Bookstein, A. (1990). Informetric distributions, part I : unified overview. *Journal of the American Society for Information Science*, 41(5), 368-375.

Bookstein, . (1990). Informetric dstributions, part ii : resilience to ambiguity. *Journal of the American Society for Information Science*, 41(5), 376-386.

Burrell Q.L. (2001). "Ambiguity" and scientometric measurement: A dissenting view, *Journal of the American Society for Information Science and Technology*, vol 52, n°12, pp. 1075-1080.

Butler, L. (2003). Explaining Australia's increased share of ISI publications - the effects of a funding formula based on publication counts. *Research Policy*, 32(1), 143-155.

Glänzel, W., & Schubert, A. (1988a). Characteristic scores and scales in assessing citation impact. *Journal of Information Science*, 14(2), 123-127

Glänzel, W. (2007) A bibliometric analysis of subject characteristics based on long-term citation observation, *Journal of informetrics*, 1(1) 92-102

Murugesan, P., & Moravcsik, M. J. (1978). Variation of the nature of citation measures with journal and scientific specialties. Journal *of the American Society for Information Science*, 29, 141-155.

Price, D. J. d. S. (1976). A General theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science*, 27, 292-306.

Ramanana-Rahary S., Zitt M., Rousseau R., (2007) Aggregation properties of relative impact and other classical indicators: Convexity issues and the Yule-Simpson paradox, ISSI Conf. Madrid (forthcoming)

Rousseau R. (1990), Relations between continuous versions of bibliometric laws, *Journal of the American Society for Information Science*, 4, 197-203

Schubert, A., & Braun, T. (1986). Relative indicators and relational charts for comparative assessment of publication output and citation impact. *Scientometrics*, 9(5-6), 281-291.

Schubert, A., & Braun, T. (1996). Cross-field normalization of scientometric indicators. *Scientometrics*, 36(3), 311-324.

Tijssen, R. J. W., Visser, M. S., & van Leeuwen, T. N. (2002). Benchmarking international scientific excellence: Are highly cited research papers an appropriate frame of reference? *Scientometrics*, 54(3), 381-397.

van Raan, A. F. J. (2001). Competition amongst scientists for publication status: Toward a model of scientific publication and citation distributions. *Scientometrics*, 51(1), 347-357.

Van Raan, A. F. J. (2005). Fatal attraction: Conceptual and methodological problems in the ranking of universities by bibliometric methods. *Scientometrics*, 62(1), 133-143.

Weingart, P. (2005). Impact of bibliometrics upon the science system: Inadvertent consequences? *Scientometrics*, 62(1), 117-131.

Zitt M., Bauin S., Filliatreau G. Coord.. (2002) Bibliometric indicators of research institutions 1997 (in French) Group Publications of the Cooperative production of inter-institutional indicators for science policy, OST/MR, http://www.obs-ost.fr/pub/groupeIpsir.pdf

Zitt M., Ramanana-Rahary S., Bassecoulard E. (2003) Correcting Glasses help Fair Comparisons in International Landscape: Country Indicators as a Function of ISI Database Delineation. *Scientometrics*, 56, 2, pp259-282

Zitt M., Ramanana-Rahary S., Bassecoulard E. (2005). Relativity of citation performance and excellence measures: from cross-field to cross-scale effects of field-normalisation, *Scientometrics*, vol 63, n°2, pp. 373-401.

Mapping Review Networks: Exploring Research Community Roles and Contributions

Alesia Zuccala and Peter van den Besselaar

a.zuccala@rathenau.nl; p.vandenbesselaar@rathenau.nl
Netherlands Center for Science System Assessment, Rathenau Institute Anna van Saksenlaan 51, 2509 CJ, Den Haag (The Netherlands).

Abstract
In this paper we investigate the position of a review network within a research specialty; the network of scholars who write reviews of their colleagues' work. This is one of the voluntary activities that researchers perform as a prerequisite for the functioning of the invisible college. We compare this network to other networks within the specialty, and this allows us to distinguish various roles: *stars, influentials, members, supporters* and *juniors*. As scholars are characterized by different role-configurations, the invisible college becomes stratified. We discuss the implications for the development of a referee factor and review factor, norms for refereeing and reviewing, and the development of systems-based research evaluations.

Keywords
peer review; subject specialties; invisible colleges; mapping knowledge domains

Introduction

Scientific communication systems engage scientists in formal research-related activities, informal activities, and volunteer-based activities (e.g., Garvey & Griffith, 1968; Griffith, 1990). Scientometrics focuses predominantly on formal aspects of scientific communication; measurable outputs like journal impact factors (e.g. Glanzel & Moed, 2002) and citation networks (e.g., Small, 2005; White, 2001). Informal activities are elusive and studied less frequently, but have been emphasized in research on acknowledgements (e.g., Cronin, 1995) and in some metric evaluations of co-authorship and underlying processes of informal collaboration (e.g., Laudel, 2001; Melin & Persson, 1996; Newman, 2004). Volunteer activities can be defined in terms of services that scientists undertake to support a scientific communication system – i.e., to ensure that contributions are good for the system as a whole in addition to the career of the individual scientist. Such activities include refereeing papers for publication, organizing conferences, chairing award/grant committees, and writing reviews of newly published books and papers.

Research concerning scientific communication systems usually focuses on highly cited and co-cited papers in a research field, and scholars who are research stars; however, voluntary support work is also essential to the system. Without it, certain types of research communities, namely invisible colleges, might not function effectively. An invisible college is a communication system comprised of approximately 80 to 100 scientists who are part of the social "in-group" of a subject specialty (Crane, 1972; Price, 1986). Invisible colleges normally grow when scientists from subject specialties share similar interests, interact with one another at select conferences, and communicate new knowledge both formally and informally. Over time, the social network of the invisible college can become more "visible" due to the published output of its scholars (e.g., White, Wellman & Nazer, 2004; Zuccala, 2006).

In this paper we present research results based on an explorative study concerning support work. Our primary focus is on the role of the *reviewer* – i.e., the scientist from an invisible college network who writes an evaluative summary of a colleague's book or research paper. The chosen specialty is an area in mathematics known as Singularity Theory. When a new paper is published in mathematics, a researcher may cite it in his own work, but prior to this, he may also be asked by a Mathematical Reviews editor to write a formal review. The function of the review is to give colleagues in the specialty (or broader field) a brief idea of its significance so that they can decide whether or not to read the original work. To be asked to review a particular paper means that a mathematician has

acquired the respect of his peers: he has the reputation of being careful, reliable and knowledgeable in the specialized area to which the paper belongs (D. Trotman, personal communication, November 3, 2006). Given the importance of mathematical reviews, our research objective is to investigate who the reviewers are in Singularity Theory research, and what their role is vis-à-vis other types of roles within the invisible college.

The Dialog MathSci® database of the American Mathematical Society (AMS) maintains a record of review contributions; therefore 85 prominent Singularity Theory authors have been selected for the purpose of constructing a review network. MathSci® covers international publication data from 1940 to the present. In addition to journal articles, "roughly 10,000 monographs, conference proceedings, theses and technical reports are reviewed annually" (MathSci® Bluesheet). Review work in mathematics is formal, but does not need to be extensive: a few lines to 600 words are written to explain main results in a paper. The AMS guide states that a review can sometimes be evaluative; however "negative critical remarks [are expected to be] objective, precise, documented and expressed in good taste." If the reviewer thinks that the item "duplicates earlier work, [he/she] must cite specific references" and if the reviewer also thinks "that the item is in error, the errors should be described precisely" (Mathematical Reviews Database, 2006).

Research Methods and Findings

Data for this study were collected using both the Dialog SciSearch® citation index and the MathSci®: index. Table 1, in the Appendix, indicates how we categorized the data before it was used for separate mapping procedures. First, we mapped the Singularities specialty based on the author cocitation procedure outlined in White and Griffith (1981), using Cosine as the similarity measure (Ahlgren et al., 2003; Leydesdorff, 2005). For a list of 85 authors, cocitation counts were retrieved from SciSearch® for the period 1974 to 2006 [i.e., 85(85-1)/2=3570 pairs]. Figure 1 shows the map produced using the SPSS-11 multidimensional scaling and cluster routine. With the SPSS cluster procedure we have identified three sub-fields of Singularities research (*A-Real and Complex Analytic Geometry*; *B-Topology of Complex Algebraic Singularities*; *C-Singularities of Differentiable Maps*) and the authors attributed to these sub-fields. SciSearch® was used again to retrieve directed citation counts between the 85 Singularity Theorists, for a NetDraw (Borgatti, 2002) mapping of a citation network (see Figure 3). The authors at the centre of this spring-embedded network have received the most citations from colleagues within the invisible college.

With MathSci® we retrieved information concerning each author's total publication count, and produced a ranked list of mathematicians who have written signed reviews. For example, Figure 2 shows the number of publications for MARIA A. S. RUAS, the year of her first publication and a ranked list of mathematicians who have reviewed her work. The names highlighted in the ranked list are authors from her invisible college/specialty network. With the MathSci® data concerning reviewers and reviewed authors, we have created another NetDraw map, using a principle components layout, to illustrate the key contributors to this invisible college's directed review network (see Figure 4). Reviewers have a distinct role in a research community as research supporters. Again, we will examine who the main supporters are in Singularities research, and how they relate to other roles within the invisible college.

In our first analysis, we examined the ranked (ascending) publication output for each author and compared it to their individual review contributions. Figure 5 shows that review contributions tend to be less frequent than publications, yet some authors have reviewed as many articles as they have published (e.g., HOUSTON, FUKUI, TROTMAN); while others have actually published less, and contributed more to the communication system as reviewers (e.g., STEVENS, CHILLINGWORTH, GIBLIN). With the ranked reviewer data, we then examined how many reviews have been written by authors 'inside' the Singularity Theory specialty and how many have been written by 'outsiders' or colleagues from neighbouring subjects (see Figure 6). This particular research area seems to be open to external reviewing: approximately 86% of this community's published articles have been reviewed by authors from other specialties.

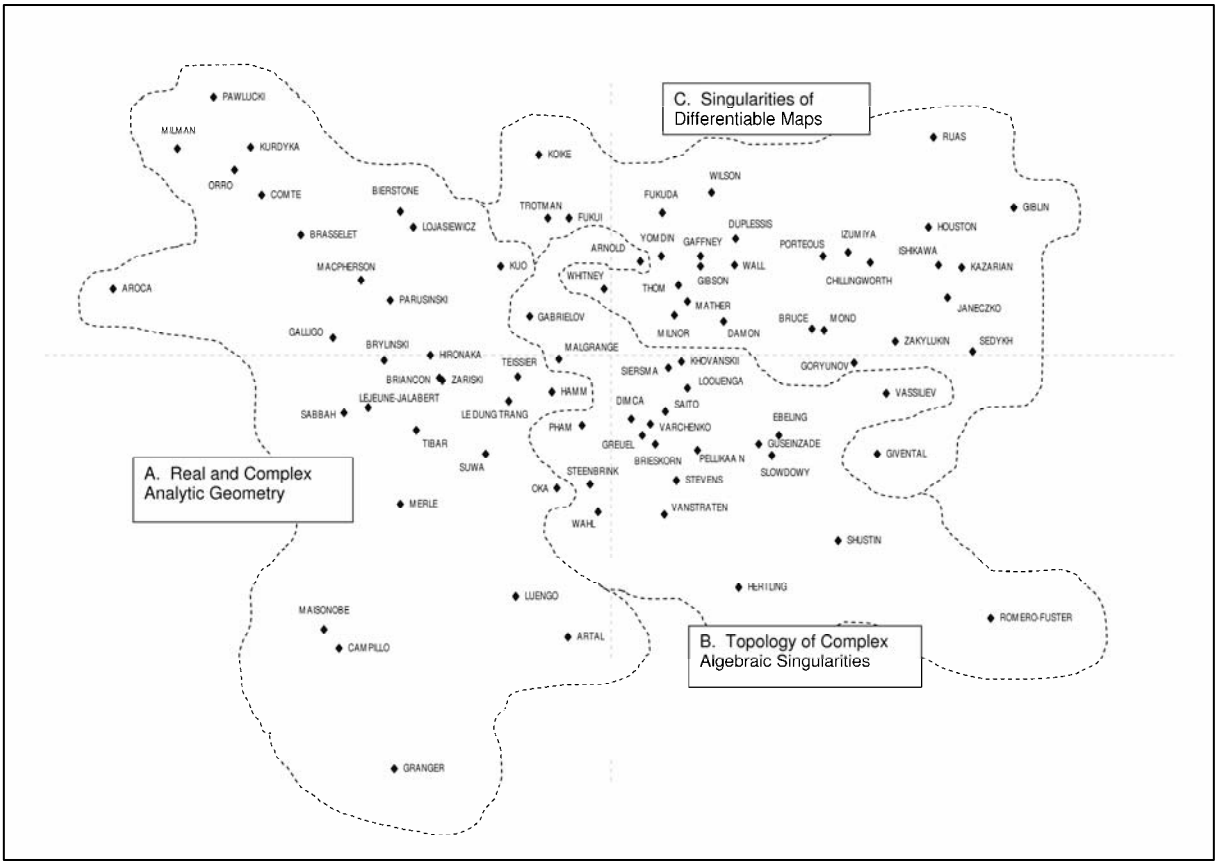


Figure 1. Cocitation map of 85 Singularity Theory authors (1974-2006), and research subfields (SciSearch®).

RUAS, MARIA A.S.? = 33 publications (First publication in 1986)

[Reviewers]

RANK	No.	Items	Term
1	4		ANDRICA, DORIN
2	3		OUTERELO DOMINGUEZ, ENRIQUE
3	3		WILSON, LESLIE CHARLES
4	2		LI, YANG CHENG
5	2		TIBAR, MIHAI
6	1		BEEM, J. K.
7	1		CHILLINGWORTH, D. R. J.
8	1		FUKUI, TOSHIZUMI
9	1		GOMOZOV, EUGENI P.
10	1		HURLEY, DONAL
11	1		IBANEZ, SANTIAGO
12	1		JANECZKO, STANISLAW
13	1		JIANG, GUANGFENG
14	1		LEVINE, J. P.
15	1		NUNO-BALLESTEROS, JUAN J.
16	1		WEINER, JOEL L.

Figure 2. Number of publications for MARIA A. S. RUAS and ranked reviewers (MathSci®).

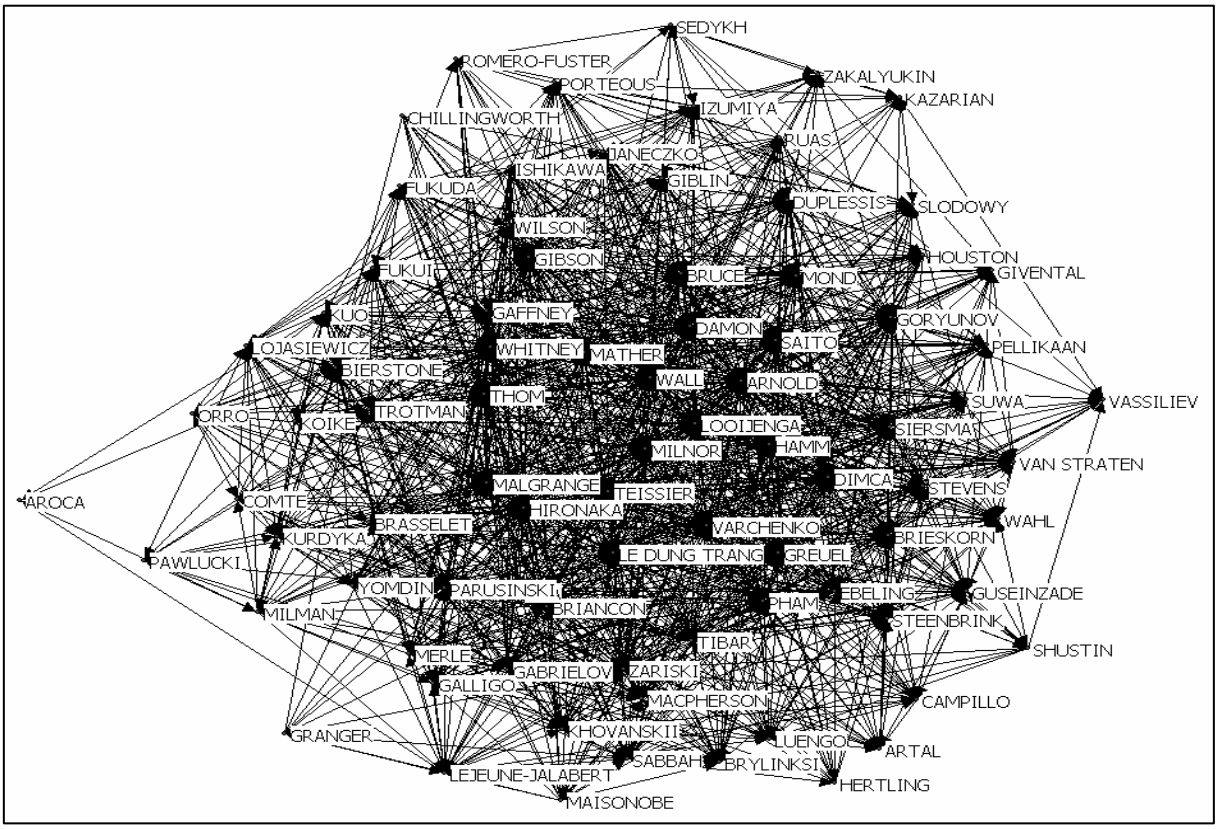


Figure 3. Directed citation network for 85 authors in Singularity Theory (1974-2006; SciSearch®).

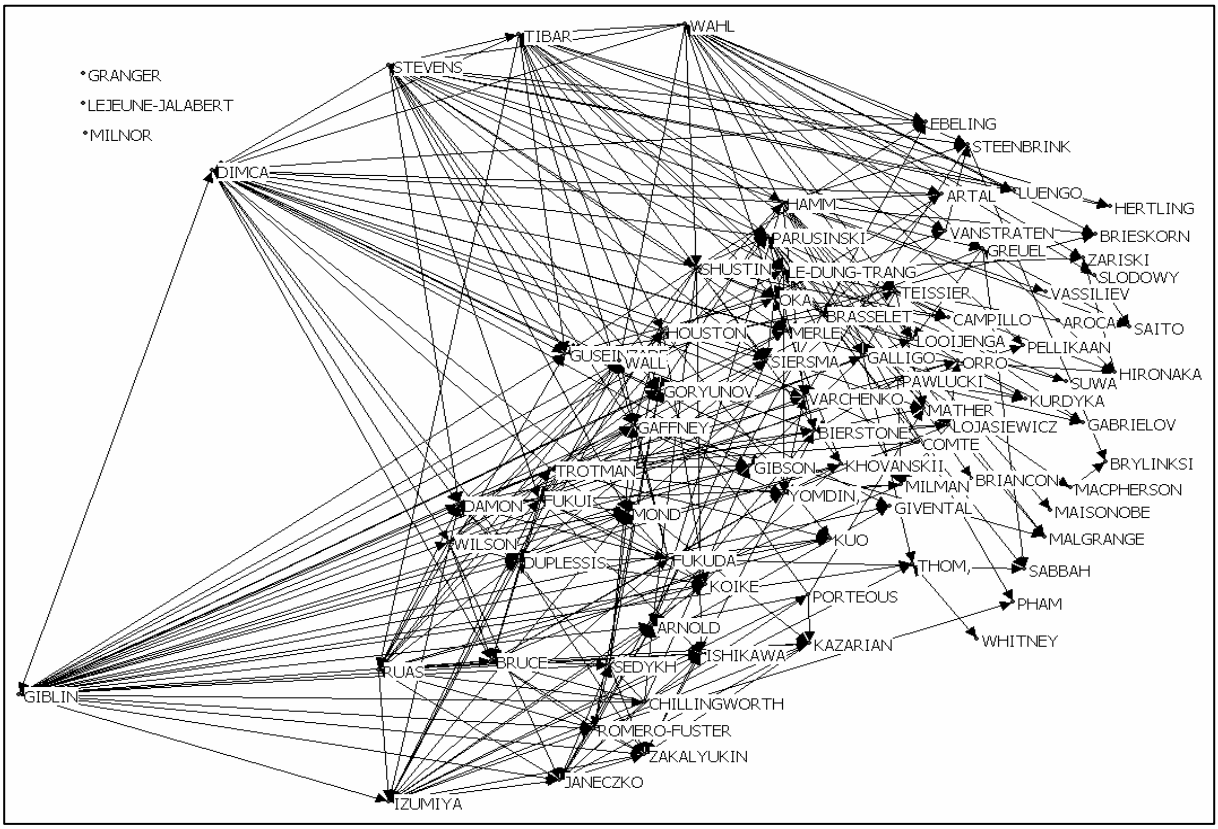


Figure 4. Directed review network for 85 authors in Singularity Theory (1970-2006; MathSci®).

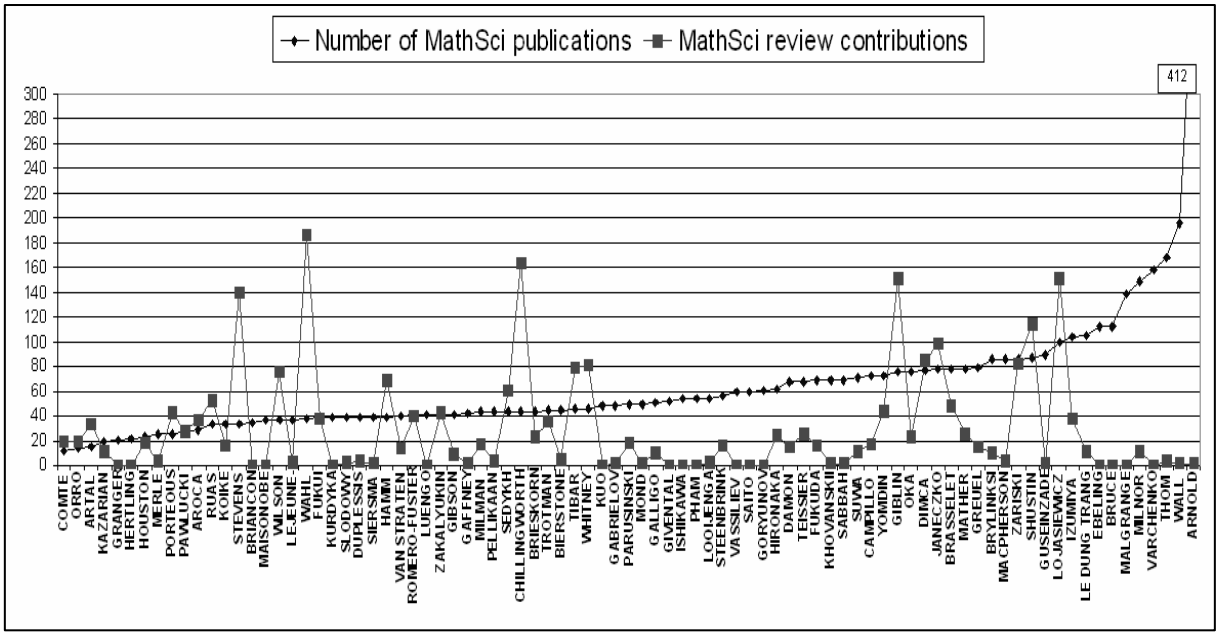


Figure 5. Publication counts for Singularity Theory authors compared to review contributions. Authors ranked by publication count (MathSci®).

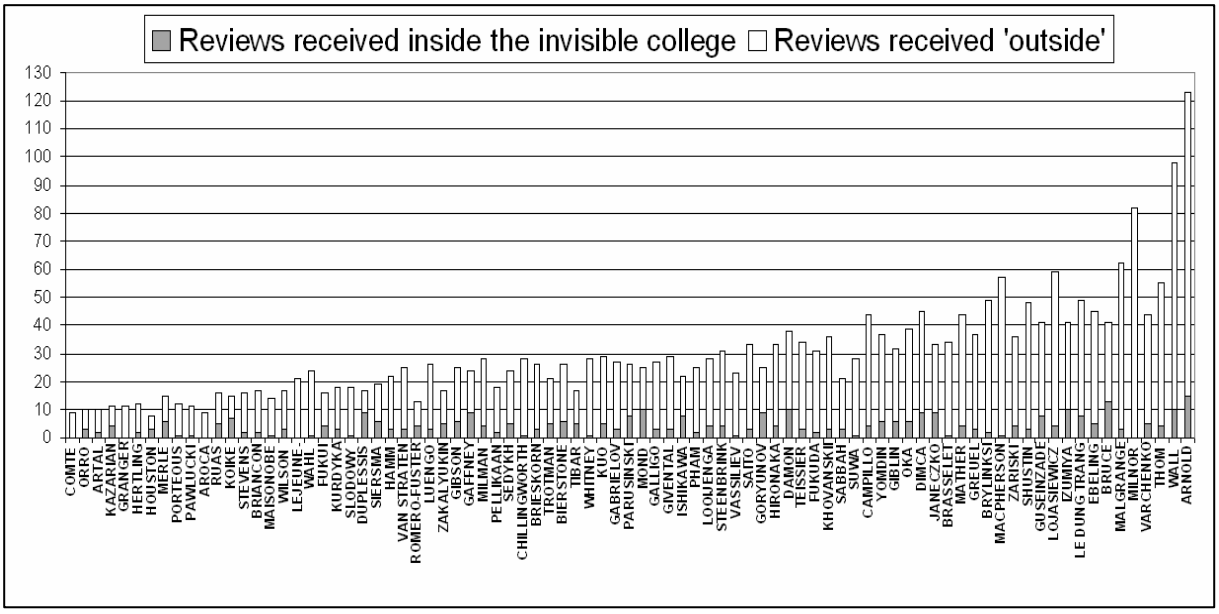


Figure 6. Reviews received by authors within Singularity Theory versus outside reviewers. Authors ranked by number of reviews received (MathSci®).

We also measured the relationship between the reviews that Singularity Theorists have written for colleagues within the invisible college, compared to reviews written for colleagues 'outside' the specialty and found a positive correlation (r=0.620). The scatterplot shown in Figure 7 shows that authors who review a lot of papers within their specialty area also tend to contribute frequently to the mathematics review system in general.

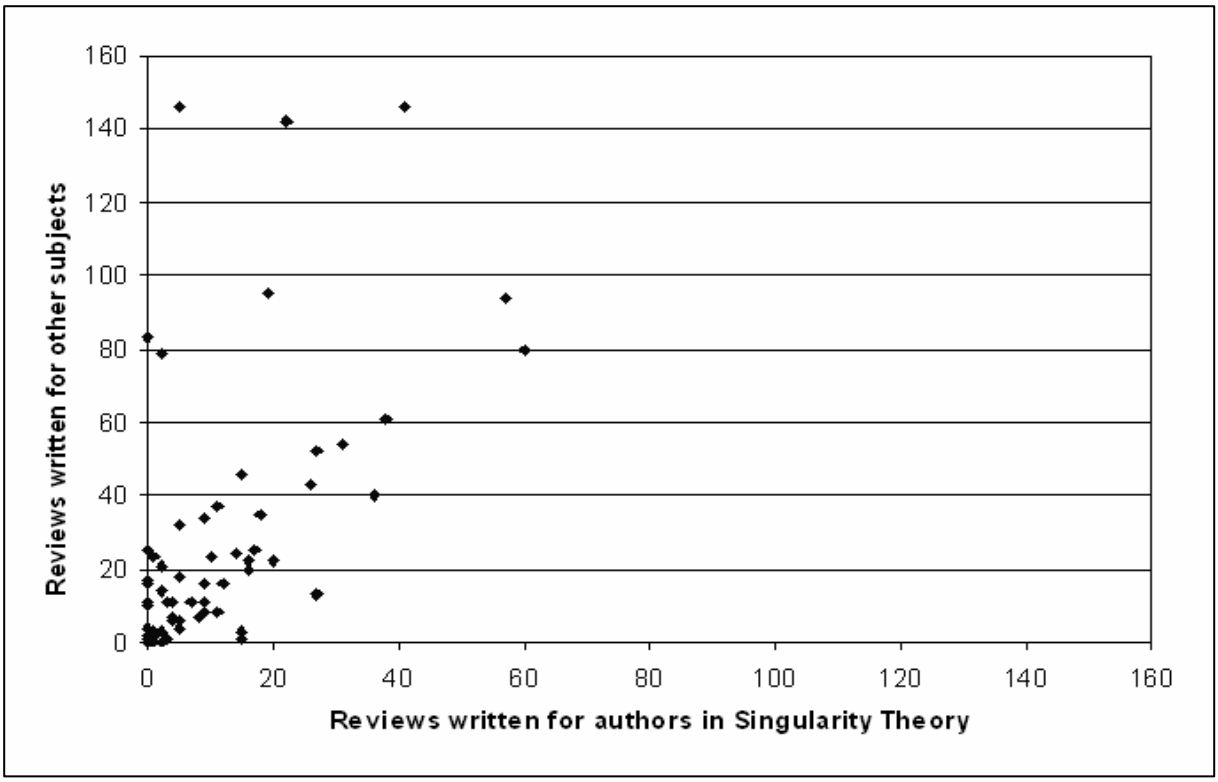


Figure 7. Reviews written for authors in Singularity Theory compared to reviews written for other subjects (MathSci®).

To determine if authors who have reviewed each other's work were also likely to cite one other (Figures 3 and 4), we used the QAP (Quadradic Assignment Procedure) matrix correlation function in UCINET (Borgatti et al., 2002). This compares the observed correlation with the average correlation of 2500 random permutations. As the latter was zero with a standard error of 0.015, the observed Pearson correlation value of 0.147 was significant (<0.00). In other words, a positive correlation exists between writing a review of someone's work and citing the same scholar.

Conclusions and Research Implications

Based on our research findings, it is clear that mathematicians will undertake review work at different stages in their career: young researchers will write reviews (e.g., COMTE; ORRO) as well as seniors, with stronger publication profiles (e.g., GIBLIN; DIMCA). Figure 5 indicates that there is also a small group of well-published mathematicians who tend not to do this type of support work (e.g., LOOIJENGA, PHAM, GORYUNOV), including an elite group clustered at the top of the publication rank (e.g., ARNOLD, WALL, MILNOR). Can we explain this review-versus-publication output imbalance? Yes, but there are perhaps more contributing factors than just one. In this study we cannot account for the motivation of an individual; therefore, if a mathematician is ambitious and wants to devote most of his time to research and publishing, he can refuse to write reviews or pass the work on to another colleague. Also, if the mathematician is academically strong – i.e., a recognized leader – he might not have time to write reviews because he is too busy mentoring Ph.D. students, chairing committees, organizing special seminars/workshops, lecturing and travelling to conferences. Junior researchers generally have more time to write reviews, and perhaps they agree to do this work to generate exposure or demonstrate to seniors where their abilities and interests lie. Senior mathematicians who write reviews might actually enjoy the process, or feel that it is a good way to keep in touch with what is happening in their research community. These mathematicians could have a reputation amongst editors for having a wide variety of interests and may also be very efficient review writers.

Review work in mathematics is not subject-specific or subject-centred. The review system tends to function in a way that is similar to the citation system: a mathematician may cite a research relevant paper of interest, just as he or she might review a paper of interest, regardless of the specialty area to which the paper belongs. Specialty areas of research grow because there is a core set of problems that mathematicians focus on at the outset, but cross-over interests with other subjects (i.e., permeable boundaries) are expected and allow researchers to build important connections. Singh (1998) reinforces this notion eloquently: "the value of mathematical bridges is enormous. They enable communities of mathematicians who have been living on separate islands…to explore each other's creations" (p. 191).

Our QAP matrix analysis of the review network and citation network points to another logical outcome, which places emphasis on familiarity: mathematicians who review each other's papers also tend to cite each other. If a mathematician becomes familiar with a piece of work and has the appropriate knowledge background to make evaluative or critical remarks, it makes sense that he might use that work to build upon new ideas in his own research. The opposite makes sense as well: a researcher who regularly cites the work of a particular colleague is also likely to agree to write reviews of that colleague's publications.

With the data that we have collected, our selected authors in Singularities research may be described and compared to each other on the basis of contribution roles. Each role is derived from the co-publication, co-citation, citation, and review data used to create Figures 1, 3 and 4. The roles also stem from our observations of the author's nodal positions on the three figures. Below, we list five types of contribution roles, which can be described separately, but need not be mutually exclusive. Multiple roles, or role configurations (including roles not identified here, e.g., supervising and mentoring) can make up an author's complete contribution profile. For instance, THOM is listed as a *star*, but he was also *influential* to many early members of the Singularity Theory community. Likewise, GIBLIN is a *member* of the Singularity community, but he has also been a strong *supporter* (*reviewer*) of many of his colleagues' work.

1) *Stars*: mathematicians who are central to the specialty area, i.e., highly co-cited with others in the specialty. Stars are also cited by researchers in all of mathematics, not just their specialty members. They have a significant reputation in mathematics as a whole, including a capacity to become award winners. Mathematicians who fit this role include HIRONAKA, MILNOR, THOM, MATHER.

2) *Influentials*: mathematicians who are well-published and highly cited or co-cited. Their work is influential to the specialty area's development, thus they are central to the invisible college's intellectual structure. Mathematicians who fit this role include ARNOLD, ZARISKI, WALL, TESSIER, LE DUNG TRANG, WHITNEY, BRIESKORN, DAMON, LOOIJENGA

3) *Members*: mathematicians with moderate-to-strong publication records who are cited by their specialty colleagues. Members often collaborate with other specialty members; hence their position is slightly more peripheral than stars and influentials, but they are still major contributors to the invisible college. Mathematicians who fit this role include BIERSTONE, MILMAN, PARUSINSKI, GUSEINZADE, VAN STRATEN, TIBAR, KURDYKA.

4) *Supporters (reviewers):* mathematicians who are members of the specialty research area, but participate often in support work – in this case, the writing of reviews. Their publication output may or may not be strong, but the amount of work that they do as reviewers is significant. Mathematicians who fit this role include JANECZKO, WAHL, GIBLIN, STEVENS, CHILLINGWORTH.

5) *Juniors*: mathematicians who are former students of senior members in the specialty; relatively young researchers who are focused on developing their research profile. Their publication output is not as strong as other members, and they tend to be peripheral to the intellectual structure of the network. Mathematicians who fit this role include COMTE, ORRO, KAZARIAN, ARTAL.

Figure 8 shows a mapping of the five different contribution roles and how they fit within Singularity Theory's basic co-citation (intellectual) structure. The extent to which these roles support the invisible college or enable it to function effectively requires further in-depth study concerning the mathematicians' behaviors and degree of investment in role-related activities.

Figure 8. Roles within Singularity Theory research.

This study of the Singularities community possesses interesting research implications concerning other forms of support work. For instance, attention has recently been given to referee work in science (another voluntary activity) and the introduction of a new impact measure termed the 'referee factor'. The 'referee factor,' defined mathematically by Rousseau (2006), "could be built in to standard assessments of performance, acting as an incentive for people to [referee] manuscripts" (Wilson & Lancaster, 2006, p. 812). According to Wilson and Lancaster, some scholars do not referee enough papers for publication; hence this is something that needs to be monitored. Unfortunately referee work is generally kept anonymous; therefore it is not so easy to measure. Now, if it is true that support work (including referee work) is a role-based activity, as we show in this study, then perhaps it is unrealistic to state the following: "for the system to be fair, all scientist should be refereeing two to three times as many articles as they submit"(Wilson & Lancaster, 2006). Our notion of roles and role-based contributions implies that a scientist's informal, formal, and volunteer-related contributions is closely tied to the type of role he/she possesses within a research community and may not change unless his/her overall profile (role configuration) changes. For example, in order for a research *star* or *influential* to referee two to three times as many articles as he or she produces, this scientist might need to minimize work associated with other areas of contribution – i.e., organizing international meetings, mentoring Ph.D. students; traveling to give special seminars, etc.

If the distinction between roles makes sense, then we may also consider the implications for developing systems-based research evaluations, that is, evaluations directed towards laboratory teams, specialty research areas, or invisible college networks. General statements about the number of reviews a scientist must write, or papers he/she must referee relative to his or her publication output are not useful if they neglect the importance of roles. A sports team, for instance, needs players to perform different functions on the field (e.g., defense; goalkeeping; forward). Not everyone can play the forward position on the field at one time; thus it is important to recognize that an invisible college operates according to the same principle. Different roles have to be performed by scientists during different times in an invisible college in order to develop and maintain this communication system, and ensure that it is operating successfully. If we evaluate researchers from this type of network on one dimension only (e.g., his/her degree of citedness) we fail to recognize the possible impact he or she can have when playing other critical roles. Future assessments at a systems level can tell us whether or not necessary roles are fulfilled in a specialty or invisible college and where changes could be made to create improvements. Moreover, a good review factor, similar to the suggested referee factor (Rousseau, 2006), might be a useful measure in this type of evaluation.

References

Ahlgren, P., Jarneving, B., & Rousseau, R. (2003). Requirement for a cocitation similarity measure, with special reference to Pearson's Correlation coefficient. *Journal of the American Society for Information Science and Technology, 54 (6)*, 550-560.

Borgatti, S. P. (2002). *NetDraw: Graph Visualization Software*. Harvard: Analytic Technologies

Borgatti, S. P., Everett, M., & Freeman, L. (2002). *UCINET 6 for Windows*. Harvard: Analytic Technologies.

Crane, D. (1972). *Invisible colleges: diffusion of knowledge in scientific communities*. Chicago,
University Press.

Cronin, B. (1995). *The scholar's courtesy: The role of acknowledgements in the primary communication process*. Los Angeles, CA: Taylor Graham Publishing.

Garvey, W. D., & Griffith, B. C. (1968). Informal channels of information in the behavioural sciences: their relevance in the structuring of formal or bibliographic communication. In E. B. Montgomery (Ed.), *The Foundations of access to knowledge* (pp. 129-146). Syracuse: Syracuse University.

Glanzel, W. & Moed, H. F. (2002). Journal impact measures in bibliometric research. Scientometrics, 53(2), 171-194.

Griffith, B. C. (1990). Understanding science: studies of communication and information. In C. L. Borgman (Ed.), *Scholarly communication and bibliometrics* (pp. 31-45). Newbury Park, CA: Sage.

Laudel, G. (2001). What do we measure by co-authorships? In M. Davis & C. S. Wilson (Eds.) *Proceedings of the 8th International Conference on Scientometrics and Informetrics* (pp. 369-384). Sydney, Australia: Bibliometrics & Informetrics Research Group.

Leydesdorff, L. (2005). Similarity measures, author cocitation analysis, and information theory. *Journal of the American Society for Information Science and Technology, 56(7),* 769-772.

Mathematical Reviews Database: Guide For Reviewers. (2006). Retrieved November 15, 2006 from http://www.ams.org/ authors/guide-reviewers.html.

Melin, G., & Persson, O. (1996). Studying research collaboration using co-authorships. *Scientometrics, 36*, 363-377.

Newman, M. (2004). Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences, 101*, 5200-5205.

Price, D. de Solla (1986). *Little science, big science....and beyond*. New York: Columbia University Press.

Rousseau, R. (2006). Biologist Rory Wilson proposes a referee factor. *ISSI Newsletter, 2(2),* 1-2.

Singh, S. (1998). *Fermat's enigma*. Toronto: Penguin Books Ltd.

Small, H. (2005). Tracking and predicting growth areas in science. In P. Ingwersen and B. Larson (Eds.) *Proceedings of the 10th International Conference on Scientometrics and Informetrics*
(pp. 13-23). Stockholm, Sweden: Karolinska University Press.

White, H. D., & Griffith, B. C. (1981). Author cocitation: a literature measure of intellectual structure. *Journal of the American Society for Information Science, 32(3),* 163-172.

White, H. D. (2001). Authors as citers over time. *Journal of the American Society for Information Science and Technology, 52(2),* 87-108.

White, H. D., Wellman, B. & Nazer, N. (2004). Does citation reflect social structure? Longitudinal evidence from the "Globenet" interdisciplinary research group. *Journal of the American Society for Information Science and Technology, 55(2)*, 111-126.

Wilson, R. & Lancaster, J. (2006). 'Referee factor' would reward a vital contribution. *Nature, 441,* p. 812.

Zuccala, A. (2006). Modeling the invisible college. *Journal of the American Society for Information Science and Technology, 57(2)*, 152-168.

Appendix. Alphabetical list of Singularity Theory authors (n=85) and data categories.

AUTHOR	Year of 1st publication	Number of MathSci publications	Average cocitation rate with other invisible college members	Citations received by indivisible college members	MathSci review contributions	Review contributions to Invisible College	External review contributions	Reviews received inside the invisible college	Reviews received 'outside' the invisible college
ARNOLD	1957	412	68	503	1	0	1	15	108
AROCA	1971	29	2	11	37	5	32	0	9
ARTAL	1991	15	3	16	33	10	23	2	8
BIERSTONE	1973	45	14	85	5	2	3	6	20
BRASSELET	1973	78	4	26	48	11	37	1	33
BRIANCON	1973	34	14	105	0	0	0	2	15
BRIESKORN	1964	44	25	134	23	2	21	3	23
BRUCE	1979	112	19	183	0	0	0	13	28
BRYLINKSI	1975	86	7	49	10	0	10	2	47
CAMPILLO	1978	72	2	15	17	0	17	4	40
CHILLINGWORTH	1967	44	2	0	164	22	142	1	27
COMTE	1996	11	1	5	19	11	8	0	9
DAMON	1973	68	18	120	15	8	7	10	28
DIMCA	1976	77	12	60	85	31	54	9	36
DUPLESSIS	1975	39	9	73	4	3	1	9	8
EBELING	1962	112	6	34	0	0	0	5	40
FUKUDA	1966	69	5	43	16	15	1	2	29
FUKUI	1985	39	3	30	38	16	22	4	12
GABRIELOV	1968	48	10	66	1	0	1	3	24
GAFFNEY	1976	42	11	84	1	1	0	9	15
GALLIGO	1973	50	5	34	10	4	6	3	24
GIBLIN	1968	76	2	16	151	57	94	6	26
GIBSON	1969	41	15	120	9	5	4	6	19
GIVENTAL	1980	52	5	38	0	0	0	3	26
GORYUNOV	1978	61	7	56	0	0	0	9	16
GRANGER	1967	21	0	2	0	0	0	0	11
GREUEL	1975	79	17	118	15	4	11	3	34
GUSEINZADE	1966	89	5	37	1	0	1	8	33
HAMM	1971	39	16	118	69	26	43	3	19
HERTLING	1993	22	1	3	0	0	0	2	10
HIRONAKA	1957	62	30	198	24	1	23	4	29
HOUSTON	1997	23	2	5	18	15	3	3	5
ISHIKAWA	1983	54	1	6	0	0	0	8	14
IZUMIYA	1978	104	4	20	38	14	24	10	31
JANECZKO	1982	78	3	14	99	38	61	9	24
KAZARIAN	1993	20	2	9	11	4	7	4	7
KHOVANSKII	1978	69	7	40	1	1	0	3	33
KOIKE	1979	33	2	20	16	0	16	7	8
KUO	1965	48	9	64	0	0	0	5	24
KURDYKA	1984	39	4	26	0	0	0	3	15
LE DUNG TRANG	1970	105	25	167	11	5	6	8	41
LEJEUNE-JALABERT	1971	37	6	47	2	0	2	0	21
LOJASIEWICZ	1950	100	15	82	151	5	146	4	55
LOOIJENGA	1971	54	24	235	2	0	2	4	24
LUENGO	1981	41	2	14	0	0	0	3	23
MACPHERSON	1967	86	7	37	4	1	3	1	56
MAISONOBE	1979	37	1	8	0	0	0	1	13
MALGRANGE	1953	139	22	145	0	0	0	3	59
MATHER	1965	78	37	301	25	0	25	4	40
MERLE	1973	25	4	14	3	1	2	6	9
MILMAN	1976	43	1	12	17	9	8	4	24
MILNOR	1950	149	63	273	11	0	11	0	82
MOND	1978	49	9	77	1	0	1	10	15
OKA	1972	76	8	44	23	5	18	6	33
ORRO	1983	14	1	9	20	9	11	3	7
PARUSINSKI	1985	49	7	48	18	7	11	8	18
PAWLUCKI	1984	26	3	24	28	12	16	1	10
PELLIKAAN	1988	43	3	11	4	0	4	2	16
PHAM	1963	54	13	83	0	0	0	2	23
PORTEOUS	1960	25	5	63	42	17	25	1	11
ROMERO-FUSTER	1983	41	0	9	40	27	13	4	9
RUAS	1986	33	1	1	53	18	35	5	11
SABBAH	1978	69	8	41	1	0	1	3	18
SAITO	1971	59	19	126	0	0	0	3	30
SEDYKH	1981	44	2	5	61	15	46	5	19
SHUSTIN	1980	87	2	11	114	19	95	3	45
SIERSMA	1973	39	14	92	1	1	0	6	13
SLODOWY	1978	39	5	25	2	2	0	1	17
STEENBRINK	1975	56	13	91	16	2	14	4	27
STEVENS	1984	33	4	21	140	60	80	2	14
SUWA	1969	71	2	8	11	0	11	1	27
TEISSIER	1970	68	26	201	25	9	16	3	31
THOM	1949	168	34	175	3	1	2	4	51
TIBAR	1984	46	4	18	79	27	52	5	12
TROTMAN	1976	45	5	33	36	16	20	5	16
VAN STRATEN	1985	40	3	17	14	3	11	3	22
VARCHENKO	1975	158	19	117	0	0	0	5	39
VASSILIEV	1990	59	4	19	0	0	0	1	22
WAHL	1974	38	10	65	187	41	146	1	23
WALL	1955	196	34	247	1	1	0	10	88
WHITNEY	1940	46	27	163	81	2	79	1	27
WILSON	1976	37	4	23	76	36	40	3	14
YOMDIN	1973	72	3	13	43	9	34	6	31
ZAKALYUKIN	1983	41	6	69	42	20	22	5	12
ZARISKI	1940	86	24	163	83	0	83	4	32

BRIC: Academic Web Contents in the Four Largest Emerging Economies

Isidro F. Aguillo, José Luís Ortega and Begoña Granadino

*{isidro,jortega,bgranadino}@cindoc.csic.es
CINDOC-CSIC. Cybermetrics Lab. Joaquín Costa, 22, 28002 Madrid (Spain)

Introduction

BRIC is a term coined (Wilson & Purushothaman, 2003) to define the four countries with the fastest growing economies: Brazil, Russia, India and China. Although there are important differences among them, they can be considered as representative of several regions, political and economic systems and also of R&D and academic structures.

Recently this "region" has attracted bibliometric attention (Zitt et al, 2006).

Methodology

The names and web addresses of the universities have been collected from the Catalogue of Universities by country of the Webometrics Ranking (http://www.webometrics.info/university_by_country_select.asp) that includes only universities with their own independent institutional web domain (10,953 in January 2007). The total number of BRIC domains is 1,674 (15.3%, or about one sixth of the total) distributed as shown in Table 1.

Table 1. Distribution by country (HE institutions with independent web domain)

Country	Universities
Brazil	196
Russia	315
India	262
China	901

Hong Kong universities (15 with .hk domain) have been excluded of the analysis as they behave differently and very prominently. In fact, as many as 6 of them appeared in Top 25 of the combined list. Some Chinese research institutions (under .ac.cn subdomain) offering higher education have been excluded as this is not their primary objective. Finally, the 131 domains appearing among the Top 2000 world list were chosen for further analysis. Details about how Webometrics Rank is build are described in Aguillo *et al.* (2006). Only 130 universities are represented as St. Petersburg University still maintains two different domains (*spbu.ru* and *pu.ru*). The irregular use of domain names by the Indian universities that changed URL addresses frequently decreasing then their link visibility can explain the low number of universities of this country in this study (Table 2).

Table 2. Population analyzed

Country	Universities
China	70
Brazil	34
Russia	21(+1)
India	5

The link analysis (Ortega *et al.*, 2007) was performed from data obtained during January 2007 from Yahoo search engine (search.yahoo.com) according to the following strategy:

linkdomain:univA site:univB

An adjacency matrix was built for the 131 institutional domains and analysed with UCINET 6.1 and graphically represented with NetDraw 2.2.

Results

Webometrics ranking (Table 3) take into account both the volume of information (number of pages, rich files and entries in Google Scholar –Table 4-) published in the webdomain and the visibility measured by the number of external inlinks.

Table 3. Top 20 BRIC universities

WR	University
97	Univ Sao Paulo
190	Univ Est Campinas
190	Beijing Univ
232	Moscow State Univ
281	Univ Fed Santa Catarina
336	Tsinghua Univ
402	Univ Fed Rio de Janeiro
421	Univ Fed Rio Grande Sul
529	Zhejiang Univ
557	Univ Fed Minas Gerais
562	Pont Univ Catol Rio de Janeiro
565	Shanghai Jiao Tong Univ
597	Univ Brasilia
633	Indian Inst Technol Bombay
635	Univ Est Paulista
665	Univ Sci & Technol China
670	Indian Inst Sci Bangalore
677	Nanjing Univ
688	Shandong Univ
711	Fudan Univ

Table 4. Number of rich files and Scholar (scholar.google.com) records

	.pdf	.doc	.ppt	.ps	Scholar
Total	418,172	254,683	59,419	14,924	131,206
Median	1330	1540	260	3	292
Min	41	128	9	0	4
Max	51600	9270	4970	2360	20200

In the Figure 1 the points shape show the nationality of each university, the size shows the percentage of English language pages over the total amount of web pages.

Conclusions

Large universities in the BRIC are among the Top ranked according to the Web indicators. For this small group there is a clear commitment to Open Access initiatives through web publication. As shown in Figure 1 there is a large overlap of this group with those universities that use extensively English in their web pages, bridging the four countries' academic webspace. If universities of English speaking countries, especially US ones, will be added as outliers, we can conjecture that a more interconnected network will be obtained.

References

Aguillo, I. F., Granadino, B., Ortega, J. L. & Prieto, J. A. (2006). Scientific research activity and communication measured with cybermetric indicators. *Journal of the American Society for Information Science*, 57, 10: 1296-1302.

Ortega, J. L., Aguillo, I. F., Cothey, V. & Scharnhorst, A. (2007). Maps of the academic web in the European Higher Education Area - an exploration of visual web indicators. *Scientometrics* (forthcoming).

Wilson, D. & Purushothaman, R. (2003) Dreaming With BRICs: The Path to 2050. Global Economics Paper, no. 99. New York. Goldman Sachs. Retrieved January 9th 2007 from http://www2.goldmansachs.com/insight/research/reports/99.pdf

Zitt, M., Laurens, P., Ramanana-Rahary, S., Thèves, J. & Filliatreau, G. (2006). Output and Visibility of "BRIC" Science: A Reassessment of Emerging Countries Positions. *Proceedings of the 9th International Conference on Science & Technology Indicators*. Leuven (Belgium), Leuven, Belgium, 7–9 September 2006.

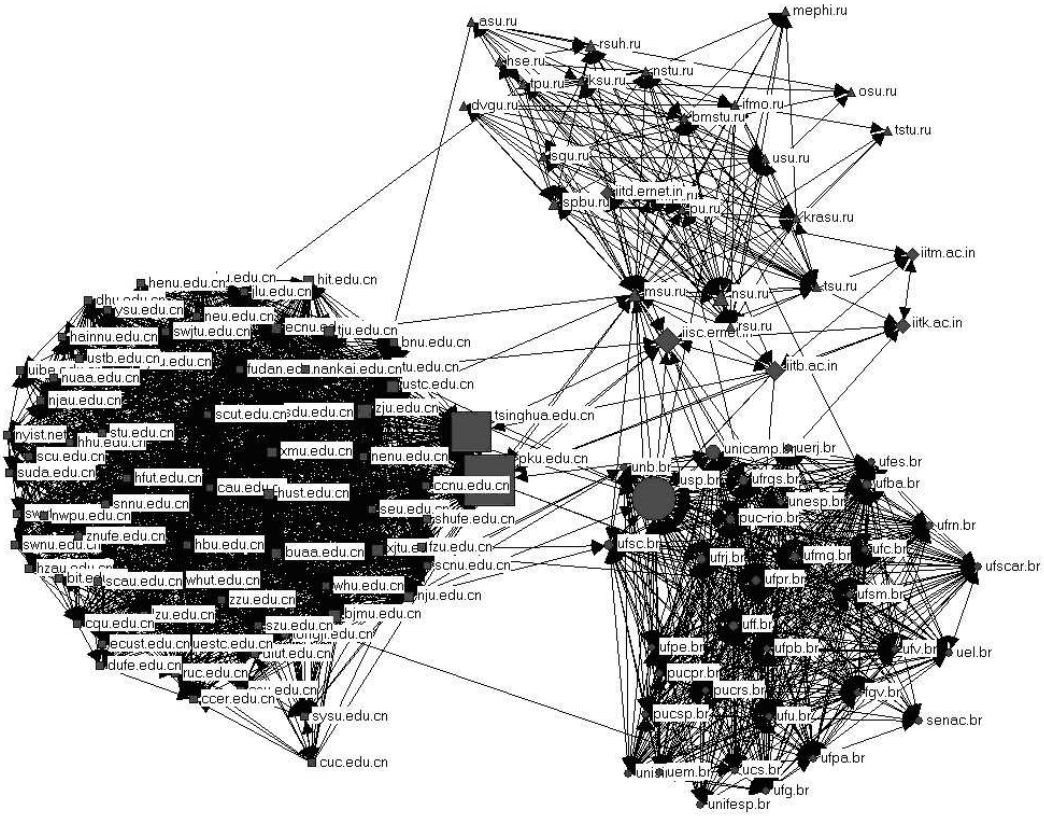


Figure 1. Link analysis of the BRIC universities

Citation Analysis and Impact Indicators of the Spanish Medical Journals (2001-2005)[1]

Rafael Aleixandre-Benavent [*], Gregório González-Alcaide[*], Alberto Miguel-Dasit[**], Miguel Castellano-Gómez[***] and Juan Carlos Valderrama Zurian[*]

[*]*rafael.aleixandre@uv.es, gregorio.gonzalez@uv.es, juan.valderrama@uv.es*
Instituto de Historia de la Ciencia y Documentación López Piñero (CSIC-Universitat de València), Valencia, Spain

[**]*amdasit@hotmail.com*
MR Section, La Plana de Vila-Real Hospital, Castellón, Spain

[***] *castellano_mig@gva.es*
Dirección General de Drogodependencias. Conselleria de Sanitat. Generalitat Valenciana, Valencia, Spain

Introduction

The SCI-JCR does not provide a complete coverage of all journals involved in scholarly research because Anglo-American journals and English language journals prevail in this data base, whereas papers published in non English-language journals tend to be less represented (Moed, 2005). Such biased coverage restricts the use of SCI-JCR as a tool for studying the complete impact of research performed outside English-speaking countries. In consequence, complementary analyses targeting domestic journals have been developed in some countries (aleixandre et al, 2007; Jin & Wang, 1999; Kuang-Hua, 2004; Sipka, 2005; Miguel-Dasit et al., 2005). Among Spanish initiatives, we point out the research project "Potential Impact Factor of the Spanish Medical Journals", supported by the Spanish Ministry of Education and Science. In this project, a citation analysis of 87 non SCI-JCR-indexed Spanish medical journals was performed to calculate their hypothetical IF value and other bibliometric indicators. The data of this study are available in http://147.156.181.37/imecitas/impacto_ime_ing.asp (figure 1). The objective of this paper is to present the main results of this project for the 2001 to 2005 period.

Method

87 Spanish medical journals included in the Spanish data base *Índice Médico Español* (IME) or in the international data bases Medline, Embase, SCI and Current Contents were compiled by using the ISI citation analysis methodology in order to correctly compare the results with those of the SCI-JCR. The bibliographic data of both the citing and the cited journals were included in a Microsoft Access database to facilitate the analysis. The impact indicators have been determined in international and national level. The national impact factor (NIF) has been obtained placing in the numerator just the citations coming from the 87 Spanish source journals while the international impact factor (IIF) has been calculated taking in account the citations of the Spanish journals as the citation coming from non Spanish journals included in SCI-JCR.

Figure 1. Home page of Potential Impact Factor of the Spanish Medical Journals

Results and discussion

73.520 citations coming from Spanish journals have been analyzed in the 2001-2005 period. The highest increase in the estimated NIF was found in the journals *Revista Española de Cardiología, Medicina Clínica, Archivos de Bronconeumología, Gaceta Sanitaria y Enfermedades Infecciosas and Microbiología Clínica* (table 1).

Regarding the IIF indicator, we found that the journals *Histology and Histopathology, International Journal of Developmental Biology*, *Revista Española de Cardiología, Medicina Clínica* and *Journal of Investigational Allergology & Clinical Immunology* were the journals with the largest increase in their calculated IIF (table 2).

Of interest, some not-SCI-JCR-indexed Spanish journals showed a highest increase in their IIF value,

compared with those obtained from other foreign journals included in the same subject category of the SCI-JCR. Nevertheless, no Spanish journal could be placed in the first quartile of the SCI-JCR.

Table 1. Journals with the biggest NIF (2001-2005)

JOURNAL	NIF*
Revista Española de Cardiología (SCI)	1.243
Medicina Clínica (SCI)	1.141
Archivos de Bronconeumología (SCI)	0.835
Gaceta Sanitaria	0.752
Enfermedades Infecciosas y Microbiología Clínica	0.714
Atención Primaria	0.691
Revista Española de Salud Pública	0.577
Vacunas. Investigación y Práctica	0.547
Anales Españoles de Pediatría	0.523
Revista Española de Anestesiología y Reanimación	0.470

* average

Table 2. Journals with the biggest IIF (2001-2005)

JOURNAL	IIF*
Histology and Histopathology (SCI)	1.911
International Journal of Dev. Biology (SCI)	1.674
Revista Española de Cardiología (SCI)	1.540
Medicina Clínica (SCI)	1.441
Journal Inv. Allergology Clin. Immunology (SCI)	1.294
Journal of Physiology and Biochemistry (SCI)	1.117
Enfermedades Infecciosas y Microbiología Clínica	1.070
Archivos de Bronconeumología (SCI)	1.017
Gaceta Sanitaria	0.928
Atención Primaria	0.733

* average

Therefore, the majority of the non-indexed Spanish journals would be included in the fourth quartile of the ranking of the SCI-JCR journals in the same subject area. In general, Spanish authors don't usually include bibliographic references of papers published in Spanish journals, and even less so if they are written in English, maybe because of their lack of spread in the Spanish libraries. This phenomena is produced in journals like *Methods and Findings in Experimental and Clinical Pharmacology, Allergologia et Immunopathologia*, *Histology and Histopathology* e *International Journal of Developmental Biology*. We have detected a slow growing in the number of citations and impact factor in the main journals (figure 2). The improvements in these indicators are most likely the result of policies implemented by the editors to enhance formal quality and quality of the contents and the rigor of the peer review process, to provide access to the full text of articles on the Internet, and to offer English-language publication. Our perspectives are to analyze the citations in the 2006-2010 period in order to determine: the most cited articles and journals and their impact factor; the coauthorship networks; and the research fronts of the Spanish medicine.

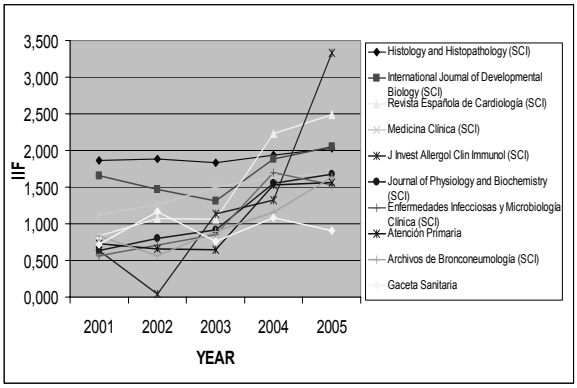


Figure 2. Evolution of the journals with the biggest IIF (2001-2005)

References

Aleixandre, R.; Valderrama, J.C.; Miguel-Dasit, A.; Alonso, A.; Castellano, M. (2007). Hypothetical influence of non-indexed Spanish medical journals on the impact factor of the Journal Citation Reports-indexed journals. *Scientometrics*, 70: 53-66.

Jin, B., Wang, B. (1999), Chinese Science Citation Database: Its construction and application. *Scientometrics*, 45: 321-332.

Kuang-Hua, C. (2004), The construction of the Taiwan Humanities Citation Index. Online Informatics Reviews, 28: 410-419.

Miguel-Dasit, A., Martí-Bonmatí, L., Sanfeliu, P., Aleixandre, R., Valderrama, J.C. (2005), Hypothetical influence of non-indexed Spanish journals on the impact factor of radiological journals. *European Journal of Radiology*, 54:321-326.

Moed, H.F. (2005), Citation Analysis in Research Evaluation: Springer, Dordrecht, 2005.

Sipka, P. (2005), The Serbian Citation Index: Context and Content. Proceedings of the 10th Conference of the International Society for Scientometrics and Informetrics. Stockholm, Sweden, p. 710-711.

[1] Supported by the Spanish Ministry of Education and Science (Estudios y Análisis 2003 & 2004 Program; Plan Nacional I+D, 2005-2007).

Research Performance in Isfahan University of Medical Sciences: Evaluation of Scientific Productivity

Farzaneh Aminpour and Payam Kabiri

aminpour@mui.ac.ir, payam.kabiri@gmail.com
Academic Evaluation & Scientometry Center, Vice Chancellery for Research,
Isfahan University of Medical Sciences, Isfahan (Iran)

Background

Research performance in Iran has increased appreciably during the recent years especially in regard to the output of scientific publication. This development can be explicitly seen in the basic and applied sciences including medicine and its subspecialties. Some of the major universities in Iran amongst them Isfahan University of Medical Sciences plays an important role in this development (aminpour, 2006). Many studies have used the number of publications produced by authors in the disciplines or in the regions and the number of citations received by those publications as measures of research performance and scientific productivity (Haiqi & Yuhua, 1997). Haiqi (1996) quantitively evaluated the scientific productivity of key medical universities in China by using CBMdisc, Medline and SCI databases. Osareh and Marefat (2005) studied the scientific productivity of Iranian researchers in Medline database from 1976 to 2003. They identified the growth and development of scientific publications in this database during this period (totally 2695 documents), highly productive authors, universities and popular journals using scientometric techniques. According to this survey Isfahan University of Medical Sciences ranked fifth regarding scientific productivity among the other medical universities in Iran. The contribution of Iran in world science production has been investigated in 2004 by Saboury and Poursasan (2005) using the indices of ISI. They found that the contribution of Iran in science production in the world in science, and social science have been 0.36 and 0.08 percent. According to this survey, universities of Tehran, Sharif, Tehran Medical Sciences, Tarbiat Modarres and Shiraz by 9.6, 7.3, 7.0, 6.9 and 6.4 percent of scientific production had the most contributions between all Iranian academic centers. The purpose of this survey is to evaluate quantitatively the scientific productivity of Isfahan University of Medical Sciences from the beginning (1985) till the end of 2006 using Medline and Web of Science (WOS).

Methods & Materials

The study was conducted according to the data recorded in the web Of Science (WOS) as one of the major databases of Institute for Information (ISI) and also Medline as the most important bibliographic database in medical sciences during the 22-year period 1985-2006. The scientific output of Isfahan University of Medical Sciences was assessed by searching PubMed Medline and WOS using "Isfahan OR Esfahan" in affiliation field. The time limitation was set on 1985-2006. Those papers which belonged to the other universities located in Isfahan have been set aside from the retrieved papers. The remained papers were considered as the scientific productions of Isfahan University of Medical Sciences in its lifetime. Then these papers were categorized according to the year of publication in Medline and also in WOS. In the second stage, the papers were classified according to the faculties which produced them. Finally the findings were compared with each other.

Results & Discussion

The findings indicate that 320 papers were indexed in Medline by the researchers of Isfahan University of Medical Sciences while 459 papers were covered by the indices of WOS during the period 1985-2006. Figure 1 shows the distribution of papers published in periodicals indexed in Medline by the researchers of Isfahan University of Medical Sciences according to the year of publication.

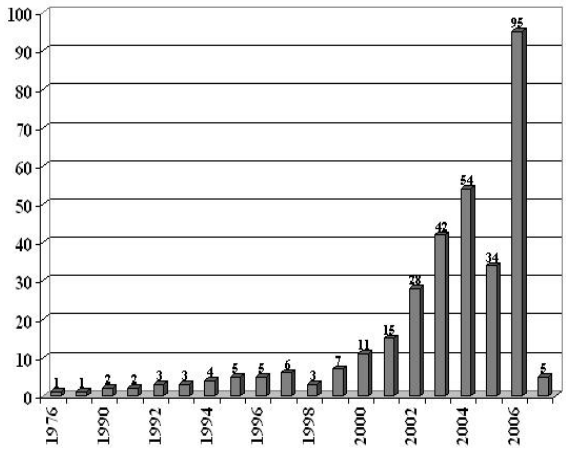


Figure 1. The number of papers published in periodicals indexed in Medline

The annul number of papers indexed in WOS is shown in figure 2. These two figures show a sharp

increase in scientific publications indexed in Medline and WOS between 2000 and 2006.

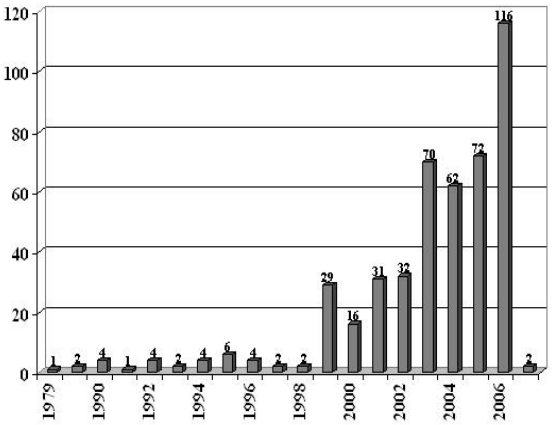


Figure 2. The number of papers published in periodicals indexed in WOS

The significant increase in scientific production indexed both in Medline and WOS after 2000 is greatly due to the motivations aroused by paper award system which has been set up by Iran Ministry of Health from this year. Some of the other reasons for this increase are: development of the university computer network, delivering online and dialup internet services to the academic staff and also to the students of higher education, accessibility of medical databases as well as electronic journals and full text collections by means of institutional subscription, and finally university advanced workshops on research methods and paper writing all started in the year 2000. Table 1 shows the number of papers indexed in Medline according to the faculties of Isfahan University of Medical Sciences while Table 2 shows the number of papers indexed in WOS also according to the faculties of this University.

Table 1. The number of papers indexed in Medline according to the faculties

Faculty	Number of Publications	Percentage
Medicine	230	70.55
Pharmacy	70	21.47
Public Health	12	3.68
Dentistry	12	3.68
Nursing	1	0.31
Rehabilitation	1	0.31
Management	0	0.00
Total	**326**	**100.00**

The results show that the faculty of medicine with 230 papers in Medline and 300 papers in WOS ranks first in terms of scientific publication. The faculty of Pharmacy stands in the second level by publishing 70 and 130 papers indexed in Medline and WOS. The contributions of faculties of Public Health and Dentistry are almost equal in both databases. Two faculties: Rehabilitation and Management do not make any significant contribution in scientific productivity of Isfahan University of Medical Sciences.

Table 2. The number of papers indexed in WOS according to the faculties

Faculty	Number of Publications	Percentage
Medicine	300	64.66
Pharmacy	130	28.02
Public Health	16	3.45
Dentistry	15	3.23
Nursing	13	0.65
Rehabilitation	0	0.00
Management	0	0.00
Total	**464**	**100.00**

Conclusion

Research performance in Iran has increased during the recent years especially in the basic and applied sciences including medicine and its subspecialties. Isfahan University of Medical Sciences plays a determining role in this development. The significant increase in scientific publications indexed in Medline and WOS after 2000 is greatly due to the paper award system set up by Iran Ministry of Health, development of the university computer network and internet services, accessibility of medical databases, electronic journals and full text collections and finally university advanced workshops on paper writing all started in the year 2000.

References

Aminpour, F. (2006). *An Introduction to Scientometrics*. Isfahan: Isfahan University of Medical Sciences Publication.

Haiqi, Z. (1996). Research Performance in Key Medical Universities in China Observed from the Scientific Productivity. *Scientometrics*, 37 (1), 177-190.

Haiqi, Z. & Zhang, Y. (1997). Scientometric Study on Research Performance in China. *Information Processing & Management* . 33 (1), 81-89

Osareh, F & Marefat, R. (2005). The Growth of Scientific Productivity of Iranian Researchers in Medline. *Rahyaft*, 35, 39-44.

Sabouri, AA & Poursasan, N. (2005). Science Production in 2004. *Rahyaft*, 34, 60-65.

Webometric Study on Iranian Universities of Medical Sciences

Farzaneh Aminpour and Payam Kabiri

aminpour@mui.ac.ir, payam.kabiri@gmail.com
Academic Evaluation & Scientometry Center, Vice Chancellery for Research,
Isfahan University of Medical Sciences, Isfahan (Iran)

Introduction

Numerous studies have been done on Webometrics especially regarding the influence of websites on each other and the reliability of Web Impact Factor (WIF) as a quantitive tool to show the visibility and impact of a website. Ingwersen (1998) investigated the feasibility and reliability of calculating impact factors for websites. He analyses a selection of seven small and medium scale national and four large web domains as well as six institutional web sites. He concluded that WIF is calculable with high confidence for national and sector domains. Smith and Thelwall (2002) published a paper entitled "Web Impact Factors for Australasian universities". They calculated the number of links into the websites of Australasian universities using a specially designed crawler and two commercial search engines, AltaVista and AllTheWeb. They mentioned that the most convenient way of measuring links between institutions is to use the advanced search facilities of commercial search engines, several of which include link data in their databases. Norouzi (2005) in his study investigated WIFs for some Iranian universities. Again**,** AltaVista was the selected search engine for counting the links to the 15 academic websites in his survey.

The academic websites of medical universities in Iran have not been exclusively taken for webometric studies. The purpose of this survey is to study the impact of all 40 universities covered by Iran Ministry of Health on the World Wide Web.

Methods & Materials

We performed the online snapshots of the websites on February 2007. The number of links into each university website was counted using the text mode of AltaVista (http://www.altavista.com/web/text). Total number of web pages indexed by the mentioned search engine was also measured. Then, the overall, inlink (internal links) and self–link WIFs for each website were calculated. Finally the universities ranked according to their inlink WIFs.

We used the following search statements in our webometric study to retrieve data for Iranian Universities of Medical Sciences. For each university the related domain name was inserted.

Total number of web pages:
domain:xxx.ac.ir/ OR domain:www.xxx.ac.ir/
The number of Total links into a website:
linkdomain:xxx.ac.ir/ OR
linkdomain:www.xxx.ac.ir/
The number of inlinks :
(linkdomain:xxx.ac.ir/ OR
linkdomain:www.xxx.ac.ir/) NOT (host:xxx.ac.ir/
OR host:www.xxx.ac.ir/)
The number of self-links:
(linkdomain:xxx.ac.ir/ OR
linkdomain:www.xxx.ac.ir/) AND (host:xxx.ac.ir/
OR host:www.xxx.ac.ir/)

Then we measured the overall, inlink and self–link WIFs as follow:

Overall WIF= The number of total links to a website divided by total number of web pages found in that website.

Inlink WIF= The number of inlinks to a website divided by total number of web pages found in that website.

Self-link WIF= The logical sum of the number of external and self-link web pages pointing to a website divided by total number of web pages found in that website.

Results & Discussion

Table 1 ranks Iranian medical universities in descending order according to their inlink WIFs. This table shows that Tehran University of Medical Sciences with 45100 pages is first in the number of web pages and also gets the highest total links (10400) but ranks 29th according to its poor inlink WIF. Yazd University of Medical Sciences which is 37th in total number of web pages ranks first due to its inlink WIF (15.89). Guilan University of Medical Sciences with 19200 web pages gets the lowest inlink WIF (0.55). Qom University of Medical Sciences by producing only 2 web pages, gets the least inlinks and total links (3&4) amongst the other universities. The highest amount of self–link WIF (0.95) in Shahrood University of Medical Sciences website shows a relatively better interconnection between its web pages comparing to the other universities. Ilam University of Medical Sciences with 6 web pages and without any self-link gets the highest overall WIF (13.50) where as Arak University of Medical Sciences with 3980 web pages and 2960 self-links gets the lowest overall WIF (0.09). The results overall suggest that Iranian universities of medical sciences are not well known internationally over the World Wide Web.

Table 1. WIF for Iranian Universities of Medical Sciences

No	University of Medical Science	URL	Total Pages	Total Links	Inlinks	Self-links	Overall WIF	Inlink WIF	Self-link WIF
1	Yazd	ssu.ac.ir	55	278	874	10	5.05	15.89	0.18
2	Ilam	medilam.ac.ir	6	81	80	0	13.50	13.33	0.00
3	Rafsanjan	rums.ac.ir	50	178	549	8	3.56	10.98	0.16
4	Mazandaran	mazums.ac.ir	152	1470	1350	117	9.67	8.88	0.77
5	Semnan	sem-ums.ac.ir	115	259	748	8	2.25	6.50	0.07
6	Kermanshah	kums.ac.ir	360	877	1630	66	2.44	4.53	0.18
7	Lorestan	lums.ac.ir	200	337	816	62	1.69	4.08	0.31
8	Zanjan	zums.ac.ir	404	967	1480	356	2.39	3.66	0.88
9	Yasuj	yums.ac.ir	167	330	607	86	1.98	3.63	0.51
10	Kermanshah	kmu.ac.ir	576	707	1560	89	1.23	2.71	0.15
11	Kordestan	muk.ac.ir	417	360	970	47	0.86	2.33	0.11
12	Shiraz	sums.ac.ir	7790	8160	15800	1490	1.05	2.03	0.19
13	Ahvaz	aums.ac.ir	687	726	1180	505	1.06	1.72	0.74
14	Jahrom	jums.ac.ir	71	180	118	59	2.54	1.66	0.83
15	Qom	qomums.ac.ir	2	4	3	1	2.00	1.50	0.50
16	Hormozgan	hums.ac.ir	1230	1700	1830	806	1.38	1.49	0.66
17	Qazvin	qums.ac.ir	246	365	306	225	1.48	1.24	0.91
18	Fasa	fums.ac.ir	677	617	719	370	0.91	1.06	0.55
19	Social Welfare	uswr.ac.ir	1820	2320	1810	1270	1.27	0.99	0.70
20	Isfahan	mui.ac.ir	5530	2570	5170	126	0.46	0.93	0.02
21	Mashhad	mums.ac.ir	1710	1800	1450	1200	1.05	0.85	0.70
22	Shahid Beheshti	sbmu.ac.ir	3610	4090	3040	2960	1.13	0.84	0.82
23	Tabriz	tbzmed.ac.ir	4370	3360	3550	1830	0.77	0.81	0.42
24	Gonabad	gmu.ac.ir	136	210	108	71	1.54	0.79	0.52
25	Hamedan	umsha.ac.ir	1640	2010	1240	1430	1.23	0.76	0.87
26	Birjand	bums.ac.ir	530	442	389	271	0.83	0.73	0.51
27	Zahedan	zdmu.ac.ir	2100	1980	1480	1280	0.94	0.70	0.61
28	Urmia	umsu.ac.ir	1080	1210	756	895	1.12	0.70	0.83
29	Tehran	tums.ac.ir	45100	10400	28000	4870	0.23	0.62	0.11
30	Bushehr	bpums.com	1010	643	506	539	0.64	0.50	0.53
31	Shahrekord	skums.ac.ir	208	249	91	156	1.20	0.44	0.75
32	Ardebil	arums.ac.ir	1020	788	294	641	0.77	0.29	0.63
33	Sabzevar	medsab.ac.ir	1440	432	405	486	0.30	0.28	0.34
34	Iran	iums.ac.ir	12600	7990	3180	10600	0.63	0.25	0.84
35	Arak	arakmu.ac.ir	3980	375	874	2960	0.09	0.22	0.74
36	Kashan	kaums.ac.ir	4140	1990	894	2940	0.48	0.22	0.71
37	Shahrood	shmu.ac.ir	98	111	17	93	1.13	0.17	0.95
38	Babol	mubabol.ac.ir	6780	2370	897	4170	0.35	0.13	0.62
39	Golestan	goums.ac.ir	706	386	91	290	0.55	0.13	0.41
40	Guilan	gums.ac.ir	19200	5490	1900	10500	0.29	0.10	0.55

Conclusion

The present Study shows that Iranian Universities of Medical Sciences are not well known internationally over the web due to the different reasons including: short history of academic websites in Iran, the few number of English web pages, poor web designs and also page viewing problems. It seems that academic webmasters must be more concerned with the factors influencing the attractiveness and visibility of academic websites.

References

Ingwersen, P. (1998). The Calculation of Web Impact Factors. *Journal of Documentation,* 54 (1), 236-243.

Noruzi, A. (2005). The Web Impact Factor:a survey of some Iranian university web sites. *Journal of Education & Psychology*, 5 (2), 105-119.

Smith A.G, Thelwall M. (2002) Web Impact Factors for Australasian universities. *Scientometrics*, 54 (3), 363-380.

.

University Groups in Flickr: Tagging for Purpose or Pleasure?

Emma Angus*, Mike Thelwall and David Stuart

*emma.angus@wlv.ac.uk, M.Thelwall@wlv.ac.uk, dp_stuart@hotmail.com
School of Computing and Information Technology, University of Wolverhampton, Wulfruna Street, WV1 1SB (United Kingdom)

Introduction

In recent years there has been a significant rise in social network websites and online applications where, 'like minded users share resources, create, tag and label content and rate it in some way' (Matthews, 2006), so-called 'web 2.0' applications.
This study investigates one of the popular new social websites, Flickr; an online image management and sharing application. Flickr allows users to upload, store and share digital images which may be 'tagged' (i.e., assigned one or more keywords) to aid in their retrieval. Flickr's 'Groups' feature allows anyone to form a public discussion and dedicated photo sharing space at Flickr in order to share images of a particular theme or topic. Groups can either be public, public (invitation only) or completely private (Bausch & Bumgardner, 2006).
There is much debate concerning both the 'usefulness' and the 'purpose' of tagging. The literature to date discusses the notions of tagging as being either for personal benefit and self organisation (Golder and Huberman, 2006; Hammond et al. 2005) or for social incentives; the opportunity to share and view other's resources (Marlow et al. 2006).

Research Questions

Using a webometric data collection, classification and informetric analysis this study investigates the tagging practices in university image groups within Flickr and aims to answer the following question: *Are the images in university groups tagged in a manner that is useful for people other than the person who has uploaded the image?*

Methods

Data Collection

Utilising Flickr's Application Programming Interface (API) a programme was written which retrieved a list of all of the groups within Flickr which had the term 'university' as part of their group ID or description (e.g., Wolverhampton University). Up to a maximum of 10 of the most recently uploaded images from each of the university groups was retrieved, from which a sample of 250 random images were taken, along with all of their accompanying tags.

Classification

Drawing on the work of the art historian, Panofsky (1983), the image indexer, Shatford (1986, 1994), and Golder and Huberman (2006), and also by visiting a random selection of tags and images, a classification scheme was developed to analyse image tags (see Table 1, column 1 in the results section). All of the 250 images were looked at alongside their tags, and tags were classified according to their perceived relationship with their accompanying image.
Mirroring the debate concerning whether tags are for 'organisational' or 'social' benefit, the classification scheme was comprised of tag categories which covered both of these concepts (see Table 1, column 2 of the results section).

Informetric Analysis

The frequency of tags used gives an indication of the range of language (tags) used and whether this corpus is following a traditional language distribution. Analysis of the number of tags used per image gives an indication of the effort put into the tagging process.

Results

A total of 2,374 university groups were identified, of which 1,983 had publicly accessible images. The initial sample of up to 10 images from each group resulted in 15,353 images, of which 12,812 had tags assigned.
The classification sample of 250 random images had a total of 2,006 accompanying tags. Table 1 shows the overall results from the investigation.

Table 1. Frequency and percentage of tag categories for 2,006 tags from random Flickr university image groups

Classification Category		Tag Motivation	Overall Percentage of Tags	Frequency & Percentage	
A1	Tag generically identifies what image is 'of'	**Useful to Flickr community as a whole** Social (Marlow et al., 2006) Altruistic (Hammond et al., 2005)	52%	309	15%
B1a	Tag specifically identifies what image is 'of' (place names/events)			289	14%
B1b	Tag specifically identifies what image is 'of' (people/animals)			43	2%
B2	Tag identifies what image is 'about'			398	21%
C1	Refining tag	**Useful only to individual/group** Organisational (Marlow et al., 2006) Selfish (Hammond et al., 2005)	39%	262	13%
C2	Self-reference tag			244	12%
C3	Task-organising tag			0	0%
C4	Tag which denotes ownership			31	2%
C5	Compound tag			233	12%
D1	Misspelling			5	0%
D3	Unable to determine relationship			5	0%
D2	Foreign word/character		9%	187	9%

Over half of all of the tags which were classified (excluding foreign words/characters) fell into categories which can be deemed to be useful for the

user community of the system as a whole, (i.e., social tags).

Number of Tags Used
Figure 1 shows the number of tags that were used per image across the university groups.

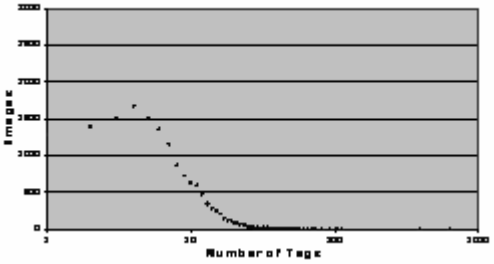


Figure 1. Number of tags used per image

A total of 53% of the images had more than four tags assigned. This suggests that image uploaders put considerable effort into tagging their images; which should hopefully aid with subsequent image retrieval.

Frequency of Terms Used
A further analysis was carried out which looked at the tag frequency for the 12,812 images which had tags assigned to them. There were 114,678 tags used in total, with 26,599 unique tags.
As Figure 2 shows, the tags used loosely follow a Zipfian distribution. The slightly skewed nature of the distribution can perhaps be attributed to the fact that language (tags) used in Flickr do not tend to be comprised of the same rich corpus of language that occurs in text based documents. Frequently occurring prepositions and conjunctions such as, 'and', 'the', and 'of' are not generally used as tags. The top two tags used were: University (0.95%) and 2006 (0.8%).

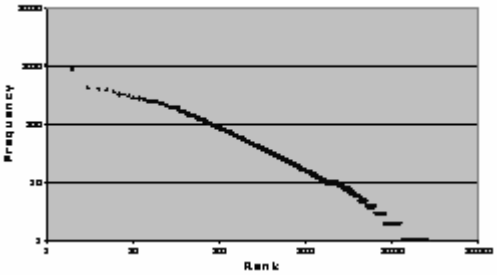


Figure 2. Frequency of terms used in Flickr university image groups

Discussion
There are obvious limitations to the classification scheme that was used for this study. Whilst the literature looked at gave a useful starting point for the development of the classification scheme, it has been found that the protocol used was not a perfect match for tags used within Flickr.

The fact that no tags were found to fall under the task-organising category (C3) can be attributed to the nature of content stored within Flickr (i.e., images rather than text-based resources).
Whilst every effort was made to ensure that the tag categories developed for the classification remained exclusive and independent of one another, it was found that many tags can lie between categories or equally well in multiple categories (i.e., an image tagged with the word 'Sarah' could be a photograph 'of' a girl named Sarah or it could 'belong' to a girl named Sarah). Looking at tags alongside their accompanying images was an attempt to minimise this variable. Nevertheless, the lines ultimately drawn reflect the classifier's own experiences, and ideas.

Conclusion
From the results presented in this investigation it seems that tags in university image groups are useful for image retrieval purposes. In support of this claim, the majority of tags are of types that are useful to the Flickr community as a whole (i.e., social tags) and image uploaders take care to assign multiple tags to their images, at least four for the majority of images. In addition, there is a broad and natural spread in the range and frequency of use of the individual tags.

References

Bausch, P. & Bumgardner, J. (2006). *Flickr Hacks: Tips and Tools for Sharing Photos Online*. Sebastopol, CA: O'Reilly.

Golder, S. A. & Huberman, A. (2006). Usage Patterns of Collaborative Tagging Systems. *Journal of Information Science*, 32(2), 198-208.

Hammond, T., Hannay, T., Lund, B. & Scott, J. (2005). Social Bookmarking Tools (I): A General Review. D-Lib Magazine, 11(4). Retrieved October 19, 2006 from: *http://www.dlib.org/dlib/april05/hammond/04hammond.html*

Marlow, C., Naaman, M., Boyd, D. & Davis, M. (2006). Position Paper, Tagging, Taxonomy, Flickr, Article, ToRead. *Proceedings of the 15th International World Wide Web Conference*, (Collaborative Web Tagging Workshop), May 22, Edinburgh, UK.

Matthews, P. (2006). The Fall and Rise of Collaborative Filtering. *Informer*, 20(Autumn), 2-5.

Panofsky, E. (1983). *Meaning in the Visual Arts*. Singapore: Peregrine Books.

Shatford-Layne, S. (1986). Analyzing the Subject of a Picture: A Theoretical Approach. *Classification Quarterly*, 6(33), 39-62.

Shatford-Layne, S. (1994). Some Issues in the Indexing of Images. *Journal of the American Society for Information Science*, 45(8), 583-588

The H-Index of H-Index and of Other Informetric Topics

Judit Bar-Ilan

barilaj@mail.biu.ac.il
Department of Information Science, Bar-Ilan University, Ramat-Gan, 52900 (Israel)

Introduction

In this poster we examine the applicability of the concept of *h*-index to topics (Hirsch, 2005), where a topic has index *h*, if there are *h* publications that received at least *h* citations and the rest of the publications on the topic received at most *h* citations. We discuss methodological issues related to the computation of *h*-index of topics (denoted *h-b* index by Banks (2006)). Data collection for computing the *h-b* index is much more complex than computing the index for authors, research groups and/or journals, and has several limitations. We demonstrate the methods on a number of informetric topics, among them the *h*-index.

The *h-b* index

How should we define the *h*-index of a topic (denoted *h-b* index by Banks (2006))? Banks does not provide a formal definition, he simply states: "The *h-b* index is found by entering a topic or compound in the Thomson ISI Web of Science database and then ordering the results in terms of citations, by largest first. The *h-b* index is defined as above in the same manner as the *h* index" (Banks, 2006, p.162).

Here, we shall provide our, more formal definition to the concept of the *h-b* index. The *h-b* index is an analog of the *h*-index, therefore:

> The *h-b index* of a set of documents is the unique number *h-b*, such that there are *h-b* items in the set that received at least *h-b* citations, and all other items in the set received at most *h-b* citations.

This is the easy part of the definition; the major question is how to obtain the set of publications on the given topic? First we see no specific reason to use the Web of Science only for the citation data; any citation database can be used for the task.

There are several options for obtaining the set of publications for which the *h-b* index is computed: The set may be compiled by experts (e.g. a bibliography for the topic), it may be the result set for a specific query retrieved from a bibliographic database, or the two methods can be combined (e.g. the set retrieved from a database undergoes peer-review by experts). Note that the database does not have to provide the citation data; this can be obtained separately from a citation database.

In case we decide to search bibliographical databases, there may be several options to choose from. Banks (2006) used the "Topic" search box on the General search page of the Web of Science (http://portal.isiknowledge.com/portal.cgi?DestApp=WOS&Func=Frame), where searches are conducted within titles, keywords and abstracts. Scopus (http://www.scopus.com) also has a comparable search option, the "search for … in article title, keywords, abstracts" option. When using this option, one has to take into account that not all items in these databases have abstracts (see Meho's comment in (Nature Newsblog, 2006)) and the keywords assignment is not uniform (they can be assigned by the bibliographical database, chosen freely by the author(s), chosen from a controlled vocabulary or thesaurus used by the publisher, etc.) and keywords are not assigned to all papers. These differences between and within the queried databases may result in missing important publications related to the topic. Note that Google Scholar (http://scholar.google.com/), the third currently available comprehensive citation database does not have an option to search within title, keywords and abstracts only, one can search either only in the title only or in the full text of the publication. Searching in the title of the publication only is too restrictive, while searching in the full text is often too noisy.

To sum up, it is highly recommended to carry out extensive searches (not only in citation databases), and then ask experts to delineate the set of documents for which the *h-b* index will be computed. In this paper we demonstrate the limitations and capabilities of the Web of Science, Scopus and Google Scholar for the computation of the *h-b* index, and provide values for several informetric topics.

Data collection

In this study we used the Web of Science and Scopus to locate articles on specific informetric topics. The following topics were selected:

- *h*-index
- Impact factor
- Citation analysis
- Informetric laws
- Link analysis
- Bibilometrics
- Scientometrics
- Informetrics

- Webometrics

The well-designed queries have to be constructed to provide good coverage of each topic without adding too much noise. Both databases (WOS and Scopus) add keywords of their own in some cases: WOS adds KEYWORDS PLUS and Scopus "index keywords".

We ran "title, abstract, keyword" searches in both databases, and report here the values of the *h-b* indices for two cases: 1) the searches were limited to title, abstract and author keywords and 2) the the searches included KEYWORDS PLUS and/or "index keywords". The usefulness of keywords added by the database and the consistency of these added keywords should be further investigated. All computations were based on publications that appeared after 1995, in order to have a fair comparison between the two citation databases. Scopus provides full citation data only from 1996 and onwards. The data was collected at the beginning of 2007.

Results

Here we only demonstrate the concept, thus in many of the cases the queries do not cover all aspects of the topic, especially for the queries informetrics, bibliometrics, scientometrics and webometrics, where we only tried to assess the relative popularity of each term. The results appear in Table 1.

Among the topics appearing in Table 1, impact factors seems to be most visible, and there are more highly-cited publications that include the term "bibliometrics" then highly cited publications with "scientometrics", "informetrics" or "webometrics" even when the searches are limited to papers that were published after 1995. In most cases there are almost no differences between WoS and Scopus. There is one exception, "link analysis" – here the *h-b* index based on Scopus is much higher, because it indexes the newsletter SIGIR Forum (http://www.acm.org/sigs/sigir/forum/), whereas WOS does not index this publication. Scopus seems to be using the index keyword "bibliometrics" rather extensively as can be seen from the difference between the *h-b* index with and without the index keywords.

Conclusion

Computing the *h-b* index for topics is a promising direction, since it seems to be a good measure of visibility. However the computation of this measure is far from straightforward.

References

Ball, P. (2005). Index aims for fair rankings of scientists. *Nature News*, 436, 900Banks, M. G. (2006). An extension of the Hirsch index: Indexing scientific topics and compounds. *Scientometrics*, 69(1), 161-168.

Banks, M. G. (2006). An extension of the Hirsch index: Indexing scientific topics and compounds. *Scientometrics*, 69(1), 161-168

Braun, T., Glänzel, W., & Schubert, A. (2005). A Hirsch-type index for journals. *The Scientist*, 19(22), 8

Hirsch, J. E. (2005b). An index to quantify an individual's scientific research output. *PNAS,* 102(46), 16569-16572.

Nature Newsblog (2006). *Top five in physics*. Retrieved November 4, 2006, from: http://blogs.nature.com/news/blog/2006/05/top_five_in_physics.html.

Table 1. The *h-b* indices of selected topics

Topic	h-WoS	h_WoS with keywords PLUS	h-Scopus	h-SCOPUS with index keywords
h index OR Hirsch-type[1]	5	5	3	3
(journal OR journals) AND ("impact factor" OR "impact factors")	27	28	29	29
"citation analysis" OR "cocitation analysis"	22	26	25	25
(Bradford OR Lotka OR Zipf) AND (law OR laws)	22	23	24	24
"Web impact factor" OR "Web impact factors"	8	9	10	10
web AND "link analysis"	8	8	13	13
bibliometric OR bibliometrics	23	24	24	30
scientometric OR scientometrics	13	15	14	14
informetric OR informetrics	13	13	12	13
webometric OR webometrics	8	10	10	10

[1] Papers on *h*-index in other senses were excluded. Hirsch-type was included to cover (Braun et al., 2005). Ball (2005) is also indexed by WOS, but it had to be added manually, since it is not covered by the query.

Bibliometric Study of Psychological Research on Tourism

Maite Barrios[*], Angel Borrego[**], Andreu Vilagines[***], Marta Somoza[**] and Candela Ollé[**]

[*]*mbarrios@ub.edu.*
Department of Metodology of Behavioral Sciences. University of Barcelona (Spain)

[**]*borrego@ub.edu, msomoza@ub.edu, cocastella@hotmail.com.*
Department of Library and Information Science. University of Barcelona (Spain)

[***]*Andreu.Vilagines@cett.es.*
University School of Tourism CETT. University of Barcelona (Spain)

Introduction and objectives

Tourism is a multidisciplinary phenomenon that has been studied from a variety of social perspectives, although it has been largely neglected by psychologists (Berno, et. al 2005). However, there has recently been an increase in the number of psychology studies in this area. The aim of the present study is to offer an overview of the psychological research in tourism using standard bibliometric techniques.

Taking 1990 as the starting-point, the temporal evolution of psychological research in tourism was analysed to determine whether the subject could be considered as a consolidated scientific discipline. Lotka's and Bradford's laws were applied to study whether the field follows the classical laws observed in other consolidated scientific disciplines. Standard bibliometric indicators such as number of authors, countries' productivity, and collaboration were analysed. The impact factor and cited half-life of the most productive journals were studied and the number of citations of the articles was also analysed.

Methods

The articles were identified using the ISI Web of Knowledge database, which includes the Science Citation Index, the Social Citation Index and the Arts and Humanities Index databases. In order to cover all the available scientific literature, the search was performed with the principal key words used in related papers (satisfaction, behaviour, motivation, attitudes, etc.). The words used to perform the search were truncated. Documents included in the study were articles or reviews. Therefore, book reviews, letters, notes, meeting abstracts, etc. were all excluded. Since data from 2006 were incomplete, documents from this year were not considered in the study. After the search, all documents were reviewed manually to eliminate any that were unrelated to the subject. The references were administered by the RefWorks software and data analysis was performed using the Statistical Package for the Social Sciences (SPSS for Windows. 12.0).

Statistical Analysis

Descriptive statistics were calculated for the variables studied. Parametric and nonparametric statistical procedures were used when appropriate for the study of the relationship between variables. Lotka's law was calculated using the least squares method. In order to verify that the observed data fitted the estimated distribution, the non-parametric Kolmogorov-Smirnov (K-S) goodness of fit test was applied (Pao, 1985).

Results

Between 1990 and 2005, 572 papers were found in the ISI Web of Knowledge database. A clear upward trend was observed from 1990 onwards, when only 2 papers were published, until 2005, with 44 papers (Figure 1). Sixty-eight per cent of the total articles in the studied period were published in the last 7 years.

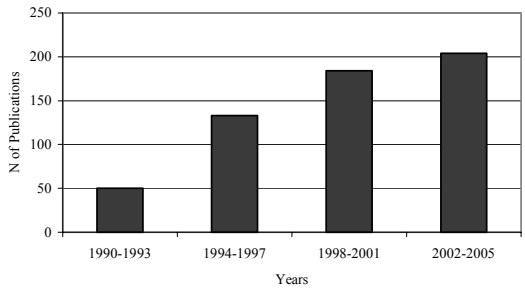


Figure 1. Temporal evolution of production

Number of authors

Eight hundred and fifty-four authors contributed to the production set. The mean number of authors per article was 1.93 (Standard Deviation (SD): 1.13, Median (Md): 2). The data showed that 40.73% of the articles had only one author, 53.34% had two or three and only 5.94% more than three. However, the evolution over time of the number of authors per article revealed an upward trend (Figure 2).

To determine whether the data fitted Lotka's law, the n value was calculated using the least squares method (n = 3.41), obtaining a C value of 0.88. The critical value was 0.055. As the maximum difference between the observed and the estimated accumulated frequencies was 0.043, below the

critical value, we concluded that the data fitted Lotka's law.

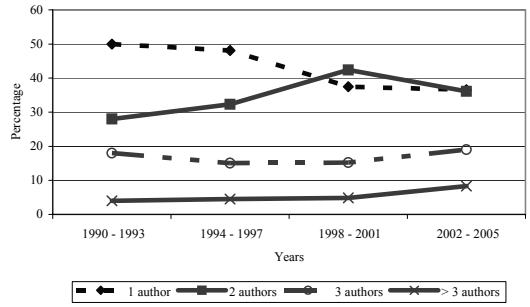


Figure 2. Evolution of the number of authors

Countries

Thirty-seven different countries contributed to the production set. The most productive country was the US (26.75%), followed by the UK (14.51%), Australia (11.54%), Canada (6.29%) and New Zealand (5.24%).

Collaboration

Taking into account only articles with more than one author (n = 339), the data show that 53.4% of the records were not performed in collaboration with other institutions (no collaboration), 26% were performed by authors from the same country (national collaboration) while 20.6% were performed by authors from different countries (international collaboration). No significant differences were found between the different kinds of collaboration and the number of authors who contributed to the articles, but in recent years an upward trend in the number of articles performed by national and international collaboration was observed (Figure 3).

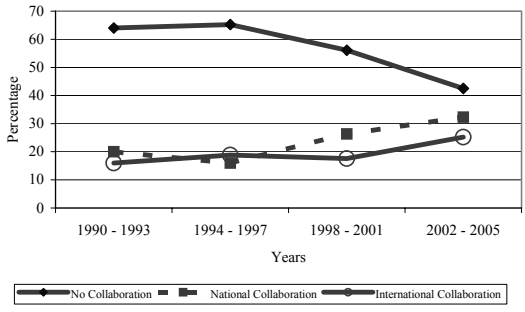


Figure 3. Temporal evolution of kind of collaboration

Journals

The articles were published in 125 different journals. The data showed that three concentric zones can be defined. The first zone showed a core journal (Annals of Tourism Research) which contained 34.1% of the articles (n = 195). A second zone also contained 34% of the articles (n = 194), published in 4 different journals. Finally, a third zone contained the rest of the articles (n = 183) published in 120 different journals. These data provide evidence that the discipline fits Bradford's law of scattering.

Impact Factor and cited half-life

Impact factor and cited half-life of the journals were studied in relation to the zones observed in the above section. The journal of the first zone had an impact factor of 0.63 and a cited half-life of 9.2. Data on the journals in the second and third zone are presented in Table 1.

Table 1. Impact factor and cited half-life of second and third zone.

	Mean (SD)	**Md**	**Mim.**	**Max.**
2n Zone				
IF	0.611 (0.12)	0.56	0.56	1.045
CHL	6.72 (1.27)	6.2	5.5	10
3rd Zone				
IF	0.94 (1.03)	0.64	0.031	8.36
CHL	7.28 (1.85)	7.25	2.7	10

Number of citations

The mean of citations of the set of records was 4.82 (SD: 6. Md: 3, Minimum value: 0 Maximum value: 41). The number of citations was also evaluated in relation to the three zones found. Articles in the first zone were cited significantly more than those in the second or in the third zone (Table 2).

Table 2. Number of citations on the three zones

	Media (DT)	**χ^2 (d.f.)**	**p**	**Dif.**	**z**	**p**
1st Zone	6.78 (7.51)	30.96 (2)	<0.001	1 - 2	4.03	<0.001
2nd Zone	3.87 (4.33)			2 -3	2.06	0.039
3rd Zone	3.75 (5.16)			1 - 3	5.11	<0.001

χ^2: χ^2 Kruskal Wallis, p: level of significance, z: z of Mann-Whitney U Test.

Conclusions

In the present study, we analysed the principal bibliometric indicators in psychological research on tourism. The results showed that this new area of science is growing and presents the classical patterns observed in other disciplines. The data fitted Lotka's and Bradford's laws and the pattern of collaboration showed that national and international collaborations are increasing.

References

Berno, T. & Colleen, W. (2005). Innocence abroad. A pocked guide to psychological research on tourism. *American psychologist*, 60, 553-600.

Pao, M.L. (1985). Lotka's law: a testing procedure. *Information Processing and Management*, 21, 305-320.

Scientometric Analysis on a Sample of Scientists in Astronomy, Particles Physics and Multidisciplinary Physics in arXiv.org (1996-2006)

Ana Isabel Bonilla*, Tony Hernández** and Isabel Gómez*

*abonilla@cindoc.csic.es; igomez@cindoc.csic.es
CINDOC-CSIC, Madrid (Spain)

** *tony@bib.uc3m.es*
UC3M, Madrid (Spain)

Introduction

According to Directory of Open Access Journals in January 2007 there are more than 2500 open access journals <http://www.doaj.org/>. The number of Institutional Repositories (IR) is growing and "It has been demonstrated that on line access improves both readership and citation impact" (Garfield, E, 2004). This can be backed up by an institutional policy to encourage or mandate authors to deposit their work in the IR (Harnad, S., 2006).

Objectives

The objective of this paper is to analyse the scientific production and collaboration in arXiv.org during 1996-2006 of a sample of the most productive Madrid authors in three disciplines of Physics, that showed high levels of co-authorship, inter-centre and international collaboration rates in SCI. Downloads and citations to these arXiv documents will be studied.

Methodology

The sample of scientists was selected considering those who published most in the following disciplines (classification based in SCI): Astronomy, Particles Physics and Multidisciplinary Physics (five authors in each one).

The documents retrieved in arXiv.org http://arxiv.org/ [Retrieved: August 15, 2006] were described by the following indicators:

- Activity indicators, based on the number of documents per author: document type, publication year.
- Collaboration: co-authorship index (CAI), defined as the average number of authors per document; Inter-centres collaboration index (IC), number of centres per document; International collaboration rate (% of documents signed by two or more countries); Large networks (six o more countries per document)
- Visibility: Signing position in the papers; Nationality of first signing position; Number of citations and downloads per document from documents covered by Citebase, a semi-autonomous citation index of University of Southampton for online research literature: <http://www.citebase.org/help/> [Retrieved: September 8, 2006]; number of citations by SPIRES HEP, a database run by the Stanford Linear Accelerator Center (SLAC) since the late 1960's as a database of Particles Physics literature <http://www.slac.stanford.edu/spires/> [Retrieved: September 8, 2006]; number of citations by Astrophysics Data System (ADS), a NASA-funded project which maintains three bibliographic databases containing more than 5.1 million records: Astronomy and Astrophysics, Physics <http://adswww.harvard.edu/> [Retrieved: September 8, 2006]

Results and Discussion

A total of 299 documents were retrieved from arXiv.org in the period 1996-2006, originated by the sample of authors analysed. The presence in arXiv has grown along this period, specially during the last three years. The most productive authors are of Astronomy (91%), followed by Multidisciplinary Physics and Particles Physics. (Table 1).

Some 81% of the documents are post-prints but there is an increase in pre-prints during the last years (2001-2006); all are in English.

Table 1. Comparison of indicators per discipline in SCI (1997-2002) and arXiv (1996-2006).

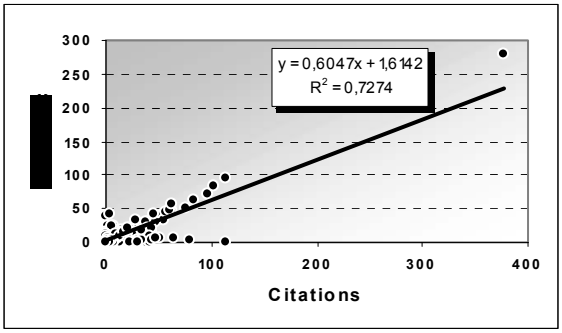
Disciplines	Nº Documents		Co-authorship Index		Average Number centres		International Collaboration Rate	
	SCI	arXiv	SCI	arXiv	SCI	arXiv	SCI	arXiv
Astronom.	234	273	9	12	5	7	81	88
ParticlesP.	229	10	36	17	10	4	59	60
Multid. P.	607	17	44	23	12	8	62	82
Total	1070	299	30	14	9	7	67	87

Note: sample=15 authors

The average yearly publications is higher in SCI in all cases, although in Astronomy the number of arXiv documents is high. In this discipline arXiv shows slightly higher co-authorship and inter-centre collaboration indexes. The opposite is the case for Particles Physics and Multidisciplinary Physics where the output co-authorship index and inter-centre collaboration are higher in SCI. The international collaboration rate is higher in arXiv than SCI in all disciplines. As to the signing

position in the papers, in 54% of the documents sample authors signed in first position and in 17% in the last one. These authors sign mostly with their Spanish address, not always, as only 44,5 have a Spanish address in first position.

Figure 1. Downloads vs Citations per document in arXiv (1996-2006) using Citebase



Citations were calculated using SPIRES HEP,

ADS and Citebase. The highest average number of citations in SPIRES HEP is in Astronomy: 32 citations document. If we use ADS; While Multidisciplinary Physics has the highest average citations per document both using ADS (34) and Citebase (40). We observed that 18% of the Citebase citations originated from SCI journals, varying per discipline (7% in Astronomy to 66% in Particles Physics). Citebase shows also data of downloads per document. Multidisciplinary Physics is the discipline presenting the highest number of downloads per document (19). The analysis of the relationship between citations and downloads in the total of documents, based on Citebase, (Figure 1) shows a positive correlation between arXiv downloads and documents citations (R=0.72). In a previous study, Kurtz et al., (2005) observed a strong correlation of the readership rate of articles (using ADS database) with the number of citations per document to astronomy journals. Moed (2005) analysed the relationship between downloads from ScienceDirect of documents in an electronic journal and citations to these documents by ISI, observing about one citation for every 100 downloads. If we compare downloads per document between post-prints and pre-prints, the average number is higher in post-prints (16 downloads in average) than in pre-prints (6) as observed in another study by Metcalfe and Fromerth (2006). Collaboration increases citations and downloads per document. There is a positive correlation between arXiv citation and co-authorship index (R=0.41) and number of authors per document (R=0.40). If we analyse documents with one foreign address, the average is 18 citations and 12 downloads, higher than he average number in total documents. In documents with at least six foreign addresses, it increases to 30 citations and 17 downloads in average.

Conclusions

The presence of documents from the sample of Spanish Physicists in this Open Access Repository has grown during the period 1996-2006. The most productive authors were those of Astronomy.

- There is a positive correlation between arXiv downloads and document citations.
- The average number of downloads is higher in post-prints than in pre-prints.
- There is a positive although moderate correlation between arXiv citation and co-authorship index and number of authors per document.
- International Collaboration and large networks increase the average number of citations and downloads per document.

We have to stress that these observations are preliminary results obtained with a small sample of physicists that will be extended to the whole population of Spanish physicists in the disciplines analysed.

References

Garfield, E., (2004). American Scientist Open *Access Forum listserv, 18 January 2004.* Available:http:www.ecs.soton.ac.uk/harnad/Hypermail/Amsci [Retrieved: October 14, 2006]

Harnad, S., (2006). Generic rationale and model for university open access self-archiving mandate.Available: http://openaccess.eprints.org/index.php?/archives/71-guid.html.73 [Retrieved: January 29, 2007]

Metcalfe, T.S. and Fromerth, M, (2006). The Citation Impact of Digital Preprint Archives for Solar Physics Papers, *Solar Physics*, 239, 549-553.

Kurtz M. J. et al., (2005) The bibliometric properties of article readership information, *ASIST,* 56, 2, 111-12

Moed, H.F. (2005). Statistical Relationship Between Download and Citations at the Level of Individual Documents Within a Single Journal. *Journal of the American Society for Information Science and Technology,* 26, 10, 1088-1097.

Spontaneity of Consciousness: V.V. Nalimov's Most Ingenious Book

Manfred Bonitz

bonitz@fz-rossendorf.de

Vassily Vassilievich Nalimov (1910–1997) is one of the greatest Russian thinkers of the XXth century.

In the community of scientometricians he is mainly known from his book (Nalimov, 1969) "Naukometriya" (Scientometrics) and he is therefore considered as one of the fathers of the discipline scientometrics.

However, V.V. Nalimov has written a lot of other books, which were partly translated into English and other languages, partly they appeared only in Russian (see the list of references below which is far from being complete. Nalimov wrote more than 300 scientific papers in the course of this life).

Without any doubt one of the most important works is "Spontaneity of consciousness" (Nalimov, 1989), which is printed only in Russian – a German translation is under way.

It is difficult to adequately describe the contents of the book, because Nalimov's personality itself is of high complexity, his knowledge of unbelievable encyclopaedicity.

Nalimov is a mathematician, his probabilistic thinking (see below) allows him to construct a model of the world, which is of high metaphoric power. But Nalimov is far more than a mathematician. He is a philosopher par excellence, in order to illustrate his philosophical views, he delivers a grandiose panorama of the philosophers from the past and from the present – from the Orphics, Pythagoras, Socrates Plato, Aristotle, Dionysius Areopagite to Leibniz, Wittgenstein and Sartre, to name just a few of those whom he is quoting extensively.

The philosophical background of all great world religions and cultures is carefully analyzed, Nalimov repeatedly returns to the fate of his native country Russia, where he was repressed under the Stalin regime for 18 years (1936–1954).

What is personality? What is the sense of life and death? Can we transcend our bodily existence? What is consciousness? What is spontaneity? These and many other questions are answered in a surprisingly consistent manner. But what is the secret of Nalimov's approach, of his model of the world?

To my opinion it is the combination of several powerful tools permitting Nalimov to eventually unite and unify realities which seemed incompatible before.

The first tool is the Bayesian formula (Bayesian syllogism) which in its probabilistic application can take away a lot of difficulties arising from Aristotelic logics.

The second tool is Nalimov's understanding of meanings. Meanings are existing axiomatically as long as people are living and using language. The infinite number of meanings can be arranged like real numbers on an axis, or "packed", thus representing a semantic vacuum. "Unpacking" of certain meanings for a certain purpose – creating a certain text – happens thru spontaneous application of a simple Bayesian formalism to the semantic vacuum.

What follows is a grandiose idea of Nalimov: according to him everything, the whole world, is a text – the physical world (the body and its brain) and the world "beyond" the physical world (soul and mind) can be handled by the same tools. We are arriving at a unified physical and semantical theory of the world. For Nalimov it is one world after all, what had been strictly divided for centuries.

Nalimov reading is delightful. He isn't a missionary, rather a friendly partner for discussion. He likes to stress that we haven't achieved more than Plato 2300 years ago – we are just expressing it in another, probabilistic language.

References

Nalimov, V.V. & Mul'chenko, Z.M. (1969). Naukometriya. (In Russian). Moscow: Nauka. 192 p.

Nalimov, V.V. (1989). Spontaneity of consci-ousness.Moscow: (In Russian). Prometej. 287 p.

Nalimov, V.V. (1994). The rope dancer. Remi-nescenses. (In Russian). Moscow: Progress. 456 p.

Nalimov, V.V. & Drogalina, Zh.A. (1995).
Reality of the irreal. (In Russian). Moscow: Mir Idej, AO Akron. 432 p.

Nalimov, V.V. (2002). The temptation of sacred Russia. (In Russian). Tomsk-Moskow: Vodolej. 288 p.

Nalimov, V.V. (2003). Probability model of language. 3rd ed. (In Russian). Tomsk–Moscow: Vodolej. 368 p.

.

Searching for Converging Research using Citations between Subject Categories: First Results and Challenges

R.K. Buter and E.C.M. Noyons

buter@cwts.leidenuniv.nl, noyons@cwts.leidenuniv.nl
Centre for Science and Technology Studies (CWTS), Leiden University (the Netherlands)

Introduction

Science is in a constant state of flux, everywhere and anywhere: internationally, locally, socially, economically, politically, cognitively, and so on. Having the ability to monitor changes in the structure of science would have obvious advantages (Pinski (1976); Urata (1990)): for example for science policy issues like prioritising and monitoring.

We are currently involved in a research project that investigates how to develop a monitor that identifies emerging and converging research areas on the basis of bibliometric methods and data. We do not focus on specific parts of science, but are looking in *all* the parts that are covered by the *Web of Science* (WoS). In the first part of this project, we concentrate on finding *converging* research.

Convergence is strongly associated with the concept of interdisciplinarity (e.g. Palmer (1999), Morillo (2003)), but it focuses on the *dynamics* of the research structure rather than the *state* or set-up of the research structure. Also, we are looking for indicators of fast growth (Tabah (1999)), as we assume that convergence is characterized by a (strong) growth in the number of citations between the converging areas. Also, we are less concerned with the *nature* of the convergence, which may be the result of e.g. collaboration between (actors from) different research areas, the incorporation of knowledge from one area into another or the creation of a multidisciplinary or interdisciplinary research area.

Since it is hard to define the extent of "a research area", we used the high-level ISI *subject categories* (Morillo (2003)). Of course, there are other ways of extracting high level structures from bibliometric data (Small (1974), Todorov (1988)), but the subject categories have the advantage that they are readily at hand. For these subject categories, we collected the citations between them for 11 consecutive years. Then, the main research question can be formulated as follows: is it possible to detect converging research from the changes in the number of citations between *all* subject categories collected over a period of 11 years?

Data

We extracted from the CWTS database, which is based on the WoS, all publications published between 1995 and 2005. We then combined these papers with their cited papers in as far as these are also covered by the WoS. The first publication year of these cited papers is 1980, the last 2005.

Growth was identified using the ratio of the citation frequency in year *t* divided by the citation frequency in the year before that, i.e. year *t-1*: *c(t)/c(t-1)*.

Results

Selection of categories

We start with all 59,536 (244 times 244) possible category pairs (i.e. citing category combined with the cited category). We applied the following process on the citation counts that were collected using a citation window of *one* year (since this seemed to capture the recent developments better).

-First, a hard limit of 200 citations over the *whole* period was applied, resulting in a reduction of 90%, leaving 5,979 pairs.
-Then, a growth ratio of above 1.0 over the last *five* years was applied, reducing the number of pairs to 185 (97%).
-Finally. for the *last two* years (using the remaining 185 pairs):
-We calculated the average growth ratio, selecting those that had a growth ratio larger than this *overall* average, resulting in 24 pairs.
- We calculated the average growth *within* each of the 185 pairs and selected those that had in these years a ratio above their own average, resulting in 22 pairs.We then combined the pairs from 3a and 3b and removed the duplicates, resulting in 38 distinct pairs.

Assessment of the selection

This research project is a collaborative effort, with participants distributed over different parts of Europe. To avoid having to duplicate the sizeable amount of the bibliometric data, we created a website that centralised the access to the bibliometric indicators and data. Using this website, the 38 distinct category pairs were assessed. Finally, four categories were selected for a more in-depth

analysis, which will also consult field experts to give an informed opinion on the nature of the increased citation traffic.

Discussion

Although our research is still on-going, we can already draw several conclusions with respect to the selection procedure.

First, we should further investigate the effect of the multiple-assigned journals on the selection process, as they tend to dominate as source of the increased levels of citation, for the reasons indicated above.

Next, we should apply a more flexible lower limit on the number of citations. The applied limit is, as illustrated in Table 1, a rather "harsh" limit: for the one year citation window, it *a priori* excludes ninety percent of the data. As a result, the current selection is dominated by "big" categories, i.e. categories that have a lot of publications and references associated. One possibility could be to divide the category pairs into groups on the basis of their total number of papers or their total number of references and apply selection procedures on each of these groups.

Also, we should consider investigating how the *distance* between the categories (as visualised in Figure 1) can be included in the selection process, since categories that are already close together in the first years of analysis tend to be less interesting as a result than categories that started further apart.

Acknowledgements

The research described in this paper is part of the EC-funded project 'New Emerging/Converging Clusters of Science and Technology' (Contract 2005/S 48-045721)

References

Small, H., Griffith, B.C. (1974). The structure of Scientific Literatures I: Identifying and Graphing Specialties. *Science Studies* 4, 17-40.

Todorov, R. (1988). Representing a Scientific Field: a Bibliometric Approach. *Scientometrics,* 15(5-6), 593-695.

Pinski, G., Narin, F. (1976). Citation Influence for Journal Aggregates of Scientific Publications: Theory, with Application to the Literature of Physics. *Information Processing & Management*, 12, 297-312.

Urata, H. (1990). Information Flows among Academic Disciplines in Japan. *Scientometrics*, 18(3-4), 309-319.

Palmer, C.L. (1999). Structures and Strategies of Interdisciplinary Science. *Journal of the American Society for Information Science*, 50(3), 242-253.

Morillo, F., Bordons, M. & Gomez, I. (2003). Interdisciplinarity in Science: A Tentative Typology of Disciplines and Research Areas. *Journal of the American Society for Information Science*, 54(13), 1237-1249.

Tabah, A.N. (1999). Literature Dynamics. *Annual Review of Information Science and Technology*. pp. 249-286.

A Review of New Web Tools for Citation Analysis and their Impact on Research

Rosaria Rosanna Cammarano, Maurella Della Seta

rosaria.cammarano@iss.it, maurella.dellaseta@iss.it
Istituto Superiore di Sanità, SIDBAE - Documentation Service, Viale Regina Elena, 299 - 00161 Rome (Italy)

Background

Citation analysis has been used since the mid-20th century as a tool to trace scholarly research, measure its impact, and, in some countries, to justify funding decisions. Eugene Garfield, from the Institute of Scientific Information of Philadelphia (ISI), in 1955 developed the idea of constructing a citation index for science, based on the assumption that the more an article is cited in subsequently published papers, the higher its impact on the scientific community (Garfield, 1972). Web of Science has been for many years the unique database allowing citation tracking. In the last three years, new tools for academic and scholarly literature ranking and evaluation analysis made their appearance on the information scene (Roth, 2005). Some of them are available free-of-charge, other upon payment.

The purpose of this presentation is to analyze the literature published on these new searching tools in the last three years, and to present the results emerged from these recent studies.

Methods

The following products were considered in this research:

Google Scholar <http://scholar.google.com> is a free-of-charge search engine aimed at finding scholarly information on the Web. Its search includes bibliographic references and the full-text of peer-reviewed papers, theses, books, abstracts, technical reports, and basically any Web page that looks like an article or an academic document.

From recent literature (Friend, 2006) it clearly emerges that an increasing number of students and scientists is consulting Google Scholar to retrieve peer-reviewed research. Today Google Scholar provides its own version of citation information, with a ranking technology that reports how often the item has been cited in other scholarly literature (Noruzi, 2005).

Scopus http://www.scopus.com is a subscription-based multidisciplinary database with full-text links, produced by the Elsevier Co. Scopus contains 29 million abstracts with citation back to 1966. Results from searches can be sorted by date, relevance, title and by the "Cited by" option. The "Cited by" feature will list the documents according to the number of citations received and list articles that cited the article examined. The Citation Tracker feature, available to all subscribers, enables users to find most highly cited authors in a field, to find and track hot topics in specific subject areas, to check the most up-to-date citation data on specific authors and articles, and, consequently, track and evaluate research trends (Bakkalbasi, Bauer & Glover, 2006), providing information about the influence of a set of articles, an author, or a group of authors.

Scirus <http://www.scirus.com> is a free-of-charge search engine produced by Elsevier that focuses on scientific, scholarly, technical and medical disciplines. Scirus searches over 300 million science-specific Web pages, Medline citations, Elsevier ScienceDirect full-text articles, BioMed Central full-text articles and patents from the USPTO. Scirus has a wide range of features to help the user searching for scientific and academic information (Felter, 2005). Search results are ranked according to relevance and they may also be sorted by date. No specific tools for citation tracking are provided.

Windows Live Academic Search <http://academic.live.com> is a free search service produced by Microsoft as a competitor of Google Scholar (Payne, 2006). It helps students and researchers to perform searches on academic and scholarly journals. The search results are free but users must have a subscription to a journal to access the full text of an article. This search service allows users to sort the search results by author, date of publication, conference in which the paper was presented, and the journal in which it was published (Soellner, 2006). It does not use citation count as a factor in determining relevance.

CiteSeer <http://citeseer.ist.psu.edu>, developed by the NEC Research Institute, Princeton, New Jersey, is a freely available, web-based resource, that provides access to scientific and information science literature (Connor, 2005). It's useful to find bibliographic citations of papers, articles, books and technical reports, create an index of citations and place them in context with the works cited. It's possible to know the "most cited authors in computer science" (August 2006) and "most accessed documents in Citeseer" (as June 2003), although the system warns that these lists may contain errors.

Procedures

Scientific literature concerning the above mentioned searching tools was retrieved by bibliographic

databases, such as PubMed, Embase, SciSearch, Pascal, and Biosis. Using as a limit the year 2004, a total of sixteen citations relating to new tools was retrieved. Moreover, a web search was performed by means of search engines such as Google, Yahoo!, Altavista; the names of the engines or databases and the words "citation analysis" and citation track*" were used as the input for searching. The retrieved papers were analyzed for comparing results.

Results

Effective options for citation analysis are available in Google Scholar, Scopus and Web of Science. Many of the articles retrieved (e.g.: Jacso, 2005; Bauer & Bakkalbasi, 2005; Della Seta & Cammarano, 2006) performed a comparison between new tools and the well-known Web of Science database. Results show that all these searching tools offer some unique scientific information, due to different contents and documents analyzed. Consequently, no resource may be considered sufficient for the citation counting of an article or an author. Web of Science offers the most comprehensive coverage back in time, because it goes back to 1945, and is still a primary, although expensive, resource for citation searching. Nevertheless, for some subject areas, specialized search engines as Google Scholar may offer the best coverage, thanks to their ability to retrieve documents that do not appear in academic journals, but are part of the growing collection of scholarly information on the Web. Moreover, since resulting citations do not match exactly, due to the difference in type of documentation considered by search algorithms, the new tools should be used to integrate findings from Web of Science. All these instruments can be seen as complementary, although a good percentage of overlapping material results from many studies.

References

Bakkalbasi, N., Bauer, K., Glover, J. & Wang, L. (2006). Three options for citation tracking: Google Scholar, Scopus and Web of Science. [Electronic version]. *Biomedical Digital Libraries*. 3(7), 1-19

Bauer, K. & Bakkalbasi, N. (2005). An examination of citation counts in a new scholarly communication environment. [Electronic version]. *D-Lib Magazine,* 11(9), 1-10

Burnham, J.F. (2006). Scopus database: a review [Electronic version]. *Biomedical Digital Libraries,* 3:1-8

Butler, D. (2004). Science searches shift up a gear as Google starts Scholar engine. *Nature, 25; 432* (7016), 423.

Connor, E. (2005). Searching for Science: a descriptive comparison of CiteSeer, FirstGov for Science, and Scirus. [Electronic version]. *Journal of Electronic Resources in Medical Libraries*, 2(2):35-47

Della Seta, M. & Cammarano, R.R. (2006). Google Scholar and Web of Science: Similarities and Differences in Citation Analysis of Scientific Publications. [Electronic version]. In: *10. Conference of the European Association for Health Information and Libraries. Proceedings*; *September 11-16, 2006; Cluj-Napoca*.

Felter, L. (2005). Google Scholar, Scirus, and the scholarly search revolution. [Electronic version] *Searcher,* 13(2), 43-48

Friend, F.J. (2006). Google Scholar: potentially good for users of academic information. [Electronic version]. *Journal of Electronic Publishing,* 9(1).

Garfield E. (1972). Citation Analysis as a Tool in Journal Evaluation. *Science*, (178), 471-479, Retrieved January 23, 2007 from: http://www.garfield.library.upenn.edu/ essays/V1p527y1962-73.pdf

Giustini, D. (2005). How Google is changing medicine. [Electronic version]. *BMJ*, Dec 24; 331(7531):1487-8.

Henderson, J. (2005). Google Scholar: A source for clinicians? [Electronic version]. *CMAJ*, 172 (12): 1549-50

Jacso, P. (2005). As we may search. Comparison of major features of the Web of Science, Scopus, and Google Scholar citation-based and citation enhanced databases. [Electronic version]. *Current Science,* 89(9), 1537-1547.

Noruzi, A. (2005). Google Scholar: The New Generation of Citation Indexes [Electronic version]. *LIBRI*, 55(4):170-180.

Payne , D. (2006, April 25). Microsoft launches Google Scholar rival. *The Scientist. News.* Retrieved January 23, 2007 from: http://www.the-scientist.com/news/display/23346

Roth, D.L. (2005). The emergence of competitors to the Science Citation Index and the Web of Science. [Electronic version]. *Current Science*, 89(9), 1531-1536

Sherman, C. (2006, April 12). Microsoft launches Windows Live Academic Search. Retrieved January 23, 2007 from: http://searchenginewatch. com/showPage.html?page=3598376

Soellner, K. (2006). Google Scholar and Windows Live Academic search – current developments with scholarly search engines. *Bibliotheksdienst,* 40(7), 828-837

Tarantino, E. (2006). Troppo o troppo poco? Web of science, Scopus, Google scholar: tre database a confronto (un caso di studio). *Bollettino AIB,* 1/2; 23-32

Vine, R. (Reviewer) (2006). Google Scholar. [Electronic version]. *Journal of Medical Library Association*, 94(1), 97-99.

Exploiting the Potential of Researchers CV Databases for Policy Making: Evidence from the Ramón y Cajal Programme in Spain

Carolina Cañibano[*,***] Javier Otamendi[*] and Inés Andújar[**]

[*]*carolina.canibano@urjc.es, franciscojavier.otamendi@urjc.es*
Universidad Rey Juan Carlos. Paseo de Artilleros s/n. 28032 Vicálvaro, Madrid (Spain)

[**]*ines.andujar@uam.es*
[**]Universidad Autónoma de Madrid. Campus de Cantoblanco. 28049 Madrid (Spain)

[***]Instituto de Investigaciones Económicas y Sociales "Francisco de Vitoria". Ctra. M-515, Pozuelo, Majadahonda, Km.18.00. 28223 Pozuelo de Alarcón, Madrid (Spain)

Introduction

This paper results from the interest in exploring alternative inexpensive ways to manage public funding research programs more efficiently. The study on which the paper relies delves into the researchers' curriculum vitae (CV) as a rich source of longitudinal data with great potential, alike studies that have been conducted in the United States (Dietz et al., 2000). In particular, it focuses on one of the CV collections which are available at the Spanish Ministry of Education and Science (MEC): the Ramón y Cajal Program (RyC). The Ministry has several of these CV collections, as researchers are compelled to submit a standard CV when they apply for Governmental funding. Up to the current date and to our knowledge, CVs are used for candidate selection purposes. But they could be managed in a better way, speeding up candidate selection processes and providing new tools for policy making and evaluation.

Spain is currently increasing its efforts to encourage the training and employment of researchers. The MEC received in 2004 around 20.000 applications linked to the different human resources training and mobility programs. The funding of grants and contracts is growing, but the amount of personnel available for their management remains stable. This implies that the Ministry's staff needs to devote more time to the management of calls, applications, grant reports, etc. and that scarce resources are available to evaluate the programs in some detail. However, the fact that the MEC receives so many applications to the different human resources programs every year means that it collects the same number of very complete and standardized CVs.

We present here the result of a methodological exercise aimed at exploring ways in which some more resources can be devoted to policy evaluation at the MEC, without increasing cost. One possibility is to try to speed-up the selection of candidates by using software and statistical tools. The question is to determine whether a quick and reliable computer program could help in the acceptance or rejection of applicants. Ideally, if an electronic CV format was designed properly, a computer program should be able to transform the CV into a set of variables, and execute a mathematical model to discriminate the applicants in two groups. At least, if not the final decision, a screening decision could be taken quickly, reliably and cheaply.

The data

The data we have used to explore the potential of the MEC CV format has been extracted from CVs of researchers who applied for the 2005 RyC grants. The access to the data was framed by a confidentiality agreement signed between the research team and the MEC, and approved by the National Agency of Data Protection. 1771 researchers applied that year, out of which 246 were granted. As major differences in professional trajectories and research outputs appear between fields, 3 areas were selected to perform the analysis: Physics (100 coded applicants in the sample out of which 17 were granted), Molecular Biology (100 applicants/13 granted) and Philosophy & Philology (66 applicants/ 6 granted). Physics and Molecular Biology were the fields receiving a higher number of applications. Including Philosophy & Philology responds to our interest in performing also the analysis for a human sciences area, which is assigned a lower number of grants within the RyC program.

Although our CV collection has the important advantage of relying on a standardized very complete format, the coding exercise showed that many issues remain unclear concerning how the CV template must be filled out and the meaning and content of some of the CV sections, particularly the "mobility" section.

The data analysis

The analysis addresses the following questions: are there systematic characteristics within each field between selected and rejected applicants? Could we

determine out of the information contained in the CV, the probability that an applicant is granted by the RyC program?

Different models are developed for each field separately. Table 1 presents the success ratio of the different models, showing the probability obtained in each one of assigning an applicant to the proper group (granted/ not granted).

Table 1. Success ratio within classifying models

SUCCESS RATIO	DA_LINEAR	LOGIT	PROBIT	GOMPIT	ALL CORRECT
TOTAL	76%	89%	88%	89%	71%
PHYSICS	87%	89%	88%	92%	83%
MOLECULAR BIOLOGY	85%	93%	92%	94%	81%
PHILOSOPHY AND PHILOLOGY	89%	n.a.	64%	97%	52%

Out of the discriminant methods, binary logistic regression has proven to be better than pure Linear Discriminant Analysis, with the Gompit model being the most consistent (above 89% success ratio in all the situations). However, Discriminant Analysis is quicker and always reaches a solution, even with small samples like for the Philosophy field in this case.

Table 2 shows however that only few variables determine the discrimination of applicants within each field.

Table 2. List of significant variables

	TOTAL		PHYSICS		MOLECULAR BIOLOGY		PHILOSOPHY & PHILIO.	
	T-TEST MEANS	LOGIT	T-TEST MEANS	LOGIT	T-TEST MEANS	LOGIT	T-TEST MEANS	LOGIT
YEARS SINCE OBTAINING PHD				*				
LONG PREDOCTORAL STAY				**				
LONG POSTDOCTORAL STAY				*				
VISITING COUNTRIES	*	*			**			
VISITING CENTRES IN USA	**	**	**		*			
CONFERENCES	*	*	**	**	**		*	
BOOKS WITH ISBN IN POSTDOCTORAL PERIOD		**					**	
CHAPTERS WITHOUT ISBN IN POSTDOCTORAL PERIOD				*				
ARTICLES WITH IMPACT INDEX IN POSTDOCTORAL PERIOD	**		**					
ARTICLES WITHOUT IMPACT INDEX IN POSTDOCTORAL PERIOD				*				
NATIONAL GRANTS			**		**	*		
SUPERVISED PHD THESIS		*	**	*				

* P-VALUE<0.10
** P-VALUE<0.05

Table 3. List of non significant variables

gender; short predoctoral stay; short postdoctoral stay; visiting centres; books with isbn in predoctoral period; books without isbn in predoctoral period; books without isbn in postdoctoral period; chapters with isbn in predoctoral period; chapters with isbn in postdoctoral period; chapters without isbn in predoctoral period; articles with impact index in predoctoral period; articles without impact index in predoctoral period; patents; international grants; research contracts

It is clear that discrimination tools must take fields into account since the characteristics of selected applicants differ between fields.

Discussion

Further research is undoubtedly needed. The relevance of the performed exercise does not yet rely on the particular significance of each discriminatory variable within each field. The major outcome is methodological. Discriminant models built out of variables extracted from applicants' CVs prove to be rather successful in assigning candidates to the correct group (granted /not granted). This might be considered as a first step towards designing tools that help selection committees in taking objective quick decisions by speeding up a first screening phase.

Several national and regional agencies are currently collaborating in Spain towards the implementation of a standard electronic CV database which would include CVs from all researchers applying for public funding[1]. This would prevent us in the future from undertaking the coding of the variables and will provide large samples of CVs that can be used to improve the accuracy and effectiveness of this sort of discrimination tools.

References

Corley, E; B. Bozeman and M. Gaughan (2003) "Evaluating the impacts of grants on women scientist' careers: the curriculum vita as a tool for research assessment". Shapira, P & Kuhulman, S (Eds). *Learning from science and technology policy evaluation: experiences from the United States and Europe*. Cheltenham, UK: Edward Elgar. pp. 293-315.

Dietz, J.; I. Chompalov, B. Bozeman, E. O'Neil Lane and J. Park (2000) Using curriculum vita to study the career paths of scientists and engineers: an exploratory assessment. *Scientometrics*. Vol.49, No.3. pp.419-442

Gaughan, M. and B. Bozeman (2002) Using curriculum vitae to compare some impacts of NSF research grants with research center funding. *Research Evaluation*. Vol.11, No.1, April, pp.17-26.

Mangematin, V. & Robin, S. (2003) "The double face of PhD Students : The example of Life Sciences en France". *Science and Public Policy*, 30/6, 405-414.

[1] For information concerning this project visit: http://www.fecyt.cvn.org/

Contribution to the Impact Factor of Academic Journals in the Field of Education and Educational Psychology of Citations to Articles Authored by Editorial Board Members

Mª Angeles Coslado* and Juan Miguel Campanario**

*coslado@uah.es
UAH (Universidad de Alcalá de Henares), Departamento de Física, Edificio de Ciencias, 28871 Alcalá de Henares, Madrid (Spain)

**juan.campanario@uah.es
UAH (Universidad de Alcalá de Henares), Departamento de Física, Edificio de Ciencias, 28871 Alcalá de Henares, Madrid (Spain)

Introduction and purpose

The impact factor, developed by Eugene Garfield, is a fundamental citation-based measure of the significance and performance of scientific journals. It is perhaps the most popular bibliometric product currently in use (Glänzel & Moed, 2002). In this work, we study the structure of the impact factor of academic journals. In a previous study, we studied the citations that contribute to impact factor coming from papers authored by editorial board members and published in any journal (Campanario, González & Rodríguez, 2006). Now, our new approach is based on calculation of citations to articles published in a given journal by the members of the editorial board. Editorial board members sometimes publish articles in their own journals (Campanario, 1996) and these articles could attract citations that contribute to impact factor. We try to ascertain how these citations contribute to impact factor.

Method

A total of 44 journals were studied from the groups *Education and Educational Research* and *Psychology, Educational* from the Social Sciences Citation Index. We selected the documents labelled by the ISI as "articles" published in 2002. Articles are usually peer reviewed and they usually report original research. We discard other document as, for example, "editorial material" or "letters" because these materials are not usually considered so valuable as original articles. For each journal, we divided the articles in two sets: those that had at least one author being an editorial board member of such journal and those that were written by external authors. According the definition of impact factor, given article published in 2002 can contribute only to the impact factor of 2003 and 2004. Thus, we computed the mean of citations that contributed to the impact factor (CIF) of 2003 and 2004 received by both sets of articles.

Results

Table 1 shows the results. We can classify the journals according the mean of citations that contribute to the impact factor obtained by each set of articles. There are six journals in which CIF to articles by external authors are much more higher than CIF to articles by editorial board members. In these cases, the mean of CIF to articles by editorial board members is zero. There are other six journals in which CIF to articles by editorial board members are at least two times higher than CIF to articles by external authors. In the remaining journals, citations to both sets of articles contribute in a similar amount to the impact factor.

Conclusions

This method can be used to compute the structure of impact factor. There are different kind of journals according the contribution to the impact factor of citations to articles by editorial board members. These authors are usually among the more prestious and productive in their fields, so, it is not rare that their papers tend to attract citations. However, this pattern is not the only observed. We are now studying a three years window in order to obtain more general conclusions.

Acknowledgments

This research was supported by Junta de Comunidades de Castilla-La Mancha FEDER, research project PAI-05-011.

References

Campanario, J.M. (1996) The competition for journal space among referees, editors and other authors and its influence on journal's impact factors. *Journal of the American Society for Information Sciences*, 47, 184-192

Campanario, J.M.; González, L.; Rodríguez, C. (2006) Structure of the impact factor of academic journals in the field of Education and Educational Psychology: Citations from editorial board members. *Scientometrics*, 69, 37-56.

Glänzel, W. & Moed, H.F. (2002), Journal impact measures in bibliometric research. *Scientometrics*, 53: 171–193.

Table 1. Mean of citations that contribute to the impact factor (CIF) of 2003 and 2004 received by articles published in 2002 by external authors and editorial board members in the journals studied.

Journal	External authors	Editorial board members	Total
APPLIED MEASUREMENT IN EDUCATION	0,60	0,67	0,61
BRITISH JOURNAL OF EDUCATIONAL TECHNOLOGY	0,48	0,00	0,46
BRITISH JOURNAL OF SOCIOLOGY OF EDUCATION	0,96	1,22	1,03
COMPARATIVE EDUCATION	0,38	1,00	0,41
CONTEMPORARY EDUCATIONAL PSYCHOLOGY	1,60	2,00	1,69
DISCOURSE PROCESSES	1,91	1,00	1,83
EDUCATION AND URBAN SOCIETY	0,39	0,25	0,37
EDUCATIONAL ADMINISTRATION QUARTERLY	0,54	1,71	0,95
EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT	1,30	1,50	1,32
EDUCATIONAL EVALUATION AND POLICY ANALYSIS	3,27	0,00	3,06
EDUCATIONAL POLICY	0,45	1,00	0,47
EDUCATIONAL PSYCHOLOGIST	3,33	4,33	3,50
EDUCATIONAL PSYCHOLOGY REVIEW	1,70	3,00	1,92
EDUCATIONAL REVIEW	0,27	0,50	0,29
INSTRUCTIONAL SCIENCE	0,80	0,00	0,76
JOURNAL OF ART & DESIGN EDUCATION	0,12	0,33	0,14
JOURNAL OF CREATIVE BEHAVIOR	0,64	1,00	0,69
JOURNAL OF ECONOMIC EDUCATION	0,30	1,00	0,34
JOURNAL OF EDUCATIONAL AND BEHAVIORAL STATISTICS	0,71	0,80	0,74
JOURNAL OF EDUCATIONAL MEASUREMENT	1,27	0,00	1,12
JOURNAL OF EXPERIMENTAL EDUCATION	1,08	3,50	1,65
JOURNAL OF GEOGRAPHY IN HIGHER EDUCATION	1,00	1,50	1,13
JOURNAL OF HIGHER EDUCATION	1,24	1,00	1,22
JOURNAL OF LITERACY RESEARCH	1,27	0,80	1,13
JOURNAL OF PHILOSOPHY OF EDUCATION	0,50	0,75	0,53
JOURNAL OF PSYCHOEDUCATIONAL ASSESSMENT	0,85	0,63	0,76
JOURNAL OF RESEARCH IN SCIENCE TEACHING	2,03	2,58	2,18
JOURNAL OF SCHOOL PSYCHOLOGY	2,21	0,00	1,91
JOURNAL OF TEACHING IN PHYSICAL EDUCATION	0,81	1,00	0,85
LEARNING AND INSTRUCTION	2,72	3,00	2,73
MEASUREMENT AND EVALUATION IN COUNSELING AND DEVELOPMENT	0,71	1,67	0,85
PSYCHOLOGY IN THE SCHOOLS	1,02	1,69	1,18
READING RESEARCH QUARTERLY	1,50	2,60	1,87
READING TEACHER	0,51	0,74	0,57
RESEARCH IN HIGHER EDUCATION	0,85	0,80	0,82
SCHOOL EFFECTIVENESS AND SCHOOL IMPROVEMENT	1,00	0,00	0,88
SCIENCE EDUCATION	1,09	1,67	1,14
STUDIES IN HIGHER EDUCATION	1,48	2,00	1,50
TEACHERS COLLEGE RECORD	1,26	1,50	1,27
TEACHING AND TEACHER EDUCATION	0,61	0,80	0,63
TEACHING OF PSYCHOLOGY	0,56	1,14	0,66
TEACHING SOCIOLOGY	1,11	1,00	1,10

Changes and Development of Key Indicators in JCR on the Web

Dang Yaru, Wang Liya, Gao Feng, Li Kan, Huang Yue and He Fenglan

dangyaru@sina.com
Civil Aviation Science and Technology Educational Evaluation Research Center,
Civil Aviation University of China, Tianjin (China)

Introduction and results

We use the data included in JCR on the web from 1997 to 2005, such as the number of articles, the cited frequency, the impact factor, immediacy index and so on, to calculate the total indicators, the average indicators and the maximum indicators in the science and social version respectively. The total indicators include the total number of journals, the total number of articles and the total citation frequency. The average indicators contain the average articles, the average citation frequency, the average impact factor and the average immediacy index. The maximum indicators consist of the maximum number of articles, the maximum number of IF, the maximum number of the citation frequency and the maximum number of the immediacy index. Besides, we calculate the quality indicators of every country and each discipline from 2003 to 2005, comprising the rate of self-citation and citation, citing journal number, cited journal number, total cited peak, total citing peak, self-citation peak, diffusion factor and impact index and so on, and analyze them respectively.

The total indicators show the overall features of JCR. In terms of such data and our calculation we can conclude that during the nine years total number citation increased faster than that of countries publishing journals and the total number of journals. The average indicators can reveal the use and the degree of emphasis of the journal, and its role and status in scientific exchanges. Through the data included in the science version of JCR and our calculation, we can conclude that the growth rate of immediacy index is faster than the citation frequency and impact factor; In the social science version, it shows a completely different situation. Although the citation speed has increased and the average citation frequency is great, the growth rate of immediacy index and that of impact factor are not synchronized relatively. And compared with the growth rate of citation, it is a little bit slower. Maximum indicator reflects the strength, the degree of advantage and the degree of uniformity distribution of all the indicators.

In the science version of JCR, from 1998 to 2003 J BIOL CHEM maintained the maximum number of articles and the maximum citation frequency, ANNU REV IMMUNOL maintained the maximum impact factor and CA-CANCER J CLIN maintained the maximum immediacy index; In the social science version, with the exception of 1997, FORBES maintained the maximum amount of articles in those years, but its amount of citation and its impact factor have decreased. The cited half-life is the indicator of the rate of aging journals. The number of journals is the statistic object, and the range of cited half-life has been divided into four regional range: ＞10, [10, 5), [5 ,0) and 0 (none). There is great difference between social science journals and science journals in the distribution of half-life, but both still have some things in common: almost 50% of the journals whose half-life distribute is between 5 and 10 (including 10).

Due to the limit of data source, we only calculated the quality indicators from 2003 to 2005, and analyzed them. In term of the changes of the self-citation and citation during the three years, the total self-citation frequency and the average self-citation frequency gradually increased, but the self-citation ratio gradually declined and citation ratio gradually increased. According to the themes given by JCR, we united them and got 13 the first level subjects in the science version and 12 in the social science version. Then we calculate the number of cited journals, the number of citing journals, the diffusion factor and the impact index.

We calculate the distribution of cited peak, citing peak and self-citation peak in terms of the data of JCR, as shown in Figure 1 and Figure 2. Moreover, we provide the peak of each journal, which can be shown in Figure 3 and Figure 4.

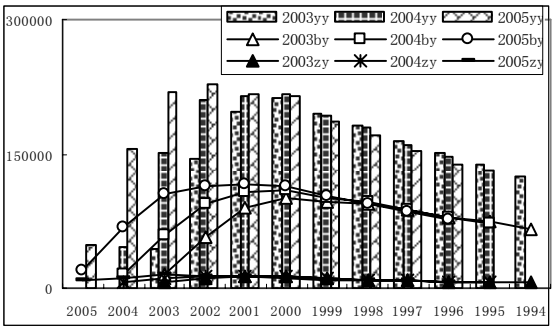


Figure 1. Distribution of peak in science

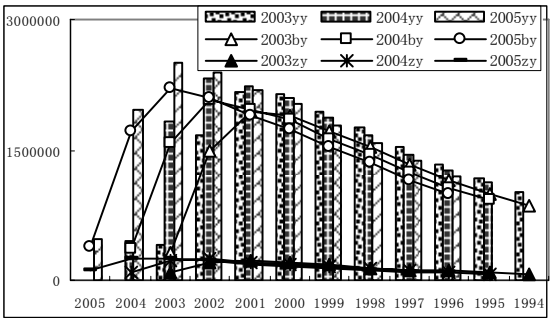


Figure 2. Distribution of peak in social science

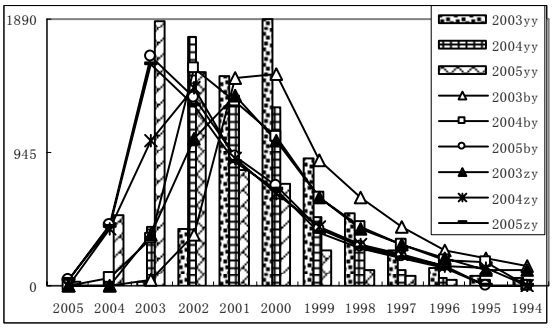


Figure 3. Distribution of the number of peak in science

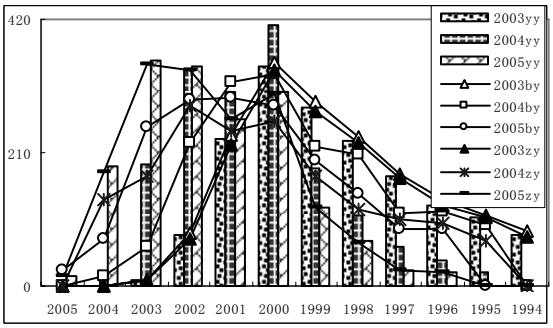


Figure 4. Distribution of the number of peak in social science

From Figure 1 and Figure 2, the distributions of the citing peak in social science version and science version are completely different: in science the distribution follows the Price law and there is a big difference between peaks; but in social the life cycle of citation is longer than in science, and the peak year is the time when the paper published after three or four year, moreover, there is a little difference between peaks. And the same in the cited peak , but there is some thing in different: in science there is a little difference between citing frequency and cited frequency, but in social science the cited frequency is far less than the citing frequency. In the social science there is a little difference between self-citation peak journals, but in science it is obvious.

In fact, the cited frequency and the citing frequency of each journal don't reach the peak after it published for two or three years. As Figure 3 and Figure 4 shown, although the cited frequency, the citing frequency and self-citation frequency of the most of the journals can reach the peak within the scheduled time, it will cost longer time to reach the peak for that of the almost half. In the science version the curves of the three kind of peak form the distribution of Price, however, in the social science version the curves of those are tousy.

Conclusions

In term of our analysis on JCR we get the following conclusions:

Firstly, the great growth of academic articles and citation reflects that the researcher pay more attention to the standardization of citation. Accordingly, the transmission speed of research increase, and the self-citation and cross-citation of discipline become more frequent.

Secondly, the development of networks allows scientists to obtain information more easily, making the speed of research increase and the new disciplines have more growth points. Information technology accelerates the development of science.

Thirdly, the acceptability and spread speed of science research are higher than those of social science research. The social science research contains more historical and cultural accumulation, which can be reflected in citation where the average immediacy index of social science is far below that of science.

Fourthly, due to half-life increase, the lifecycle of science research results has been prolonged, which indicates that the degree of substitution of research declined, while the continuity strengthened on one hand; the difficulty of the current research increased on the other.

Finally, the reduction rate of the self-citation shows the way using citation becomes more scientific and the current scientific research becomes more extensive, exoteric and cooperative.

References

http://portal.isiknowledge.com/portal.cgi?DestApp=JCR&Func=Frame，2006-09-20.

Eugene Garfield,Alexander Pudovkin. Algorithmic Procedure for Finding Semantically Related Journals. *Journal of the American Society for Information Science and Technology*, 2002, 53(13):1113-1119.ISI. http://www.isinet.com. 2006-05-20

The Relationship of Price Index and Median Citation Age

Lixin Chen*, Zeyuan Liu* and Liming Liang**

*lynnchenlixin@gmail.com, liuzy@vip.163.net
Dalian University of Technology, WISE LAB, School of humanities & social sciences, Dalian, 116023 Liaoning (P.R. China)

**pllm@public.xxptt.ha.cn
Henan Normal University, Institute for Science ,Technology & Society, Xinxiang, 453002 Henan (P.R. China)

Introduction

Citing half life and Price index are two important indicators of literature obsolescence. In 1958, J. D. Bernal borrowed the radio term "half-life" to describe the literature aging process, named as citing half life of documents. Later, R. E. Burton and R. W. Kessler (1960) defined citing half life as: for the all references, how long the newer publications were cited more than half in a long period of time. This definition has been called median citation age. Price (1970) proposed the ratio of the references published no more than five years in the all references as an indicator to measure the speed of literature obsolescence. The index is also known as the Price Index, the N-year period Price Index was marked as P (N).

Although citing half life and Price Index is two different concepts, but there is certain relationship between them (Egghe 1997; Cunningham 1995).

Format Data and Methods

In 2004, we downloaded data from Web of Science (Thomson-ISI). We choose 106 journals about mechanics according to the category of Science Citation Index-Expanded (SCIE). In SCIE database we got more than 200,000 documents from 1945 to 2003. Total citation data are more over 2.9 million, removing about 50,000 of the non-standard data (including all in-pressed documents). With the remaining citations (total 2,869,099) data, we established our original citations database.

There are two methods to research literature obsolescence, synchronous and diachronous, we choose the former, synchronous.

Analysis and Results

According to the above data and research methods, we calculated median citation age and Price index P (1) - P (10) of the mechanical journal from 1945 to 2003, then drew points chart (see Figure 1).

Except for large fluctuations in the early years, from 1951 to 2003, the median citation age was constantly extended in past 53 years. From 1951 to 2003, Price Index P (1) - P (10) curves were decreasing as time went by, in the past 50 years.

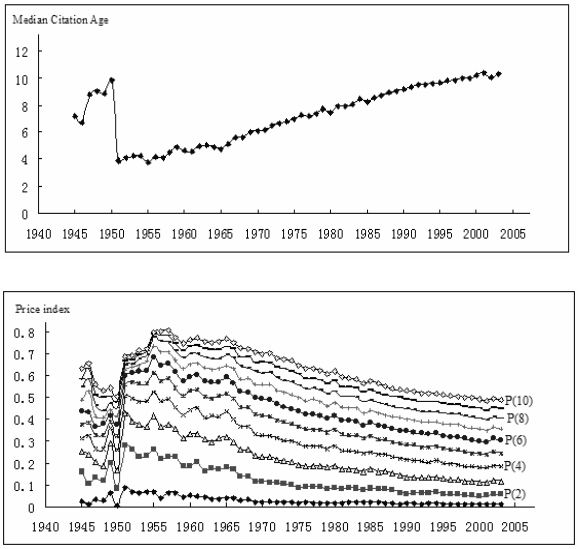


Figure 1. Median Citation Age Curve & Price Index P (1) - P (10) Curve (1945-2003)

For further analysis of the quantitative relationship of them, we calculated correlation coefficient of median citation age and Price index (1) ~P (46) from 1951 to 2003 (see Figure 2). In Figure 2, with the change of Price index P (N), correlation coefficient changed, and shows that median citation age and Price index P (3) - P (8) have strong negative correlation, the peak value is -0.9893 with Price index P (4).

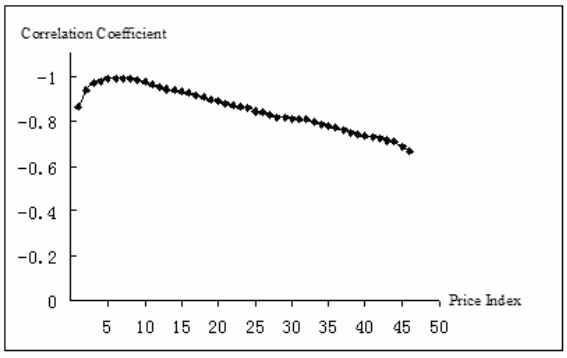


Figure 2. Correlation coefficient of median citation age and Price index P (1) ~P (46) (1951-2003)

From Figure 1 and Figure 2, we saw that median citation age and Price index maybe have a negative power function relationship:

$P(N) \times H = C$

in which, P (N) is Price Index, H is median citation age, C is constant.

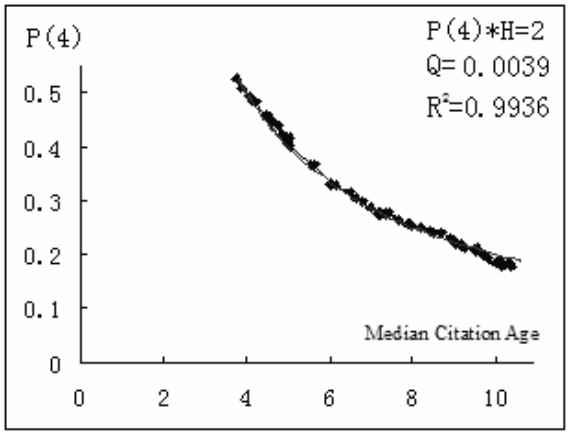
After calculation, we obtained the six negative power function regression equation (see table 1).

Table 1. Six negative power function regression equation

Equation	C	Q	r^2
P(1)*H=C	0.167	0.008	0.619
P(2)*H=C	0.726	0.049	0.786
P(3)*H=C	1.395	0.027	0.941
P(4)*H=C	2.002	0.004	0.994
P(5)*H=C	2.532	0.023	0.967
P(6)*H=C	2.987	0.104	0.856
P(7)*H=C	3.387	0.254	0.642
P(8)*H=C	3.736	0.46	0.317

Note: C is constant value; Q is the residual sum of squares of Price index, r is correlation coefficient.

From Table 2 and Figure 3, we found the best regression equation is P (4) *H=2, with the lowest residual sum of squares, only 0.004, and relevant coefficient is 0.9936, close to 1. We used our data do a test for P (4) *H=2; the estimated value of median citation age is less than 10% of the observed value of that.



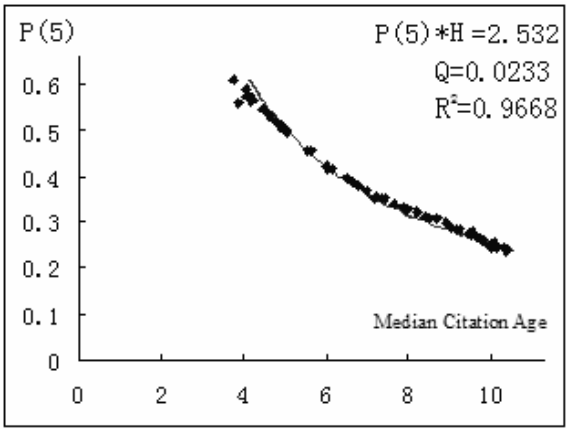


Figure 3. Regression equations of median citation age and Price index P (4), P (5)

Conclusions and discussion

In the field of mechanics, median citation age and Price index have simple quantitative relationships: P (4) *H=2. So we can use Price index P(4) to calculate median citation age rather than Price Index P (5), it is simple and with high precision.

In mathematically (Egghe 1993) and practically (Wallace 1986), it is proposed that with the increasing of literature, median citation age become shorter and short and the speed of obsolescence become faster and faster. But our case is opposite.

Acknowledgments

The work done for this paper was partly funded by National Natural Science Foundation of China (NSFC Project No. 70373055).

References

R. E. Burton, R. & W. Kebler (1960). “Half-life” of some scientific & technical literature. *American Document,* 1960 11: 18-22.

D. S. Price (1970). Citation measure of hard science, soft science, technology & nonscience. Communication among scientists & engineers, Heath Lexington, 1970, 3~22.

L. Egghe (1997). Price index and its relation to the mean and median reference age. Journal of the *American Society for lnformation Science*, 1997, 48(6): 564~573.

S. J. Cunningham, D. Bocock (1995). Obsolescence of computing literature. *Scientometrics*, 1995, 34(2): 255~262.

L. Egghe (1993). On the influence of growth on obsolescence. *Scientometrics*, 1993, 27(2): 195~214.

D. P. Wallace (1986). The relationship between journal productivity and obsolescence. *Journal of the American Society for Information Science*, 1986, 37(3): 136~145.

.

Capacity Building in Specialties towards National Health Research through Research Funding

Ms. Sandhya Diwakar* and Dr. KK Singh**

*sandhyadiwakar@yahoo.com
Indian Council of Medical Research, Ansari Nagar , New Delhi (India)

**singhkeshari@yahoo.com
Indian Council of Medical Research, Ansari Nagar , New Delhi (India)

Background

India faces the twin epidemic of continuing/emerging infectious diseases as well as chronic degenerative diseases. All though the death rate has been declining but health care structure in the country is over-burdened by increasing population. Medical research can help reduce the disease burden of the country by providing us with very valuable footholds. Research capacity strengthening is one of the most powerful, cost effective and sustainable means of advancing Medical Research.

Indian Council of Medical Research as an apex body in India for promotion of medical research started Financial Assistance to Post Graduates pursuing MD/MS/DM/MCH with the primary purpose of providing the evidence base for policy making, practice guidance and programme development for health care. This programme forms a powerful instrument to arrive at a selection of funding priorities that takes into account National/International trends but also enhance linkages between the spectrum of key stakeholders.

Objective & Methodology

Does India perform medical research in areas where it is most needed? This paper attempts to map India's Disease Burden to the medical research being done in the country. The former reflects the needs while the latter reflects the areas in which research is being done.

This paper is based on Disease Burden reviews by National Commission on Macroeconomics and Health, information obtained from the internet, unpublished material, presentations , abstracts and reports of WHO. ICMR's MD/MS/DM/MCH financial assistance program reflects the preferred medial research areas in India.

Observations & Key Findings

Available Statistics on Disease burden shows that ARI, Prenatal Illness, Cardiovascular and Mental health disorders are the major medical problems faced by India.

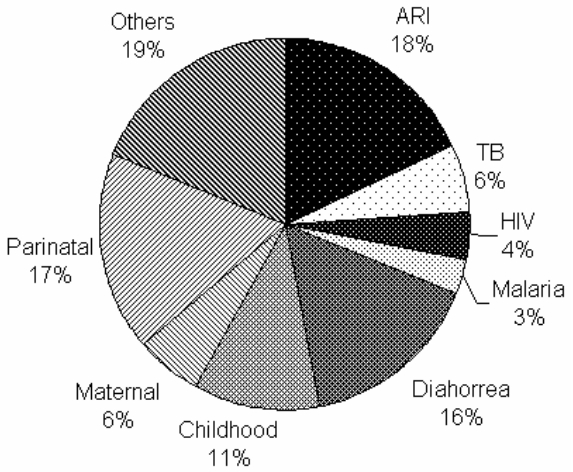


Figure 1. Communicable diseases, % share in Disease Burden

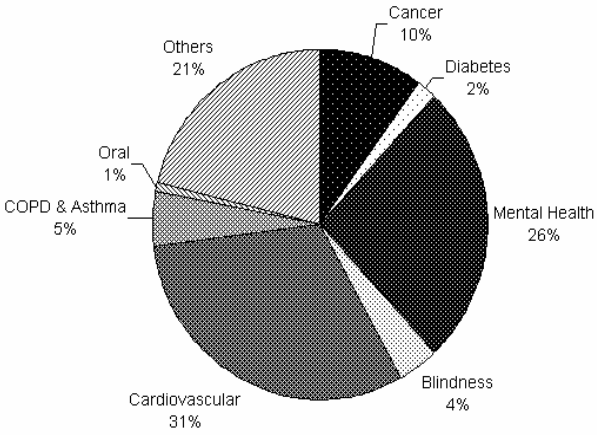


Figure 2. Non-communicable diseases, % share in Disease Burden

As per the Report of WHO the projected disease burden of India for 2020 is maximum for Malaria (29%) followed by Measles (27%), Suicide (19%), Prenatal Problems (16%), and Diarrhea (11%). Analysis of data from National Sample Survey, undertaken by the National Commission on Macroeconomics and Health (NCMH) suggests that the out-of-pocket expenditure by individuals hospitalized on account of heart disease was roughly Rs 11,000 per person, or 120% of the average annual per capita expenditure of the households they belonged to.

In order to study the preferred research areas, Medical research areas of 76 Post graduates under ICMR's MD/MS/DM/MCH financial assistance program were analyzed.

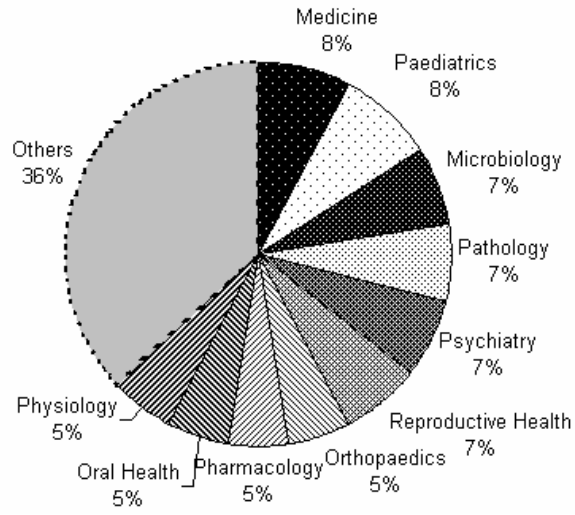


Figure 3. % Preferred Medical Research areas, based on the MD/MS/DM/MCH program

This analysis reveals that
Medicine and Pediatrics are the most preferred research areas with a total of 16% of the total researchers opting for these areas.
Microbiology/Virology,Pathology, Psychiatry and Reproductive Health/ Obstetrics & Gynecology account for a whopping 26 % of the preferred research areas, with each area accounting for 7 %
Each of the areas Orthopedics, Pharmacology, Oral Health and Physiology account for 5 % of total preferred research areas.
Other preferred research areas are Biochemistry, Community Medicine, ENT, Hematology and Respiratory Medicine, Endocrinology, Nephrology, Surgery and Anesthesiology.

Conclusion & Key Findings

Data on the observed Indian research in different subfields of medicine show that the work done in India in general is not integrated well with the need or the disease burden of the country. The comparative study with disease burden certainly indicates that researchers have to be motivated to carry out research in areas of interest of Public Health. Therefore, there is a need to intensify research activities in these priority subject areas to fill the gap.

India could be much more purposive in her research priorities and probably should invest much more in priority medical research identified by the Disease Burden study.

It is our hope that these small research exercises / findings may help in generating larger research questions and initiating large-scale studies. If the required infrastructure and manpower are made available, these studies can then be used and significant data can be scrutinized and improved. For in this way, the question of whether the estimates are right or wrong will become secondary to the more fundamental issue of whether or not they prove useful for promoting public health.

References

Health Information of India 2004,Government of India, Central Bureau of Health Intelligence, DGHS, MOH & FW, Nirman Bhavan, New Delhi-11 website:www.cbhidghs.nic.in

The Global Burden of Disease Christopher J.L.Murray & Alan D. Lopez, WHO

Bull Health Organ, Volume 84, Number 2, February 2006

National Commission on Macroeconomic and Health

http://www.who.int/macrohealth/action/NCMH_Burden%20of%20disease_(29%20Sep%202005)

Impact of the Research Stays on the Scientific Output

Daniela De Filippo*, Sanz-Casado, Elías **, Isabel Gómez*

*dfilippo@cindoc.csic.es, igomez@cindoc.csic.es
CINDOC (CSIC), Madrid (Spain)

**elias@bib.uc3m.es,
Department of Library Science and Documentation. Carlos III University of Madrid (Spain)

Introduction

During the last decades the interdisciplinarity and internationality of science have produced a great increase of scientific collaboration. At the moment, collaboration is one of the central elements of scientific activity.

The importance of the evaluation of R+D activities has originated the development of diverse methodologies to study scientific activity and its results, and in this context collaboration indicators acquire a central importance. On this line, the collaboration between groups is an important aspect of research, since research stays in other centers can favour the relation between institutions and foster the development of common projects and joint publications. According to Kyvik and Larsen (1994), the informal communication between scientists is fundamental in the process of creation of new knowledge and is expressed through electronic communications, conversations in congresses and research stays. These ways of communication have as main motivation the search of greater interchange and links with other researchers, that might well originate collaboration. Therefore the study of the research stays can shed light on the analysis of scientific collaboration.

Objectives and hypothesis

Considering the research stays as an important element to enhance scientific collaboration, the present work analyzes the relationship between the mobility of the researchers and their scientific output. The hypothesis is that those groups with researchers who have made research stays increase their production in collaboration and also tend to increase the quality and the impact of their publications.

Sources and methodology

The study is centered in the analysis of the researchers of the Carlos III University of Madrid (UC3M). This center has been selected because it is a multidisciplinary institution -what enables to analyze the behaviour of collaboration in different fields of knowledge-, because its medium size facilitates the study of the whole of its researchers. Besides, we could have access to institutional sources to complement the results obtained.

In order to know the characteristics of research stays, in a first stage we studied the Annual Reports of UC3M of academic courses 1997/1998-2002/2003.
To quantify the impact of this mobility on the scientific production the international Thomsom Scientific Web of Science (WoS) (SCI, SSCI, A&HCI) in the period 2001-2005 was used.
The documents were obtained through a search in the field "Address" of the name of the institution and its different variables. Indicators were produced following the methodology successfully used at CINDOC (Fernandez et al: 1993). All the departments of origin of the researchers have been identified and, in those cases in which this information was not recorded, the names of the authors have been consulted in the institutional data bases to assign the corresponding department.

The documents were classified into two groups: documents “with mobility” (signed by authors with at least one research stay or that have received foreign researchers) and documents “without mobility”. We have detected that the intra-institutional collaboration (documents signed by authors of two or more departments of UC3M) is almost nonexistent; those documents in co-authorship correspond to members of the same group. Therefore, when the group has a researcher with mobility, it is considered that this document corresponds to the category "with mobility" since the bonds of this researcher with the external center favour the whole group, both as to collaboration and publication. The, results of the researchers "with mobility" were compared to the rest of documents of UC3M.

Results

Results of the total production of UC3M have been obtained to determine their characteristics and collaboration profiles. Next we analysed the documents of researchers “with mobility” versus these of researchers “without mobility”. The main results obtained are the following.

General results of the UC3M output

- In the period 2001-2005 a total of 1535 UC3M documents were recorded in ISI databases.

- The scientific areas of greater production are: Engineering and Technology (48% of doc), Social

Sciences, Physics and Mathematics (main areas of specialization of the University). Humanities publish mostly in national journals and show a low number or international publications.
Mathematics has an Impact Factor (IF) higher than the average of Spain and Social Sciences shows a number of cites per document larger than the national average.

- The greater production is that of the Polytechnic School (71% doc) and specially its departments of Mathematics, Physics and Computer Science. The department of Economy of the Faculty of Social and Legal Sciences have also a relevant output.

- The collaboration patterns show some 33% of documents in international collaboration, especially with the USA, United Kingdom, France, Germany and Italy. The national collaboration occurs mainly with other centers of Madrid.

Research stays and scientific output
- In average, 22% of the professors in UC3M have made stays in other centres.

- The major ratios of stays per professor are in the Faculty of Humanities, Communication and Documentation (table 1).

Table 1. Research stays by faculty and professor

Faculty	N Prof.	%	N Stays	%	Stay/ Prof.
Polytechnic	746	41	251	37	0.34
Soc.& Law	902	50	343	51	0.38
Human. & Docum.	170	9	75	11	0.44
Total	**1818**	**100**	**675**		**0.37**

- The groups “with mobility” have published around 62% of documents of UC3M

- As to the document type, articles predominate with a higher percentage than article average of UC3M. The language is mostly English.

- Publications in Engineering/Technology, Physics and Mathematics predominate. In these areas some 60-70% of documents of UC3M correspond to groups “with mobility”. We point out that the average IF of Engineering and Technology and the Activity Index (AI) of Physics and Mathematics are higher than the average of UC3M. Also the documents "with mobility" of Engineering and Technology show a number of cites per document, higher than the average of the University. Humanities journals have not IF, although some of them can belong also to Social disciplines.

- The distribution by centers shows the high output of the Polytechnic School and evidences that in most cases, the documents "with mobility" present better indicators than the rest (see table 1).

- At the departmental level it is observed that in 3 of the 4 more productive departments (Mathematics. Physics and Economy) more than 70% of documents correspond to groups “with mobility”.

- The international collaboration rate is 38%, five points over the average of UC3M.

- There is a strong relationship between the main destinity of the research stays and the co-authorship with countries and centers. Most frequent mobility and co-authorship collaboration take place with USA, United Kingdom, Germany, France and Italy. At the level of centers, the more relevant are: Oak Ridge National Laboratory (U.S.A.), CNRS (France), the Max Planck Institute (Germany) and the Catholic University of Lovaine (Belgium).

Table 2. Comparative indicators

Faculty	Research Stays	N Doc	%	IF	% non Cited doc	Cites /Doc	Int. Coll. Rate
Polytech	Yes	679	62.6	1.406	52.6	1.5	37.1
	No	405	37.4	1.338	65.9	1.4	22.0
Soc.& Law	Yes.	224	64.9	0.959	58.9	1.2	44.2
	No	121	35.0	0.871	72.7	0.7	22.3
Human & Doc...	Yes.	8	36.4	0.764	100.0	0.0	0.0
	No	14	63.6	0.396	92.9	0.0	0.0
GISC*	Yes.	39	97.5	3.329	30.8	4.8	48.7
	No	1	2.5	0.513	100.0	0.0	0.0

*Interuniversity group in complex system

Conclusions
The results show that a positive relationship exists between mobility and scientific production and they produce positive effect on the quality and the visibility of publications. This points out need the interrelation with other centers and the benefits of these relations that derive from research stays.

In order to continue the studies on this subject, the differences of production by areas and disciplines in each department will also be analyzed. We will continue with this research line following Mahlck & Persson (2000) methodology. The differences of visibility, quality and impact between the groups “with mobility” and “without mobility” will be studied.

References
Kyvik, S. & Marheim-Larsen, I. (1994). International Contact and Research performance... *Scientometrics*, 29, Nº 1, 161-175
Mahlck, P. & Persson, O. (2000) Socio-bibliometric mapping of intra-departmental networks. *Scientometrics*. 49, Nº 1, 81-91.

The Spanish Presence in Editorial Boards of International Scientific Journals: A Tool for the Promotion of Spanish Science

E. García-Carpintero, B. Granadino, N Sastre and L.M. Plaza

bgranadino@cindoc.csic.es, lplaza@cindoc.csic.es
Centro de Información y Documentación Científica (CINDOC), Consejo Superior de Investigaciones Científicas (CSIC), Joaquín Costa, 22-28002 Madrid (Spain)

Introduction
The main objective of this work is to determine the correlation between the Spanish scientific output covered a set of SCI journals ranking with highest impact factor, and the presence of Spanish researchers in the editorial boards of these journals.

The presence of scientists of a country in the editorial boards does not imply a higher scientific output in these journals for that country, although it can contribute to improve the visibility of its scientific output of that country.

Methods
We have analyzed the composition of the editorial boards of the first 100 journals included in the 2004 Journal of Citation Report - Science Edition by ISI, according to their impact factor. It has analyzed the composition of editorial boards of 20 first journals, with higher impact factor, of 15 subject categories. The studied subject categories are Astronomy, Mathematics, Physics, Chemistry, Engineering, Telecomunica-tions, Psychiatry, Neurosciences, Materials Sciences, Food Science, Biochemistry, Biotechnology, Cell biology, Medicine, Immunology. This methodology will allow to determine the participation of Spanish scientists in the most relevant scientific journals. The analysis of the Spanish scientific output in these journals during the period 2000-2004, will allow us to determine if the Spanish scientific output is higher in journals in which Spanish researchers are present in their corresponding editorial boards.

Results
The total number of members in the 100 editorial boards of the 100 SCI journals with highest impact factor is 3473 (34.7 members per journal as average). From these 0.46% are Spanish memberships and they are active in the editorial boards of 9 journals. A 9% of the journals studied have at least one Spanish researcher in their editorial boards.

The number of Spanish researchers in editorial boards (16) is low when compared with other countries with similar scientific characteristics, e.g. Italy and The Netherlands which have 54 and 41 members respectively. Their researchers included in the editorial boards participate in a higher number of journals covering a wider spectrum of scientific fields. As expected, United States is the country with a major presence in the editorial boards with 2292 members (66.0% of the total number of members). United Kingdom, Germany and France are the European countries with a higher number of members in the editorial boards with 361 (10.4%), 174 (5.0%) and 83 members (2.4%), respectively (Figure 1).

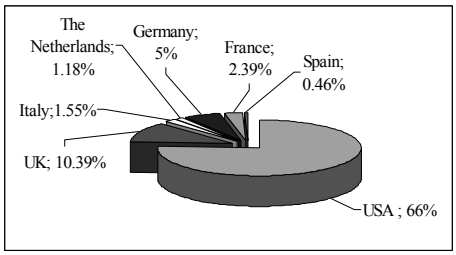


Figure 1. Percentage of members in Editorial boards of first 100 studied journals.

A 91% of journals among the first 100 studied journals are edited in the countries with major presence in the editorial boards. A 70% of journals are edited in United States and a 21% are edited in United Kingdom (figure 2).

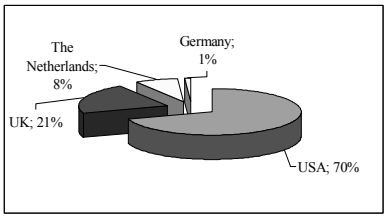


Figure 2. Percentage of countries publisher of 100 first journals.

We have identified 877 European editorial boards memberships of 20 different countries. Analyzing the editorial boards and the scientific output of each country, it can be observed a positive correlation between both indicators (figure 3).

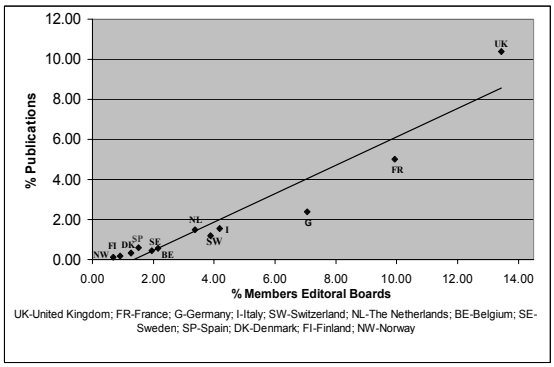


Figure 3. Presence in editorial boards vs scientific output in the top 100 journals.

Considering the 20 first journals corresponding to the 15 subject categories previously cited, a simple distribution can be observed; United States (54%), UK (24.91%), The Netherlands (7.83%) and Germany (3.91%) (figure 4).

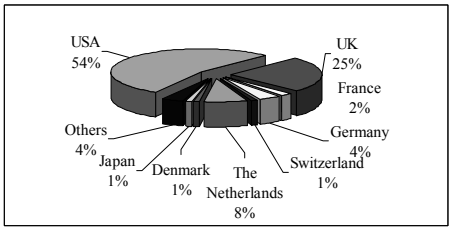


Figure 4. Percent of countries publishers of journals included in the 15 selected subject categories.

Results from the analysis corresponding to the 15 subject categories, show that concerning to European countries UK, Germany and France are those with a major number of memberships in their editorial with 9.84%, 5.53% and 3.5% respectively. It can be noted that Italy that represents a 2.55%, has a significant number of members in the editorial boards of telecommunications (17.9%) and cell biology (11.6%). In the case of Spain with a low representation of 1.01%, only food sciences, material sciences and immunology are significantly represented.

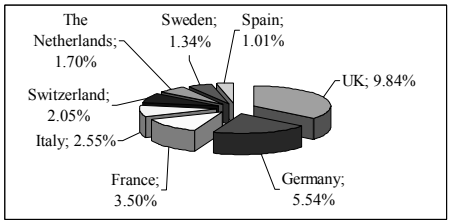


Figure 8: Number of members in editorial boards of journals included in the 15 selected subject categories.

Conclusions

Indicators studied are useful for the discussion and the proposal of measures to enhance the Spanish presence in the editorial boards of the most important scientific journals by comparing the situation with other countries.

A higher presence of Spanish scientists in the editorial boards of the 100 journals with higher impact factor will probably contribute to an increase of the Spanish Scientific output covered by these journals. However, it is necessary not only to increase the number of members in these editorial boards but also to increase the number of manuscripts submitted by Spanish authors for a higher visibility of the Spanish research in order obtain a higher presence and visibility at world-wide level.

References

Albert A.; Plaza L.M.; Granadino B. (2006). El flujo de conocimientos desde el Sistema público español de I+D a las industrias biotecnológicas. En Radiografía de la investigación pública en España. Red CTI/CSIC, *Editorial Biblioteca Nueva, Madrid, 373-392.*

Barron J.P., Harrison B., Iijima K., Breugelmans R., Yamamoto K, (2005). A survey of medical journal publishing in Japan: editorial board structure, review systems and ethics. *European Science Editing* 31 (3), 79-81.

Braun T., Dióspatonyi I, (2005). The counting of core journal gatekeepers as science indicators really counts. The scientific scope of action and strength of nations. *Scientometrics* 62(3), 297-319.

Campanario J.M. (1996). The competition for journal space among referees, editors and other authors and its influence on journals´ impact factors. *Journal of the American Society for Information Science*, 47, 184-192.

Plaza L.M.; Bordons M. (2006). Proyección internacional de la ciencia española. *Anuario del Instituto Cervantes 2005. Madrid*

.

40 Years Discussion on the Counting of Publications

Marianne Gauffriau[*], Peder Olesen Larsen[**], Isabelle Maye[***], Anne Roulin-Perriard[***], Markus von Ins[***, 1]

[*]*mga@dtv.dk*
Technical University of Denmark, Technical Knowledge Center of Denmark, D'ARC-DTU Analysis and Promotion Center, DK-2800 Lyngby (Denmark)

[**]*pol@webspeed.dk*
Marievej 10A, 2. DK-2900 Hellerup (Denmark)

[***]*isabelle.maye@cest.admin.ch, aroulinp@bluewin.ch, markus.vonins@bluewin.ch*
Center for Science and Technology Studies (CEST), Effingerstrasse 43, CH-3003 Bern (Switzerland)

Large scale counting of publications became a possibility in scientometrics with the advent of databases in the late 1960es. However, from the very beginning of publication counting based on databases different methods were used. This was partly due to the properties of the data available. But a major reason was the lack of clear definitions and a clear and consistent terminology. A major problem was that the distinction between the basic units of analysis and the objects of study was not generally made. The basic units are the author names, institution names and country names (derived from the institutional addresses) given in publications. The objects of study can be authors, institutions, countries, or other objects for which results of the counting is presented in tables and graphs.

Different counting methods give different results. The reason is multiauthored publications and institutional and international cooperation. A survey of 40 years use and discussion of counting methods for publications shows that different terms are used for the same method, that the same name is used for different methods and that many methods in the literature are not precisely defined. 44 different names found in the literature for the main counting methods will be presented. Often no information at all is given in publications about the counting methods used. This causes difficulties in using data from the old literature and in making time series. In 1994 it was stated that scientometrics/bibliometrics was a field in crisis (Glänzel & Schoepflin). In 1995 a workshop was convened in River Forest to discuss the need for the development of both conceptual and technical standards in bibliometric research and technology. The proceedings of the workshop were published in Scientometrics (Glänzel et al, 1996). The participants in the workshop agreed on the need for standardization, that ambiguous terms and definitions should be avoided, and that all bibliometric publications should include adequate information about methodology. The proceedings of the conference and the general literature indicate consensus about three main counting methods, whole counting, fractional counting, and first author counting (the terms whole counting, full counting and total counting are often used for the same method). The term complete counting (Gauffriau et al., 2007) was not mentioned in the discussion, although some of the participants included complete counting in the discussion but used the terms integer and total counting. In 2000 it was stated that total counting is the de facto standard method used for scientometric evaluations. However, the total counting method described is an example of complete counting and not identical with the whole counting method mostly used (Egghe et al., 2000). It is evident from the literature that complete counting is not the de facto method for evaluation of countries and institutions.

Only complete counting gives the number of acts of cooperation at the national, international, inter-institutional and intra-institutional level. Therefore the Center for Science and Technology Studies (CEST) for several years has produced results based on complete counting at the level of institutions, sectors and countries (CEST, 2004).

In 2004 the need for standardisation was stated again: "Although basic indicators like absolute publication and citation counts are widely accepted as useful tools in measuring research performance, their uncritical use can result in incorrect interpretations. A proper normalization of standard measures and the use of relative indicators are, therefore, indispensable in trend analyses or medium- or long-term studies [in order to] guarantee the validity of conclusions drawn from bibliometric results." (Persson & Danell, 2004). But no common terminology, precise definitions, or standardization of methods for counting publications and citations have been reached, and there is not sufficient knowledge and agreement on the theoretical and technical limitations connected with the different methods. Our literature survey shows that the problem is neglected. Thus in a rather recent review on bibliometrics the problem is treated in one sentence:

"Higher rates of collaboration are usually associated with higher productivity, although counts will vary based on the method of allocating authorship (one credit for each publication vs. partial credit based on number of authors, etc.)." (Borgman & Furner, 2002). In a comprehensive Handbook of quantitative science and technology research. The use of publication and patent statistics in studies of S&T systems from 2004 (Moed et al., 2004) counting methods and the problems associated with counting methods are not mentioned. In a recent book on citation analysis in research evaluation (Moed, 2005), the problems of counting methods are briefly discussed. Most of the discussion is based on unpublished material. The conclusion is that "an analysis per publishing author … is most useful". However, this conclusion is only for studies assessing the trend in a single country's publication output and is not supported by references to the literature.

The problems connected with different counting methods have been present for more than 40 years but there is limited factual knowledge about the differrences in their results. More work is required on the comparison of counting methods (Gauffriau & Larsen, 2005). It has repeatedly been stated that the differences are small, but this assumption is unfounded. But whereas there is a lack of knowledge about the size of the differences, it is beyond doubt that the differences are increasing with time. This is because the number of authors per publication and the extent of institutional and international cooperation are steadily increasing. It is also beyond doubt that whole counting gives double counting so that for example two or more countries get credit for the same publication. In rankings based on whole counting some countries are winners, other countries losers. Among the winners are countries with a small science base and/or with a high level of international cooperation. Among the losers are countries with a low level of international co-operation, including Japan and USA. USA has a comparatively low level of international cooperation but this is not because scientists there do not cooperate. Cooperation between different institutions in USA substitutes for international cooperation. Furthermore the results for whole counting are non-additive. This means for example that whole counts for the European Union cannot be obtained by adding the whole counts for the member countries in the European Union. However, this mistake is repeatedly made in the literature. There is virtually no knowledge about the effects on citation countings. The differ-rences here must be larger, because it is well-esta-blished that publications with many authors are ci-ted more than publications with few authors, and that publications based on international cooperation are cited more than national publications.

When the different counting methods give different results it is not because some methods are correct, others wrong. The methods provide answers to different questions. Elsewhere we present a formal analysis of counting methods (Gauffriau et al., 2007). We propose a terminology and a classification for counting methods for publications. On this basis we display the properties of several counting methods, interpret the results obtained by using different counting methods on the one hand and the information on scientific cooperation and productivity available by combining them .

Support from the Carlsberg Foundation to one of us (POL) is gratefully acknowledged.

1. The author sequence is alphabetic and does not reflect relative contributions to the work.

References

Borgman, C. L. & Furner, J. (2002). Scholarly communication and bibliometrics. Annual Review of Information Science and Technology, 36, 3-72.

CEST (2004), CEST Scientometrics Scoreboard. Indicateurs-clés de la place scientifique suisse (1981-2002). http://www.cest.ch/Publikationen/2004/scoreboard_juillet_2004.pdf

Egghe, L., Rousseau, R. & Van Hooydonk, G.(2000). Methods for accrediting publications to authors or countries: consequences for evaluation studies. Journal of the American Society for Information Science, 51, 145-157.

Gauffriau, M. & Larsen, P. O. (2005). Counting methods are decisive for rankings based on publication and citation studies. Scientometrics, 64, 85-93.

Gauffriau, M., Larsen, P. O., Maye, I., Roulin-Perriard, A. & von Ins, M. (2007). Publication, cooperation and productivity measures in scientific research. Scientometrics, in press.

Glänzel, W., Katz, S., Moed, U. & Schoepflin, U. (1996). Preface, Scientometrics, 3, 165-166.

Glänzel, W. & Schoepflin, U. (1994). Little scientometrics, big scientometrics … and beyond? Scientometrics, 30, 375-384.

Moed, H. F. (2005), Citation Analysis in Research Evaluation. Springer, Dordrecht.

Moed, H. F., Glänzel, W. & Schmock, U. (Eds.), (2004), Handbook of quantitative science and technology research. The use of publication and patent statistics in studies of S&T systems. Kluwer Academic Publishers, Dordrecht (the Netherlands).

Persson, O., Glänzel, W. & Danell, R. (2004). Inflationary bibliometric values: The role of scientific collaboration and the need for relative indicators in evaluative studies. Scientometrics 60, 421-432.

Comparison of two counting houses in the field of pharmacology and pharmacy: Web of Science versus Scopus

Juan Gorraiz[*], Christian Schlögl[**]

[*]*juan.gorraiz@univie.ac.at*
Vienna University, Library and Archive Services, Central Library for Physics, Boltzmanngasse 5, A-1090 Vienna (Austria)

[**]*christian.schloegl@uni-graz.at,*
University of Graz, Institute of Information Science, Universitätsstraße 15/F3, A-8010 Graz (Austria)

Background and purpose

With the launch of Scopus in November 2004, the monopoly of ISI's multi-disciplinary citation indices was ended. By now Scopus has improved so far that it has become a serious competitor of the Web of Science.

There are already several publications which provide more or less extensive evaluations of Scopus with particular emphasis on a comparison with the Web of Science. However only a few of them (e.g. Bauer & Bakkalbasi, 2005; Bosman, et al., 2006; Jasco, 2006) compared the two products with regards to scientometric analyses. The study presented here also focuses on the latter by investigating the following research questions:

- To which extent are high impact SCI journals covered by Scopus?
- Is the impact factor usually lower for an SCI journal than the corresponding indicator computed for that journal in the Scopus database?
- How reliable are the data in these two databases?

Methodology

Our analysis deals only with pharma journals. For this reason our study is descriptive and exploratory. In order to get the pharmacy journals with the highest impact, we used the JCR of the year 2005, subject category "pharmacology & pharmacy". We downloaded the data for the impact factor and complemented them with the number of 2003 and 2004 articles and the number of 2005 citations to 2003 and 2004 articles. Since Scopus does not provide any citation statistics comparable to the JCR, we had to retrieve the corresponding data from this database manually. But as Scopus provides a sophisticated search interface, data collection could be performed with relatively little efforts. When computing the impact factor for the Scopus journals, we used the same formula as in the JCR. All data were collected in November 2006.

Findings

Coverage of JCR high impact pharmacology and pharmacy journals in Scopus

A comparison of journals from a certain subject category between the two databases is not without problems for various reasons. Normally each database uses its own classification system. Furthermore the different size of the databases must be considered. While there were 193 journals included in the subject category "pharmacology & pharmacy" (and 75 partly overlapping "toxicology" journals) in the JCR, Scopus covered altogether 266 active titles. As a consequence the coverage of SCI journals in Scopus is higher than contrariwise. However it was not expected that each of the top-100 SCI pharmacy journals (ranked by impact factor) was included in Scopus. Finally the indexing procedure is not consistent in different databases. 96 journals were indexed as pharma journals in both citation indices. The remaining four journals were assigned to other categories in Scopus ("medicine" and "biochemistry, genetics and molecular biology").

Differences in impact factor

Since there are more journals included in Scopus than in WOS, a journal in Scopus has a higher chance to get cited in general. As a consequence the values for the impact factor should also be higher in Scopus. This assumption turned out to be true for most journals in our journal sample. The impact factor was higher for 82 journals in Scopus, 18 journals had a higher impact factor in the JCR. A journal with a high impact factor in the JCR usually has a high impact factor in Scopus, vice versa (see Table 1, which lists only the top-10 journals). This goes along with the correlation (Pearsons's r=0.96) of the impact factors of the top-100 pharma journals. The size of the differences (in %) between the impact factor values follows a skew distribution (see Table 2). For most journals the difference is relatively small (median = 10.3%). Yet two journals ("Reviews of Physiology Biochemistry and Pharmacology" and "AAPS Pharmsci") attracted attention because of their remarkable discrepancies of 74 and 146 percent respectively.

Table 1. Top 10 JCR (2005) pharma journals (ranked by impact factor (IF)): SCI vs. Scopus.

Rank JCR	Journal title (abbreviated)	IF JCR	IF Scopus	Rank Scopus
1	ANNU REV	19.83	20.81	1
2	NAT REV DRUG	18.77	17.88	2
3	REV PHYSIOL	17.05	9.824	5
4	PHARMACOL REV	15.68	15.95	3
5	TRENDS	10.37	10.50	4
6	PHARMACOL	8.357	9.086	6
7	MED RES REV	7.964	8.636	7
8	DRUG DISCOV	7.755	5.418	18
9	CLIN	7.526	7.772	9
10	ADV DRUG	7.189	8.144	8

Reliability of data

In contrast to the citations received, there should occur only small – in the ideal case no – differences in the number of articles between the two databases. As can be seen in Table 2 the difference in the number of (research and review) articles was below 5% for two thirds of all journals in 2003 and 2004. Nevertheless we computed a greater difference of at least 10% for one fifth of all journals. There were even a few journals in each year, in which the article counts differed more than 50% (2003: "Reviews of Physiology Biochemistry and Pharmacology", "Pharmacogenomics", "Drug Discovery Today", 2004: "Reviews of Physiology, Biochemistry and Pharmacology", "Drug News & Perspectives", "AAPS Pharmasci"). These discrepancies raise severe questions with regards to data quality.

Table 2. Distribution of the % differences between Scopus and JCR in the impact factor and the number of articles in 2003 and 2004.

Differences between Scopus and the JCR	Impact factor	2003 articles	2004 articles
0-5%	25	71	68
5-10%	23	10	11
10-20%	28	7	11
20-30%	16	7	5
30-50%	6	2	2
>=50%	2	3	3
Total	100	100	100
Values in Scopus greater	82	65	65
Values in SCI greater	18	14	17

In order to analyse these big divergences, we checked the entries in the two databases. For most journals the mismatch in the article counts was due to a different assignment to document types. Missing issues and volumes are a further reason for a mismatch. Interestingly this occurred in the WOS in which one issue of "Reviews of Physiology, Biochemistry and Pharmacology" was not fed into the database. Changes in journal titles which are followed at different points in time are a further source for discrepancies. This holds true for "AAPS Pharmsci" in our journal sample.

Conclusion

Our analysis showed that JCR and Scopus have a good coverage of high impact journals in the field of pharmacology and pharmacy. We computed higher values for the impact factor for most journals in Scopus. This is mainly due to the fact that Scopus includes more journals which generate higher citations in this database. The difference in the number of research and review articles was tolerable for the majority of the investigated journals. Nevertheless we calculated greater differences for several journals which were caused by both Scopus and JCR. In our study Scopus turned out to be suitable for scientometric analyses. However the practicability of such studies could be strongly increased if Scopus would add an bibliometric "infrastructure" comparable to the Journal Citation Reports but with several enhancements (e.g. field-normalized journal impact indicators without mixing different document types).

Acknowledgments

The authors would like to thank Ms. Helen J. de Mooij and Mr. Camiel Gielkens from Elsevier for their support.

References

Bauer, K. & Bakkalbasi, N. (2005). An examination of citation counts in a new scholarly communication environment. D-Lib Magazine, 11 (9). Retrieved November 30, 2006 from: http://www.dlib.org/dlib/september05/bauer/09bauer.html.

Bosman, J., van Mourik, I., Rasch, M., Sieverts, E. & Verhoeff, H. (2006). Scopus reviewed and compared. Utrecht University Library. Retrieved March 14, 2006 from: http://tinyurl.com/3a4m63

Jacso, P. (2006). Evaluation of citation enhanced scholarly databases. *Journal of Information Processing & Management*, 48 (12), 763-774.

S&T Research in India: An Overview of its Research Output and Quality

B.M.Gupta[*], S.M.Dhawan[**] and R. P.Gupta[***]

[*]*bmgupta1@yahoo.com, bmgupta@nistads.res.in*
National Institute of Science, Technology & Development Studies
Dr K.S.Krishnan Marg, New Delhi

[**]*smdhawan@yahoo.com*
Library & Information Consultant & Former Scientist F
National Physical Laboratory, New Delhi (India)

[***]*guptarp@yahoo.com*
Office of the Principal Scientific Adviser to the
Government of India,Vigyan Bhawan Annexe, Maulana Azad Road, New Delhi (India)

Introduction

Despite India's glorious record of achievements in S&T and the progress it has made in setting up and expanding R&D infrastructure, and building capacity and capability in a number of S&T areas, its pace of growth has been slow compared to many other developing countries, as reported in number of studies based on SCI 1-12. Most of these studies have looked at India's performance based on SCI data covering select and short periods. This study seeks to examine India's performance based on SCI-Expanded (Web of Science) data covering select but longer period from 1985-86 to 2001-02. SCIE is much more comprehensive than SCI. Besides, the study also compares India's publications output in databases such as SCI, SCIE and SCOPUS.

Methodology and Data Source

India's S&T publications output (comprising 85383 records and their cumulated citations data as on February 2004) during 1985-86, 1993-94, and 2001-02 were downloaded from the SCIE edition of Web of Science database. After cleaning the raw data, and making value additions, tables were generated for analysis using well known absolute indicators.

Analyses & Results

Publications data from SCIE edition of Web of Science during 1985 - 2005 reveal that India annualized publications growth is accelerating from 2.51% during 1985-2002, 5.96% during 1996-2005, and to 10.5% during the recent five years, 2000-05. In absolute numbers, its output rose from 13634 to 28603 publications during 1985 - 2005. The statistics on India's growth rates (computed for quenniums 1996-2000 and 2001-05) do not differ significantly in databases such as SCI (32.40%), SCIE (33.17%) and SCOPUS (32.85%). However, they do differ significantly on publications count. India's quinquennial output (2001-05) in SCIE was 46% greater than SCI, and 17.9% lesser than SCOPUS. As per SCIE. India published 23153 publications (coming from 1734 institutions) in 1985-86, 27088 publications (coming from 2223 institutions) in 1993-94 and 35142 publications (coming from 3443 institutions) in 2001-02. Only 10% institutions (181 institutions in 85-86, 215 in 93-94, and 310 in 01-02), account for 80% output by the country. It shows that excellence in S&T is confined to select few institutions only. A significant number of institutions are low in publications productivity.

The Universities & Colleges sector contributed the largest share, but its share is declining (52.2% to 46.6% during 1985-86 to 2001-02). The Mission-oriented R&D sector has shown rise (from 28.37% to 37.93%), followed by Institutes of National Importance (17.23% to 20.42%), and Industry (1.78% to 2.01%) during the corresponding periods. The institutional participation is the largest in Universities & Colleges sector (48-50% share), followed by Mission-oriented R&D institutions (22-30%), Industry (14-20%), and Institutes of National Importance (less than 1%) during 1985-86 to 2001-02. Publications growth is the fastest in Mission-oriented R&D sector (102.9%), followed by INI (79.8%), and Industry (72.2%). The Universities & Colleges sector, in spite of its largest institutional participation, showed the slowest growth rate (35.6%) and even well below the country average growth rate (51.7%).Collaborative research has shown must faster growth (729.6%) than the country growth rate in S&T (51.78%). The country output through collaborative research is rising (7.34% to 45.10% during 1985-86 to 2001-02). Mission-oriented R&D sector has shown the fastest growth rate in collaborative research (1017%), followed by Universities & Colleges (726%), Institutes of National Importance (620%) during the corresponding period. India's collaborative share at national level is greater (4.3% and 23.07%) than its share at international level (3.04% and 17.06%) during 1985-86 and 2001-02. India's partnership

with world countries for collaborative research in S&T has been spreading, growing from 70 countries to 113 during in 1985-86 to 2001-02. India's share of collaborative publications with USA has declined from 43.8% to 35.1%, and with Canada it has also declined from 9.52% to 4.91% during 1985-86 to 2001-02. With other countries it is rising. A significant share of India's research output in S&T (75 - 80%) is published in low impact journals (IF between 0.001 and 1.999). Its publications share in medium impact journals (IF between 2.0 and 3.999) is small (8% to 15%) and in high impact journals (IF 4.0 and above) it is still smaller (0.96% to 5.12%) during 1985-86 to 2001-02. Besides, 3.5% to 10.8% share of India's output is reported in zero impact journals. Mission-oriented R&D sector showed the highest average impact factor per paper (1.454), followed by INI, Universities & Colleges, and Industry sectors in 2001-02.

India's share of highly-cited publications (those receiving 20 or more citations) is small (4.8% to 5.45%) Its share of medium-cited publications (receiving 5 to 19 citations), is moderate (24.4% to 26.8%) during 1985-86 and 1993-94. Its share of papers with no citations is the largest (30 to 44.28%). This analysis pertains only to papers, which appeared during 1985-86 and 1993-94 and cumulated citations they received till February 2004. Chemistry, physics, and engineering have been the leading areas of research in India. Their respective publication shares have been (69.08%, 68.4% and 77.7% respectively) and above the country's average (51.8%) during 1985-86 to 2001-02. Agriculture, biology, basic life sciences, clinical medicine, biomedical sciences and earth & environment sciences have been its medium productivity areas. Mathematics and computer science have been its low productivity areas. The publication growth has been the fastest in clinical medicine (141.6%), followed by basic life sciences (81.6%) and biomedical sciences (76.4%) during 1985-86 to 2001-02. In contrast, the country witnessed negative growth in biology (–32.9%), slow in agriculture (18.3%), and earth & environmental sciences (51.1%). Mathematics is also running out of steam due to its slow growth. Computer science, though, witnessed high growth rate (118.5%) but its national share is still small (1.3% and 1.8%) during 1985-86 and 2001-02.

Conclusion

The country must draw out long-term and short-term plans for R&D developments in the country. It needs to organize goal-oriented and need-based programs at the national and institutional level, encourage greater institutional participation, greater collaboration in research at national and international level, attract young talent into science, organize in-service training of staff for creativity and innovation, and set up sophisticated instrument facilities for S&T education and research in the country. Universities & Colleges sector needs to be given special attention to enhance its quantity & quality of research output. A stronger monitoring and evaluation system is needed for assessing performance of Ph.D. work and for projects funded by national S&T agencies. Higher budget allocation at the institutional level, for hiring doctoral and post-doctoral students may be considered to increase research output. Linkages between low, medium and high productivity institutions needs to be strengthened to enable low and medium productivity institutions to benefit from the rich experience of high productivity institutions, and from their modern equipment facilities. India's collaborative share in research with countries like the USA, UK, Russia, and Canada is on decline and with countries like Germany, France, the Netherlands and Spain is on the rise. New initiatives are needed to strengthen the India's collaboration both with developed and developing countries.

References

Chidambaram, R Measures of progress in science and technology. Current Science March 2005, 88(6), 856-60

Garg, K.C. and Dutt, B. Bibliometrics of Indian science as reflected through Science Citation Index. Journal of Scientific & Industrial Research 1992, 51, 329-40

Arunachalam, S.;Srinivasan, S. and Raman, V.Science in India: A profile based on India's publications as covered by Science Citation Index, 1989-1992.Current Science 10 March 1998, 74(5), 433-441

Basu, Aparna. Science publication indicators for India: Questions of interpretation. Scientometrics March 1999, 44(3), 347-360'

Basu., Aparana and Nagpaul, P.S. National mapping of science – India: Multidisciplines Information Today & Tomorrow 2000, 19(3), 3-7.

Basu, Aparna and Nagpaul, P.S.National Mapping of Science. New Delhi; NISTADS. 1998. (Report 248/98)

Basu, Aparna and Aggarwal, Ritu, Indian scientific literature in Science Citation Index: A Report .Information Today & Tomorrow 2001, 20(4), 3-8.

Garfield, Eugene. Mapping science in the third world. Part 1. Science & Public Policy June 1983, 10(3), 112-127.

Raghuram, N. and Madhavi, Y. India's declining ranking. Nature 1996, 383(6601), p.572

Arunachalam, S.Is science in India on decline? Current Science 2002, 83(2), 107-108

May, Robert M. Scientific wealth of nations. Science 1997, 275, 793-796.

King, David. The scientific impact of nations. Nature 15 July 2004, 430, 311-316.

Visualizing Library and Information Science from the Practitioner's Perspective

Yusef Hassan-Montero and Víctor Herrero-Solana

yusefhassan@gmail.com, victorhs@ugr.es
Scimago Research Group, University of Granada, Faculty of Library and Information Science,
Colegio Máximo de Cartuja s.n., 18071 Granada (Spain)

Introduction

The knowledge domain visualization studies are mostly based on scientific publishing data, such as citation and co-citation data. These studies provide a valuable view of the knowledge domains' intellectual structure and evolution. However, scientific data do not enable one to visualize the other side of the coin: how a knowledge domain is viewed by its practitioners.

In this work we present an overview of the Library and Information Science profession. For this purpose, professional subjects and their relationships are visually represented as a network.

Materials and Methods

The data sample was downloaded from 'Exit' web site (http://www.directorioexit.info/) by means of a crawler, on November 24th, 2006. 'Exit' is a free online directory of 'experts in information handling', where the admission of each new member is evaluated by a committee. The data sample contains 646 member records; 72.5% of the members are from Spain.

Each record is self-filled by the experts. The record contains no more than ten fields (name, institution, country, email, web page, subject, etc.). The expert can select at most five subjects from a predefined list with 48 different subjects. Figure 1 shows the distribution of experts for each subject, from four (*Statistical Information*) to 204 (*Knowledge Management*).

Co-occurrence analysis enables us to calculate relationships between subjects. That is, the relation between two subjects would be determined by the number of members that have selected both subjects in their professional profiles. Co-occurrence relationships among subjects were represented as a NxN matrix, where N is the number of different subjects (i.e. N=48).

The matrix was visually represented as a network display, where each subject is shown as a node, and their co-occurrence relationships as links between nodes. In order to show a clearer and more comprehensible graph, the less significant links were pruned. The pruning method used was Pathfinder algorithm (Schvaneveldt; 1990), with the parameters q=N-1 and r=∞. Also, Kamada & Kawai's algorithm (1989) was used to situate the nodes in space. In the final network display (Figure 2), link thickness indicates the relationship weight (co-occurrence frequency), and the node's label size reflects the number of members.

Discussion

The network display (Figure 2) shows a prominent root node with the most popular subject (*Knowledge Management*). From this node emerge five long branches: an Information System Design branch (*Information Architecture, Usability, Web Design,* etc); a Digital Libraries branch (*Open Access, Electronic Journals, Academic Library,* etc); an Information Retrieval and Bibliometrics branch (*Bio-medical Information, Patents, Databases,* etc); an Archives branch (*Historical Archives, Company Archives and Records Management*); and finally, a multifaceted branch that enclose *Planning and Management, Marketing and Library*.

Also, there are a few nodes without branches, connected directly with the root node (e.g. *Law Information, Journalistic Information, Document Management Software*, etc.). The nodes in the same branch have strong relations among them, but in many cases these relationships are remarkable. For example, while *Academic Library* is in the Digital Libraries branch, other kinds of libraries (*Public, National and School Libraries*) are in the Marketing-Management branch.

Another noteworthy node is *Ontologies and Thesauri*, because it is not connected with *Classification and Indexing*. Similar is the case of *Competitive Intelligence*, because it is not in the Marketing-Management branch. Another interesting case is *Search Engine Optimization*, which is strongly connected to the Usability branch instead of to Information Retrieval.

Conclusions

The LIS domain shows at least five major groups of subjects: 1) Information Retrieval & Bibliometrics, 2) Digital Library, 3) Information Architecture & Usability, 4) Marketing & Management, and 5) Archives & Record Management. Knowledge Management is the most popular subject and the root of the network.

The presented visualization provides a means of knowledge acquisition and understanding of the LIS domain from a professional perspective. Also, this visualization could be used as a graphic interface for the 'Exit' directory.

References

Kamada T, Kawai S. (1989). An algorithm for drawing general undirected graphs. *Information processing letters*, 1989, 31(1), 7-15.

Schvaneveldt R. (Ed.) (1990). *Pathfinder associative networks: studies in knowledge organization*. Norwood, NJ : Ablex; 1990.

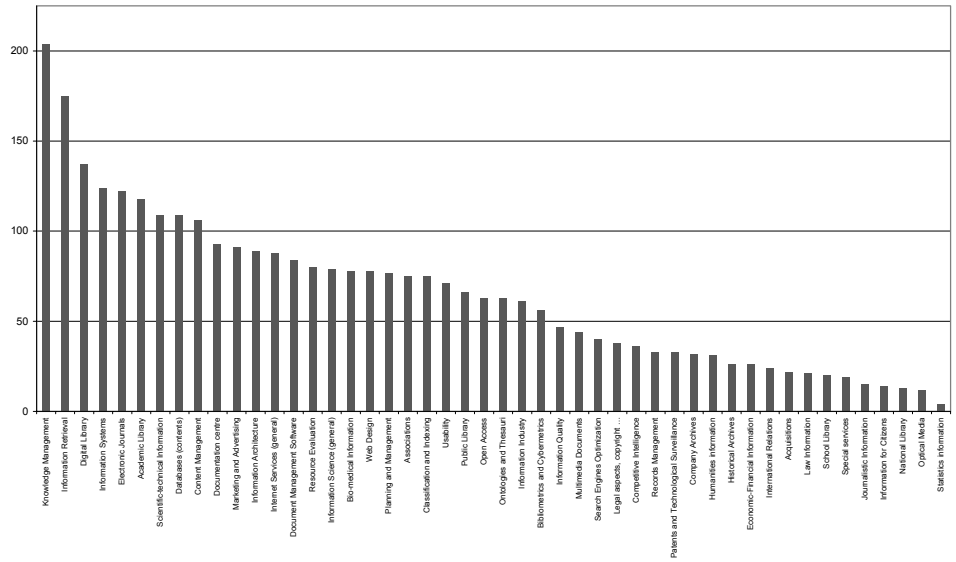


Figure 1. Number of members for each subject.

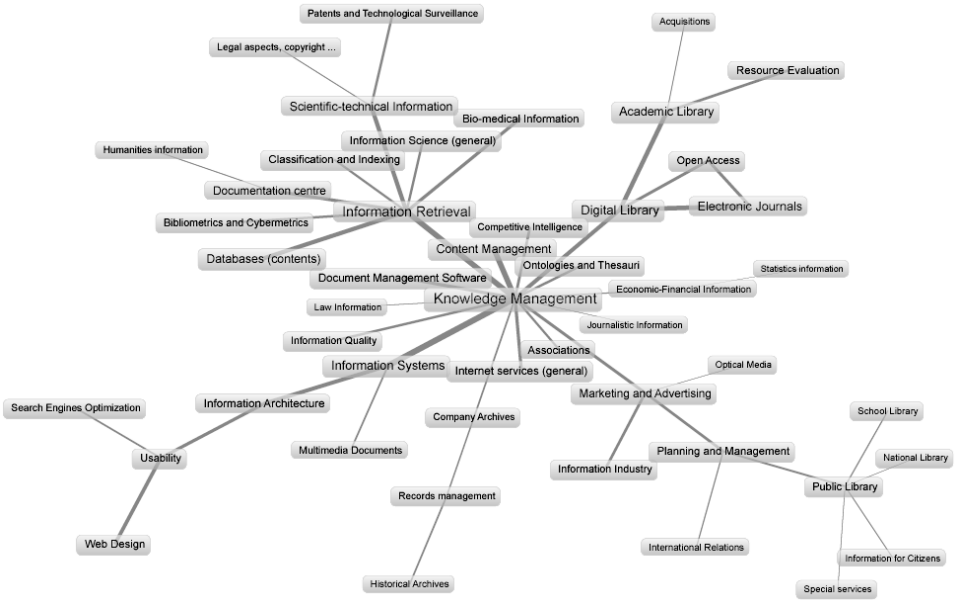


Figure 2. Network display.

Measuring Diversity of Research in Bibliographic-Coupling Networks

Frank Havemann[*], Michael Heinz[*], Marion Schmidt[**], and Jochen Gläser[***]

[*]frank.havemann@ib.hu-berlin.de, michael.heinz@ib.hu-berlin.de
Departement of Library and Information Science, Humboldt-Universität zu Berlin, Berlin (Germany)

[**]*mschmidt@cbs.mpg.de*
Max Planck Institute for Human Cognitive and Brain Science (Library), Leipzig, (Germany)

[***]*jochen.glaser@anu.edu.au*
Res. Eval. & Policy Proj., Res. School of Social Sciences, Australian National University, Canberra (Australia)

Introduction

Inspired by the hypothesis that diversity of research might decline as a result of new science policy measures we search for bibliometric methods that render changes in the diversity of research measurable and therefore comparable in time series and between countries or organisations. No convincing measurement of research diversity has been provided so far. Opinions of scientists on the subject cannot be considered as reliable evidence because scientists may be biased. Furthermore, quality and marginality of a scientific enterprise are often empirically indistinguishable. The majority may perceive nonconformist approaches as bad science. Conversely, scientists might rationalise insufficient recognition of their work as being due to specificity rather than quality. Testing the above-mentioned homogenisation thesis requires measures of diversity that do not depend on the perception of scientists. Bibliometric indicators can be used to construct these measures because they are unobtrusive and objective, i.e. they neither affect the behaviour they measure nor depend on scientists' opinions about the attribute that is measured. Bibliometric methods have been used to measure cognitive diversity, albeit under the heading of interdisciplinarity (Bordons *et al.* 2004; Rafols & Meyer 2006). The measurement of cognitive diversity at higher levels of aggregation has not yet been attempted. In this poster we discuss measures of biodiversity that can be adopted for the measurement of research diversity and present results of a feasibility study in a field that shows relatively clear natural boundaries, namely Electrochemistry.

Measures of Biodiversity

There is a broad discussion about appropriate measures of diversity in ecology (Purvis & Hector, 2000). Ecological studies mainly use four aspects of diversity, namely the number of species in the unit of study or *species richness*, the variation of species abundance or *evenness*, and the *difference* between species in a unit (Shimatani, 2001; Magurran, 2004). When genetic diversity of a population of one species is to be measured, mean differences between genotypes have been used as a measure of diversity (Shimatani, 2001). These measures can be used alone or combined in diversity indices. Among the various diversity measures that combine species richness and evenness in a single measure the Shannon index is one of the most enduring.

Diversity of Research

The concept of diversity has rarely been used in science studies and has not yet been defined with sufficient rigour. To our knowledge, the first one to propose a bibliometric approach to measuring diversity was Hariolf Grupp (1990). Schmidt *et al*. (2006) conducted a feasibility study of diversity measures at the country level based on co-citation clusters, albeit with a negative result. Ordinary single-linkage co-citation clustering gives no adequate picture of the structure of a speciality due to the well-known chaining effect of single-linkage clustering. Rafols & Meyer (2006) operationalised interdisciplinarity as cognitive diversity and measured it for a small sample of papers using a method similar to our approach. Departing from the approach of Schmidt *et al.* our study conceptualises a specialty as a species (consisting of genetically different individuals – papers) rather than a biotope (consisting of subspecialties – collections of papers citing co-citation clusters). A paper's references link it to its scientific ancestors as an individual's genes link it to its biological ancestors. As two closely related individuals have a lot of genetic information in common, two papers that have references in common are considered related. This is the well-known bibliographic coupling of papers. A major difference between genetic information and reference lists is that members of a species have most of their genetic information in common. This is paralleled by the information represented by paradigms and all other knowledge shared by members of a specialty, most of which does not appear in reference lists of research papers. Thus, each of two bibliographically uncoupled papers can be coupled to a third paper, a relation that cannot occur between genotypes of a species. We have therefore taken into account that the bibliographic information of a paper is not complete

enough to reflect the real relatedness to other papers. This deficit could be overcome by including the references of the cited papers, the references of the references of the cited papers and so on. This would mean not only considering the parents but all ancestors when determining the relatedness of two individuals. We will test this idea in a later study. In the study presented here we tried a simpler way. We assessed the bibliographic difference between two papers by the length of the shortest path between them in the network of bibliographically coupled papers. Distances between papers are weighted with the Jaccard measure of bibliographic coupling. This approach is similar to that used by Botafogo *et al.* (1992) for measuring the compactness of hypertext. Egghe & Rousseau (2003) generalised this BRS-compactness to weighted networks and applied the generalised definition also to a bibliographic-coupling network. They used the number of shared references of two papers as the measure of bibliographic coupling. The distribution of the Jaccard index J of bibliographic couples is skewed but the distribution of its logarithm is largely symmetric. Therefore we use $d = -\log(J)$ as our distance measure between two coupled papers. We have estimated the bibliographic difference of two papers by the length of the shortest path between them in the network of papers weighted with distance d.

Results of the Study

To test the proposed measure of research diversity we used a sample of 4209 papers from 14 electrochemical journals indexed in the CD-ROM version of the SCI 1998. We treated this sample as a network of bibliographically coupled papers. A total of 3784 papers (ca. 90%) were linked in the main component, which was used to calculate the research diversity of electrochemistry in selected countries.

Table. Indicators of research diversity in electrochemistry for selected countries (see text).

country	n_i	p_i	m_i	z_i	rank
USA	652	89.9	13.35	–26.14	6
Japan	577	88.1	13.34	–24.03	5
Germany	339	86.9	14.21	39.62	2
France	325	91.3	12.96	–36.18	7
Russia	268	84.5	14.41	41.27	1
UK	220	90.2	13.26	–12.28	4
World	3784	89.9	13.57	0	3

Country distributions were obtained from the main component by selecting all links between papers that had at least one author from this country. Means m_i of the countries' distributions of links between n papers are given in the Table. Means cannot be directly compared because of the size differences between the country sub-samples. Comparability was achieved by transforming the means m_i to $z_i = [m_i - m(\text{all})] / [s(\text{all}) / \sqrt{N_i}]$, where s denotes the standard deviation and $N_i = n_i(n_i - 1)/2$. With p_i we denote the percentage of all of a country's papers that were found in the main component. The standardisation is necessary to give big samples a better chance to escape from the population mean. Using $|z_i|$ as argument of the standard normal distribution we get the probability that the deviation from population average is produced by chance which in all cases is less than $1/10^{32}$.

Discussion

The fact that 10% of papers are not related to the main network indicates that the delineation of the specialty itself remains a challenge. The two countries with higher then average diversity, Russia and Germany, have also the highest percentages of papers outside the main component (15.5% and 13.1%), whereas France with lowest diversity value has also the least percentage outside the main component (8.7%). The countries' means differ significantly from the world average at a very high confidence level. These findings support our hope that indeed average shortest distances in bibliographic coupling networks are reliable measures of research diversity. The results of case studies recently conducted by Rafols and Meyer (2006) with a similar approach are also encouraging. Further tests will include the calculation of time series and the discussion of results with experts in Electrochemistry.

References

Bordons, M., F. Morillo & I. Gómez, 2004: Analysis of cross-disciplinary research through bibliometric tools. In: H.F. Moed, W. Glänzel & U. Schmoch (eds.), *Handbook of quantitative science and technology research.* Kluwer, 437-456.

Botafogo, R., E. Rivlin, & B. Shneiderman (1992). Structural analysis of hypertexts: identifying hierarchies and useful metrics. *ACM Transactions on Information Systems* (TOIS) 10 (2), 142-180.

Egghe, L. & R. Rousseau (2003). A measure for the cohesion of weighted networks. *J. Amer. Soc. Inform. Sci. Technol.* 54 (3), 193-202.

Grupp, H. (1990). The concept of entropy in scientometrics and innovation research. *Scientometrics* 18 (3-4), 219-239.

Magurran, A. (2004). *Measuring Biological Diversity.* Blackwell Publishing.

Purvis, A. & A. Hector, 2000: Getting the measure of biodiversity. *Nature* 495, 212-219.

Rafols, I. & M. Meyer (2006). Diversity measures and network centralities as indicators of interdisciplinarity: case studies in bionanoscience. Working paper (to appear).

Shimatani, K. (2001). On the measurement of species diversity incorporating species differences. *Oikos* 93 (1), 135-147.

Schmidt, M., J. Gläser, F. Havemann, & M. Heinz (2006). A Methodological Study for Measuring the Diversity of Science. In: *International Workshop on Webometrics, Informetrics and Scientometrics & 7th COLLNET Meeting*, Nancy, pp. 129-137. INIST.

Comparing H-indices for Scientists in Different ISI Fields

Juan E. Iglesias and Carlos Pecharromán

*jeiglesias@icv.csic.es, **cpg@icmm.csic.es
Instituto de Cerámica y Vidrio and **Instituto de Ciencia de Materiales, Consejo Superior de Investigaciones Científicas, Cantoblanco, 28049, Madrid (Spain)

Introduction

The *h* index (Hirsch, 2005a,b), a measure of the significance of the scientific output of a researcher, is the highest integer *h* such that *h* among this person's N_p papers have collected at least *h* citations, the remaining N_p-*h* papers having less than *h* citations each. Hirsch argued that *h* is a better measure than any other single index, but it seems inadequate to compare research workers in different areas, since the most highly cited biologists have *h* values twice those of the most cited physicists. In this paper we present a rational method to account for this, introducing simple multiplicative corrections to the *h* index which depend on the worker's ISI field.

The distribution function of citations

Let the function $N(x)$ give the number of papers which have been cited a total of *x* times. Laherrère & Sornette (1998) studied the 1120 most cited physicists over the period 1981-1997. They ranked them according to their total number of citations, *C*, and plotted this vs. rank, *k*, and got a good fit for the Zipf function $C(k)$ assuming it to be a stretched exponential:

$$C(k) \propto \exp\left[-\left(k/k_0\right)^{\beta}\right] \quad (1)$$

with $\beta \approx 0.3$. Redner (1998) studied the distribution of the 6,716,198 citations between 1981 and mid-1997 (Small & Pendlebury, 1997) to 783,339 papers published in 1981. Interestingly, 47% of these papers were uncited, and only 1% had more than 105 citations. Redner notes that while the curvature this data shows in a log-log plot would indicate a distribution for $N(x)$ of the type (1), this function does not adequately fit the large-*x* data. Instead, he finds that a good fit can be obtained for the Zipf function $C(k)$ with a power law,

$$C(k) = k_0 k^{-\alpha} \quad (2)$$

with $\alpha \approx 1/2$, and similarly for $N(x)$, with $\alpha \approx 3$.

Calculation of the *h*-index

a) Power-Law distribution

We assume the citation distribution for an average research worker in a given field is of type (2). If C_{total} is the total number of citations, and $N_p >> 1$,

$$\int_1^{N_p} C(k)\mathrm{d}k = \int_1^{N_p} k_0 k^{-\alpha}\mathrm{d}k = C_{total} \quad (3)$$

and, after some manipulation,

$$h = \left[(1-\alpha)\chi N_p^{\alpha}\right]^{\frac{1}{(\alpha+1)}} \quad (4)$$

where

$$\chi = C_{total} N_p^{-1} \quad (5)$$

is the average number of citations per paper for that emblematic research worker. Using $\alpha = 1/2$,

$$h = \sqrt[3]{\frac{N_p}{4}}\,\chi^{2/3} \quad (6)$$

Other authors (Egge & Rousseau, 2006; Glänzel, 2006) have used similar formalism to study the dependence of *h* on N_p for this distribution, but drew no conclusions for inter-areas comparison.

b) Stretched- exponential function

The analytic variation of *h* is more intricate here than it is in the previous case. We assume

$$C(k) = k_0 e^{-\eta k^{\beta}} \quad (7)$$

where k_0 and η are coefficients to be determined, and put $\beta = 0.3$ (Laherrère & Sornette, 1998; Redner, 1998).

The total number of citations, C_{total} is given by:

$$C_{total} = \sum_1^{N_p} C(k) \approx \int_0^{\infty} C(k)\mathrm{d}k = \int_0^{\infty} k_0 e^{-\eta k^{\beta}}\mathrm{d}k \quad (8)$$

Solving for *h*, we get the function

$$h = C_{total}\frac{\eta^{1/\beta}}{\Gamma\left(1+\frac{1}{\beta}\right)}e^{-\eta h^{\beta}} \quad (9)$$

where η only depends on C_{total} and β, and $\Gamma(z)$ is the gamma function. The *h* index depends both on N_p and on χ_c, and is obtained numerically, but a explicit dependence is not easy to determine for the whole range of values, and we cannot assume a slow variation with N_p.

Normalization Data

We assume for the average worker in a given science field the world average of citations/paper which corresponds to that particular field. We use ISI (Essential Science Indicators, Baselines, 2005) data for the mean number of citations a paper in each field and year has received since publication. The value is naturally larger for older papers, but a tendency to stabilize is visible in all fields after a few years. The values for each year are strongly dependent on the field, and can vary substantially: the quotient between the values corresponding to "Molecular Biology & Genetics" and those corresponding to "Mathematics" is 8.3 for papers published in 1995.

a) Power-Law Distribution

We have tabulated in the first column of Table I the recommended f_i values for normalization of the h index, computed under the assumption that the distribution function of the citations is given by Eq.(2).

Table 1. Normalization factor for the ISI Fields of Science, relative to the field "Physics"
To put h-indices of different fields in a common scale, multiply by f_i, the tabulated value (see text for details).

ISI Fields	*Power Law*	*Stretched Exponential*			
		100 papers	**200 papers**	**500 papers**	**1000 papers**
Agricultural Sciences	1.27	1.20	1.24	1.30	1.35
Biology & Biochemistry	0.60	0.77	0.73	0.68	0.64
Chemistry	0.92	0.95	0.94	0.93	0.92
Clinical Medicine	0.76	0.86	0.83	0.80	0.77
Computer Science	1.75	1.97	—	—	—
Economics & Business	1.32	1.23	1.28	1.36	1.42
Engineering	1.70	1.79	—	—	—
Environment/Ecology	0.88	0.93	0.92	0.90	0.88
Geosciences	0.88	0.93	0.91	0.89	0.88
Immunology	0.52	0.73	0.68	0.63	0.58
Materials Science	1.36	1.29	1.35	1.44	—
Mathematics	1.83	—	—	—	—
Microbiology	0.63	0.79	0.75	0.71	0.67
Mol. Biology& Genetics	0.44	0.68	0.64	0.57	0.53
Neuroscience &Behavior	0.56	0.75	0.71	0.66	0.62
Pharmacol.&Toxicology	0.84	0.90	0.89	0.86	0.85
Physics	1.00	1.00	1.00	1.00	1.00
Plant & Animal Science	1.08	1.05	1.06	1.07	1.08
Psychiatry/Psychology	0.88	0.93	0.91	0.90	0.88
Social Sciences, general	1.60	1.58	1.72	—	—
Space Science	0.74	0.85	0.82	0.79	0.76

They have been calculated by averaging the number of citations/paper for each field, for years 1995-1998, and normalizing to the corresponding values for the field "Physics", then raising to the 2/3 power (Eq.(6)). Comparison of two people having very different values of N_p requires also the cubic root of the ratio of these numbers. The normalizing factors for the five fields which could be labeled "(molecular) life sciences": "Microbiology", "Biology & Biochemistry", "Neuroscience & Behavior", "Immunology" and "Molecular Biology & Genetics" (Table I) agree well with Hirsch's observation that h indices in these fields are about twice those in Physics for Nobel prize laureates.

b) Stretched-Exponential Distribution

The h-index in this model (7) is strongly dependent on N_p and so is the correction factor. The last four columns on Table I contain the factors relative to the field "Physics" for all fields, meant to compare h indices of authors having 100, 200, 500 and 1000 papers. It can be seen that this model predicts correction factors more conservative than those of the power law model. It appears that the values predicted converge to those of the power-law model in

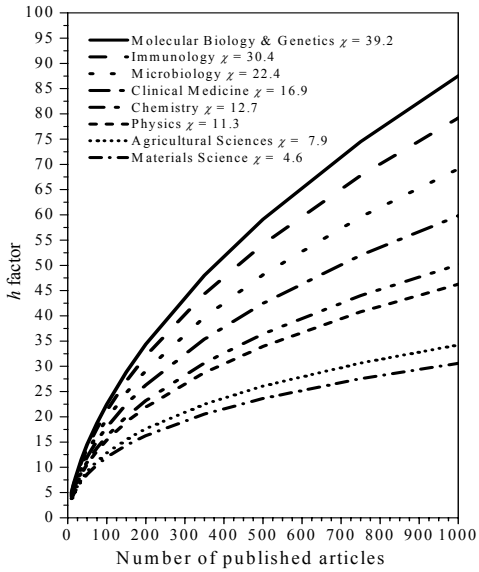


Figure 1.- The h-index vs N_p for selected ISI areas, in the stretched exponential model.

the limit of an high number of papers. Some caution must be exercised with ISI fields with citations/paper rates smaller than 8. (Mathematics, Computer Sciences, Engineering, Social Sciences and Material Sciences), because the numerical method used to calculate the correction factor may produce some inaccuracies.

The results presented here are available in extended preprint form (Iglesias & Pecharromán, 2006).

References

Egge, L. & Rousseau, R. (2006): *Scientometrics*, **69**, 121-129.

Glänzel, W. (2006): *Scientometrics*, **67**, 315-321.

Hirsch,J.E.(2005a):http://arxiv.org/abs/physics/0508025

Hirsch, J. E. (2005b): PNAS, **102**(46), 16569-16572.

Iglesias, J. E. & Pecharromán, C. (2006): http://arxiv.org/abs/physics/0607224

ISI(2005):http://portal.isiknowledge.com/portal.cgi?DestApp=ESI&Func=Frame,

Laherrère, J. & Sornette, D. (1998): European Physical Journal B **2**, 525525-539.

Redner, S. (1998): European Physical Journal B **4**, 131-134.

Small, H. & Pendlebury, D. (1997): http://physics.bu.edu/~redner/projects/citation/isi.html.

A National Research Profile-Based Immediacy Index and Citation Ratio Indicator for Research Evaluation

Peter Ingwersen[*], Jesper W. Schneider[*], Morten Scharff[**] and Birger Larsen[*]

[*]{pi, jws, blar}@db.dk
Royal School of LIS, Department of Information Studies, Birketinget 6, DK 2300 Copenhagen S (Denmark)

[**]msc@fist.dk
Danish Agency for Science, Technology & Innovation, Bredgade 40, DK 1260 K (Denmark)

Introduction

In this study we analyzed the behaviour of Field-based Crown Indicatorsiii (here named FCI) for two smaller EU countries, Denmark and Finland, as well as for the European Union and globally calculated over three periods, 2001-2005, 2004 (cited 2004-05) and the most recent year, 2005. The purpose was to observe the robustness and stability of the indicator as a National Immediacy Index during the latter two periods compared to the five-year window.

The motivation is to provide national research funding agencies with information on the relative citation impact and national citation visibility of the most recent output in central research areas. To this end FCI (Moed et al., 1995) is applied based on data from the National Science Indicators database (NSI, Thomson Scientific). It is calculated based on the publication profile for each country, as defined by 19 broader scientific fields. The citation visibility is defined as the ratio of received citations by the country over the number of expected citations.

With respect to the citation visibility two calculations are done: First, for each field its global citation impact is multiplied by its national publication volume, providing the number of expected citations for that field. Essentially, this value represents a 'shadow-field' of the original national research field. When compared to the actual number of citations obtained by that field it is possible to observe the number of citations gained or lost. Secondly, the aggregated values over all the 'shadow-fields' signify the expected citation volume of the country in question. Finally, when calculating the mean citation visibility ratio the underlying national publication shares of the research fields (here 19) are acting as weights.

The present study assumes that the recent five-year citation and publication window, 2001-2005, provides a robust set of indicators. However, one would like to obtain assessment information on research as recent as possible – in the form of a kind of National Immediacy Index (NII). We compared the citation performance during the three snapshots by diachronic citation analyses. The aim was to observe if 1) it makes sense to apply broader research fields, but in very short and recent time frames (robustness); 2) there exist a significant correlation between impact and visibility results obtained from the three snapshots (stability).

Humanity fields were excluded but social science fields were included as one of the 19 central fields forming the profile taken from NSI (2006). Publication types covered journal and review articles, notes and proceedings papers from journals and selected central research conferences. The 'Multidisciplinary' field in NSI adheres mainly to the sciences but does not contain field-specific articles from Nature, PNAS, Science, etc., which are placed in their proper fields. The Public Health sub-field was included in Clinical Medicine (extracted from Social Science, general). The NSI feature "one-year period cited-to-present" was used for the data collection. National self-citations are included.

In order to be robust we considered that each field in a NII must demonstrate at least 75 publications as a threshold for the most recent year, corresponding roughly to 15 citations received in a central field over the most recent year, with an impact of .20.

Analysis Results

The citation and publication volume for Denmark was 300,716 citations to 45,930 publications (c/p = 6.55) for 2001-2005; for 2005 alone Denmark obtained 6,528 citations to 10,167 publications (c/p = .64). In 2005 Finland received 4,663 citations to 9,065 publications (c/p = .51).

Table 1 provides for each country, EU and globally 1) the research profile in percent-shares of each field for the period 2001-05; 2) the standard FCI scores per field for all three periods; 3) the mean absolute impact per period; and 4) the mean citation (visibility) ratios. In the latter case an index value of 1.0 signifies that the national citation visibility is as expected compared to the global one.

We observe the high Danish absolute impact (NII) for 2005 (.64; FCI = 1.49) and that the Danish profile 2005 differs significantly from the Global, EU and Finnish ones, with foci on Biology & Biochemistry, and less on Chemistry and Engineering. Pearson's r for DK/Global profiles = .93; FIN/Global profiles = .96; and EU/Global profiles = .997. But we observe also that FCI as well as the citation (visibility) ratios decrease slightly in Denmark along with the extension of the citation (and publication) windows. In contrast, all Finland's

index scores increase, probably because their local citation behavior differs from that of Denmark (and the overall EU, which seems stable over the periods). Quite notably, already in 2005 many national fields turn out substantial FCI scores (values in bold signify scores above 1.25): 15 fileds in Denmark – ten more than Finland. However, a closer look at the Table discloses that very often the high impact does not persist over the next analysis periods. In Denmark only Neurosciences, Molecular biology & Genetics and, in particular, Immunology contributes negatively. Although Materials Sc. scores past 2.0 in 2004(05) it decreases in the longer time frame, as does Economics. Similar ups and downs can be observed for Finland and EU.

In terms of robustness one field in Denmark displayed a FCI score in 2005 far below the threshold (Computer Sc., FCI = .08; 7 citations,); in Finland several fields were below the FCI threshold, e.g., Computer Science (FCI=.17); Economics (FCI = .04; 5 citations) and Mathematics (FCI = .08; 8 citations). One may hence conclude that simply to apply the most recent year (here 2005) as basis for NII does not make sense for smaller countries.

van Raan (1999) has proposed for institutions that a FCI index score of 2 is Very Good and 3 is Excellent. Owing to 'flattening' by the averaging we suggest that in the NII a FCI score of 1.25 is Good; 1.50 is Very Good; and 2.0 is Excellent. The highest score so far observed for the three periods was found in 2005 for Physics & Space science produced in Switzerland (2.09).

The stability over the three analysis periods for the FCI scores per country was measured by means of Pearson's r in pairs. For both Denmark and Finland the best correlation was identified between the two periods 2004(-05) and 2001-05 (DK, r = .72; FIN, r =.78; Critical Value = .57). The correlation for Denmark between the other period pairs was very low (.37). In contrast the large-volume region of EU demonstrated the best correlation between the two recent periods (.92), but all its correlations were very sumilar.

Conclusion

We find it meaningful only to apply the National Immediacy Index calculated for the penultimate year cited diachronicly up to the present for countries with an annual research output of approx. 8-9,000 ISI-indexed publications and the 19 broader central fields. Our results do not indicate the necessary robustness of NII nor stability for the most recent year's research probably due to local citatioin behavior – both at individual field and national levels. We also propose simultaneously to apply annually our suggested NII compared to a performance analysis of the most recent five years. There are several benefits: one may immediately observe changes in the national publication profiles, research foci, FCI, citation visibility ratio and aspects like citedness, compared to the current research policy decisions, since the research done leading to the NII is young (only 3-4 years old). The two-year citation window in our NII balances between the request for recency in assessment and the ranges of citation peaks for the different research fields involved, an issue to be further investigated.

References

Moed, H. de Bruin, R.E. & van Leuven, Th.N. (1995). New bibliometric tools for the assessment of national research performance: database description, overview of indicators and first applications. *Scientometrics*, 33, 381-442.

van Raan, A.F.J. (1999). Advanced bibliometric methods for the evaluation of universities. *Scientometrics*, 45(3),417-423.

[iii] The field-specific Crown Indicator is named *CPP/FCSm* by Van Raan (1999), i.e., 'citations per publications' (CPP) over the 'mean field citation score' (FCSm) at a global scale. Self-citations are commonly removed and the same time windows and document types apply to both CPP and FCSm.

Table 1. National and global research profiles (in percent), Field Crown Indicators (FCI), citation (visbility) ratios and absolute impact scores. Gray cells indicate top-4 research foci per country; bold and italics scores indicate FCI scores above index 1.25 and below 1.0. (NSI, 2006).

	DK - pro-	Denmark FCI			FIN - pro-	Finland FCI			EU - pro-	EU FCI			World pro-	Global Impact		
Scientific Fields	file 2001-05	2005	'004(-05)	2001-05	file 2001-05	2005	2004(-5)	2001-05	file 2001-05	2005	2004(-5)	2001-05	file 2001-05	2005	'004(-5)	2001-05
Agriculture & Plant Sc.	11,33	**1,57**	**1,31**	**1,42**	8,88	**1,35**	**1,30**	1,23	7,57	1,16	1,14	1,13	7,82	0,26	1,40	2,89
Biology & Biochemistry	11,02	**1,31**	1,17	1,12	7,72	1,07	1,00	1,08	7,04	1,03	1,01	*0,99*	6,89	0,64	3,71	7,56
Chemistry	8,62	**1,34**	**1,44**	**1,50**	9,03	1,02	*0,98*	1,02	12,91	1,09	1,05	1,07	13,18	0,44	2,36	4,28
Clinical Medicine	26,20	**1,57**	**1,45**	**1,48**	28,16	**1,38**	**1,25**	**1,43**	24,30	1,05	1,03	1,01	23,14	0,48	2,62	5,27
Computer Science	0,79	*0,65*	**1,52**	**1,42**	1,30	**1,33**	*0,98*	1,08	1,10	1,15	1,08	*0,99*	1,23	0,13	0,66	1,51
Ecology/Environment	3,98	**1,69**	**1,46**	***1,29***	4,95	**1,34**	1,22	**1,31**	2,48	1,10	1,09	1,07	2,57	0,29	1,66	3,59
Economics & Business	1,55	**1,58**	1,03	*0,85*	1,23	*0,27*	*0,91*	*0,86*	1,36	*0,84*	*0,91*	*0,84*	1,38	0,15	0,70	1,82
Engineering	4,75	**1,45**	**1,86**	**1,49**	5,99	**1,57**	**1,27**	**1,26**	6,66	1,10	1,07	1,07	7,58	0,14	0,84	1,78
Geosciences	3,39	**1,32**	**1,45**	**1,45**	2,59	**1,55**	**1,52**	**1,35**	2,92	1,16	1,11	1,10	2,77	0,35	1,62	3,44
Immunology	2,16	*0,87*	*0,80*	*0,83*	1,56	1,07	*0,99*	*0,96*	1,59	*0,96*	*0,95*	*0,95*	1,50	0,90	5,34	10,62
Materials Science	1,25	**1,45**	**2,08**	***1,30***	2,67	1,01	*0,93*	1,02	3,12	1,07	1,05	1,06	3,65	0,21	1,26	2,54
Mathematics	1,18	**1,87**	**1,36**	**1,39**	1,28	*0,61*	*0,83*	1,08	2,24	1,14	1,11	1,09	1,91	0,12	0,61	1,32
Microbiology	3,25	**1,29**	1,04	1,22	2,24	*0,76*	*0,76*	*0,93*	2,34	1,02	*0,96*	1,02	2,06	0,60	3,54	6,90
Mol. Biol. & Genetics	2,65	*0,91*	1,13	1,17	2,82	*0,74*	1,00	1,05	2,74	1,00	*0,98*	1,01	2,64	1,12	6,32	12,63
Multdisciplinary	1,16	**1,62**	**1,47**	1,10	1,11	*0,59*	*0,67*	*0,99*	1,25	1,09	1,00	1,06	1,40	0,66	2,81	4,48
Neuro Sc.& Behavior	3,40	*0,94*	*0,94*	*0,93*	4,77	*0,91*	*0,82*	*0,90*	3,81	*0,99*	*0,96*	*0,96*	3,56	0,58	3,60	7,88
Pharmacology	2,12	**1,47**	1,03	1,11	2,13	*0,92*	1,08	1,08	1,87	1,12	1,08	1,07	1,99	0,41	2,51	5,01
Physics & Space Sc.	9,96	**1,65**	**1,76**	**1,62**	10,25	**1,25**	**1,66**	**1,30**	13,31	**1,25**	1,23	1,17	12,90	0,47	2,36	4,12
Social Sc. general	1,25	**1,49**	**1,31**	***1,32***	1,31	*0,64*	*0,68*	1,08	1,40	1,17	1,06	1,04	1,83	0,16	0,64	1,50
Mean Field Crown Indicator	100	***1,49***	***1,40***	***1,41***	100	***1,19***	***1,18***	***1,25***	100	***1,10***	***1,08***	***1,06***	100	0,43	2,36	4,65
Citation ratio: absolute / expected		1,40	1,37	1,31		1,16	1,15	1,20		1,08	1,05	1,04				
Absolute Citation Impact		0,64	3,32	6,55		0,51	2,78	5,82		0,48	2,54	4,93				

Institutional Evaluation in Moroccan Higher Educational System: A Proposed Model

Abdelali Kaaouachi

kaaouachi@est.univ-oujda.ac.ma
EST, Université Mohamed Premier, Oujda (Morocco)

Introduction

The Higher Educational Systems through the world know changes without parallels, caused by several factors: an increase of students numbers, a diversification of the institutions and their teaching and research schemes, a reduction in the financial resources, a strong need for the economic sector to qualified employees and a performance research, consequences of the globalization of the economy and those of the development of information and communication technologies...

The quality improvement of teaching and research becomes more and more a growing concern of all the actors, owing to the higher education is closely related to the economic and social development of the countries. Thus, the actual reforms of higher educational systems all over the world have stressed the introduction of new demarches of quality assurance...

In this presentation, we will discuss the question of the institutional evaluation in the Moroccan higher educational system.

Presentation of the Moroccan higher educational system (MHES) and its principal indicators

The MHES is composed of public and private institutions. There are:

- 14 universities include 83 institutions;
- 61 institutions that are not related to universities but regulated by other ministries;
- 35 training centers in pedagogy;
- 73 institutions related to the professional training;
- 213 private institutions and one private university.

Some indicators of MHES are:

- The rate of schooling in higher education is 14%;
- The budget of the department of the higher education and the scientific research is 4300 Millions of Dirhams;
- The number of teachers is 10 413 ;
- The number of students is 364 039.

Description of the new reform of this system

Since the year 2003, the MHES has been engaged in a new reform, following the orientations of the "reform Charter" and the promulgation of "law 01-00" bearing on the organization of the higher educational sector. This reform relates to the following components:

- The undertaking of a new teaching architecture LMD (Licence, Master, Doctorat) caracterised by a new organization: modular system, semesters, multidisciplinarity, flexibility, continuous evaluation, interactivity with the economic and social environment...;
- The installation of quality assurance models by the introduction of an accreditation system of teaching disciplines and their academic programs;
- The adoption of an evaluation system and the creation of evaluation and regulation authorities: National Commission of Coordination of Higher Education, Council of Coordination of Institutions that are not related to the universities, Commission of Coordination of Private Higher Education...

Evaluation in MHES and detection of its methodological limits

The evaluation is in the heart of modernization in the MHES, it is the principle quoted in the texts of laws and voluntarily expressed by the decision makers of the sector. The efforts making in evaluation procedures show that

- The evaluation is adopted solely at the time of the big reforms;
- The two levels of evaluation are: Institutional (some restricted institutions: EST, FST, ENCG), Teaching disciplines and their Programs (Applied Licence, Training in NTIC);
- The accreditation of third cycle programs;
- The accreditation of disciplines and modules of the new reform;

Some special commissions were created to be in charge of the evaluation operations, for example, the National Commission of Accreditation and Evaluation (CNAE). In the future, the evaluation operations will be conducted by the National Instance of Evaluation which is in the creation phase.

However, measurements taken to ensure quality are in phase of recent installation, and the first

operations carried out show important methodological limits:

- The function of evaluation is diagnostic but not formative;
- The evaluation concerns specific programs and institutions;
- The evaluation concerns inputs not the processes or the outputs;
- The lack of a battery of quality standards;
- The absence of self evaluation procedures;
- The absence of follow-up operations.

Proposition of an institutional evaluation model

The institutional evaluation may be defined as a continuous process of analysis and appreciation of the institution working. The goal is to maintain the quality and to improve it.

There is an actual emergence over the world that the institutional evaluation is based on the following two steps:

- Internal evaluation: an institution collects and analyses data from criteria of evaluation guide. In this step, strengths, weaknesses, opportunities and threats are detected. As a result, internal evaluation report is produced;
- External evaluation: experts visit the institution and review the internal evaluation report.

The model that serves to a guide of evaluation is based on criteria of various activities of an institution: teaching, research, society services and management strategy.

Mission and objectives

- Clarity and pertinence.

Teaching

- Programs (philosophy, objectives, evaluation, adequacy with the needs of the work market, existence of periodical regulations, methods of teaching and learning);
- Facilities (classrooms, laboratories, equipement, computing and information infrastructures, library);
- Students (admission requirement, orientation and advising systems, support the attainment programs objectives, assessment, integration);
- Faculty (number of teachers, competencies, continuous training, evaluation).

Research

- Resources and policies of the research (number of researchers, funds, materials, structure, evaluation, cooperation);
- Research outputs (number of publications, citation index, prizes, research impact).

Society services

- Professional advising, expertise offer, support to keep scientific and cultural heritage, support to protect the environment, involvement to implant the big values in the society (peace, justice, dialogue, cooperation, equality, human rights...).

Management strategy

- Financial management (academic and administrative expenditures, planning, annual audit);
- Human resources management (recruitment, allocation to working, professional development practices, evaluation);
- Materials management (inventory, maintenance, planning);
- Communication practices (transparency, publication of periodical reports).

The interest of the suggested model

The interest of the suggested model is large in several dimensions:

- The creation of an evaluation culture in the MHES;
- The use of the evaluation model in accreditation operations;
- The improvement of the quality of institution outputs;
- The information of public on the value of institutions.

References

COSEF (COmmission Spéciale Education Formation) (1999). La Charte Nationale d'Education et de Formation. *Edition Spéciale (Morocco)*.

Kaaouachi, A. (2001). Les indicateurs de performance : outils d'une gestion rationnelle de l'enseignement. *Prospective Pédagogique*, 24, 32-34.

Kaaouachi, A. (2006). An Evaluation Model of Engineering Faculties in the Arab World. *International Conference on Engineering Education (ICEE) (Saudi Arabia), Conference Proceedings*, 331-337.

Kaaouachi, A. (2006). Un cadre général pour l'implantation d'un système qualité dans les universités des pays du Sud. *XXIIIème congrès de l'AIPU (Monastir)*.

Madaus, G.F., Scriven, M. & Stufflebeam, D.L. (1983). *Evaluation models: Viewpoints on educational and human services evaluation.* Hingham, MA, U.S.A.: Kluwer Academic Publishers.

Stufflebeam, D.L. (2001). *Evaluation models*. San Francisco: Jossey-Bass.

Vidal, J. & Quintanilla, M.A. (2000). The teaching and research relationship within an institutional evaluation. *Higher Education*, 40, 2, 217-299.

Iranian Medical Research Performance: Does it Comply with the National Needs?

Payam Kabiri* and Farzaneh Aminpour**

*payam.kabiri@gmail.com
Isfahan University of Medical Sciences (Iran)

**aminpour@mui.ac.ir
Academic Evaluation & Scientometry Center, Vice Chancellery for Research,
Isfahan University of Medical Sciences, Isfahan (Iran)

Background

Health & Medical researchers of each country should target the potential problems & diseases of that country. *Burden of disease* is a well defined indicator to specify and rank major health problems in each country. The term is first introduced by Murray and Lopez (1996) in their work entitled "The Global Burden of Disease" published in Harvard. Subsequently, the World Health Organization provided a report on the Global Burden of Diseases through the world in 2001. Burden of disease consists of the years lost for each person in his life due to a disease added by the disability adjusted life years due to that disease. It shows the importance of each disease/condition categories in a society which research should focus on. So the Health & Medical researches of a country should comply with the disease rankings in that country. The scientific publications & their relevance to estimated disease burden show how the scientific researches in a country comply with the real needs.

In the study done by Dandina (2004) in India, the majority of disease/health categories of the Indian burden of disease receives only 0.3.6% of all Indian published papers where as the other papers focused on non-communicable diseases.

Rochon and his colleagues (2004) studied the randomized clinical trials published in six leading medical journals throughout the world in 1999. He found that about 44% of these publications studied only one of the 35 leading causes of global burden of disease, and of these 90 randomized clinical trials 32% studied the top 10 causes of Burden of Disease. According to this survey, the three most commonly covered conditions were ischemic heart disease (33 papers, 12%), HIV/AIDS (18 papers, 6%), and cerebrovascular disease (10 papers, 4%).

Other researchers (Swingler, Volmink & Ioannidis, 2003) show that the systematic reviews published in Cochrane database were modestly correlated with the related burden of disease although the total number of systematic reviews is low.

Methods & Materials

In Iran, Burden of Diseases Project was done in 2003 for the first time by Naghavi & his colleagues and then reported by Iran Ministry of Health. The first 20 high ranked diseases were identified this report.

In the present study, PubMed Medline database was selected as the source for retrieving scientific papers published by Iranian health & medical researchers. The database was searched for the term "Iran" in the affiliation field and the time was limited to 2003 till the end of 2006. Each of the disease categories specified as the 20 top ranked diseases in the mentioned report was searched using both free text search & also MESH term search methods combining together with the operator "OR". Finally the number of retrieved papers for each category compared with the related ranking given by the mentioned report.

Results

Of the disease/condition categories reported by 2003 report by Iran Ministry of Health, the first most burden belonged to traffic accidents, natural disasters, Ischemic heart diseases, depression & addiction.

As it shows in Table 1 All the Iranian papers indexed in PubMed Medline after 2003 was 5383 from which only 15 (0.27 %) were devoted to the first ranked burden of disease which was traffic accidents. The number and percentage of published papers related to other disease/condition categories are also shown in Table 1.

Table 1. The number & percentage of published papers related to five top disease categories in Iran

Top Ranked Burden of Disease	Number of Papers	Percentages from 5383 papers
Traffic accidents	15	0.27 %
Natural disasters	14	0.26 %
Ischemic heart disease	59	1.09 %
Depression	69	1.28 %
Addiction	13	0.24 %

According to this table natural disaster with 14 papers (0.26 %), ischemic heart disease with 59 papers (1.09 %), depression 69 papers (1.28 %), and

addiction with 13 papers (0.24 %) ranked after traffic accidents.

Discussion & Conclusion

The burden of disease shows the volume & size of each health problems in a society which should got the most attention for research. To cope with these major health problems in each country, scientific researches and publications of that country should comply with the occurrences of disease categories to a large extent. However, the study of the research outputs & published papers of different countries shows low compliance with the burden of diseases in each country. Unfortunately, research performance & output in Iran doesn't comply with the national needs either. There is inordinately low relative research outputs in several diseases or conditions that cause major disease burden in Iran.

We conclude that the coverage of the most conditions covered by Iranian papers is not representative of their importance in the global burden of disease. It seems that Iranian health & medical researches does not targeted the current problems. Iranian researchers should consider the importance of diseases in their scientific researches and be more concerned with research priorities in healthcare.

Although, globally quantity of published papers is correlated with their importance of burden of disease, in developing countries like Iran there is no significant relationship between numbers of papers published in each category & its burden. The low publication rate, few research projects, distortion & nonconformity of medical researches performed and internationally published for each disease/condition category in the country are the basic reasons.

References

Dandona, L., Sivan, Y., Jyothi, M., Bhaskar, U. & Dandona R. (2004). The lack of public health research output from India. *BMC Public Health*, 4,55.

Murray, L. & Lopez, AD. (1996). *The Global Burden of Disease*. Boston: Harvard University Press.

Naghavi, M. (2003). *The Iran burden of Disease report*., Iran Ministry of Health.

Rochon, PA., Mashari, A., Cohen A. et al. (2004). Relation between randomized controlled trials published in leading general medical journals and the global burden of disease. *CMAJ*, 170, 1673–1677.

Swingler, G., Volmink, J., Ioannidis, J. (2003). Number of published systematic reviews and global burden of disease: database analysis. *BMJ,* 327, 1083-1084.

The World Health Report. (2001). *G*eneva: world Health organization.

Distribution of Co-Author Pairs Frequencies of the *Journal of Biological* Chemistry Explained as Social Gestalt

Hildrun Kretschmer [*,**] and Theo Kretschmer [**]

kretschmer.h@onlinehome.de

[*] WISELAB, Dalian University of Technology, Dalian, 116023 (China)

[**] COLLNET Center, Borgsdorfer Str. 5, D-16540 Hohen Neuendorf (Germany)

Introduction

The original Lotka's Law refers to single scientist distribution, i.e. the frequency of authors A_i with i publications per author is a function of i: $A_i = f(i)$. However with increasing collaboration in science and in technology the study of the frequency of pairs or triples of co-authors is highly relevant. Starting with pair distribution well-ordered collaboration structures of co-author pairs will be presented, i.e. the frequency of co-author pairs N_{ij} between authors with i publications per author and authors with j publications per author is a function of i and j : $N_{ij}=f(i,j)$ using the normal count procedure for counting i or j. We have assumed that the distribution of co-author pairs can be considered to be the reflection of a social Gestalt and therefore can be described by the corresponding mathematical function based on well-known general characteristics of structures in interpersonal relations in social networks. We have shown that this model of social Gestalts can better explain the distribution of co-author pairs than by a simple bivariate function in analogy to Lotka's Law.

Gestalt Theory and Hypothesis

In the wake of a tangible change of paradigm in science a number of holistic theories have emerged which are operating on the idea of holographic interacting entities in the world, with several of them also implying a *field concept.* For example: magnetic field in physics, morphogenetic field of living organisms in evolutionary biology, psychological fields in psychology or sociology (Gestalts), etc. *General assumption: A force which emanates from a field generates a balanced evenness among all the individual components in their totality.* This is called *conciseness tendency* in Gestalt Psychology, i.e. there is a *´tendency towards a good Gestalt´*. Several authors take the view that *these fields can be mathematically described*.

Based on Gestalt theory the existence of field-force equilibrium is assumed for explanation how according to the conciseness principle mathematically precise Gestalts could exist in social networks. The development of a simple mathematical function is based on well-known *general characteristics* of structures in interpersonal relations in social networks. These general characteristics are already partly identifiable in groups of higher vertebrates. General characteristics are for example: "Birds of a feather flock together" or "Opposites attract". Both opposing proverbs give rise to reflect on the notion of complementarity. (e.g. also particle/wave by Niels Bohr). The modern notion of complementarity had existed already in a clear-cut manner in old Chinese thought, in the Yin/Yang teaching. Yin and Yang have to be seen as polar forces, as complementary tendencies interacting dynamically with each other, so that the entire system is kept flexible and open to change.

It is conclusively consistent with all above considerations to seek a simple mathematical function (conciseness principle) for the description of ´Gestalt´ (Kretschmer 2002):

$$Z=\text{constant} \cdot (A+1)^{\alpha} \cdot (A_{COMPLEMENT}+1)^{\beta} \cdot (B+1)^{\gamma} \cdot (B_{COMPLEMENT}+1)^{\delta}$$

The interpersonal relations between persons in social groups (variable Z), for example friendship, marriage or co-authorship, etc., will be recorded from the point of view of every individual person with value X of the variable of personality characteristics (for example, age, education or productivity) to all the other authors with value Y of this variable.

Dissimilarity = $A=|X-Y|$

Similarity=$A_{COMPLEMENT} = |X-Y|_{max}+|X-Y|_{min} - |X-Y|$

In analogy: $B= X+Y$ and

$B_{COMPLEMENT} = |X+Y|_{max}+|X+Y|_{min} - |X+Y|$

Accordingly, we have obtained a function with four parameters and one constant. This function can encompass the complementary tendencies (Yin and Yang) in their dynamic interplay on two dimensions each (A and B).

The diversity of behavioural Gestalts is expressed by the variation of the parameters of this function, with the diversity being dependent upon the conditions causing these Gestalts (e.g. men and environment). These many Gestalts are classified into prototypes (Fig. 1). The shapes of the Gestalts can change between the prototypes.

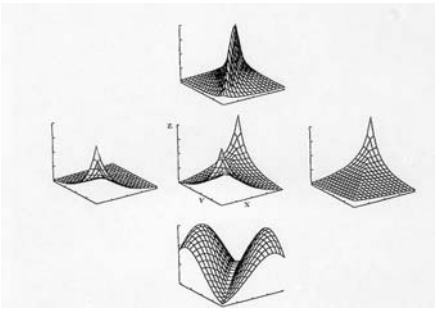


Fig. 1: Prototypes of social Gestalts (Kretschmer 2002, p. 485

Hypothesis:
- The distribution of the frequency of co-author pairs N_{ij} can be considered to be the reflection of a social Gestalt with X=log i, Y=logj and Z= N_{ij}
-The model of social Gestalts can better explain the distribution of co-author pairs than by a simple bivariate function in analogy to Latke's Law.

Data and Results
Co-authorship distribution of pairs of collaborators obtained from the *Journal of Biological Chemistry (JBC)* presented in Fig. 2. The articles from 1980-1998 of the journal *JBC* were studied with 96,136 authors and the total sum of N_{ij}=779,236 co-author pairs. The data are obtained by SCI. Figure 2 shows the distribution of the frequency of co-author pairs N_{ij} between authors with i publications per author and authors with j publications per author. For clarity´s sake and optimum visualization the presentation of data are restricted to authors with at most ten articles i(i=1,2…10). The figure on the right is rotated by 90°. The left pattern in Fig. 2 is most similar to the left prototype of the second row in Fig. 1.

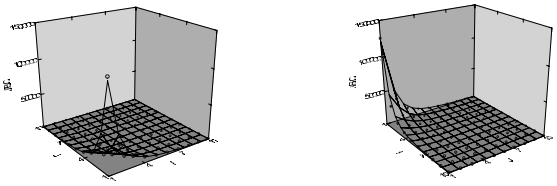


Fig. 2: Distribution of the frequency of co-author pairs N_{ij} of the journal *JBC*

The empirical distribution of the co-author pairs (cf. Fig. 3, first row, left) is rather equal to the Gestalt won by regression analysis (Fig. 3, first row, right) but different from the pattern in the second row won by regression analysis on the basis of the bivariate function in analogy to Lotka's Law (N_{ijT} =constant/$(i \cdot j)^a$).

This is the first *proof for the assumption the distribution of the co-author pairs N_{ij}.can be considered to be the reflection of a social Gestalt.*

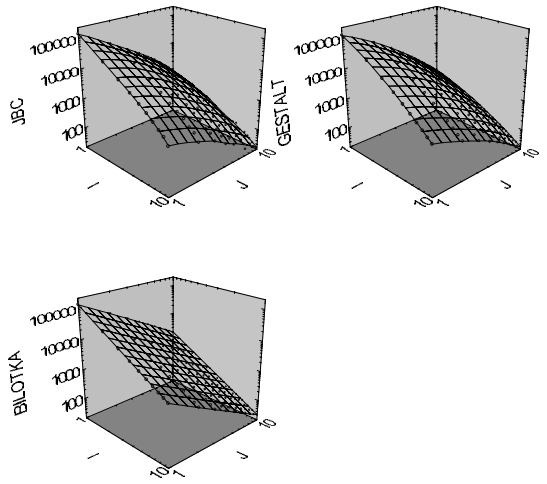


Fig. 3: Triple logarithmic presentation of the empirical pattern (first row, left) in comparison with Gestalt (first row, right) and bivariate function (second row). After regression analyses the correlation between theoretical Gestalt and 469 empirical values of log N_{ij} is R=0.99, F=5945.6.

Discussion and Conclusion
Independent on the social Gestalts there is a long list of other different phenomena we have to study regarding collaboration in science. For example, it is well known the co-authorship depends on different aspects, for example on field, time, social, economic, political conditions, etc.

However, in this paper we have selected a special basic research question in analogy to Lotka's Law and we assume maybe the result is valid in bibliographies of journals in general independent on the above mentioned different phenomena.

References
Kretschmer, H.(2002): Similarities and Dissimilarities in Co-authorship Networks; Gestalt Theory as Explanation for Well- ordered Collaboration Structures and Production of Scientific Literature. *Library Trends*. Vol. 50 No.3, 474-497.

Author Cocitation Analysis of Polyamides Specialty 1974–1999 and the Modification of Multivariate Statistics

Mojca Kotar

mojca.kotar@ntf.uni-lj.si
University of Ljubljana, Faculty of Natural Sciences and Engineering, Department of Textiles,
Snezniska 5, SI-1000 Ljubljana (Slovenia)

Introduction
The research presented is part of a wider study into formal communication in the polyamides specialty, starting from the invention of fibre-forming polyamides (Kotar, 2005). The purpose of the presented part of the study was to reveal a possible model of the intellectual structure of the polyamides specialty in the period 1974 to 1999 as is made evident by author cocitation analysis.

Author Cocitation Analysis as the Method Used
The procedures described by McCain (1990) were followed. All 40 prominent authors were frequently cited, some of them very frequently (top five authors, by number of citing articles, are listed in Table 1).

Table 1. Polyamides authors for author cocitation analysis 1974–1999

No.	Author	No. of citing articles	No. of cited publications
1.	Flory PJ	23,347	2,062
2.	Paul DR	3,745	834
3.	Peterlin A	3,249	794
4.	Treloar LRG	2,589	349
5.	Ward IM	2,473	996

High citedness resulted in an uncommonly wide range of numbers of cocitations in a matrix (from 0 to 1,072 cocitations) and in the compression of the majority of authors into an indiscernible group around the multidimensional scaling (MDS) map centre (Fig. 1). Consequently, a modification of multivariate statistics procedures was needed.

Modification of Multivariate Statistics
The procedure of a modified cluster analysis started with transformation of raw cocitation data into a matrix of correlations. Another correlation was performed before applying the clustering routine. Using the SPSS syntax this means: CORRELATIONS → PROXIMITIES (*measure correlation*) → CLUSTER (*method complete*). The modified multidimensional scaling procedure consisted of a transformation of raw cocitation data into a matrix of correlation coefficients, from which Euclidean distances were calculated and author points placed on the MDS map. In terms of SPSS syntax this means: CORRELATIONS → PROXIMITIES (*measure euclid*) → ALSCAL (*model euclid*) (Fig. 2). Doreian, Batagelj, and Ferligoj (2005, p. 148) state that the minimum and maximum clustering methods are invariant under all transformations of the (dis)similarity measure which do not change the ordering of pairs of units, as was the case in our modification. By using the modified multidimensional scaling, the effect of magnitude of numbers of cocitations was successfully diminished, thus allowing for detection of authors' profile similarities with Euclidean distances.

Conclusions
The clusters of authors in the dendogram (not shown here) and the placements of authors on the MDS map (Fig. 2) correspond well to the expert knowledge of the polyamides specialty. There is no core group of authors with Flory being the most central figure. Five distinct groups of authors are discernible. The MDS map (Fig. 2) shows the possible model of the intellectual structure of a polyamides specialty in the period 1974 to 1999.

Acknowledgment
The author wishes to thank Dr Franci Sluga, Assoc Prof, Department of Textiles, Ljubljana, for his help with the interpretation of results.

References
Doreian, P., Batagelj, V., & Ferligoj, A. (2005). *Generalized blockmodeling* (Structural Analysis in the Social Sciences No. 25). Cambridge: Cambridge University Press.

Kotar, M. (2005). *Bibliometrijska analiza informacij v tekstilstvu in njihovo razširjenje* [Bibliometric analysis of information in textiles and its dissemination]. Unpublished doctoral dissertation, University of Ljubljana.

McCain, K. W. (1990). Mapping authors in intellectual space: A technical overview. *Journal of the American Society for Information Science*, *41*, 433–443.

.

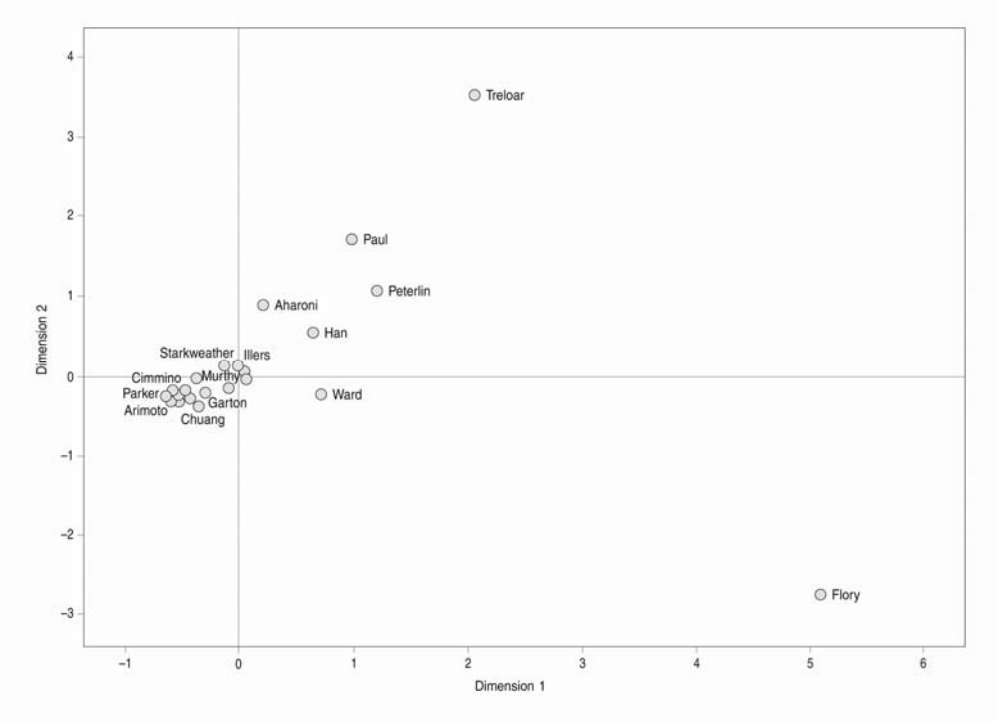


Figure 1. MDS map of 33 authors from polyamides specialty 1974–1999
(classical multidimensional scaling, stress 0.050, R^2 0.996)

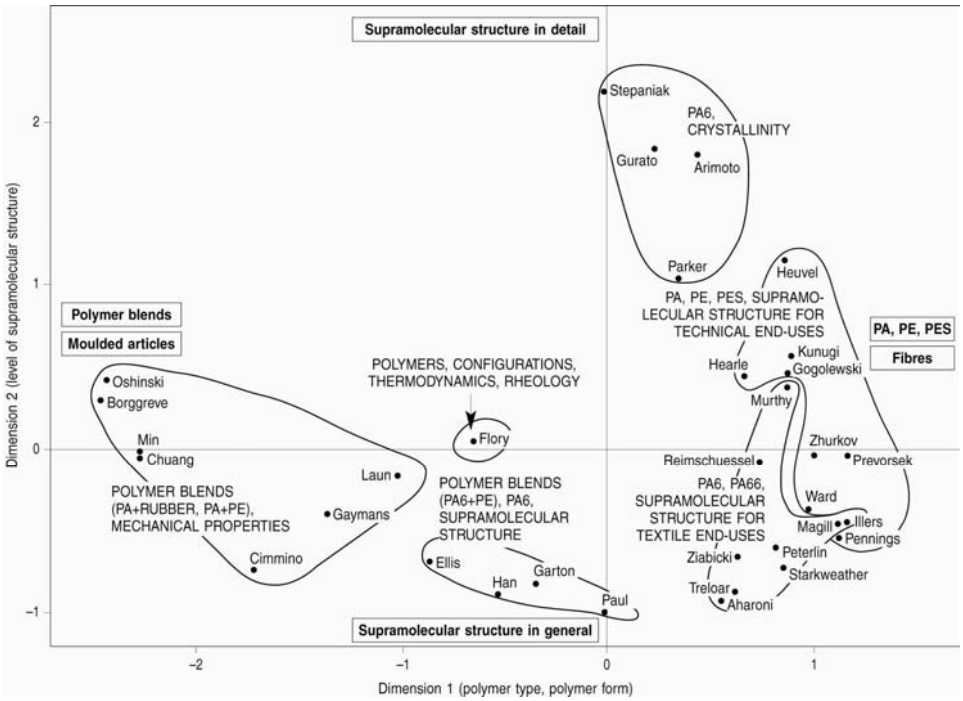


Figure 2. MDS map of 33 authors from polyamides specialty 1974–1999
(modified multidimensional scaling, stress 0.106, R^2 0.953)

Productivity and Prestige among Brazilian Scientists

Paula Leite and Jacqueline Leta

plmelo@bioqmed.ufrj.br, jleta@bioqmed.ufrj.br
UFRJ (Universidade Federal do Rio de Janeiro), Instituto de Bioquímica Médica. Av. Brigadeiro Trompowsky, s/n, Cidade Universitária - Ilha do Fundão. Rio de Janeiro-RJ (Brazil)

Introduction

Although publishing is closely correlated to science, there are scientists who publish a large number of papers and others who publish very few. The motivations for publishing have been an important issue in the sociology of science. According to Fox (1983) there are many reasons leading a scientist to publish and to choose a specific field. Among these reasons are personal characteristics such as personality or age, the scientific environment in which the scientist has carried out his/her studies and accumulative advantages. A specific issue behind the last reason is the Matthew effect, which "consists in the accruing of greater increments of recognition for particular scientific contribution to scientists of considerable repute and the withholding of such recognition from scientists who have not yet made their mark." (Merton, 1968). In other words, according to Merton, the more recognition scientists have (which for some is a synonym for prestige) the more access to research facilities scientists get.

The aim of this study is to investigate the academic and scientific outputs of productive and non-productive Brazilian scientists in order to identify whether or not the trends found for this community are in accordance with both the accumulative advantage theory and the Mathew effect principle.

Method

An extensive database was built with information about scientists from CNPq (National Council for Scientific and Technological Development) and ISI (Institute for Scientific Information) within the period of 1997-2004. All the scientists included in the group under study had a PhD and were registered in research directories by the time of sampling. Scientific, academic and personal characteristics, such as awards received, mentorship of graduate students, degree year, geographic location, science field of his/her research, institution where he/she is affiliated, gender and also publication data were analyzed, and some possible relationships were suggested.

Our database encompasses information about 51,223 scientists. We selected a sample corresponding to the most productive science fields and institutions supported by federal government in the international scenario, which resulted in 2,111 scientists from our original database.

Having defined both groups of scientists of high and low productivity, the number of scientists in our sample was cut down to 240 (120 in each group) for a more detailed analysis.

Results and discussion

The institutions selected by our sampling were universities, one in each of the five Brazilian geographic locations, namely UFPA (Federal University of Pará), UFPE (Federal University of Pernambuco), UFRGS (Federal University of Rio Grande do Sul), UFRJ (Federal University of Rio de Janeiro) and UnB (University of Brasília) and the science fields were biology, engineering and earth sciences.

Prestige and productivity

As an indicator of prestige we chose a particular type of fellowship, entitled "productivity in research" (PQ-1A, PQ-1B, PQ-1C, PQ-1D and PQ-2), a kind of extra salary awarded by CNPq to a very tiny fraction of Brazilian scientists. This fellowship is renewed every year through a peer review process. The data on the award of these fellowships in our sample were obtained from the CNPq database.

There is a gradual hierarchy in terms of prestige from non-awardees to awardees of PQ-1A fellowship. As can be seen in the graph (Figure 1), the prestige of this fellowship is closely related with productivity. However, there is a fraction of scientists of low productivity that are recipients of these types of fellowships. This finding suggests that factors other than publication may be associated with prestige.

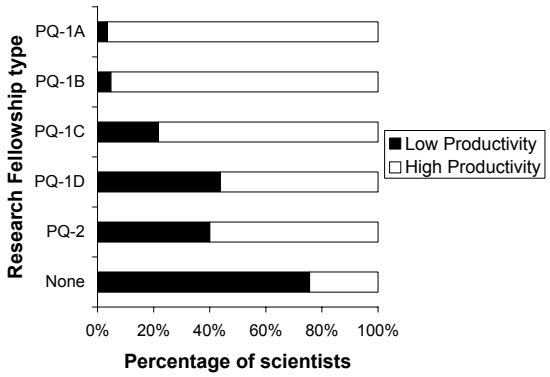


Figure 1. Scientists of high and low productivities relative to fellowship awards.

Prestige and institutions

Comparing the productivity of each university in the three science fields selected, we could see that the number of papers varies significantly among the universities (data not shown). Many factors, beyond doubt, contribute to the different profiles observed. Southeast region, where the most productive university (UFRJ) is located, have some characteristics indispensable for scientific progress. This region is much more industrialized, with more job opportunities, a higher number of universities and research institutions. As a consequence of these social features, a more competitive research environment is found in such region, contributing to its higher productivity, which appears to be consistent with the accumulative advantage theory.

As our results show that the distribution of fellowship grants is related with productivity, we looked at how the awarding of this "productivity in research" fellowship would be granted among different geographic locations. We thus compared productivity with fellowship grants awarded in each university (data not shown). Our results show that scientists from UFRJ receive a disproportionate larger fraction of these fellowships, possibly compromising the number of fellowships granted to scientists from other universities. This finding points to the Matthew effect principle: there are some intrinsic advantages in being from a particular institution. In this specific case, being affiliated to UFRJ, one of the most traditional and most productive in the federal system (Leta & col, 2006), with a large number of scientists, may represent an advantage for career progress, even if the scientist does not fit in a high productivity profile.

Prestige and other factors

According to the literature (Alliston & Stewart, 1974), senior scientists tend to be more productive. Is it true for our sample? Estimating the seniority of scientists in our sample considering the time they concluded their PhD, we can see an increase of productivity along the career stage (figure 2).

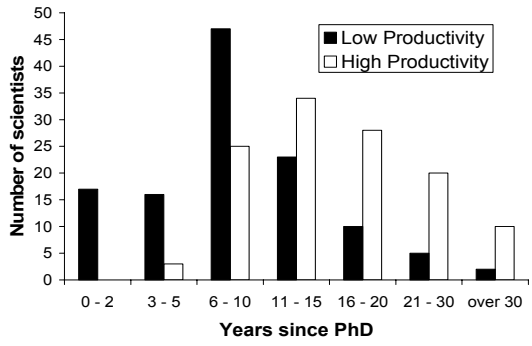


Figure 2. Number of scientists of high and low productivity according to career stage.

The same profile is observed when we compare productivity with research fellowships (data not shown).

Another factor discussed in the literature (Schiebinger, 2001) is the gender of the scientists. Usually, male scientists tend to be more productive than female. Comparing the productivity of male and female in our sample, there is a slight tendency for the former to be more productive than the latter (data not shown). However, if the analysis focuses on research fellowship grants, the differences are more pronounced: male scientists seem to be more likely to gain this benefit, suggesting that there are reasons other than professional competency involved in the granting of the fellowship.

Conclusions

According to the profile of this particular scientific community, we can suggest that the distribution of research fellowships is related to scientific productivity, which is, in turn, associated with career stage. On the other hand, gender comparisons indicate that productivity is not the only factor underlying fellowship awards. Also, our results show that geographic location and some institutional features bring some advantages for these scientists to get published and obtain other benefits. Although our results are consistent with the accumulative advantage theory, a more detailed comparison between the two groups of scientists under study is underway. Some other indicators of prestige, such as number of students, participation in editorial boards, and authorship of patents are being collected from the CNPq database. As a final step, we will interview the scientists from our sample as well as graduate students and members of scientific committees about their concept of prestige in science since this concept is a social construct that depends on others' views and assumptions.

References

Alliston, P. D., Stewart, J. A. (1974) Productivity differences among scientists: evidence for accumulative advantage. *American Sociological Review*, 39 (4), 596-606.

Fox, M. F. (1983). Publication productivity among scientists: a critical review. *Social Studies of Science*, 13 (2), 285-305.

Leta, J., Glanzel, W., Thijs, B. (2006) Science in Brazil. Part 2: Sectoral and institutional research profiles. *Scientometrics*, 67 (1) 87-105.

Merton, R. K. (1968). The Matthew effect in science – the reward and communication systems of science are considered. *Science*, 159, 56-63.

Schiebinger, L. (1999) *Has Feminism Changed Science?* Cambridge: Harvard University Press. Foreign Translations: Portuguese (Editora da Universidade do Sagrado Coração, 2001).

Two New Indicators Derived from the h-Index for Comparing Citation Impact: Hirsch Frequencies and the Normalised Hirsch Index

Jonathan M. Levitt* and Michael Thelwall

j.m.levitt @ wlv.ac.uk, m.thelwall @ wlv.ac.uk
School of Computing & Information Technology, University of Wolverhampton, Wulfruna Street, Wolverhampton WV1 1SB (UK)

Introduction

The h-index is defined to be the largest number (h) of documents that are cited h or more times (Hirsch, 2005). Several studies, including Hirsch (2005), Batista, Campiteli, and Kinouchi (2006), Braun, Glanzel and Schubert (2006) and Van Raan (2006), use the h-index for informetric investigations, or to report h-values for a set of scholars (Cronin & Meho, 2006; Oppenheim, 2007), testifying to the importance of this new measure.

We introduce two new indicators, *Hirsh Frequencies* and *Normalised Hirsch Index*, designed to compare citation impact, and apply them to the annual publications of Library and Information Science. The two new indicators are both related to the h-index. Hirsh Frequencies are designed to reveal the spread of highly cited documents by enabling the distribution of citations to be expressed in terms of the number of documents that exceed a specified multiple of the h-index. In contrast, the Normalised Hirsch Index enables h-indexes to be compared across document collections of different sizes. As described in the Methods, both indicators are readily calculated using Web of Science online, therefore retaining the ease of calculation of the h-index.

We define the *Hirsch k-Frequency*, f(k), as the number of documents in the collection that are cited at least k * h times, where h is the h-index of the collection. For a set of documents we define the *Normalised Hirsch Index*, h_{norm}, by

$h_{norm} = 100\ h^2/n$

where *h* and *n* are the h-index and number of documents of the set. We use h-squared rather than h because there is squaring inherent in the h-index; for example an h-index of 20 requires at least 400 citations (20 documents with 20 or more citations), whereas an h-index of 10 only requires 100 citations. We use the term 'Normalised' to indicate adjusted for sample size, rather than to indicate lying between fixed norms.

Methods

In order to illustrate and assess the techniques we use a case study approach, selecting the field of Library and Information Science (for a comparative analysis over time. This field was selected as one that is known to the authors so that the findings are more interesting and the value of the information produced can be more readily assessed.

Prior to investigating these indicators we systematically identified the complete set of journals and other publications that contain one or more documents in 'Information Science & Library Science' (IS&LS) from the Web of Science. We used this data to isolate all documents in IS&LS and to calculate that the overall h-index, h_{all}, for IS&LS for reviews and articles between 1970 and 2005 is 107.

We readily calculated the indicators using the Web of Science online. We obtained the h-indexes and Hirsch Frequencies by using the 'Times Cited' facility that ranks the citations of search results in order of decreasing citation. We obtained the number of documents in the set from Web of Science's account of the number of documents found.

Findings

Some of our findings are presented in Figures 1 and 2 (all indicators to the 26th November 2006).

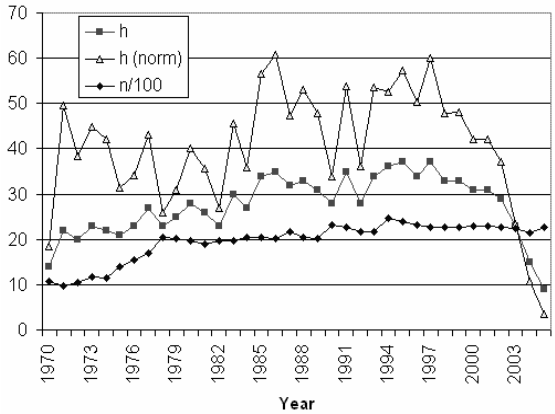


Figure 1. The standard and normalised h-index for IS&LS together with the number *n* of articles and reviews published per year from 1970 to 2005.

In Figure 1, note that the h-index and normalised versions both decline significantly beginning around 1998. This is not surprising and reflects the fact that articles take a while to accrue large numbers of citations. However, this finding indicates that longitudinal comparisons of recent h-indexes (for

departments, authors or journals) may need to adjust for declines in h-indexes. We suggest using h_{norm}.

Figure 1 indicates that, apart from 1998 and after, h_{norm} does not show a clear trend, whereas h tends to increase with n. The absence of a clear trend indicates that, apart from recent periods, the trends in h and n might to some extent cancel one another in h_{norm}. Interestingly, the h_{norm} could be read as a record of the lasting significance of research of each year. A possible interpretation of Figure 1, then, is that whilst 1970 was not a particularly significant year for IS&LS research, the few years after 1971 were, in addition to the mid-to late 1980s and the late 1990s. It is too early to say whether the 2000s are providing lasting contributions to the field. Of course, these findings are all preliminary and do not take into account the dynamic nature of the size of the citing population.

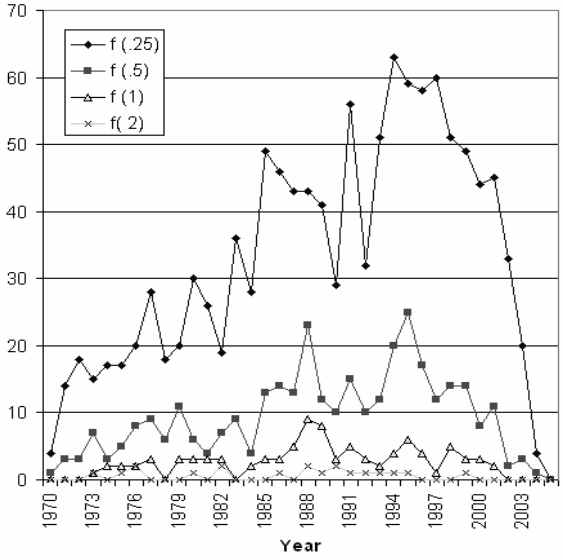


Figure 2. Hirsch k-Frequencies for IS&LS articles and reviews from 1970 to 2005, relative to the h-index h_{all}=107, for the whole collection.

In Figure 2, f(.25), f(.5), f(1) and f(2) provide considerable information on the distribution of the citations that cannot be inferred from the values of h. For example: (1) although the value of h in 1995 is identical to that in 1997, 6 documents in 1995 were cited at least 107 times (h_{all} =107) yet only 1 document in 1997 was cited at least h_{all} times, and (2) although the value of h in 2002 is higher than that in 1992, 2 documents in 1995 were cited at least 54 times (h_{all}/2=53.5) yet 10 documents in 1992 were cited at least 54 times. Note, however, that k-frequencies tend to be dependant upon the size of a document collection, and hence the strongly increasing trend for f (.25).

Conclusions

We seek to demonstrate some of the advantages of using additional indicators to complement the h-index for the task of comparing document collections. We compared different years in Library and Information Science, but the indicators could be used more widely, for instance when comparing the citation impact of departments, authors or journals. We suggest that our Normalised Hirsch Index offers more accurate comparisons between different collections from the same subject than the h-index and the Hirsch Frequencies provide information on the number of documents within different citation ranges. The Hirsch Frequencies f(.25), f(.5), f(1) and f(2) were presented in Section 3, but a study can examine other Hirsch Frequencies and need not restrict itself to four Hirsch Frequencies.

The indicators introduced here need to be interpreted with care. Although the Normalised Hirsch Index provides a rough indication of the average level of citation, sometimes this indication could be very imprecise. In principle the Hirsch Frequencies do allow comparisons between different subject categories as the frequencies are on multiples of the h-index of the subject category; however we recommend treating comparisons between subject categories with particular care.

References

Batista, P.D., Campiteli, M.G., & Kinouchi, O. (2006). Is it possible to compare researchers with different scientific interests? *Scientometrics* 68(1), 179-189.

Braun, T., Glanzel, W. & Schubert, A. (2006). A Hirsch-type index for journals, *Scientometrics* 69(1), 169-173.

Cronin, B. & Meho, L. (2006). Using the h-index to rank influential information scientists, *Journal of the American Society for Information Science & Technology* 57(9), 1275–1278.

Hirsch, J.E. (2005). An index to quantify an individual's scientific research output, *Proceedings of the National Academy of Sciences of the United States of America* 102(46), 16569-16572.

Oppenheim, C. (2007). Using the h-index to rank influential British researchers in information science and librarianship, *Journal of the American Society for Information Science & Technology* 58(2), 297-301.

Van Raan, A.F.J. (2006). Comparison of the Hirsch-index with standard bibliometric indicators and with peer judgment for 147 chemistry research groups, *Scientometrics* 67(3), 491-502.

Breast Cancer Research, Death Rates and Life Expectancy in Iceland: are they Linked?

Grant Lewison

glewisonxx@aol.com
Evaluametrics Ltd, 50 Marksbury Avenue, Kew, Richmond, Surrey, TW9 4JF (England)

g.lewison@ucl.ac.uk
School of Library, Archive and Information Studies, University College London, Gower Street, London WC1E 6BT (England)

Introduction
Because the links between biomedical research and better health are complex, it is difficult to show that more research actually helps. Why then do small countries carry out such research? Several reasons can be posited, including:
-to underpin the training of healthcare personnel
-to allow the import of new techniques into medical practice
-to attract high quality medical staff, who will be better able to argue for more resources
-to provide good advice to government
-to develop clinical guidelines for the country.

Some previous studies (Lewison et al., 2001; Rippon et al, 2005) have shown a positive correlation between disease burden and the amount of research, but others have shown none, or a negative correlation. What is actually needed is a time series, with disease burden trends moving oppositely to research outputs, or case studies from a country that can demonstrate an association between research and clinical benefit, and that it is causal and not co-incidental.

A particular difficulty for such studies is that there are many confounding factors that affect disease rates, including local socio-economic conditions, and research institutions sited in relatively poor areas, perhaps for historical reasons. Thus studies on breast cancer prognosis in south-east England (Thames Cancer Registry, 2001) showed that the worst outcomes occurred in northwest and southeast London, which were close to the two major research centres in the region, and that the best outcomes, in the leafier parts of Surrey, were in a breast cancer research desert.

We need an indicator that is relatively independent of the general health status of the local population. One such measure is differential life expectancy, or differential survival rate, between those who die from the given disease and deaths in the general population. This can occasionally be positive, as for prostate cancer, but for breast cancer it is negative, and Figure 1 shows the situation for Danish women.

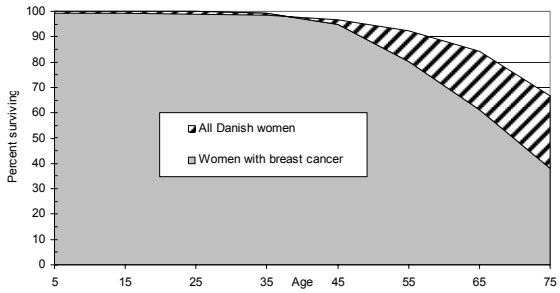


Figure 1. Survival curves for all Danish women (diagonal shading) and those dying from breast cancer, 1995-99 (grey tint). Source: WHO

The five Scandinavian countries may be compared as all have good public health systems and a strong tradition of biomedical research. Death rates from breast cancer range from 3% to 6%, with Iceland much the highest in the early 1990s, even after mammograms were introduced in 1987.

Breast cancer research outputs
Data were obtained from counts of articles, notes and reviews in the Science Citation Index (SCI) that were within the sub-field of cancer research (Cambrosio et al, 2006), defined by means of cancer journals and title keywords, and also contained title words indicative of work relevant to breast cancer, e.g., BRCA, breast*, mammar*, or were in *Breast Cancer Research and Treatment*. World output is about 3000 papers per year, just under 9% of all cancer research. Within this total, the relative commitment (RC) of the Scandinavian countries, calculated as the ratio of their integer-count percentage presence in breast cancer research to that in all biomedical research in the SCI, has been between about 1 and 2, except for Iceland, whose RC exceeded 10 for a few years in the mid-1990s, see Figure 2.

The increase in Icelandic research output may have resulted from new methods in molecular biology for

the analysis of tissue samples from breast cancer patients and their family members.

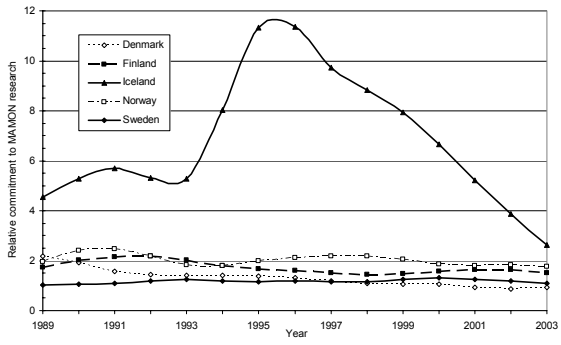


Figure 2. Variation in relative commitment to breast cancer research in five Scandinavian countries, five-year running means.

Changes in Icelandic breast cancer death and survival rates

Figure 2 showed that Iceland increased its RC to breast cancer research massively in the 1990s, and the actual number of papers quadrupled between 1988-94 and 1995-2001. There was a parallel drop in the death rate, which went down from 5.5% of all female deaths in the mid-1990s to 3.5% in 2000; no other Scandinavian country achieved such a reduction, all remaining between 3 and 4%. The drop was particularly large among younger women, see Figure 3.

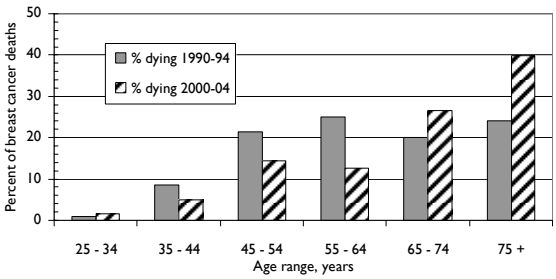


Figure 3. Distribution of ages at death for Icelandic women dying from breast cancer in 1990-94 (tinted bars) and in 2000-04 (striped bars). Source: WHO

Discussion

Since mammographic screening was introduced in 1987, and patients had rapid access to clinical treatment if fine needle cytology provided a diagnosis of breast cancer, one might have expected breast cancer death rates to have declined in the late 1980s or early 1990s. The reverse was the case.

The temporal correlation between the subsequent rise in research output and the fall in deaths is striking, but much of the research was focussed on genetics, with Icelanders participating in the discovery of the BRCA1 and BRCA2 genes. Moreover, its research level was much more basic than the average for all breast cancer research in 2000.

Possible contributing factors that may have speeded up the transfer of knowledge from the lab to the bedside were the unusually close co-operation between the university, the university hospital and the Icelandic Cancer Society, and the fact that many of the leading researchers had dual appointments. Some of the research involved the use of genetic typing to identify women at increased risk of breast cancer, so the main causative mode may have been prevention and very early diagnosis rather than cure. Icelandic breast cancer researchers also had a high level of international collaboration, with almost half the papers having a foreign co-author and 32 other countries involved in total, so they would have been well aware of current developments world-wide.

Although the case for Icelandic breast cancer research having had a positive influence on disease outcomes is superficially attractive, it has to be admitted that it rests on a rather small number of papers. Moreover, there may be other explanations for the improvements of the late 1990s, such as a greater availability of new drugs through the Icelandic health service – though these may have been made more readily available in an atmosphere that was increasingly positive towards fighting breast cancer. However, no corresponding rise in research or drop in death rates occurred in the other Scandinavian countries.

Acknowledgments

The MS Excel macro used for geographical analysis was developed by Dr Philip Roe. The breast cancer filter was developed by Mr Robert Carpenter of Bart's Hospital and Dr Lynne Davies of Cancer Research UK. Dr Jorunn Eyfjord of the Icelandic Cancer Society and Dr Sigurdur Ingvarsson of the Iceland University Hospital made helpful comments.

References

Cambrosio A, Keating, Mercier S et al (2006) Mapping the emergence and development of translational cancer research. *European Journal of Cancer,* vol 42 (18), pp 3140-3148

Lewison G, Grant J and Jansen P (2001) International gastroenterology research: subject areas, impact and funding. *Gut*, vol 49, pp 295-302.

Rippon I, Lewison G and Partridge MR (2005) Research outputs in respiratory medicine. *Thorax*, vol 60, pp 63-67.

Thames Cancer Registry (2001) Information pages – data-area of residence-major cancers-breast cancer-survival tables. Available at: http://www.thames-cancer-reg.org.uk/data/documents/report2001/breast_rs_2-001.htm

WHO (2006) Statistical tables of causes and ages of death in member countries.

.

Counting Citations: Fractionation of Addresses and "World-Scale", a New Scalar

Grant Lewison

glewisonxx@aol.com
Evaluametrics Ltd, 50 Marksbury Avenue, Kew, Richmond, Surrey, TW9 4JF (England)

g.lewison@ucl.ac.uk
School of Library, Archive and Information Studies, University College London, Gower Street, London WC1E 6BT (England)

Introduction

Research evaluation commonly rests on citation counts. There is by now an extensive literature on the theory of citations – what do they really mean and is it fair to construct the edifice of evaluation on their foundation? There are also many papers that purport to rank nations, or research institutions, on the basis of such counts. However most do not define the fields of research accurately, or allow for the differences between basic research and applied or clinical work, and some don't even use a standard citation time window. It is hardly surprising that there is widespread scepticism about the trustworthiness of such supposedly quantitative evaluations.

In this paper, attention is paid to the need to allow for the contribution made by other entities to the research being evaluated – either countries or institutions – and to the distribution of citation counts, specifically, how many of an entity's papers score highly enough to be in the top x% of the world on this measure. Account is also taken of the need for statistical realism when comparisons are being made, so that one can be reasonably sure that any conclusion that A is better than B is not because of mere chance.

Last year at Leuven, Aksnes and Sandström (2006) showed that fractional counting gave a much fairer account of the relative merits of Swedish and Norwegian research. This method will almost always depress a nation's apparent citation performance relative to that shown by integer counts because it has been shown several times that internationally co-authored papers are more highly cited than ones from a single country, and they will figure less prominently in a country's tally if fractional counts are used. Several fractional counting methods exist, including the assignment of extra credit to the first author, or the corresponding author, but equal partition among all the addresses seems more justifiable and is used here.

Many papers that evaluate research on the basis of citations use mean citation counts, sometimes scaled to account for differences between fields. But most evaluators are more interested in performance at the upper end of the scale, and want to know how many papers are published by an entity that score above a certain benchmark. This can conveniently be fixed in relation to the distribution of citations for all papers in the given subject, perhaps the number of citations (in a given time window) needed for a paper to be in the top 20%, top 10%, top 5% or even higher. The choice of centile will depend on the output of the entity. We may need to compare the observed and expected numbers of papers meeting the criterion, and these must be big enough for the difference to be statistically significant if it is to be concluded that A is better or worse than the world average – or better or worse than B.

Such an indicator is similar to the well-known "activity index" that determines an entity's commitment to a subject area as the ratio between its percentage presence in the subject to its percentage presence in all science, multiplied by 100. However, the new scalar is calculated from the percentage of the entity's output cited at or above the x^{th} centile to x, again multiplied by 100. We can call it "world-scale" by analogy with the charter market for oil tankers. Values can be calculated for several values of x, and then either used individually to give a composite picture of performance, or averaged.

Methodology and results

This paper illustrates the above argument with a worked example. The numbers are real, but they are introduced purely to make the working of the new approach to citation analysis clear rather than to give definitive indicators of countries' relative merit.

The sample used for analysis consists of 1000 papers randomly chosen from the field of cancer research from the year 1998, cited from 1998 thru 2002. The original selection was based on a cancer "filter" (see Cambrosio et al., 2006) applied to the Science Citation Index, and contained title words as well as a list of specialist journals. Citation counts in the five year window ranged from zero to 187. The top 50 papers (5%) all had 43 citations or more; the top 100

(10%) had 31 cites or more and the top 200 (20%) had 17 or more. However because citation counts are integers, there were actually 104 papers (10.4%) with 31 cites or more, so comparisons at this level need to be made with x = 10.4, not 10.0.

For each paper, the fractional contributions from 12 leading countries were calculated and put in a column of the spreadsheet in which the papers were sorted in descending order of citation counts. It was then easy to perform either a fractional or an integer count of the numbers of papers from each country with respectively at least 43, 31 and 17 cites, and to compare these numbers with the total number in the sample. Data are given in Table 1.

Table 1. Numbers of papers from 12 leading countries in cancer research with 43 cites (top 5%), 31 cites (top 10%) and 17 cites (top 20%) in 5 years.

	Integer counts				Fractional counts			
Cites	***43***	***31***	***17***	***0***	***43***	***31***	***17***	***0***
AU	0	1	3	19	0.0	0.1	0.7	13.9
CA	1	3	8	39	1.0	3.0	7.6	34.0
CH	1	1	4	20	0.1	0.1	1.8	14.5
DE	3	8	18	82	1.4	5.3	14	64.3
ES	2	4	8	27	2.0	3.3	6.7	24.7
FR	4	7	15	70	2.6	5.2	12	60.7
IT	2	6	14	71	0.9	3.5	9.1	55.2
JP	6	8	19	122	4.5	5.0	13	111
NL	2	3	4	26	1.5	2.5	2.8	18.1
SE	1	2	4	21	1.0	1.4	2.2	17.7
UK	5	11	20	83	3.0	7.0	12	65.1
US	30	66	12	413	28	59	104	365
Other	5	12	30	186	4.3	8.4	19.5	151
W'ld	50	104	206	1000	50	104	206	1000

World-scale values can now be calculated for each of the countries at each of the three centiles. For example, Spain has 8 papers at the 20% level out of 27 on an integer count basis, so its world-scale is (8 x 100)/(27 x 0.206) = 144 and 6.7 papers out of 24.7 on a fractional count basis, giving a world-scale of 132. But even on an integer count basis, the expected value is only 5.56 papers, so the difference between the observed and expected values is not statistically significant at the 5% level.

Clearly, the sample size is too small for it to be validly concluded that Spanish cancer research is superior. On the assumption that exactly similar results would be obtained with larger samples, how big would these have to be for us to conclude safely that Spanish cancer research was in fact superior? The answers are shown in Figure 1.

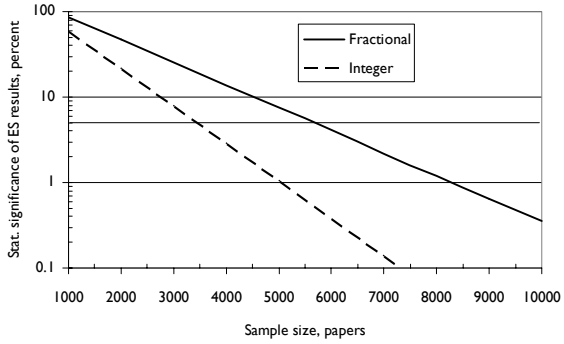


Figure 1. Statistical significance of apparently superior Spanish citation performance in cancer research at the 20th centile as a function of sample size, on integer and fractional counts.

Comparison with mean cites per paper

Figure 2 shows the mean integer count world-scale values for the 12 countries of Table 1 plotted against mean cites per paper. There is clearly a close correlation between the two indicators, but the world-scale values discriminate rather more between the best (USA) and worst (Australia, AU). Points above or below the trend-line represent countries with a higher or lower impact at the higher centiles. The two indicators are similar, but measure different aspects of performance and each has its merits.

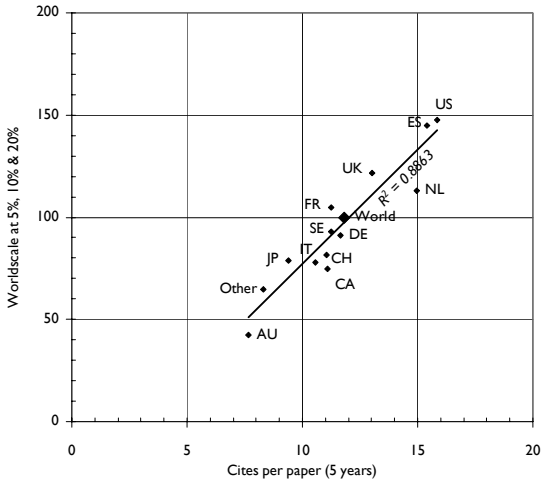


Figure 2. Comparison of citation world-scale values for cancer research papers from 12 nations with their mean cites per paper.

References

Aksnes DW and Sandström U (2006) National citation indicators, their methodological foundation and political interpretation. *Abstracts of 9th International Conference on Science and Technology Indicators*, Leuven, Belgium, pp 3-5

Cambrosio A, Keating, Mercier S et al (2006) Mapping the emergence and development of translational cancer research. *European Journal of Cancer,* vol 42 (18), pp 3140-3148

.

New Visual Interfaces for Author Co-Citation Mapping

Xia Lin

xlin@drexel.edu
College of Information Science and Technology Drexel University, Philadelphia, PA 19104 (USA)

Introduction

Visualization is an essential tool for citation analysis and bibliometic studies. In recent years, many citation visualization systems have been developed. Most of them are centered on their unique visual displays. For example, the Butterfly system described in Mackinlay, Rao & Card (1995) called attention to citation visualization with its innovative "butterfly display" design. SCI-MAP developed by Henry Small (1999) uses a set of hierarchically nested circle-maps to represent co-citation clusters and document pathways in a scientific domain. VxInsight by Sandia National Laboratories (Boyack, et al. 2002) uses the landscape metaphor to produce a terrain view of documents and citation relationships. CiteSpace by Chaomei Chen (2006) combines network and cluster displays with a timeline design to visualize patterns and trends in scientific literature.

All these innovative systems apply information visualization technology to citation data uniquely. Each offers useful insight for the underlying citation data from a different perspective. And all the interfaces are designed differently. There has been little discussion in the literature on what visual elements are most appropriate for citation data visualization, and what design principles should be followed in developing a citation mapping and visualization system. This is an important question that needs to be explored before any of the systems can become a practical tool. In this poster presentation, we will discuss several useful interface design principles that we identified during the redesign of our author co-citation mapping system, AuthorMap. The purpose is to start a discussion on visual interface design for citation data visualization.

AuthorMap

AuthorMap (previously called AuthorLink, Lin, et al. 2003) is an experimental system built with the ISI Arts & Humanities Citation Index (AHCI) from 1988 to 1997. A main feature of the system is its capability to generate instant author co-citation mapping from a very large citation database. The system generates two types of maps: a link-map generated by the Patherfinder algorithm (Schvaneveldt, 1990) and a region-map generated by Kohonen's self-organizing Feature Map (Kohonen, 1997). When a seed author is provided by the user, the system quickly identifies a list of authors co-cited most often with the seed author, and then generates an author co-citation matrix for all the authors in the list before mapping the data to a visual interface. The interface allows the user to explore author relationships and author co-citation patterns through various interactive functions. For example, the user can click on an author's name on the display to start a new mapping for that author. The user can also drag an author's name to the search box to initialize a search. When multiple authors' names are added to the search box, a Boolean search will be conducted automatically. The user can also show or hide the number of co-citations on the screen, or request more information on specific authors.

Recently, the AuthorMap interface has been modified to work with all the ISI citation databases through the DIALOG service. Figure 1 is an example of mapping author "Howard D. White" in the ISI Social Science Citation Indexing (SSCI). The mapping clearly shows three focus areas of White's intellectual work: they are represented by author groups from the "citation analysis" area, the "library and information science" area, and the "general knowledge theories" area. Their relationships are shown through links and co-citation counts on the displays.

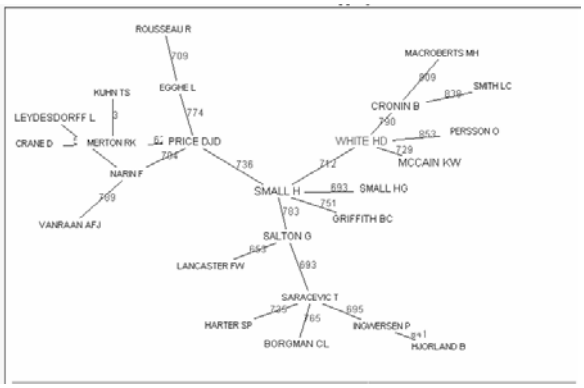


Figure 1. An Author co-citation map

Interface Design Principles

During the conference, we will demonstrate the AuthorMap system and its new interfaces. The new interfaces have been implemented with a Rich Internet Application (RIA) tool with the graphical component designed in Macromedia Flash. With this advanced visual design environment, it becomes clearer that the real issue in designing a citation-

based visualization system is the lack of understanding in what visual elements, styles, and representations can best be used to present the underlying citation data and their relationships. Many visual styles of various visualization systems are dominated by the mapping algorithms, rather than user need or system functionality.

According to Bruce "Tog" Tognazzini (2003), "Effective interfaces are visually apparent and forgiving, instilling in their users a sense of control. Users quickly see the breadth of their options, grasp how to achieve their goals, and do their work." Tognazzini then provides a list of "first principles of interaction design," which are fundamental for any visual interface design. In particular, for citation data, we understand that some of the principles need to be emphasized and others need to be added or expanded. Following is a brief discussion of the four most important principles we considered during the redesign process.

Readability

– It is important to consider readability for both text and graphics. Even on a visual interface, text is often necessary to help users understand the displays. The color, the background contrast, and the font size are essential considerations for text readability. Text and graphics need to be integrated visually and made easy to read and interpret.
– The amount of information provided on the screen will significantly impact the graphics readability. Many visualization interfaces provide too much information, which only makes the graphics more difficult to understand.

Validity

– It is important to visualize the underlying data/knowledge structures, not the structures created by mapping or layout algorithms.
– Graphics and visual displays show only a summary of the underlying data. It is unavoidable that some information will be lost due to the process of mapping high-dimensional citation data to the two-dimensional visual space. The user thus should be provided with tools that can help them explore and understand details of the display to validate the structures shown on the screen.
– Keep in mind that people often interpret graphics differently. It is important not to mislead the viewer as it can be easily done on a visual display.

Usability

– "Overview first, zoom and filter, then details-on-demand" (Shneiderman, 1996). Many interactive functions should be designed based on this principle. However, an overview of a very large amount of citation data might not be very useful.
– The interface needs to keep the user informed at all times. Many citation searching and mapping tasks are computationally expensive, yet the system response time has to be kept reasonably short (often suggested as less than one or two seconds). A balance has to be found.

Usefulness

– The purpose of visualization is insight, not pretty pictures. This is particularly important for citation data mapping.
– Visual citation mapping system is only useful if it will help users in their intellectual endeavor with citation data -- whether it is citation analysis, knowledge discovery, or searching and browsing. The system must allow users to do their own mapping and help them see things that are otherwise difficult to see.
– The system needs to connect to some significant citation databases or allow users to provide their own data for mapping and analysis.

References

Boyack, K. W.; Wylie, B. N.; & Davidson, G. S. (2002). Domain visualization using VxInsight for science and technology management. Journal of the American Society for Information Science and Technology, 53(9), 764-774.

Chen, C. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. Journal of the American Society for Information Science and Technology, 57(3), 359-377.

Kohonen, T. (1997). Self-Organizing Maps. Berlin: Springer-Verlag, 2nd ed.

Lin, X.; White, H. D.; & Buzydlowski, J. (2003). Real-time author co-citation mapping for online searching. International Journal of Information Processing & Management. 39(5), 689-706.

Mackinlay, J.D.; Rao, R.; & Card, S.K. (1995). An organic user interface for searching citation links. InProceedings of CHI'95 (Denver CO, May 1995), ACM Press, 67-73.

Schvaneveldt, R. W. (Ed.), (1990.) Pathfinder associative networks: Studies in knowledge organization. Norwood, NJ: Ablex.

Shneiderman, B. (1996). The Eyes Have It: A task by data type taxonomy for information visualizations. In Proceedings of the IEEE Symposium on Visual Languages, pages 336-343, Washington. IEEE Computer Society Press, 1996.

Small, H. (1999). Visualizing science by citation mapping, Journal of the American Society for Information Science, 50 (9), 799-813.

Tognazzini, B. (2003) First principles of interaction. Design.. Retrieved January 30, 2007 from: http://www.asktog.com/basics/firstPrinciples.html

Patent Indicators of Genetic Engineering Research of EU Countries - Overseas Performance: Preliminary Study

Szu-chia Scarlett Lo

losu@dragon.nchu.edu.tw
Graduate Institute of Library and Information Science, National Chung Hsing University, 250 Kuo-kuang Rd., Taichung 402, Taiwan, R.O.C. (Taiwan)

Introduction

The genetic resources and technologies maintain being important parts in the economic framework of the 21st century. The researchers from both public and private sectors continue flooding into this field. Not only the scientific research, the development of genetic technologies still carry on the roles as driving forces of bringing new resources and global commerce (Rifkin, 1999). In this study, the author applies the idea of technological merit by taking the patent bibliometrics approach, examining USPTO patents granted to EU countries in Genetic Engineering from 1981 to 2004 to depict the achievement of the Genetic Engineering research of European Union (EU) countries.

Data and Research Methods

The data source used in this study is USPTO Patent database, one of the most exhaustive patent sources. The patents analyzed in this study were selected by the International Patent Classification (IPC) numbers and assignee country was added to the criteria to identify the patents granted to EU countries. The patents, which have the principal IPC numbers of section of "Mutation or genetic engineering" (C12N 15/00) are defined as basic Genetic Engineering patents. For setting apart the patents granted to EU countries, the assignee country was used in the strategies. Two patent bibliometrics methods were used in the study, Patent Count and Citation Analysis. Patent Count was used for productivity analysis and Citation Count was used for impact analysis. Besides two counting techniques, Bradford's model (Narin & Moll, 1977; Garfield, 1980) was used to identify the core assignees that hold a substantial portion of the Genetic Engineering techniques.

Results

Productivity

919 USPTO were granted to 14 out of EU countries in Genetic Engineering Research, 39 in average annually. The records showed that the earliest patents were granted in 1981, which were granted to France, German and Hungry. Before 1989, less than 50 patents were issued and it was until 2001, the number of patents granted annually exceeded 100. As indicated above, it was till 1981 before any patents granted. For the next 8 years (including 1981) there were 2 to 8 patents granted to EU countries. The period of 1981 to 1988 was identified as 'Initial' stage. During the years of 1989 and 1995 more patents were granted comparing to the 'Initial' stage. 20 patents in average were granted annually and it could be seen as 'Developing' stage. During the period of 1996 to 2001, the number of patents granted grew rapidly. The numbers of patents were increased by 1.5 and 2 times every year. It could be identified as 'Raising' stage. Figure 1 gives a visual of the result of patent count.

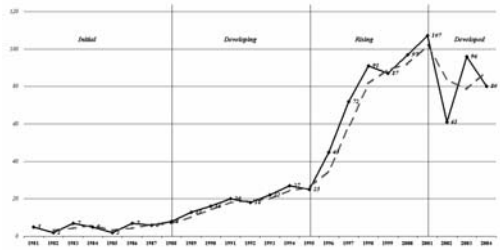


Figure 1. Annual Patent Count

919 patents granted to EU countries were most located in two techniques groups, "Mutation/Genetic engineering" and "Recombinant DNA-technology". Further examining the patents in "Mutation/Genetic engineering", it was found that the techniques of those patents mainly deal with the process of Recombinant DNA Technology, especially the techniques of introducing the foreign genetic material by using vectors and the application of the recombinant techniques to produce the proteins and enzymes. The patents in "Mutation/Genetic engineering" were most about Modified DNA or RNA Fragments and Introduction of foreign genetic material using vectors techniques. Figure 2 shows the distribution of technologies.

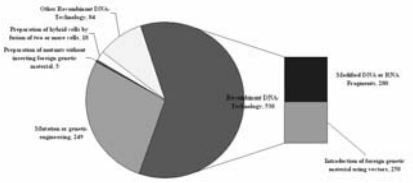


Figure 2. Distribution of Patent Technologies

Germany, France and United Kingdom were the most productive countries among 14 EU countries granted USPTO patents, owned more than two third of the total patents (68,34%). 295 assignees were identified and it was found that most of the assignees owned limited numbers of patents. Close to 55% of the assignees owned one patent. Only 9.15% of the assignees were granted more than 10 USPTO patents. Probing the distribution of the productivity of the assignees by Bradford's model, 9 assignees that own more than 20 patents were identified in the core zone. The author analyzed the assignees by 4 zones and the mathematic relation is '$1:a:a^3:a^4$'. The value of 'a' is 2. Figure 3 is the visual presentation of the results of Bradford's Law analysis.

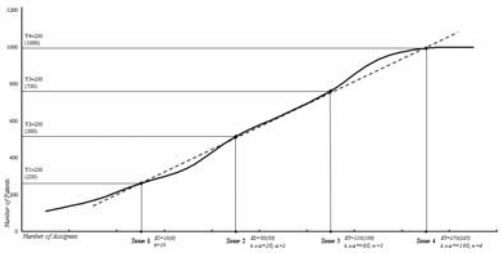


Figure 3. Distribution of Assignees Productivity - Bradford's Law Analysis

Table 1 lists the top 5 productive assignees. Most of the productive assignees specialized or had patenting the techniques on modifying the DNA and RNA fragments, such as Centre National de la Recherche Scientifique and Institut National de la Sante et de la Recherche were granted more patents in modified form of genes encoding animal proteins, and Plant Genetic, Zeneca, Bayer and BASF specialized in modified form of genes encoding plant protein.

Table 1. Top10 Productive Assignees

Rank	Assignee	Patents
1	Institute Pasteur	49
2	CNRS	31
3	Novo Nordisk A	30
4	Plant Genetic Systems	29
5	INSR	27
6	SmithKline Beecham plc	26
6	Zeneca Limited	26
8	Bayer Aktiengesellschaft	22
8	Gist-Brocades N.V.	22
10	BASF Aktiengesellschaft	18
10	Transgene S.A.	18

Besides recombinant DNA technology, Institut Pasteur also held the techniques on DNA sequences, and patents granted to Nordisk and Gist-Brocades were more related to production of Enzymes. Transgene concentrated on the technologies of preparing vector and expression systems of introducing foreign genetic materials. Among the core assignees, the collaboration strategies could be detected by identified the co-owned patents. For example, Institut Pasteur, Centre National de la Recherche Scientifique and Institut National de la Sante et de la Recherche, co-owned several patents and some of them were even international effort. However, there was not particular technological area shown in those co-owned patents.

Research Impact

919 EU countries owned patents were cited 3,328 times, 3.62 in average. 43% patents were not cited by later issued patents. Closed to 84.8% of patents were cited less than 5 times. A few patents were cited more than 20 times. Among the most productive countries, UK, France and Netherlands were most cited EU countries. The top 10 productive assignees were cited 1,130 times, 3.79 in average. Institut Pasteur was the most cited assignee, 77% patents were cited. Institut Pasteur was cited 180 times, 3.67 in average. Centre National de la Recherche Scientifique was listed the second place by the citation count. 88% of granted patents were cited by other patents, 178 in total and 5.74 in average. Other productive assignees also showed high impact, such as Gist-Brocades. Gist-Brocades was granted 22 patents, listed 8th place by the patent count, but demonstrated high research impact comparing to other productive assignees, Gist-Brocades was cited 139 times, 6.32 times in average.

Conclusion

919 patents were granted to 14 EU countries from 1981 to 2004. Four developing periods were identified through the issued date analysis, initial, developing, raising and developed. As entering year 2007, the declining starts to show, except the bouncing force in Germany. Recombinant DNA-technology, especially the modification of DNA and RNA fragments and introducing genetic materials were the techniques focus. The productive indicator also showed the research effort of applying the mutation techniques on production of protein and enzymes. This study tried to apply the patent indicators to show the performance of genetic engineering research. The primary results showed that the patent bibliometrics might be used to assess and present the research the search results. Further correlation analyses could be done to reveal more details in depth.

References

Garfield, F. (1980). Bradford's law and related statistical patterns. *Current Contents*, 19: 5-12.

Narin, F. & Moll, J. K. (1977). Bibliometrics. Annual Review of Information Science and Technology, 12, 35-58.

Rifkin, J. (1999). The Biotech Century: Harnessing the gene & remaking the world. New York: Jeremy P. Tarcher/Putnam..

A Validation Study of HistCite™: Using the Discoveries of *Fullerenes* and *Nanotubes*

Diana Lucio-Arias

dlucioariasfmg.uva.nl
ASCoR (Amsterdam School of Communication Research), University of Amsterdam (the Netherlands)

Introduction

The purpose of this paper is to analyze the (re)construction of scientific knowledge through citations. Algorithms from social network analysis and information measures can be used to identify different characteristics of the citation network that are associated with the process of knowledge construction. Two developments relevant to the nanoscience field of research are considered as cases: Fullerenes, discovered in 1982 and the fullerene-like structures Nanotubes, discovered in 1991 while researchers were investigating fullerenes.

The relations represented by citations in scientific literature enable us to reconstruct scientific developments historically. The antecedents and consequences of a scientific discovery can be accurately traced through the diagram of relations drawn from citations. However, this diagram does not yet inform us about cognitively significant linkages among the citations. The software HistCite™ creates chronological network of citations from any set of documents retrieved from the ISI Web of Science. Is it possible to diagram the dynamics of a particular scientific development over time using references? Algorithms from social network analysis and information measures are added to analyze structures and dynamics of these citation networks. Different stages of development can be expected to show different citation patterns.

HistCite™ was used to create a citation diagram for two sets of documents: those retrieved (March 2006) from the ISI Web of Science with "fullerene*" in their titles and those with "nanotubes*". The sets contained respectively 7,696 and 9,672 documents. Using HistCite, the citation index created can be visualized as a citation network and exported into the Pajek-format for further analysis. I used for both cases the citation network created with the 30 most cited documents, mainly because my interest is on emergence and diffusion of knowledge. The most cited documents would represent the intellectual base of the sets (Price, 1965; Persson, 1994) and therefore, the claims that have already been validated by the community of scientists. Being a very coherent and consistent subset of documents enables an easier interpretation of the results of applying network algorithms and information measures, before considering further scaling.

The analysis of the citation weights trough the search path count available in Pajek identifies the main path through which early and later documents are connected. This allows a structural analysis (Carley *et al.*) of the historical reconstruction. In Figure 1 and Figure 2, the nodes and links forming this main path for each set of documents are highlighted.

Because each scientific text adds new information, the citation network can also be analyzed in terms of the expected information content that the distribution has changed when the next scientific text was introduced. For this case, the distributions considered were of the references made in each of the texts. Because a posteriori citations ("citing") give information about expected recognition in subsequent texts, the citing dimension was also considered. Distribution of the cited references may measure the convergence of references to previous scientific developments. Title words distributions and their combination with cited references will be considered in a follow-up of this approach (Van den Besselaar & Heimeriks, 2005).

Computing this expected information measure estimates a value for each of the links of the HistCite figure. This value depends on distributions and is expressed in bits of information. However, it has a meaning and interpretation different from the transversal weights provided by Pajek and used to determine the main path highlighted in blue. Main path analysis enables a structural interpretation of a network, it identifies the most relevant or influential set of nodes (Hummon & Doreian, 1989). The expected information measure enables us to analyze path dependent transitions in the set (Frenken & Leydesdorff, 2000). Both of these indicators add value to the citation diagram provided by HistCite, where the links represent a one directional relation given by the citations with each a value of one.
The expected information value was computed for all possible combinations in the fullerene set of documents. The negative transitions represent critical paths and have been dashed in Figure 1 and Figure 2. Critical paths are related to complex system dynamics and indicate the extent by which the next state of the distributions will be determined by the previous one. In principle, such an evolving

system can also be expected to "forget" parts of its history; the critical paths identify nodes that become "repetitive" to understand the evolution of the scientific development (fullerenes) in terms of the common references to previous literature. In the other hand, main path analysis chooses from the citation network a sequence of papers that best represent the historical development within a particular set.

Discussion

HistCite is a powerful tool that allows, among many other things, to identify the most cited documents in any set or combination of sets of documents, and builds the inner-citation-matrix for them. The algorithm used to identify the most relevant documents is based on frequency of citations over time and therefore all the links have a value of one. HistCite visualizes the most relevant papers and the relations among them and is, therefore applicable to the historical reconstruction of scientific developments. Main path algorithms available in Pajek identify, among the network of citations, which papers are the most relevant in the overall flow of citations. The algorithm used in HistCite bases its estimations on frequencies; main path algorithms gives priority to highly cited papers that have high citations to previous papers identifying papers that can be associated with transitions (similar to the betweeness algorithm). HistCite can be complemented with results from main path analysis and information measures in three ways: first, in a different identification of relevant papers for the set of documents retrieved; second, identifying the main path among those most cited papers; and third, giving a different value to each of the links in terms of bits of information associated with the transition.

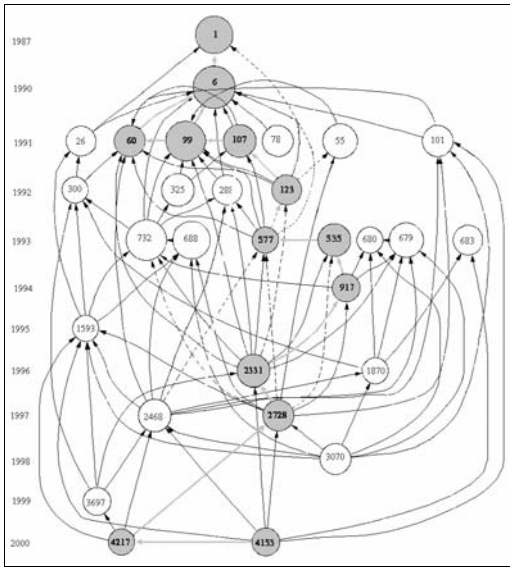


Figure 1. Citing relations among the thirty most cited documents with "fullerene(s)" in their titles, HistCite

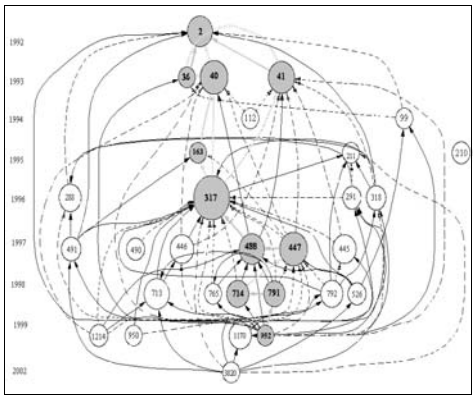


Figure 2. Citing relations among the thirty most cited documents with "nanotube(s)" in their titles, HistCite

Acknowledgements

I would like to thank Eugene Garfield, Alexander Pudovkin, and Soren Paris for all their insight into the use of HistCite™. Acknowledgments also go out to Loet Leydesdorff for valuable comments on earlier drafts.

References

Carley, K.M.; Hummon N.P. & Harty, M. (1993). Scientific Influence: An analysis of the main path structure in the Journal of Conflict Resolution. *Knowledge: Creation, Diffusion, Utilization*, 14(4): 417-447.

Chen, C. (2005). CiteSpace II: Detecting and Visualizing Emerging Trends and Transient Patterns in Scientific Literature, *Journal of the American Society for Information Science and Technology*, 57(3): 359-377.

Frenken, K. & Leydesdorff, L. (2000). Scaling Trajectories in Civil Aircraft (1913-1997), *Research Policy* 29(3): 331-348

Garfield, E. (1973). Historiographs, librarianship, and the history of science. *Essays of an Information Scientist*, 2: 134-135.

Garfield E, Pudovkin AI, Istomin VS. (2002) Algorithmic Citation-Linked Historiography–Mapping the Literature of Science, *Proceedings of the 65th Annual Meeting of the American Society for Information Science & Technology (ASIS&T)*, 39: 14-24

Hummon, N.P. & Doreian, P. 1989. Connectivity in a citation network: The development of DNA theory. *Social Networks*, 11:39-63.

Persson O (1994). The intellectual base and research fronts of JASIST 1986-1990, *Journal of the American Society for Information Science and Technology,* 45(1):31-38.

Price, DJD. De Solla (1965). Networks of scientific papers. *Science,* 149 (3683) : 510-515

Van den Besselaar, P. & Heimeriks, G. (2006). Mapping research topics using word-reference co-occurrences: a method and an exploratory case study. *Scientometrics* 68 (3): 377-393.

Information Science Network – ISN: Social Network Analysis of Scientific Production of LIS Field in Brazil

Renato Fabiano Matheus

rfmatheus@yahoo.com.br
Núcleo de Estudos em Tecnologias para Informação e Conhecimento – Netic (Brazil)
Federal University of Minas Gerais – UFMG, Information Science School (Brazil)

Introduction

This paper synthesizes previous results of the analyses of the network of co-authorship in scientific papers published by the Library and Information Science (LIS) research community in Brazil's main journals and annals of events. Data from 1991 to 2005 obtained from the Information Science Network (ISN) database – denominated RedeCI project – were analysed by means of statistical and network analysis techniques. Results show a low-density in the co-authorship network, few connections among different postgraduate programs, and a large number of sporadic authors. It is believed that this paper, and the RedeCI project, can stimulate further scientific network analysis studies, being a contribution to scientometric research. Finally, a general picture of the present reality of the Brazilian LIS research community is presented.

Methodology

Social network analysis is a methodology with a wide range of applications (Wasserman, Faust, 1994), many of which are used in the analysis of scientific production (Otte, Rousseau, 2002; Matheus, Silva, 2006). ISN ("RedeCI") (Parreiras *et al.*, 2006) is a project that was conceived to give support to network analysis of data related to scientific collaboration. In order to achieve this goal, a database and a web site were set up (http://ww.redeci.netic.com.br) to allow the storage and retrieval of information about the network of authors of papers published since 1991 in the most important journals and annals of events in the LIS field in Brazil The technological infrastructure adopted by RedeCI project is based on open software. Data about the scientific profile of the authors, such as the academic affiliation of professors, were obtained from their curriculum available at the Lattes platform (http://lattes.cnpq.br/index.htm), a governmental initiative to keep track of information about Brazilian scientific production. The selection of the most important publications was based on the classification of journals provided by the Capes evaluation system – named Qualis (http://servicos.capes.gov.br/webqualis). Data about authorship and the title of the selected papers were obtained from the journals' Web sites, from CDs of the events, and, in some cases, from a formatted file generated specifically for the RedeCI database. The resulting scientific network of co-authorship was analysed using both UCINET (Borgatti, Everett, Freeman, 2002) and LOTKA (Rousseau; Rousseau, 2000) software. ISN adopts social network metrics, including centrality and power law degree distributions. Some statistical analysis, including numeric regression, is also used.

Network database

The following table contains information about data sources, their period and total numbers of papers and authors.

Table 1. Journals and annals of events in the RedeCI co-authorship network database

Data Source*	Type (Event; Journal)	From	To	Nr. Paper	Nr. Author
CINFORM	E	2003	2005	127	195
ENANCIB	E	1994	2003	290	363
CI	J	1995	2005	337	469
DGZ	J	1999	2005	160	167
EB	J	1996	2005	72	96
I&S	J	1991	2005	219	256
PCI	J	1996	2005	156	195

Source: Parreiras *et al.*, 2006

* Data sources are the annals of events and main journals of the LIS community in Brazil, i.e.: CINFORM - *Encontro Nacional de Ensino e Pesquisa da Informação* (National Meeting of Teaching and Researching of Information); ENANCIB - *Encontro Nacional de Pesquisa em Ciência da Informação* (National Meeting of Research in Information Science); CI – *Ciência da Informação* (Information Science); DGZ – DataGramaZero; EB – Encontros Bibli; I&S – *Informação & Sociedade: Estudos* (Information and Society: studies; PCI – *Perspectivas em Ciência da Informação* (Perspectives in Information Science)

Table 2 shows the number of authors by type of publication. It is indicated if they have just one paper published (i.e., sporadic author) or more.

Table 2. Sporadic and prolific authors

Type of publication	Nr. of Sporadic Authors (A)	Nr. of Prolific Authors (B)	Relation (A/B)
Journals	615	329	1,869
Annals	239	178	1,342
Both	854	507	1,684

Source: Parreiras *et al.*, 2006

Data analysis

Considering all the data available in the Rede CI, the scientific production was analysed using LOTKA software. The results show that 72% of the papers have just one author, which indicates a low level of collaboration.

Table 3. Power Law Degree Distribution Estimate (Lotka's Law)

Parameter	Value
C	0.7233
Beta	24.017

Source: Parreiras *et al.*, 2006

Centrality measures were made using UCINET for 949 authors that have at least one paper in co-authorship. Pearson correlation shows that different types of centrality measure provide different information about the network, with indicates that they should be used carefully.

Table 4. Pearson correlation for centrality measures

Centrality Measure	Degree	Betweenness	Closeness
Degree	1.000		
Betweenness	0.356	1.000	
Closeness	0.178	0.275	1.000

Source: Parreiras *et al.*, 2006

Taking the group of authors with institutional academic affiliation, 116 teachers from 9 institutions were selected. Among them, 54 had papers in co-authorship with other teachers.

Table 5. Pearson correlation for centrality measures

Institution*	Nr. Of Professors	Nr. of Professors With Co-authorship	%
PUCCAMP	8	4	50
UFBA	12	4	33
UFF/IBICT	18	14	78
UFMG	20	11	55
UFPB	7	0	0
UFSC	11	3	27
UNB	13	7	54
UNESP	11	5	45
USP	16	6	38
Total	116	54	47

Source: Silva *et al.*, 2006a

* Institutions that have postgraduate programs in Information Science in Brazil. See more at http://www.capes.gov.br/.

Comments

The bulk of the papers that had just one author (72%). The LIS postgraduate professors consist of a small group (116) scattered in nine academic centres across the country. A small community such as this is bound to produce few papers in co-authorships either in or between institutions (Silva *et al.*, 2006a). Beyond such group, there is a larger group of sporadic authors, indicating that postgraduate students do participate in the scientific production of the field. Such results may indicate that LIS is not a mature area in Brazil. Future developments for the RedeCI project should include additional studies and developments such as: studies of different research branches (Silva *et al.*, 2006b); additional actor attributes; new importing schemes for data, as exemplified by OAI (Open Archive Initiative) protocol (http://www.openarchives.org); translation of RedeCI's web site; expansion of the network of papers to other knowledge areas; use of additional advanced network analysis techniques. This study also exemplifies an application of data about Brazilian researchers' profile made available through Lattes (http://lattes.cnpq.br/index.htm).

Acknowledgments

I want to thank Fernando Silva Parreiras, Antonio Braz de Oliveira e Silva, and Wladmir Cardoso Brandão for their partnership as coordinators of RedeCI (ISN). I thank my colleague Luiz Henrique Gomes for his careful review.

References

Borgatti, S. P., Everett, M. G., Freeman, L. C (2002). Ucinet for Windows: Software for Social Network Analysis. Harvard, MA: Analytic Technologies.

Matheus, R. F., Silva, A. B. de O. (2006). Análise de redes sociais como método para a Ciência da Informação, DataGramaZero - Revista de Ciência da Informação, 7, 2.

Otte, E., Rousseau, R. (2002). Social network analysis: a powerful strategy, also for information sciences. Journal of Information Science, Thousand Oaks, 28, 6, pp. 441-453.

Parreiras, F. S., Silva, A. B. de O., Matheus, R. F.; Brandão, W. C (2006). RedeCI: colaboração e produção científica em ciência da informação no Brasil. Perspectivas em Ciência da Informação, 11, 3, pp. .

Rousseau, B., Rousseau, R (2000). LOTKA: A program to fit a power law degree distribution to observed frequency data. Cybermetrics: International Journal of Scientometrics, Informetrics and Bibliometrics, 4, 2, pp.1-13.

Silva, A. B. de O., Matheus, R. F.; Parreiras, F. S.; Parreiras, T. S (2006a). Análise de redes sociais como metodologia de apoio para a discussão da interdisciplinaridade na ciência da informação. Ciência da Informação, 35, 1, pp. 72-93.

Silva, A. B. de O., Parreiras, F. S., Matheus, R. F., Brandão, W. C. (2006b). Redes de co-autoria dos professores da Ciência da Informação: um retrato da colaboração científica dessa disciplina no Brasi. In: *Enancib*, 7, Marília, São Paulo, November, 2006.

Wasserman, S., Faust, K (1994). Social Network Analysis: methods and applications. Cambridge: Cambridge University Press

.

The Relationship between Influence and Image: Two Views of the *Oeuvre* of Conrad Hal Waddington using Historiographic Mapping and Author Tri-Citation Image Analysis

Katherine W. McCain

kate.mccain@ischool.drexel.edu
College of Information Science & Technology, Drexel University, 3141 Chestnut Street, Philadelphia, PA 19104 (USA)

Introduction

Conrad Hal Waddington (1905-1975) was a well-known evolutionary and developmental biologist, geneticist and systems theorist whose contributions from 1929 to 1977 span the natural sciences, social sciences, and humanities. A companion paper (McCain, 2007) examines the influence of Waddington's work over the 60 years covered by the Web of Science in three 20 year segments (1945-1964, 1965-1984, 1985-2004) through historiographic mapping of his work in the context of contemporary research. Overall, Waddington's work provided the basis for a number of distinct research streams in each 20 year period, demonstrating that his influence existed both during and after his lifetime. His theories and experimental confirmation of the related concepts of canalization and genetic assimilation were important throughout the entire 60 year period while his work on embryonic induction essentially disappeared in the 1965-1984 period. This "rediscovery" of his work is associated with the emergence of Evolutionary Developmental Biology (aka Evo-Devo)—a field which stresses the importance of the developmental process on the phenotype (physical, physiological, behavioral traits) of the organism and thus on the processes of natural selection and evolution.

The research reported in this poster looks at the most recent decade—1995-2004—and compares the evidence of Waddington's influence on current research with his citation image (White 2000*).*

Methods

Influence data were obtained by an analysis of almost 1500 papers in the Web of Science citing Waddington over the period 1995-2004 together with all of Waddington's major cited *oeuvre*. Garfield's HistCite software (2004) was used to extract a citation network of 58 highly-cited books and articles receiving 15 or more citations within the network. The Newman-Girvan algorithm, available through UCINet/Netdraw, was used to identify important subnetworks that corresponded to highly visible research themes. Citation Image data were gathered via Dialog for the 47 authors most frequently co-cited with Waddington across the three ISI databases (files 34, 7,434, 439). A tri-citation approach (Marion, 2002) was used to collect data on the number of papers citing each pair of authors along with Waddington as a third cited author. Thus, although he doesn't appear in the analysis, Waddington's *oeuvre* forms the context in which the author co-citation data are analyzed. The results are displayed as a cluster-enhanced MDS map and as a PFNet based on the highest tri-citation counts between authors.

Results & Discussion

Waddington's Influence—HistCite analysis

Figure 1 shows the network of Waddington's work and highly cited contemporary papers.

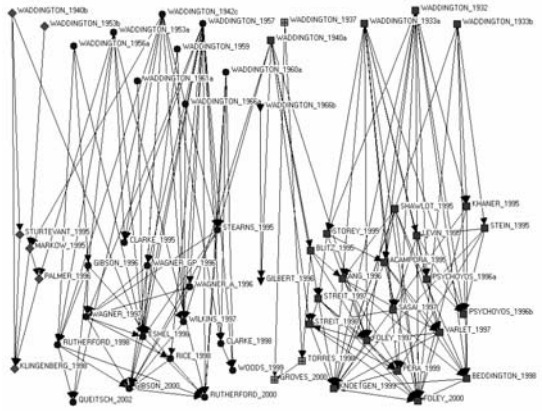


Figure 1. Influence Network: Historiograph of Waddington-based research themes, 1995-2004

The HistCite & NetDraw analysis produced one large network consisting of two large and three small subnetworks (and 9 additional Waddington works that were not connected to any contemporary highly-cited papers). The two halves of the network appear to correspond to two different topic areas associated with Waddington's major research themes–Canalization / Genetic Assimilation and Embryonic Induction. The two halves are linked only by a few citations to Waddington's *Organizers & Genes*—otherwise all links are within one subnetwork or the other. The identified subgroups include: Fluctuating Asymmetry (diamonds), Canalization/Genetic Assimilation (circles),

Evolutionary Developmental Biology (downward triangles), Auditory Placode Induction (boxes), Embryonic Induction (squares).

Waddington's Image—Tri-Citation Analysis

Figures 2 and 3 show the results of the tri-citation image analysis. Figure 2 includes the results of the cluster analysis at the 6 cluster level (a dendrogram showing the complete analysis is available from the author).

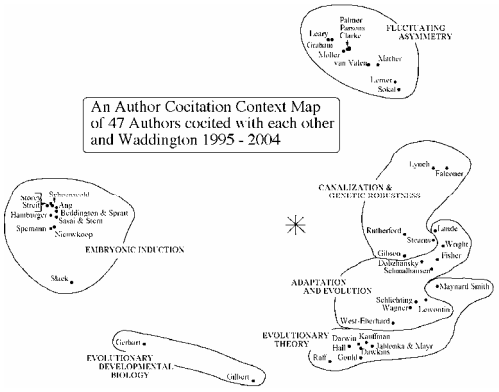


Figure 2. Waddington's Citation Image as a 2-dimensional cluster-enhanced map

The mapped author set includes authors publishing across more than a century of work – from Darwin, Spemann, Dobzhansky, Fisher and Sewall Wright to contemporary developmental biologists and evolutionary theorists. The names of the clusters are impressionistic, based on the content of the citing papers that share citations to the authors in the clusters as well as the content of the cited authors' works.The topics are certainly those discussed in the emerging literature in Evolutionary Developmental Biology. In this contextual author cocitation analysis, the Gilbert and Gerhart dyad serves as a bridge between Embryonic Induction and the interrelated set of clusters representing evolutionary theory, adaptation, and fluctuating asymmetry. Other well-known Evo-Devo authors, such as Raff, Hall, Wagner and others, are contained in clusters relating to their broader perspectives.

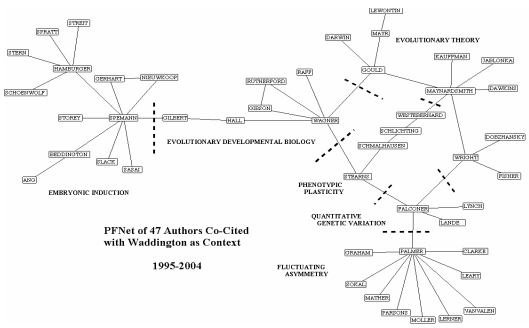


Figure 3. Waddington's Citation Image as a PFNet.

The PFNet algorithm creates a "skeletal" structure based on the highest raw tri-citation counts between author pairs. The dashed lines delimit subject associations among author subsets and are not based on the cluster analysis used in Figure 2. Authors frequently linked with others appear in the center of star patterns – for instance Hamburger and Spemann in Embryonic Induction, and Palmer in Fluctuating Asymmetry. Known Evo-Devo authors are more visible as a group in the PFNet than in the MDS map (which is based on overall profile similarity). They form a bridge between Spemann in Embryonic Induction and both Gould (linking to evolutionary theory) and Stearns (one of a group of authors with a focus on Phenotypic Plasticity).

Summary

The HistCite Influence Network demosntrates Waddington's influence and illustrates the construction of research themes (as represented by citation linkages) that draw on older work. Waddington's research and ideas have clearly retained their importance in current research. Waddington's Citation Image (map and PFNet) is made up of the authors with whom he is linked by current researchers, many of whom also have highly cited papers in the Influence Network. The Citation Image is more grounded in an historical context than the Influence Network, since it can explicitly include non-contemporary authors (generalized as their *oeuvres*) whereas HistCite necessarily focuses on "within-time-period" individual cited works (unless older work is forced into the network as in this study). In HistCite, historical authors of interest outside the time period and authors of works outside of ISI's indexing coverage are listed as cited works in the "Outer References" display and do not participate in the network that shows intellectual influence.

References

Garfield, E. (2004). Historiographic mapping of knowledge domains literature. *Journal of Information Science,* 30(2), 119-145.

Marion, L.S. (2002). A tri-citation analysis exploring the citation image of Kurt Lewin. Proceedings of the ASIST Annual Meeting, 39, 3-13.

McCain, K.W. (2007) Analysing Influence over Time: An Historiographic Mapping of the Research of Conrad Hal Waddington (1905-1975) Accepted for presentation, ISSI 2007.

White, H.D. (2000). Toward ego-centered citation analysis. . In B. Cronin & H.B. Atkins (Eds.), *The Web of Knowledge* (pp. 475-496). Medford, NJ:Information Today [ASIS Monograph Series].

Using Historiographic Mapping to Trace Persistent Highly Visible Research Themes in Medical Informatics

Katherine W. McCain and Scot Silverstein

kate.mccain@ischool.drexel.edu, scot.silverstein@ischool.drexel.edu
College of Information Science & Technology, Drexel University, 3141 Chestnut Street,
Philadelphia, PA 19104 (USA)

Overview

The research presented in this poster examines trends in the field of Medical Informatics through its published literature. We focus on the twelve years since the debut of *Journal of the American Medical Informatics Association (JAMIA*)—1994-2005. The goal is to identify topics that have persisted in the MI literature and those that are under-represented or missing, although called for in critical appraisals of the field.

Methods

We retrieved all articles citing at least one article in a set of core MI journals over the period January 1994-October 2005 in the Web of Science. The cited journals included were: *JAMIA, Computers & Biomedical Informatics, International Journal of Medical Informatics Journal of Biomedical Informatics, Methods of Information in Medicine* and *Medical and Biological Engineering and Computing*); and *Medical Decision Making.* We used Garfield's HistCite software (Garfield, 2004) and social network software (UCINet and NetDraw) to identify highly cited articles within and outside the data sets and to visualize strings and networks of articles cited within the data sets above a selected threshold (LCS, the Local Citation Score).

Results and Discussion

We found in initial analyses that the MI literature fell into two very distinct journal subsets—a general subset and a subset focused on topics predominantly covered in *Medical Decision Making*. The differences in authors and topics were sufficiently striking that the two subsets are analysed and discussed separately.

General MI

The general MI dataset contained a total of 9307 articles. At a citation threshold (Local Citation Score) of 55+ citations, HistCite produced a single, loosely-connected network of 45 articles along with 2 article pairs and 4 articles which were sufficiently well-cited to "make the cut" but which were not connected to any other highly cited MI articles. We captured the co-citation matrix for the networked and paired articles and re-analysed it using the Newman-Girvan procedure in UCINet/NetDraw. This procedure identifies coherent subnetworks based on the network structure; these subnetworks are highly congruent with the topics of articles as represented in titles and abstracts. Figure 1 shows the network and dyads, shape-coded by subnetwork.

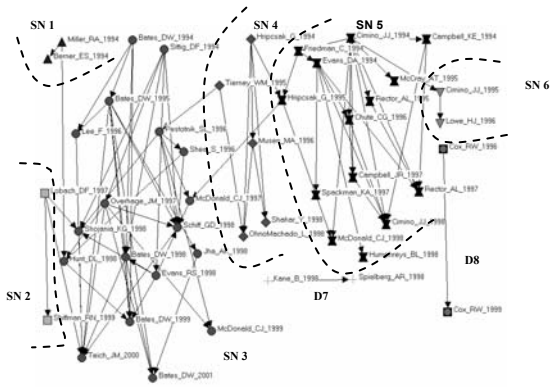


Figure 1. Research topics in general MI

The topics represented, from left to right, are:
SN1: Medical Diagnostic Systems
SN2: DSS Evaluation
SN3: CPOE/Medical Records/Error Prevention
SN4: Clinical Guidelines Computerization
SN5: Clinical Classifications & Vocabularies
SN6: Internet/WWW
D7: E-mail/Physician-Patient Communication
D8: MRI Imaging

Medical Decision Making

The dataset citing articles in *Medical Decision Making* contained 4737 articles. At a LCS threshold of 40 citations, HistCite extracted 51 highly cited articles comprising 2 large and 4 small article networks and 11 unconnected but highly cited articles. Figure 2 shows the networks, shape-coded by topic.

The topics represented, from left to right, are:
SN1: End of Life Care Preferences
SN2: Cost-Effectiveness Analysis Issues
SN3: Patient Willingness to Pay
N4: Patient Involvement in Clinical Decision Making
N5: Basics of Medical Decision Analysis
N6: Patient Understanding of Preventive Care
N7: unclear
N8: Uncertainty in Clinical Care

Both the general MI and MDM networks included a mix of articles from MI and general medical journals. *JAMIA* was represented in all but the last topic in Figure 1 while, by design, *MDM* was ubiquitous in the Figure 2 topics.

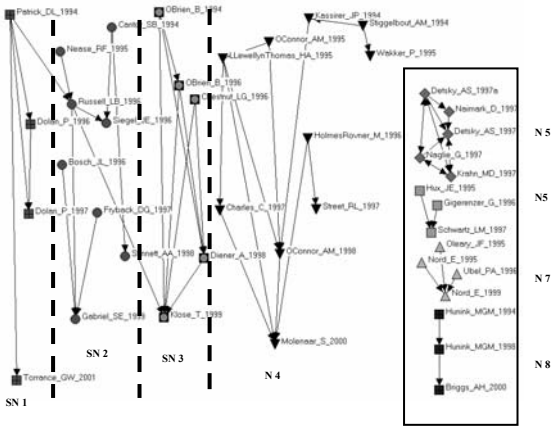


Figure 2. Research topics in medical decision making

Conclusions

Two groups of topics predominate. Many of these topics are of high visibility in both the clinical and consumer literature—CPOE (Computerized Practitioner Order Entry), medical error prevention and cost issues are examples. Other topics such as clinical classifications and vocabularies and medical decision analysis are of significant interest to both MI professionals and clinicians. Some themes that, in our experience, are of growing interest to healthcare informaticians were not visible. These include ethics and legal issues in medical informatics, information technology & health professions education, and telemedicine and social informatics, most especially, sociotechnical issues related to clinical information technology (seehttp://www.ischool.drexel.edu/faculty/ssilverstein/medinfo.htm).

The absence of these topics represents a gap in research and authorship in areas important to the success of health informatics initiatives, especially national EMR initiatives such as have been initiated in the US and the UK supporting culture change and more effective, evidence-based practice. This suggests that medical informatics research and training, while facilitating the improvement of healthcare, may itself benefit from an evaluation of these themes. (see Friedman et al. JAMIA 2004; 11: 167-172).

References

Friedman, CP et al. (2004) Training the next generation of informaticians: The impact of BISTI and bioinformatics—A report from the American College of Medical Informatics. *Journal of the American Medical Informatics Association* 11:167-172.

Garfield, E. (2004). Historiographic mapping of knowledge domains literature. *Journal of Information Science*, 30(2), 119-145.

Silverstein, S. (no date) Sociotechnologic issues in clinical computing: Common examples of healthcare IT failure. Available at: http://www.ischool.drexel.edu/faculty/ssilverstein/medinfo.htm. Last accessed on January 30, 200

Institutional Health Research Collaboration in Mexico. A Bibliometric Study

C.A Macías-Chapula, J.A. Mendoza-Guerrero, I.P Rodea-Castro and A. Gutiérrez-Carrasco

chapula@data.net.mx
Hospital General de México, Research Unit.
Algeciras 43 A Col. Insurgentes Mixcoac. Delegación Benito Juárez 03920 México D.F (Mexico)

Introduction

Health is an essential factor and consequence of development. Health research is indispensable for improving health and health equity and contributing to overall development (Nuyens, 2005; Sadana, et al, 2004).

In Latin America and the Caribbean (LAC) the State is the principal source of financing health research, contributing more than two-thirds of the resources. This is the reverse of the situation in the United States, and Canada where some two-thirds of the funding comes from the private sector (Pellegrini, 2000).

International collaboration in science has been an important issue of discussion among scientists. It has been shown that this type of collaboration provides among other benefits, better visibility and major citations to those publications derived from this type of communication (Lucas, 2005; Wagner, et al, 2001; Narvaez-Berthelemont, 1994).

Recently, Sancho, et al, (2006) have published indicators of LAC collaboration in science. There remain however some research lines to be explored; for example, studies regarding the behavior of specific disciplines and institutions at the micro-level of analysis.

Purpose

The purpose of this work is to present the results derived from the analysis of the collaboration indicators of ten leading Mexican hospitals and national health research institutes. The final goal of our study is to pilot our approach in Mexico and to extend the model to other regional countries.

Method

This work was based on a previous analysis of the scientific output of health research institutions in Mexico. In that study, over 8000 Mexican affiliations were identified for the period 1999-2003, in MEDLINE.

Our current study included only those institutions conducting mainly clinical medicine research, such as hospitals, centers and institutes.

A selection of the ten (out of 500) leading institutions in terms of scientific production in clinical medicine was conducted in order to identify the patterns of collaboration of such institutions. The Web of Science (WOS) was used to identify the affiliations of participating institutions; and the study was limited to the period 1990-2005. Our thesaurus of Mexican health research institutions was used to identify all the permutations to an institution. (Macías-Chapula, et al, 2006; 2007).

For each selected institution, the following data was obtained:

-Number of publications produced in the period 1990-2005.

- Number of authors and co-authors to a paper. This was classified in one-author per paper; 2-5 authors per paper; 6-10 authors per paper; and 11 or more authors per paper.

-When papers were published by two or more authors, the collaboration was classified as follows: (a) intra-institutional, that is, all authors belong to the same institution; (b) national, that is, the inter-institutional collaboration within Mexico; (c) regional, that collaboration within Latin America and Caribbean countries; and (d) international, when the collaboration belongs to a country other than the American continent. The collaboration with USA and Canada was excluded from the study since this type of collaboration is being analysed separately.

-Number of collaborating countries and institutions.

Microsoft Excel 2000; Microsoft Access 2000; and Bibexcel 2001 were used in the different processes involved in the validation and organization of data. Results were tabulated, analyzed and interpreted accordingly**.**

Results

A total of 73,110 records were retrieved for Mexico in the Web of Science for the period 1990-2005. From this amount, only 6,649 (9.09%) records corresponded to the ten leading research institutions selected in this study. Table 1., provides a list of these institutions with their corresponding amount of records. Table 2., provides a distribution of co-authorship by institutions.

Table 1. Leading health research institutions in Mexico in the field of clinical medicine (WoS)1990-2005)

Rank	Institution	No. of Records	%
1	National Institute of Nutrition, Salvador Zubiran	1,995	30.00
2	Century XXI National Medical Centre	767	11.54
3	National Institute of Cardiology, Ignacio Chavez	740	11.1
4	National Medical Centre La Raza	481	7.23
5	Mexico´s General Hospital	478	7.19
6	National Institute of Pediatrics	488	7.34
7	National Institute of Neurology and Neurosurgery, MVS	462	6.95
8	National Institute of Respiratory Diseases	447	6.72
9	National Institute of Cancer	396	5.96
10	National Institute of Psychiatry, Juan Ramon de la Fuente	395	5.94
	Total	6,649	100

Table 2. Distribution of coauthorship by institutions.

Institution / AU	1	2 to 5	6 to 10	11 +	Total
National Institute of Nutrition, Salvador Zubiran	133	938	828	96	1,995
Century XXI National Medical Centre	20	388	337	22	767
National Institute of Cardiology Ignacio Chavez	19	343	350	28	740
National Institute of Pediatrics	32	237	204	15	488
National Medical Centre La Raza	6	231	221	23	481
Mexico´s General Hospital	14	240	205	19	478
National Institute of Neurology and Neurosurgery, MVS	22	296	126	18	462
National Institute of Respiratory Diseases	14	200	228	5	447
National Institute of Cancer	7	178	188	23	396
National Institute of Psychiatry Juan Ramon de la Fuente	21	289	78	7	395
Total Authors	**288**	**3,340**	**2,765**	**256**	**6,649**
%	**4.33**	**50.26**	**41.60**	**3.85**	**100**

Over 90% of the production in each institution was published by the collaboration of two or more authors. 93.10% of the collaboration was national (54.30% intra-institutional and 38.80% inter-institutional); 1.68% was regional (within the LAC); and 5.22% was international (outside the LAC regions). Table 3., describes the type of collaboration found.

Table 3. Distribution of type of collaboration by institutions.

Intitution	Intra-inst	National	Regional	International	Total
National Institute of Nutrition, Salvador Zubiran	2,275	1,194	61	250	3,780
Century XXI National Medical Centre	742	723	15	56	1,536
National Institute of Cardiology Ignacio Chavez	849	510	60	85	1,504
Mexico's General Hospital	500	618	25	56	1,199
National Medical Centre La Raza	562	463	14	67	1,106
National Institute of Pediatrics	557	381	16	15	969
National Institute of Respiratory Diseases	431	471	10	51	963
National Institute of Neurology and Neurosurgery, MVS	596	306	17	35	954
National Institute of Cancer	471	359	7	68	905
National Institute of Psychiatry Juan Ramon de la Fuente	404	253	4	27	688
Total	7,387	5,278	229	710	13,604
%	54.30	38.80	1.68	5.22	100.00

Over half of the collaboration was reported as intra-institutional by all institutions; with the exception of Mexico's General Hospital with 500 collaborations; and the National Institute of Respiratory Diseases with 431.

Overall, regional collaboration was identified with 17 countries and international collaboration with 45. Main regional countries were led by Venezuela (29.82%); Brazil (15.35%); Argentina (10.96%) and Cuba (9.21%). On the other hand, International collaboration was led by Spain (21.69%); France (16.06%); and England (11.83%).

Conclusion and discussion

The results of our study are providing light into the positioning of national health research institutions; their collaboration patterns and benchmarking in their field of interest. Further results will provide data on the behaviour over time of the institutions analysed; the type of sub-disciplines found and more important, the institutions within each country that collaborate with the Mexican health research institutions.

We can conclude that while Mexican health research institutions maintain a high pattern of collaboration, this is mainly intra-institutional and national. Further studies need to be conducted in order to complement our results with that of the behaviour of universities and high technology institutes of research.

References

Lucas, A. O. (2005). International collaboration in health research. *Bull World Health Organ,* 83, 482.

Macías-Chapula, C.A., Mendoza-Guerrero, J.A., Rodea-Castro, I.P., Gutiérrez-Carrasco, A., Juárez-Sánchez, E. (2006). Normalización de Instituciones mexicanas en ciencias de la salud. Hacia la construcción de un catálogo de autoridad. En *III Seminario Internacional sobre Estudios Cuantitativos y Cualitativos de la Ciencia y la Tecnología*. "Prof. Gilberto Sotolongo Aguilar" La Habana, Cuba.

Macías-Chapula, C.A., Mendoza-Guerrero, J.A., Rodea-Castro, I.P., Gutiérrez-Carrasco, A. (2007). Health research collaboration in Mexico. Paper presented in *The International Workshop on Webometrics, Informetrics and Scientometrics* & Eighth COLLNET Meeting, 6-8 March, New Delhi, India.

Narváez-Berthelemont, N. (1994). An index to measure the international collaboration of developing countries basedon the participation of national institutions: The case of Latin America *Scientometrics,* 34**,** 37-44.

Nuyens Y. (2005) No development without research: Cahallenge for capacity strengthening. Geneva: Global Forum for Health Research. Retrieved November 20, 2006 from: http://www.globalforumhealth.org/filesupld/RCS/RCS_Nuyens.pdf

Pellegrini F. A. (2000). Ciencia en pro de la salud: notas sobre la organización de la actividad científica para el desarrollo de la salud en América Latina y el Caribe. *Rev Panam Salud Publica,* l7, 345-349.

Sadana, R., D'Souza, C., Hyder. A.A., Chowdhury, A.M.R. (2004). Importance of health research in South Asia. *British Medical Journal*, 328, 826-30.

Sancho, R., Morillo, F., Fillippo, D., Gómez, I., Fernández, M.T. (2006). Indicadores de colaboración científica Inter-Centros en los países de América Latina. *Interciencia,* 31, 284-292.

Wagner, C.S., Brahmakulam, I., Jackson, A., Wong, A., Yoda, T. (2001). *Science and technology collaboration: building capacity in developing countries?* Santa Monica, CA: RAND Science and Technology

Characterization of research groups in the cardio-cerebrovascular field. Spain 1996-2004.

Raúl I. Méndez-Vásquez[*], Eduard Suñén-Pinyol[**], Ginés Sanz[***], Jordi Camí[**,****]

*rmendez@prbb.org
Agència d'Avaluació de Tecnología i Recerca Mèdica, Barcelona (Spain).

**esunyen@prbb.org
Parc de Recerca Biomèdica de Barcelona, Barcelona (Spain).

***gsanz@cnic.es
Centro Nacional de Investigaciones Cardiovasculares, Madrid (Spain).

****jcami@prbb.org
Universitat Pompeu Fabra, Barcelona (Spain)

Introduction

A bibliometric report of research groups in the cardio-cerebrovascular field is presented. Research groups were identified by co-author analysis as used by others authors (Calero 2006, Bordons 1995, Newman 2000, Peters 1991). This methodology groups authors' bibliographical names based on their frequency of co-authorship, which makes purging of synonyms and homonyms names essential. Co-author analysis was performed using an in-house-developed algorithm (Suñen 2001) optimized in previous studies (Camí 2003, Noguer 2006). To identify groups, the algorithm builds co-authorship networks and filters irrelevant connexions. The main characteristic of the algorithm is that it allows setting minimal values to a set of four parameters that control the identification of research groups. These parameters are: number of authors per group (A), number of published documents per author (B), number of co-authored documents (C), and number of co-authorships per author (D). The aim of the present study was to identify and characterize the research groups working in the cardio-cerebrovascular field by using a set of MeSH terms (NIH, USA) and the co-author analysis algorithm.

Material and methods

The National Citation Reports for Spain (NCR) a product from Thomson-ISI, was the source of citable documents (articles, *reviews* and *proceedings*) published from Spanish centres during the period 1996-2004. Documents having in their title and/or keywords any of the 3000 different MeSH terms approved by an expert cardiologist were selected for the study (study collection). The MesH terms were classified in 23 areas in order to rank research groups by topic. Twenty one terms and wildcards were used to exclude documents from nephrology, hepatology, neurotransmitters and cancer. All documents from authors having less than half of their publications in the study collection were excluded from the study. Research groups were identified by the in-house-developed algorithm calibrated as follow: A=4, B=4, C=3 and D=3. Identification and examination of the resulting groups was carried out repeatedly until approval from the expert. The algorithm was executed six times. Bibliometric analysis was performed on the study collection as well as on the set of documents published by every member of the groups previously identified (group collection). Research groups were assigned to the most frequent autonomous region (CA), institutional sector, affiliation centre and topic appearing in the group collection. Bibliometric indicators obtained for these aggregation levels were: number of citable documents (Docs), number of citations (Cit), average citations per document (CD, Cit/Docs), cardio-cerebrovascular Spanish citation rate (MCCE, the average of the yearly quotients between the CD of an area and the CD of the study collection), percentage of documents not cited in the period (%NC), percentage of documents co-published with foreign authors (%Int).

Results

Seventy nine research groups including 691 authors affiliated to Spanish centres and 301 affiliated to foreign centres were identified. The group collection (3144 documents) accounts for 44.9% of the study collection, a percentage that showed a wide variation according to CA (see Table 1). The group collection showed CD and MCCE values greater than the study collection, and lower %NC and %int values. Three quarters of the research groups belonged to the health sector (75.8%) and a third (27.8%) to the university sector. Two thirds of the total number of groups was located in the CA of Cataluña and Madrid. Research groups belonging to these CA showed the highest CD, MCCE and %int values (see Table 1). Also, the CA of Navarra showed MCCE values higher than the average. The largest groups were located in Murcia. Although the study collection includes documents from the CA of Aragon, Cantabria, Canarias, Castilla-La Mancha,

Extremadura, Illes Balears and La Rioja, no research groups were assign to these CA. Research groups were assigned to 46 research centres and hospitals, which represent 44,2% of the total number of centres publishing in the cardio-cerebrovascular field. Hospital Clínic, Hospital Vall d'Hebron and Hospital Sant Pau, all from Barcelona, showed the highest number of research groups in all research centres included in the study. Hospital Clinic and Hospital Sant Pau from Barcelona and Hospital San Carlos from Madrid showed the highest CD, MCCE and %int Values. Most of the research groups were assigned to five out the 23 topics: clinic cardiology (49.4%), coagulation-platelets and thrombosis (20.3%), diagnosis technique (24.1%), ischemic cardiopathy (16.5%) and arterial hypertension (12.7%). Vascular surgery showed the highest CD, MCCE and %int values.

Discussion

Research groups in the cardio-cerebrovascular field are located mainly in Cataluña and Madrid, and in the clinic cardiology topic. The observed asymmetric distribution of research groups in the cardio-cerebrovascular field is identical to that observed in biomedicine in terms of production (Camí 2005). The opinion from the expert ensured the proper selection of MeSH terms and the identification of the authors whose main research area was indeed the cardio-cerebrovascular field. However, the difference in the number of documents of the study and group collections for a given CA, topic or research centres, indicates that there are some research groups that publish in topics related to the cardio-cerebrovascular field, but not as their main activity area.

Table 1. Bibliometric indicators for the group and study collections and by CA. Spain 1996-2004. (only CA with assigned research groups are shown)

	Group collection										Study collection					
Autonomous Community	**Grp[1]**	**%tot[2]**	**Memb[3]**	**Docs[4]**	**%cov[5]**	**Cit[6]**	**C/D[7]**	**MCCE[8]**	**% NC[9]**	**% Int[10]**	**Docs[4]**	**Cit[6]**	**C/D[7]**	**MCCE[8]**	**% NC[9]**	**% Int[10]**
CA de Cataluña	29	36,7	9,2	1.390	61	16.621	12,0	1,17	22,7	23,2	2.266	24.165	10,7	1,10	25,4	26,4
C de Madrid	21	26,6	9,7	938	44	11.208	12,0	1,27	26,2	19,0	2.150	22.720	10,6	1,09	29,3	23,9
CA de Andalucia	7	8,9	5,7	157	20	825	5,3	0,55	30,6	7,0	772	4.626	6,0	0,65	33,8	16,1
C Valencia	9	11,4	9,3	290	40	1.573	5,4	0,69	33,1	12,1	734	4.358	5,9	0,74	34,1	21,1
CA de Galicia	3	3,8	10,3	174	42	801	4,6	0,73	25,9	10,3	415	2.730	6,6	0,84	28,4	16,9
CA de Castilla y León	3	3,8	7,7	65	24	310	4,8	0,48	30,8	9,2	270	2.048	7,6	0,65	33,7	11,9
C Foral de Navarra	4	5,1	7,8	138	52	1.268	9,2	1,04	22,5	12,3	266	1.795	6,8	0,92	28,2	13,5
CA de la R de Murcia	2	2,5	11,0	120	52	698	5,8	0,62	27,5	4,2	231	1.316	5,7	0,55	32,5	12,6
CA del Pais Vasco	1	1,3	4,0	37	17	180	4,9	0,48	32,4	0,0	216	1.745	8,1	0,84	31	15,3
CA del P de Asturias	1	1,3	5,0	28	25	218	7,8	0,91	17,9	3,6	111	954	8,6	0,77	29,7	14,4
TOTAL	**79**		**8,8**	**3.144**	**45**	**31.701**	**10,1**	**1,05**	**25,8**	**18,1**	**6.999**	**59.162**	**8,5**	**0,91**	**30,2**	**22,2**

1, number of groups; 2, percentage to the total number of groups; 3, average number of member per group; 4, number of citable documents; 5, percentage of documents to the total number of documents; 6, number of citation; 7, average citations per document; 8, cardio-cerebrovascular Spanish citation rate; 9, percentage of documents not cited in the period; 10, percentage of documents co-published with foreign authors.

Reference

Bordons M., Zulueta M. A., Cabrero A., Barrigon S., Koenig Michael E.D., Bookstein Abraham. (1995). Identifying research teams with bibliometric tools. Proceedings of the fifth biennial international conference of the International Society for Scientometrics and Infometrics (pp 83-92).

Clara Calero, Renald Buter, Cecilia Cabello Valdés, Ed Noyons. (2006). How to identify research groups using publication analysis: an example in the field of nanotechnology. Scientometrics, 66(2);365-376. DOI: 10.1007/s11192-006-0026-z

Camí J, Suñén-Piñol E, Mendez-Vasquez RI. (2005) Mapa bibliométrico de España 1994-2002: Biomedicina y Ciencias de la Salud. Med Clin (Barc);124:93-101. Available at: http://www.prbb.org/bac/PCE1994-02/MEDCLIN-2005.pdf

Camí J, Suñén E, Méndez-Vásquez RI. (2003). Caracterización bibliométrica de grupos de investigación biomédica en España (1994-2000). Barcelona. Available at: http://bibliometria.prbb.org/MapaBiomedicoGrupos.

Mark E.J. Newman. (2000). Who Is the Best Connected Scientist? A Study of Scientific Coauthorship Networks. Phys.Rev (pp 1-17). Retrieve October 15, 2004 from: http://arxiv.org/abs/cond-mat/0011144/.

Noguer Carmona S, Gallo de Puelles P, Parada Martínez A, García-Duran Huet M, Mendez-Vasquez RI, Suñen Pinyol E. (2006). Estudio bibliométrico sobre esquizofrenia en España (1999-2004): Productividad, impacto y colaboración (2ª Edición). Barcelona. Working Paper Series Senyfundació. Barcelona 2006.

Peters, H.P.F., Van Raan, A.F. J. (1991). Structuring scientific activities by co-author analysis. Scientometrics, 20(1); 235-255. DOI: 10.1007/BF02018157.

Suñen Pinyol E. Algorithms for automatic research group detection. (2001). Facultat d'Informàtica de Barcelona, Universitat Politècnica de Catalunya. Barcelona. Available at: http://www.prbb.org/bac/sunen/Memo.pdf.

Full report available at: http://www.cnic.es/estudio.html

Acknowledgements: This study was supported in part by the Centro Nacional de Investigaciones Cardiovasculares.

Analysis of the Interdisciplinar Collaboration in Spanish Scientific Production on Radiology and Diagnostic Imaging

Alberto Miguel-Dasit*, Gregorio González-Alcaide**, Juan Carlos Valderrama-Zurián**, Adolfo Alonso-Arroyo**, Rafael Aleixandre-Benavent**

*amdasit@hotmail.com,
MR Section, La Plana de Vila-Real Hospital, Castellón (Spain)

**gregorio.gonzalez@uv.es, juan.valderrama@uv.es, adolfo.alonso@uv.es, rafael.aleixandre@uv.es
Instituto de Historia de la Ciencia y Documentación López Piñero. Universitat de València-CSIC, Valencia (Spain)

Introduction

Radiology is a medical discipline related to many other medical disciplines (Miguel-Dasit et al, 2004). On the other hand, the scientific collaboration is a critical point for the progress and advance of science. Nevertheless, the analysis of the collaboration between the different medical disciplines has received little attention in the literature. A possible reason for this is the difficulty of determining the diverse typologies of collaboration (Klein, 2004). The objective of the present study is to estimate the possible multidisciplinarity of the Spanish radiology science, by means of an analysis of collaboration patterns in scientific publications written in collaboration between Spanish radiology investigators and researchers from other medical disciplines.

Material and methods

a) Study search: specific search profiles were designed to recover published papers between 2001 and 2005 in journals included in the "Radiology, Nuclear Medicine & Medical Imaging" section of the Journal Citation Reports (JCR) database. All the published papers in JCR-indexed journals that were not original research articles were excluded from this study. Also, articles submitted by Spanish institutions were recovered including the term "Spain" in the "Address" field of the JCR.

b) Identification of medical disciplines and normalization methods: medical specialties were determined by analyzing the institutional signatures of papers, because the medical disciplines include both the "macroinstitutions" (such as universities or hospitals) as well as the different departments, services and units according to the medical specialties of the authors of papers. The normalization of the possible different variants by which a same medical specialty can be denominated was performed according to the list of specialties that is included in the thesaurus Medical Subject Headings (MeSH), elaborated by the National Library of Medicine, in Bethesda, United States.

c) Descriptive and quantitative analysis: for each one of the identified specialties, both the number of total papers by specialty and the total number of papers published in collaboration with other specialties were estimated. When all the collaborations between specialties were identified, a collaboration's network regarding the different disciplines was constructed. All the items were introduced in a Microsoft Access database to facilitate the analysis. Visual representations was obtained by using the specific software "Networks-Pajek" (available at http://vlado.fmf.uni-lj.si/pub/networks/pajek).

Results and Discussion

A total of 1,030 documents were identified, which includes 2,914 institutional signatures. In 967 papers (94%) and in 2,232 signatures (76,6%), a collaboration between investigators from at least two different specialties was identified. A total of 72 different specialties were identified, the Radiology specialty having the highest number of papers (n=478), followed by Physics and Radiotherapy (n=132, both) (table 1).

The most intense collaborations between disciplines were established with Radiology, in the following order: Histology (68 papers in collaboration), Surgery (n=51), Pediatrics (n=38), Biophysics (n=25), Neurology (n=23), Gynecology (n=22), Gastroenterology (20), Radiotherapy (n=20) and Psychiatry (n=20). Of interest, we found other four disciplines not related to the Radiology which showed a high number of collaborations with other disciplines or specialties, which were: Nuclear Medicine, Engineering, Oncology and Physics. Figure 1 shows the specialties network related to the disciplines with 5 or more published papers in collaboration.

Table 1. Productivity and collaboration according to specialty (2001-2005)

Specialty	No. of papers	No. of specialties with papers in coll*	No. of coll*	Mean coll*/ paper
Radiology	478	60	525	1,1
Physics	132	28	154	1,17
Radiotherapy	132	41	214	1,62
Nuclear Medicine	94	37	130	1,38
Surgery	86	28	131	1,52
Histology	85	29	167	1,96
Gynecology	67	14	59	0,88
Biophysics	62	24	121	1,95
Oncology	50	29	96	1,92
Engineering	50	32	79	1,58

*coll: Collaboration

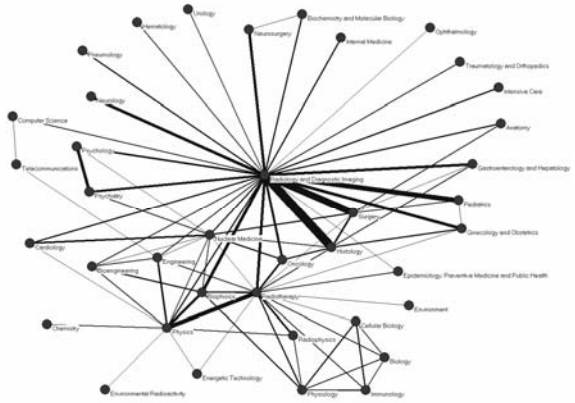


Fig. 1. Collaboration network according to specialty (2001-2005) (n >=5 collaborations)

Conclusion

On the basis of our method of quantifying contributions from a variety of disciplines to the Spanish radiology science, the results of this study indicate that the multidisciplinarity of this field has been established. A limitation of this study was the consideration of the institutional affiliations as the main indicator for delineating specialties. Therefore, future research on interdisciplinary fields must include more sophisticated methods for field delineation, such as cocitation or co-word analysis (Small & Griffith, 1974; Noyons & Van Raan, 1998; McCain KW, 1998).

Acknowledgements

This work has been supported by the I3P Program, cofinanced by the European Social Foundation of the CSIC (Consejo Superior de Investigaciones Científicas).

References

Klein, J.T. (2004). Interdisciplinarity and complexity: an evolving relationship. E:CO Special Double Issue, Vol. 6 Nos. 1-2, pp. 2-10.

McCain K.W. (1998). Neural networks research in context: A longitudinal journal cocitation analysis of an emerging interdisciplinary field. Scientometrics 41: 389-410.

Miguel-Dasit A., Marti-Bonmatí L., Aleixandre R., Sanfeliu P. & Valderrama J.C. (2004). Spanish production of research articles on diagnostic imaging in Cardiology and Radiology (1994-1998). Revista Española de Cardiología 57(9):806-814.

Noyons .E.C.M. & Van Raan A.F.J. (1998). Monitoring scientific developments from a dynamic perspective: Self-organized structuring to map neural network research. Journal of the American Society for Information Science 49: 68-81.

Small H. & Griffith B.C. (1974). The Structure of Scientific Literatures I: Identifying and Graphing Specialities. Science Studies 4: 17-40.

The Role of Madrid in the Spanish Regional Collaboration

Fernanda Morillo and Daniela de Filippo

* *fmorillo@cindoc.csic.es*
CINDOC, CSIC (Spanish Research Council), Madrid (Spain)

Introduction

An important characteristic of the scientific task is the increasing trend to work in group. The motivation to cooperate is influenced by the goals of those providing the funding, the need to access knowledge or equipment, or the opportunities of researchers to work together (e.g., geographic proximity). Indeed, the Spanish National R&D Plan 2004-2007 (2003) has an effect on the promotion of collaboration in some areas. Madrid has an important role as a centre of attraction. Other regions of Spain are taking advantage through collaboration, whereas the evolution of Madrid share in Spanish output is decreasing throughout time in a similar way as the share of USA in the world output (Leydesdorff and Wagner, 2006). This could be due to the creation of new centres and universities in other regions and it seems that there is a tendency to a more homogeneous distribution of the Spanish scientific output.

The first question that arises is related to the institutional sectors responsible for this homogeneous distribution, that is to say, which sector shows the most outstanding increase in other regions during the studied period. The hypothesis is that the University will be responsible of the decentralized trend towards other regions because of its high output. This first question is linked to others: is this increment related to a more collaborative characteristic? Which part of this collaboration is related with Madrid? Finally, if the hypothesis turns to be right, those universities that collaborate and the links between them are studied. Also, to better understand those links, the connections considering the area of research are analysed.

Methods

Spanish output is analysed through international Thomson Scientific Web of Science (WoS) databases (SCI, SSCI y A&HCI) in the period 2001-2005. From this dataset Madrid and other regions' output are identified to compare them with the behaviour of Spain as a whole. As it is detected that Madrid loses weight in the Spanish output during the period, other regions' output is also studied using total assignment of documents. Salton formula is applied to normalize relations between institutional sectors or universities. UCINET and NETDRAW software were used to draw results. To check if University is responsible for decentralization, the evolution of the number of publications by sector and region is studied. Sectors considered are: University, CSIC (Spanish Research Council), Joint CSIC centres (mainly with University and Public Administration), Public Administration, Other Research Organizations, Non-profit institutions, Companies, and Health Sector. To answer the question related with the collaborative characteristics, the inter-sectoral collaboration was studied, as well as in which part of it Madrid participates. Finally, the connections between universities in all areas and in some selected areas are studied.

Results

Although Madrid is losing part of the Spanish share, it shows an increasing trend in the national collaboration rates. Madrid is highly connected with Cataluña and in a lesser extent with Andalucía and C.Valenciana. When analysing the differences between 2001 and 2005 concerning the shares of Spanish output, Madrid loses more than 1 percent and Cataluña and C.Valenciana gain 2 and 1 points of percentage respectively. Part of these increments is explained by collaboration practices mainly with Madrid.

Decentralization

Although in total number of documents the higher contribution comes from the University sector, considering percentages, the greatest increments are observed in Non-profit Institutions (most of them coming from combined efforts of successful institutions like University and Health Sector). They are followed by Companies and Joint CSIC centres. Considering regions with greater impact in the decentralization trend of Madrid, similar results are observed but, in Cataluña, Joint CSIC centres have less importance in the overall situation and Non-profit Institutions have proportionally more importance, as they rank fourth by total output (mainly because of very big public medical institutions). For C.Valenciana the greatest difference is produced by Health-Sector and CSIC, while Non-profit Institutions and Companies have very little importance.

Collaboration between sectors

University is the most productive and most collaborative sector. Its stronger relation is with the Health Sector, followed by CSIC and Joint CSIC

centres. The Health Sector (the second one in number of documents) is linked first with University and secondly with Non-profit Institutions. The latter are connected also with University (in the first position). Normalizing sector links we can see that the stronger ties are University-CSIC and Health Sector-Non-profit Institutions.

Collaboration with Madrid
University and Health Sector are those with the largest collaboration rates with Madrid, as expected. Considering the increment of the collaboration throughout time, the greatest increase is performed by Companies, although their output is scarce. They are followed by Non-profit Institutions and Health Sector. The data are similar to those of the decentralization study of publications, that is to say: some of the output from peripheral regions is enlarged through their collaboration with Madrid.

Universities
Considering relations between Universities, their situation in the map of Spain is very important. For more than 100 documents shared, intra-regional collaborations can be found. If links of 50 or less documents are considered, Madrid comes back to be the centre of Spain and, its Complutense University and other big important Universities show links with Galicia, Valencia, etc. To get a clearer view, this information was normalized and more details of smaller centres were obtained, like private Universities and those coming from small regions.

Universities per area
To check for thematic variations, some areas were selected: those with high national collaboration rates (e.g. Clinical Medicine) and those with low national collaboration rates (e.g. Mathematics). All the links have been normalized to enlighten important nodes.

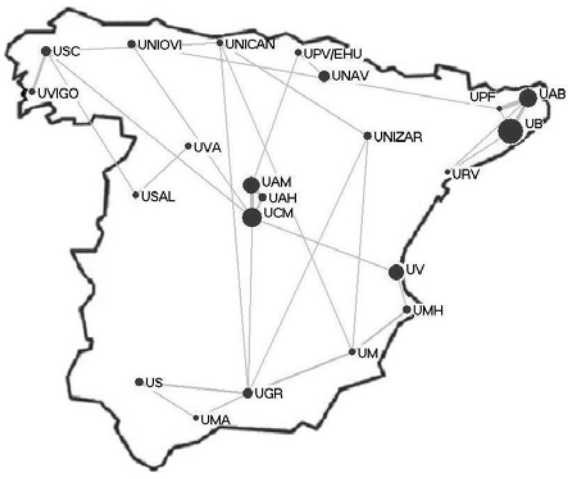


Figure 1. Normalized collaboration links between Spanish Universities in Clinical Medicine (> 250 documents and > 0.015 Salton index).

In Clinical Medicine, the stronger links are inside Cataluña region, followed by those created within Madrid. These can be explained by their size or their geographic situation. On the other hand, in Mathematics, Technical Universities establish connections between them mainly because of their thematic interests.

Conclusions
Although Madrid is losing gradually the central role to give place to other regions, it does play an important part in collaboration practices in Spain. This collaboration benefits the involvement of less developed regions. The first position is occupied by the University sector but other institutions coming from the combination of different sectors have an increasing importance. The significance of University in collaboration practices was also observed by Yamashita and Okubo (2006). In their work the main participants of Japan in collaboration practices are universities. On the other hand, in France, universities and public research institutions have the same importance, mainly due to joint laboratories. The case of Spain is very similar to that of France. Increasingly, all over the world, joint research centres are being established to address complex research problems (Hara et al., 2003).

Looking closely to Universities, the most important links are coming from the centre, but the geographic situation is more important when the number of documents shared increases. In general, important connections can be explained by the size of University (e.g. large Universities) or the thematic interest (e.g. technical Universities). This step by step method can help to understand the complex relations based on geographic proximity or thematic interests. To continue this research, a deeper insight on collaboration patterns and their influence in general output will be obtained, as well as impact and visibility.

References
Hara, N., Solomon, P., Seung-Lye K. & Sonnenwald, D. H. (2003) An Emerging View of Scientific Collaboration: Scientists' Perspectives on Collaboration and Factors that Impact Collaboration. *Journal of the American Society for Information Science*, 54 (10): 952–965.

Leydesdorff, L. & Wagner, C. (2006) Is the United States losing ground in science? A global perspective on the world science system. *<http://users.fmg.uva.nl/lleydesdorff/US_Science/US_Science.pdf>*.

Yamashita, Y. & Okubo, Y. (2006) Patterns of scientific collaboration between Japan and France: Inter-sectoral analysis using Probabilistic Partnership Index (PPI). *Scientometrics*, 68 (2): 303–324.

Ways of Adequacy for Evaluation of Brazilian Scientific Production: National Impact versus International Impact

Rogério Mugnaini[*], Elías Sanz-Casado[**] and Carlos García-Zorita[**]

[*]Latin-American and Caribbean Center on Health Sciences Information (BIREME/PAHO/WHO) (Brazil)

[**]*elias@bib.uc3m.es*
Universidad Carlos III de Madrid. Departamento de Biblioteconomía y Documentación (Spain)

Introduction

Many bibliometric indicators have been used to evaluate the scientific performance of researchers, universities and countries. The present study aimed to put forward possible forms of analysis to Brazilian scientific journals to improve the scientific production evaluation criteria in the Brazilian framework. The exclusive use of international indicators leaves the Brazilian scientific policy out of context of its local reality, disconsidering specific particularities of each country in research priorities (Davyt & Velho, 2000; Bordons, Fernández & Gómez, 2002).

The research was conducted according to two objectives: (1) To investigate if SciELO (Scientific Electronic Library Online)-indexed Brazilian scientific production could provide impact indicators to appropriately contribute to the national science evaluation (Packer, A. L. et al., 1998). (2) To find more indicators and new methodologies for science assessment in the context of the Brazilian science reality.

Methods and materials

An exploratory quantitative study was conducted, based on qualitative and quantitative characteristics of SciELO scientific journals: (1) A comparison between the national and international impact of twelve Brazilian journals was conducted based on citations received from journals in SciELO and ISI (Thomson- Scientific – Institute for Scientific Information) databases. (2) A methodology of journal analysis was presented applying the multivariate statistical analysis techniques to a set of impact (national and international citations) indicators of 39 SciELO journals.

SciELO data recorded in ISIS (Integrated Scientific Information System) format was transfeered to Microsoft Access, in which ISI data (from the unmatched file) was delivered by its Research Services Group. SciELO full text database – around twenty of 134 SciELO indexed journals in 2004, some also ISI-indexed, what made possible a comparative citation analysis, between two groups of six most cited Life Sciences journals:

Group 2: SciELO and ISI indexed journals;
Group 1: SciELO but non-ISI indexed journals.

Findings and discussion

The first comparison between the group's impact in both databases shows that being indexed in ISI databases is a condition to be cited in the international context.

Table 1. Citations in SciELO and ISI databases, from 2000-2004, to the SciELO journals groups.

Citation database- Group	**2000-2004 citations**
SciELO - Group 2	8,981
SciELO - Group 1	7,519
ISI - Group 2	23,886
ISI - Group 1	8,842

The discrepancy above indicates that the comparison from national to international impact (specially in relation to Group 2) can not be limited to quantitative aspects, once the "size of the audience" (MacRoberts & MacRoberts, 1996) represented by the set of citing journals in ISI databases is extremely large. This has suggested comparisons of affiliation information of the citing articles from ISI (foreign) journals and respective cited articles.

Table 2 shows that approximately 26% of the citations from foreign journals to Group 2 comes from national collaboration articles (for Group 1 journals it goes near from 31%, see Table 3). On the other hand, the collaboration articles are the most cited for Group 2, while Group 1 receives more citations to single authorship and national collaboration articles.

The multivariate analysis of SciELO journals indicators allowed the identification of different groups of journals (Figure 1), separated according to different practices of scientific communication.

Three impact indicators based on SciELO and ISI citations (Impact Factor, Influence Index and Year Citations – source year 2004) associated to 39 SciELO journals, were used in Factor Analysis by the method of principal components. The correlation matrix was submitted to a varimax rotation. Two factors were extracted, representing 88.9% of the measurement variations (Factor 1 = 47.1%, Factor 2 = 41.8%). Factor 1 represents International impact, while Factor 2 national impact.

Finally, the Cluster Analysis technique permitted to identify four different groups through the trend of national or/and international impact.

Conclusions

The criteria adequacy used in the evaluation of the national scientific production can be met considering impact indicators measured by citations to national journals. Criteria should value the publication in national journals with renowned quality. This will allow the publication of important articles in Portuguese, and encourage the process of quality improvement of national journals.

References

Bordons, M.; Fernández, M. T.; Gómez, I. (2002) Advantages and limitations in the use of impact factor measures for the assessment of research performance in a peripheral country. *Scientometrics*, 53, 195-206.

Davyt, A.& Velho, L. (2000) A avaliação da ciência e a revisão por pares: passado e presente. Como será o futuro? *História, Ciências, Saúde – Manguinhos*, 7, 93-116.

MacRoberts, M. H. & MacRoberts, B. R. (1996) Problems of citation analysis. *Scientometrics*, 36, 435-444.

Packer, A. L. et al. (1998) SciELO: uma metodologia para publicação eletrônica. *Ciência da Informação*, 27, 109-121. Retrieved January 20, 2007 from: http://www.scielo.br/pdf/ci/v27n2/scielo.pdf

Table 2. Citations from ISI (foreing) journals, to SciELO journals from Group 2, períod 2000-2004, crossing cinting and cited authorship.

Group 2 (cited)	**Collaboration (cited article)**								**Total**	
	National		**International**		**Single authorship**		**No identified**			
Colaboration (citing article)	**Freq.**	***%***	**Freq.**	***%***	**Freq.**	***%***	**Freq.**	***%***	**Freq.**	***%***
National	1,356	*36.3*	340	*11.0*	69	*16.3*	271	*48.9*	**2,036**	***26.1***
International	2,336	*62.6*	2,720	*88.4*	342	*80.7*	277	*50.0*	**5,675**	***72.9***
Single authorship	40	*1.1*	17	*0.6*	13	*3.1*	6	*1.1*	**76**	***1.0***
Total	**3,732**	***100***	**3,077**	***100***	**424**	***100***	**554**	***100***	**7,787**	***100***

Table 3. Citations from ISI (foreing) journals, to SciELO journals from Group 1, períod 2000-2004, crossing cinting and cited authorship.

Grupo 1 (cited)	**Collaboration (cited article)**								**Total**	
	National		International		Single authorship		No identified			
Colaboration (citing article)	**Freq.**	***%***	**Freq.**	***%***	**Freq.**	***%***	**Freq.**	***%***	**Freq.**	***%***
National	153	*36.1*	20	*15.9*	158	*31.7*	134	*30.8*	**465**	***31.4***
International	268	*63.2*	102	*81.0*	328	*65.9*	294	*67.6*	**992**	***66.9***
Single authorship	3	*0.7*	4	*3.2*	12	*2.4*	7	*1,6*	**26**	***1.8***
Total	**424**	***100***	**126**	***100***	**498**	***100***	**435**	***100***	**1,483**	***100***

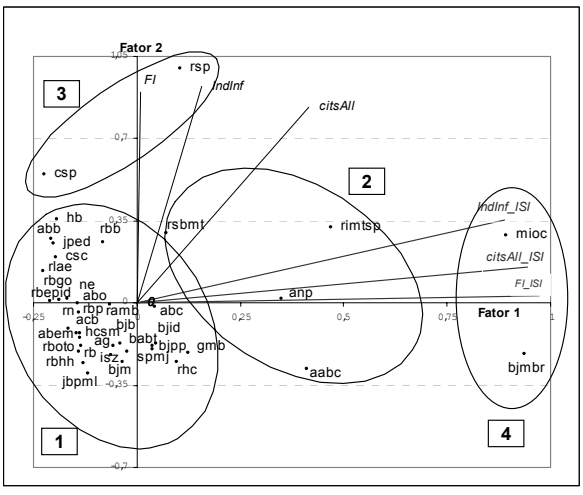


Figure 1. Factorial Analysis of impact indicators of 39 Life Sciences SciELO journals – source year 2004. Four groups resulted from Cluster Analysis

Citation Profiles in Brazilian Journals of the SciELO Database in Different Scientific Areas

Rogério Mugnaini, Rogério Meneghini and Abel Packer

rogerio@bireme.br
Latin-American and Caribbean Center on Health Sciences Information (BIREME/PAHO/WHO) (Brazil)

Introduction

The usage of citation indexes have become increasingly important for studies of the dynamics of scientific communication. In the regional context, they offer quantitative indicators of the utmost importance for scientific evaluation.

In developing countries increasing efforts are devoted to understand the scientific and technological scenery. An important point is that the ISI database and particularly the impact factor (IF)-JCR (both Thomson Scientific) are practically the only source of indicators that most countries, including the developing ones, have been using for scientific output evaluation and establishment of scientific policies.

The availability of these ISI indicators is optimized by their on time indexation. However, alternative indexes are necessary since a single measurement to evaluate journal citation profiles (Glanzel & Moed, 2002) and the often utilization of IF out of its context (Moed, 2005) create drawbacks in the conclusions.

This study compiles an analysis of the scientific communication characteristics in Brazilian journals of the SciELO database (www.scielo.br), which, seven years after its launching, encompassed in 2004, 133 multidisciplinary journals and more than one million references.

Methods and materials

The study of bibliographic references allows knowing the specificities of each area of scientific communication. The concept of reference has a retrospective sense, whereas the concept of citations is referred to the future. While the reference represents the recognition given to another document, i.e., it is based on bibliographic references and is characterized as a synchronous study, citation is the recognition by other document, therefore consisting of a diacrhonous study (Egghe & Rousseau, 1990).

A synchronous study has been carried out, from comprehensive recovery of 133 journals indexed in the SciELO database in 2004, aiming at analyzing the documental typology and aging (obsolescence) in eight areas in which SciELO journals are classified. For this purpose, a factorial analysis has been applied (SPSS 12) and the data were tabulated and operated in Microsoft Excel.

Findings and discussion

The distribution of journals and references per area are shown in Table 1, where one sees the number of health and human sciences journals.

Table 1. Number of journals and references by SciELO scientific areas – sorce year 2004.

Area	Journals*	References*
Agricultural Sciences	21	48,297
Applied Social Sci.	7	4,918
Biological Sciences	20	47,077
Engineering	11	11,339
Exact and Earth Sci.	13	28,883
Health Sciences	44	93,350
Human Sciences	36	28,289
Ling, Lett. and Arts	2	898

*Some journals are classified in more than one area.

Both documental typology and aging are constrained to the IF limitations. The former is due to the fact that areas whose journals cite mostly documents other than journals makes IF an inappropriate indicator of impact; the latter, revels which areas have no impact because their journals use to cite earlier documents, and therefore for a distinct reason poses a question in the IF utility, since it considers one and two years old citations.

For this reason, the analysis of the citation practices in different areas is useful to reveal when the IF is not an adequate indicator.

Documental Typology

The factor analysis by the method of principal components was applied to the area vs. document type percent matrix, resulting in two factors that represented 97.5% of the measurement variations (Factor 1 = 59.0%, Factor 2 = 38.5%).

In Figure 1 it is possible to observe that distinct quadrants correspond to different major areas, and different types of documents are strongly associated to each of them. In the first quadrant one sees that thesis and annals of congresses present some relation with the area of agriculture. For humanities, in the second quadrant, books prevailed over other documents in terms of percentage of citations. In the third quadrant, where life sciences are concentrated, articles are the most important source

of citations, followed by books. Finally, in the fourth quadrant, exact and earth sciences overcome other documents whereas engineering is more inclined to annals.

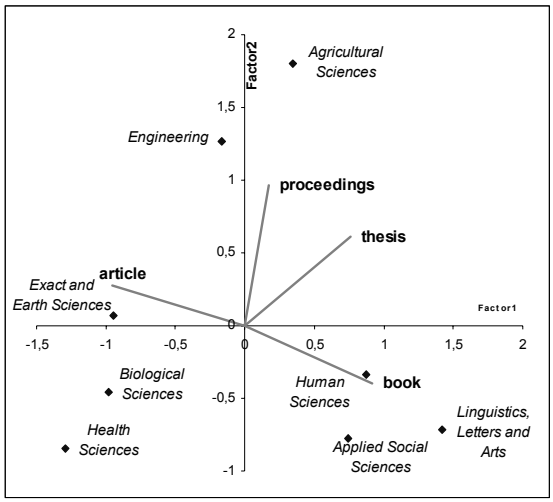


Figure 1. Factor Analysis between citing area and cited document type – sorce year 2004.

Aging (Obsolescence)

To study the obsolescence of the literature cited in the different areas (also using the percent matrix), the same method of multivariate analysis permitted to extract two factors, representing 97.1% of the measurement variations (Factor 1 = 59.7%, Factor 2 = 37.4%).

In Figure 2 one sees that the citations to recent documents are concentrated in the third quadrant; the fourth quadrant concentrates the 90's references; the first, 80's and 70's; and the second quadrant to 60's documents or earlier.

Linguistics, letters and arts showed the highest obsolescence, since around 30% of its references are to documents earlier to 1969 (this area is represented by two journals, meaning that any results must be analyzed carefully). Biological sciences and agricultural sciences presented very similar results to each other, concentrating their references to the 80's. Health sciences are discriminated by its several references to 1990-1995 documents. The other areas are grouped in the third quadrant, although located next to the center.

In biological taxonomy and systematics areas, doubts have been raised as to the ineffectiveness of ISI-IF for the purpose of evaluation studies, in which citations occur more frequently after the two years considered in the calculation of these indicators (Mello & Carvalho), which in fact can be observed in Figure 2.

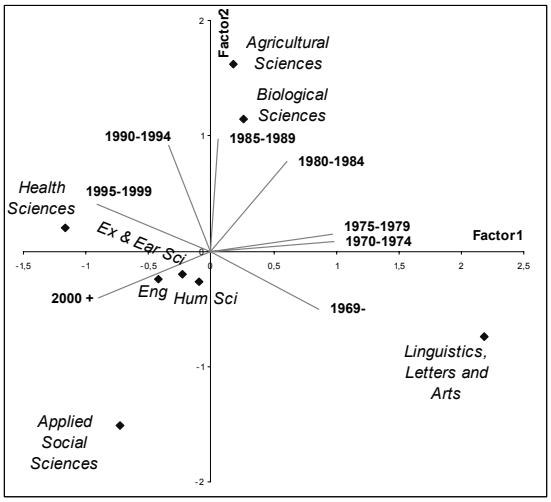


Figure 2. Factor Analysis between citing area and cited document age – sorce year 2004.

Conclusions

The diversity among areas in terms of citations can be distinguished by the documental typology and obsolescence, offering ways to alternative indicators, or to the complementation of usual indicators, which minimizes the assessment limitations, given the proper characteristics of each area.

References

Egghe, L. & Rousseau, R. (1999) *Introduction to Informetrics: Quantitative methods in library, documentation and information science*. Amsterdam: Elsevier.

Glänzel, W. & Moed, H. F. (2002) Journal impact measures in bibliometric research. *Scientometrics*, 53, 171-193.

Melo & Carvalho (2005) Impacto Profundo: o que o fator de impacto não nos revela. *JC e-mail*, 2921, december 22nd. January 20, 2007 from: http://www.jornaldaciencia.org.br/Detalhe.jsp?id=34047

Moed, H. F. (2005.) *Citation Analysis in Research Evaluation*. Dordrecht: Springer.

History Research in a Globalized World: a Bibliometric Approach

Ülle Must

ylle@archimedes.ee
Archimedes Foundation Väike-Turu 8, Tartu 51013 (Estonia)

Introduction
History as the memory of nations shapes their identity and supports future developments. It is important in this globalised world to understand better the way in which cultures and identities are formed; the relationship between national and global identities, feelings of belonging, traditions, convictions and languages. Published studies of history research are one of the most important resources for this understanding .

In current paper we investigate outcomes of the work of journal selection of two expert bodies – European Humanities Reference Index (EHRI) (more information at http://www.heranet.info) and ISI Thomson Arts & Humanities indexing staff. We will concentrate on results delivered from the fields of History, Archaeology and Anthropology.

Methods
Category 'A' journals were used in a comparative analysis from EHRI from the subject fields "History" "Archaeology" and "Anthropology". A total of 235 journal titles were analysed. Searches of ISI Web of Knowledge AHCI for the period 2000-2006 were conducted for thirty-three European countries (Country field) and by the subject History; "History", "Archaeology" and "Anthropology" (History fields). The following data were gathered: the number of papers, the total number of citations, the average number of citations, the H-index. A total of 27,944 papers were analyzed. The following data was additionally compiled from AHCI and compared with EHRI data: the top five lists of journals in which history papers from the thirty-three countries were published, and concurrencies between the two datasets.

Findings
A total of 114,579 papers were published by the thirty-three countries in AHCI during the stated period, of which 26.6% belonged to the History fields. More than half of these papers were published by researchers from UK. At the same time, English speaking countries do not have a wider audience than non-English authors. It means that the impact of language to cited-ness is not as important as it is usually thought.

Meantime we can follow differences inside History fields – in Archaeology and Anthropology. The methods used by hard sciences are applied here. This means that results will be achieved via collaborative work and with a greater number of co-authors. Archaeology and Anthropology are closely connected with evolutionary biology, geosciences etc. (Conard NJ, 2003; Karkanas P, 2000)

Majority of work in the History field is still done by individuals – on average 89% of papers are published by one author. 11% of journals, on average, create a core in which half of all the papers from a listed country are published. A characteristic common to all countries is the preference given to those journals published in that listed country or a neighbouring country; 50.9% of papers from Poland, for example, were published in Acta Poloniae Historica; 81.8% of Bulgarian papers in Revue Bulgare d'Histoire; 56.1% of Spanish papers in Hispania-Revista Espanola de Historia; and 56.6% of Norwegian papers in Historisk Tidsskrift. Journals, which belonged to the top five lists in the selected countries, were all locally oriented (more than 70% of authors belonged to the same or neighbouring country where the journal was published).

Comparisons between AHCI and EHRI data show that from 232 EHRI journal titles, 111 (47.9%) are reflected in AHCI. It is therefore possible to follow the differences between the History fields – in history, 68% of EHRI titles were listed in AHCI, in anthropology – 55.2%, and in archaeology – 29.5%. Almost one third (27.6%) of journals from EHRI list are published in USA while more than half (51.4%) of those that appear in both databases are from USA.

Conclusions
1. Historians from English speaking countries dominate the field, but at the same time they do not have a wider audience than non-English authors.
2. Performance in the History field is still an individual activity. Big international collaboration programmes (EU FP, ESF, COST) have not yet influenced historians' working habits.
3. Publication habits (in which journal to publish) are connected with authors historical and cultural background.

4. The choice of high-ranking international publications, of both expert groups, almost coincides in the history and anthropology fields but in case of archaeology it is much more episodic. Since expert assessment plays a major role in Humanities, then the lists created by European Humanities councils are definitely a good basis for further activities. Europeans are proud of their identity, and at the same time our histories have interwoven throughout the millennia; all the thirty-three countries have either been empire builders or been part of an empire as well as different cultural areas at one time or another. Therefore, general treatments in the field of History, proceeding from just one party remain one sided. Today it is finally possible to create an integral treatment of the European history and civilisation and interlink it with the globalized world.

References

Conard N. J., Bolus M. (2003) Radiocarbon dating the appearance of modern humans and timing of cultural innovations in Europe: new results and new challenges. *Journal of Human Evolution* 44 (3), 331-371.

Karkanas P, Bar-Yosef O, Goldberg P, et al.(2000) Diagenesis in prehistoric caves: The use of minerals that form in situ to assess the completeness of the archaeological record. *Journal of Archaeological Science* 27 (10), 915-929.

Perspective Factor: Past, Present and Future of Life Sciences

Ryosuke L. Ohniwa, Aiko Hibino and Kunio Takeyasu

ohniwa@lif.kyoto-u.ac.jp, *hibino@lif.kyoto-u.ac.jp, takeyasu@lif.kyoto-u.ac.jp*
Laboratory of Plasma Membrane and Nuclear Signaling, Kyoto University Graduate School of Biostudies,
Yoshidahonmachi, Sakyo-ku, Kyoto 606-8501 (Japan)

Introduction

Life sciences now cover large research areas and provide various benefits to the public through medicine, biotechnology and so on. Individual researches require a large amount of research funds from the government. Therefore it is critical for the government to grasp the research trends for timely policy planning and effective distribution of research funds.

PubMed (Schulman, 2000) is the largest database of articles in the field of life sciences, and MeSH (Medical Subject Headings) (Schulman, 2000), which is well-organized keyword thesaurus adopted by PubMed, describes the contents of individual articles. Thus, the transition of the appearance of MeSH keywords attached on the articles represents the trends of research topics in life sciences. Based on the MeSH keywords' trend (i.e., whether the area covered by specific keyword is newly developing or not), we have proposed a new index, Perspective Factor (PF) (ohniwa et al, 2004), which estimates how many newly developing topics are involved in a journal.
In this study, we evaluated the time-dependent changes of PF in the last 30 years, and found that the transition of MeSH keywords can be used as an indication of the trends in development of scientific topics. The results obtained represent an example of a dramatic transition of the research activities in life sciences.

Perspective Factor

PF represents the average number of newly developing MeSH keywords per an article in a certain journal, and, therefore, the PF value of a given journal, PFj, is defined as follows:

PFj = Aj / Bj

where Bj is the number of articles published in the journal j, and Aj is the number of MeSH keywords handled in the journal j which became more popular after the journal j was published. The degree how each MeSH keyword is newly developing in a certain year is obtained as X/Y, where X is the total number of MeSH keywords appearing in the following two years on PubMed, and Y is the total number of MeSH words appearing in the four consecutive years including the previous year and the following three years on PubMed. For example, the degree of 'MAP Kinase Signaling System', which is known to be involved in important process to regulate gene expression and closely related to cancer, in 2000 is evaluated as 0.715, in which X is 2086 (2001 + 2002) and Y is 2919 (1999 + 2000 + 2001 + 2002). Thus, developing topics gain higher value, and declining or mature topics gain lower value. To obtain the value Aj, the top 5 % of the total increment of all MeSH words appearing in a certain year is collected. In the case of 2000, the MeSH keywords filling 'X/Y>0.577' are included in the top 5%, and, therefore, 'MAP kinase Signaling System' is counted as one of the keywords for Aj. About the detail of the PF calculation, see our previous article (Ohniwa et al, 2004).

Compared with well-known 'Impact Factor (IF) ', which is an indicator for the journal quality based on the 'citation analysis' (Wade, 1975; Garfield, 1955; Brody, 1995), PF is free from various concerns about the accuracy of science citation and suspicions on the impartiality of the citation analysis, such as Biased Citation by authors, the size of the audience of journals, and 'Review effect' (MacRoberts and MacRoberts, 1989; Kostoff, 1998; Seglen, 1997).

Transition of Perspective factor

Nature, *Science* and *Proceedings of National Academy of Science in USA* (*PNAS*) have been categorized as multidisciplinary journals covering almost all disciplines. In addition, these journals have been believed to have top-level qualities reflecting trends in life sciences. By extraction of the articles related to life sciences from these journals (Ohniwa et al, 2004), the PF values of these journals in the last 30 years were traced; the absolute values of *Nature* and *Science* increased until late 1990s and decreased after that period (Fig. 1).

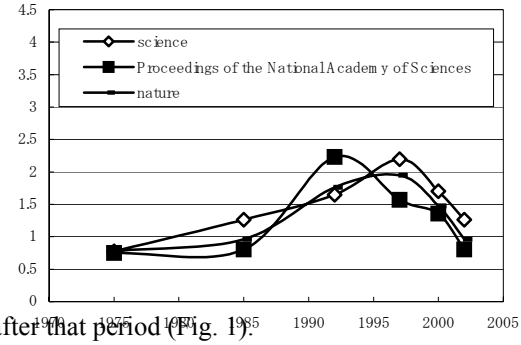


Figure 1. Transitions of PFs in multidisciplinary sciences.

The value of *PNAS* increased until early 1990s and decreased in late 1990s. This means that there are fewer newly developing topics after late 1990s than before. If the development of science can be considered as the development of scientific concepts, life sciences may have run into a stagnation period.

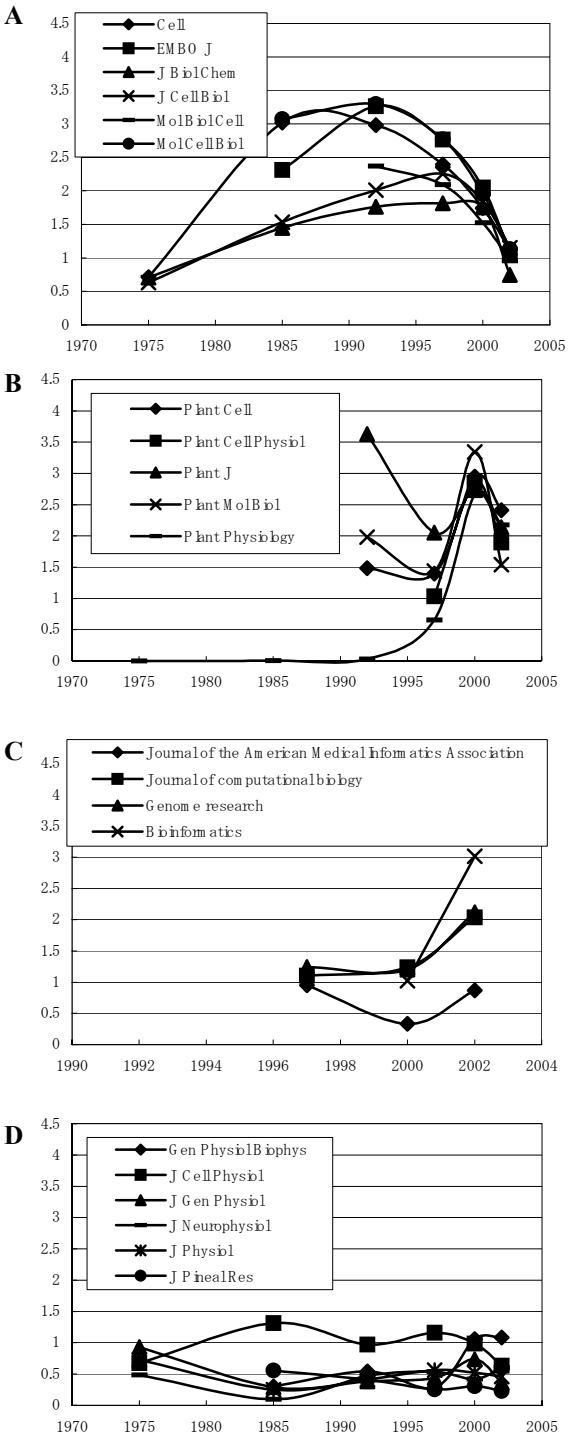


Figure 2. Trends in Perspective Factor of in the fields of (A) *biochemistry*, *molecular bio*logy and *cell biology*, (B) *plant biology*, (C) *bioinformatics* and (D) *physiology*.

To specify the trends in specific disciplines, the PF values of the representative top-level journals in each discipline were calculated. In the field of *molecular biology*, *biochemistry* and *cell biology*, the PF values increased until middle 1990s and decreased after that period (Fig. 2A). This tendency was also found in the fields of *genetics* and *development* (data not shown). The journals belonging to *microbiology* showed their PFs' peaks in early 1990s (data not shown). In the fields of *plant* and *bioinformatics*, their peaks appeared after 2000 (Fig. 2B, C). The journals in other fields (i.e., *immunology*, *neuroscience* and *physiology*) have possessed relatively low PF values over 30 years (Fig. 2D).

Concluding Remarks

The development of techniques in the field of life sciences has been focussed into 'genome project'. The declare by Bill Clinton in 2000, in which the accomplishment of making rough draft of human genome was announced, represented a great success of life sciences. However, our data suggest that, behind the proudful declare, the activities toward developing new concepts in life sciences have decreased since mid 1990s. Now, in the post-genome project era, it is necessary to re-evaluate the policy planning and grant distribution strategies for driving the life sciences to the next stage.

References

Ohniwa, R.L., Denawa, M., Kudo, M., Nakamura, K. and Takeyasu, K. (2004). Perspective Factor: A novel indicator for the assessment of journal quality*, Research Evaluation*, 13, 175-180.

Schulman, J. (2000). Using Medical Subject Headings (MeSH) to examine patterns In American medicine-Preliminary Cnsideration of Vocabulary Change as a Metric, from: http://www.nlm.nih.gov/mesh/patterns.html.

Garfield, E. (1955). Citation Index for Science, *Science*, 122, 108-111.

Brody, S. (1995). Impact factor as the best operational measure of medical journals, *Lancet*, 346, 1300-1301.

Seglen, P.O. (1997). Why the impact factor of journals should not be used for evaluating research, *BMJ*, 314, 497.

Ksotoff, R.N. (1998). The use and misuse of citation analysis in research evaluation, *Scientometrics*, 43, 27-43.

In-Depth Analysis of Wikipedia Community

José Felipe Ortega, Jesús M. González-Barahona and Gregorio Robles

felipe@gsyc.escet.urjc.es, jgb@gsyc.escet.urjc.es, grex@gsyc.escet.urjc.es
Universidad Rey Juan Carlos, Grupo de Sistemas y Comunicaciones, c/Tulipán s/n, 28933 Móstoles (Spain)

Introduction

In this paper we present an in-depth analysis of the Wikipedia author community and articles. We also describe a methodology to collect and analyze relevant information from Wikipedia databases, and apply it to the 10 largest language editions. We focus our research on authors' contributions, size and number of articles, and their evolution over time, inferring some relationships between authors' contributions patterns and Wikipedia contents.

These relationships reflect (and in part, explain) the evolution of the different language communities so far, and their trends toward the future. The growth of the whole project is affected by four different factors:

- *System infrastructure*: Currently, most of the traffic served by Wikipedia comes from a cluster in Florida, maintained by Wikimedia Foundation. In the past months, several mirror projects have been set up in Europe (France) and Asia (Yahoo! cluster in Singapore). However, Wikipedia size seems to grow up much faster than the project facilities and the threat of a serious slow down in content service agility could be evident very soon.

- *Software evolution*: MediaWiki [1] is the core software platform of Wikipedia Project. A very active community of developers supports the evolution of this software package, adding new functionalities users asked for and boosting performance improving administrative tools.

- Article and contents evolution: Articles are the

- core of Wikipedia. There is one article for each different topic, and topics are selected through consensus among user wishes. So far, article contents may include text, graphics, photographs, math formulas and multimedia. They reflect authors' interests and level of contribution (some languages gather more articles than others), so this factor is in close relationship with the last one.

- *Community changes and contributions*: Wikipedia expansion is also affected by contributions from editors and developers. Wikipedia strongly depends on the work of volunteers to maintain its current rate of growth. If, for some reason, article editors change its current behaviour, or decrease their rate of software contributions, this will definitely affect Wikipedia growth model.

Some previous work has also been made in this line specifically for the Wikipedia. In a paper presented in the last ISSI conference, Jakob Voss (Jakob Voss, 2005) introduced interesting preliminary results about contents evolution and authors, focusing in the German version. The number of distinct authors per article (which follows a power-law) along with the number of distinct articles per author (which follows Lotka's Law) were also analyzed.

Methodology

Our methodology to analyse Wikipedia follows the following steps. First, we collect the database dumps for the top 10 Wikipedia language versions (in number of articles, according to the list publicly available from Wikipedia main page [2]. We have developed a Python tool, WikiXray, to process database dumps, automatically collecting relevant information and massaging it for an in-depth statistical analysis. Quantitative results for each language database of Wikipedia can be obtained from two different points of view:

Authors' community: Relevant aspects include the total number of editors and contributions, distributed contributions for each author in a certain period of time and correlating with Wikipedia own results for active and very active authors.

Articles evolution: Most important aspects in this field include article size evolution in time, article size distribution, and correlations between articles evolution and user contributions.

Case Study: the top 10 Wikipedia Language Versions. We have tested our methodology and tools to analyse the database dumps corresponding to the 10 largest language versions of Wikipedia. As the time of writing, the largest language version is English, followed by German, French, Polish, Japanese, Dutch, Italian, Portuguese, Swedish, and Spanish versions, in descending order.

Wikipedia Authors' Community

[1] http://www.mediawiki.org/wiki/MediaWiki

[2] http://www.wikipedia.org

In Figure 1 we present the distribution of contributions made by registered authors in the top 10 Wikipedias. Then, we can compute the Gini coefficient, first introduced by Conrado Gini (Gini, 1936), and show how unequal contributions from different authors to the project are. The straight curve shows the perfect distribution among authors (ideal line) and the Lorenz curve below represents the actual distribution. The ratio between the areas below both curves gives us the coefficient. We see a great level of inequality for all language versions.

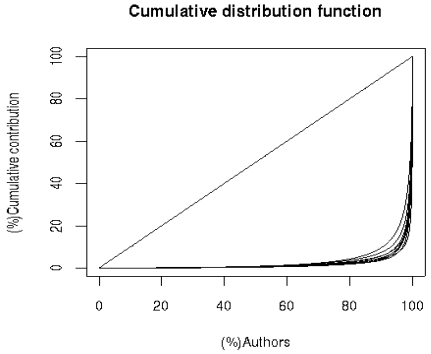


Figure 1. Cumulative distribution function for the top 10 Wikipedias.

Articles Evolution

We first consider the histograms of the articles size in each language. Figure 2 shows the histogram of the article size (log scale) in the English and Polish Wikipedias. The solid black line plotted over the histograms represents the probability density function of the article size, showing about the same information but with better resolution.

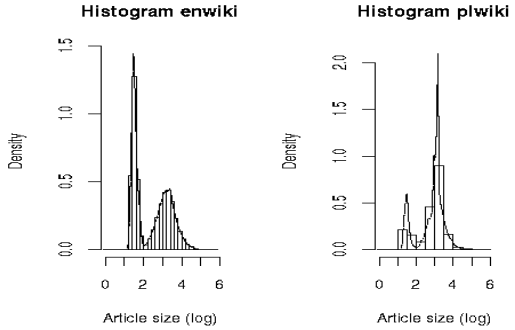


Figure 2. Histogram of the articles size in the English (left) and Polish (right) Wikipedias.

We identify two different populations: *tiny articles*, with quite a low size (left half, mostly *redirects*), and *standard articles* (right half). We can see that both groups present Gaussian distributions.

Finally, in Figure 3 we show a 3D representation of the probability mass values for each population in every language version against the average number of edits received per article in that version. There exist a certain relationship between the average number of edits articles received in a Wikipedia and the level of proximity of its probability mass values for *tiny* and *standard* articles.

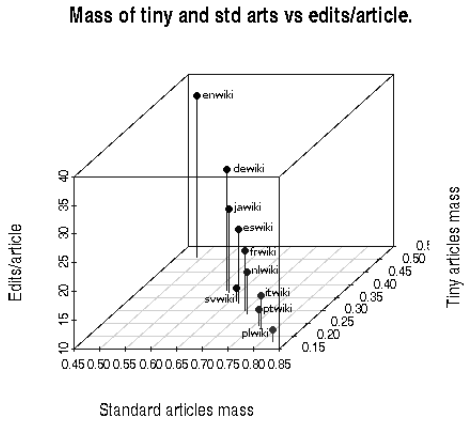


Figure 3. Probability mass values for tiny and standard articles against average number of edits per article for the top 10 Wikipedias.

Conclusions and Future Research

We first show that the Gini coefficients found for the studied languages show great inequalities in the contributions by different authors, with a small percentage being responsible for a large share of the contributions. These factors could help to model the underlying author communities. We also recognize patterns that can be used to characterize Wikipedia articles for each language version attending to its length. Two main subgroups (*tiny* and *standard articles*) represent peculiarities in contributions behaviours for each language. For each language version, there is also a certain correlation between the probability mass values of *tiny* and *standard* articles and the average number of edits received per article. This methodology therefore leads towards an integral quantitative analysis framework for the whole Wikipedia project, a very ambitious goal that we confront for the near future.

References

Gini, C. (1936). On the measure of concentration with especial reference to income and wealth. Cowless Comission.

Voss, J. (2005). Measuring Wikipedia. Proceedings of the 10th International *Conference of the International Society for Scientometrics and Informetrics, Stockholm.*

Exploring Visually the European Academic Web Space

Jose Luis Ortega*, Isidro Aguillo*, Viv Cothey**, Andrea Scharnhorst***

*jortega@cindoc.csic.es, isidro@cindoc.csic.es
Cybermetrics Lab., CINDOC-CSIC, Joaquín Costa, 22. 28002 Madrid (Spain)

**viv.cothey@wlv.ac.uk
School of Computing and Information Technology, University of Wolverhampton, Lichfield Street, Wolverhampton (United Kingdom)

***andrea.scharnhorst@vks.knaw.nl
Virtual Knowledge Studio, Royal Netherlands Academy of Arts and Sciences, Cruquiusweg 31, 1019 AT Amsterdam (The Netherlands)

Introduction

Visualization of Information (VI) (Chen, 2003) is a technique that it aims to show conceptual entities and their relationships through visual metaphors that allows us to interpret and extract conclusions about a certain complex phenomena. Inside of VI, Maps of Science (Small & Griffith, 1974; McCain, 1990) are a model of the utility that show the scientific relationships among authors or academic institutions through the citations, co-authorship, or co-word analysis. Although Noyons (1999) defines Maps of Science as "landscapes of scientific research fields created by quantitative analysis of bibliographic data", recently web data have been used as additional source of information on scientific networks.

The World Wide Web is a complex network that connects web pages and sites through hypertextual links creating a large and dense network of nodes (Scharnhorst, 2003). Network Analysis is both a suitable way to present graphically the link relationship in the web and a technique to analyze and understand the web structure and topology.

Objetives

The aim is to get insights how structured the European academic space in the web is. In particular, we ask which agglomerative aggregation of universities across European countries can be seen on the web. Do we get a random network? Which role plays geographic neighbourhood? Or will we observe a superposition of national networks?

Methods

535 universities of the 14 European countries (EU except Luxembourg) in 2004 were selected from *Webometrics Ranking of World Universities* (http://www.webometrics.info/). This set of European universities were mapped according to the link relationships among them. Two different set of web data were used: search engine data and crawler data. The search engine was used to retrieve the link relationships between web sites and the crawler was used to extract the pages hosted in each web site. The search engine data were obtained from *Yahoo! Search* with the query:

+site:{university domainA} +linkdomain:{university domainB}

On the other hand, the crawler data were extracted with the software *Blinker* (Cothey, 2004) to find the number of pages and domains of the 535 universities. Both set of data were obtained in August of 2005. This graph was laid out with the *Spring embedding* algorithm through NetDraw. Finally, multiple arcs with fewer than 50 links were removed to reveal a clearer graph of a network of 527 nodes.

Each node represents an university website. Its colour is according to the nationality of the university and the size is according to the number of pages extracted. The university websites have been also classified into five categories according to their content. These five categories (General, Technology, Social Sciences, Biomedicine and Humanities) were created ad hoc to express the main academic subject area. The shape of the nodes show these categories.

Results

From a topological point of view, the graph (Figure 1) shows the properties of a scale-free network (Albert, Jeong & Barabasi, 1999), which means a few nodes attract a huge amount of links and the rest of nodes attract only a few of them. One can also see that the universities are grouped by country. In particular, we can distinguish the German sub-network in red, the British in blue and the French in yellow (This will be appreciate in the full colour poster). Surprisingly, France shows a barely connected network and with low volume of published pages. Additionally, it is visible that there are small countries which universities do not form

an homogeneous sub-network but are spread out over other large sub-networks such as in the case of Austria where Austrian universities are connected with universities in Germany or in the case of Ireland where Irish universities connect to British universities. It is important to note that the Scandinavian countries constitute a compact and close sub-network.

Figure 1 also shows that the nodes centrally located attract more links than the rest ones. It has been shown earlier that a correlation exist between the number of pages on a web site and the in-links it attracts from other web sites (Adamic & Huberman, 2002; Thelwall, 2004; Katz & Cothey, 2006). Thus, the size of a web site is key factor to achieve a high centrality degree in the web network.

Discussion and Conclusion

The general picture of the European academic network sector allows us to see that it is built up of multiple national networks. That is, there is not a single unique network but a network of national networks that are aggregated one with another.

We can also conclude that the EU university network is made up by the aggregation of national networks. Thus an university is first linked to others within its country and then this national network is connected to other national networks.

References

Adamic, L. A. & Huberman, B. A. (2002), Zipf's law and the Internet, Glottometrics, 3, 143-150 http://www.hpl.hp.com/research/idl/papers/ranking/adamicglottometrics.pdf

Albert, R., Jeong, H. & Barabasi, A.L. (1999), Internet - Diameter of the World-Wide Web, *Nature*, 401(6749), 130-131.

Chen, C. (2003), *Mapping Scientific Frontiers: The Quest for Knowledge Visualization*, London: Springer-Verlag.

Cothey, V. (2004), Web-Crawling Reliability, *Journal of the American Society for Information Science and Technology*, 55(14), 1228-1238.

Katz, J. S. & Cothey, V. (2006), Web indicators for complex innovation systems, Research Evaluation, 15(2),85-95 http://www.sussex.ac.uk/Users/sylvank/pubs/WICIS.pdf

McCain, K. (1990), Mapping Authors in Intellectual Space: A Technical Overview*, Journal of the American Society for Information Science and Technology*, 41(6), 433-443.

Noyons, E.C.M. (1999), *Bibliometrics mapping as a science policy and research management tool*, Leiden: Leiden University.

Scharnhorst, A. (2003), Complex Networks and the Web: Insights from Nonlinear Physics, Journal of Computer-Mediated Communication, 8(4) http://www.ascusc.org/jcmc/vol8/issue4/scharnhorst.html

Small, H. & Griffith, B.C. (1974), The structure of the scientific literatures Identifying and graphing specialties, *Science Studies*, 4(1), 17-40.

Thelwall, M. (2004), L*ink Analysis: An Information Science Approach,* Amsterdam et al.: Elsevier Academic Press

.

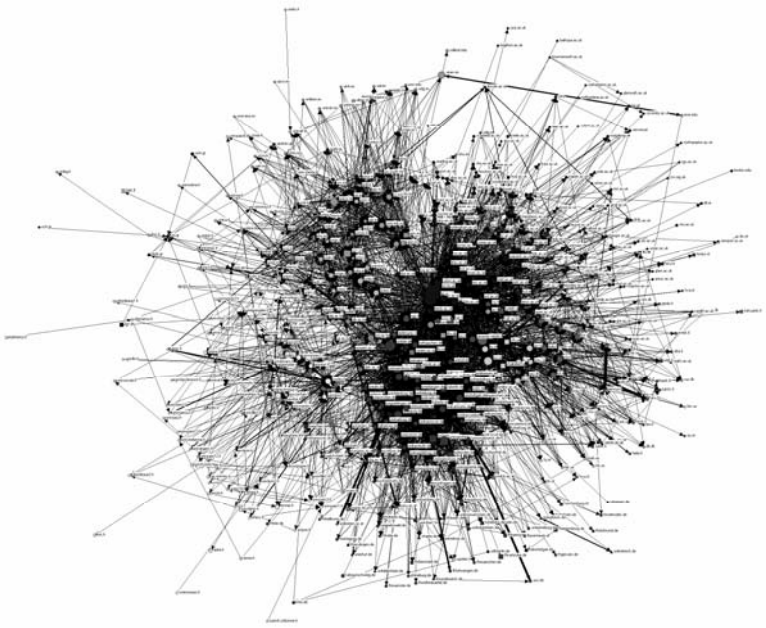


Figure 1. European Universities Network of links (527 nodes; 8028 ties)

Basic Characteristics of Cancer-related Research in Puerto Rico: An Approach to Local and International Journals

Ana Patricia Ortiz[*], William A. Calo[*], Carlos A. Suárez[**], Erik Suárez[*], Isabel Iribarren[***] and Elías Sanz-Casado[***]

[*]*aportiz@rcm.upr.edu*
[*]University of Puerto Rico, Medical Sciences Campus (Puerto Rico)

[**]*{wcalo, esuarez, csuarez}@ rcm.upr.edu*
[**]University of Puerto Rico, Graduate School of Information Science and Technologies (Puerto Rico)

[*]*elias@bib.uc3m.es, iiribar@bib.uc3m.es*
[***]University Carlos III, Madrid, Department of Librarianship and Documentation (Spain)

Introduction

Cancer is the second cause of mortality in Puerto Rico, country where the incidence and mortality rates of this disease have increased dramatically during the last decades. Even though the scientific activity has increased in Puerto Rico during the last decades the analysis of scientific production is still largely unexplored principally en cancer research. The objective of this study is to compare the cancer-related publications trends in Puerto Rico's local scientific journals (not indexed in the Thomson databases) and international journals indexed by the SCI from 1982 to 2005.

Methodology

Journal sources: The Puerto Rico Health Sciences Journal (PRHSJ) and the "Boletín de la Asociación Médica de Puerto Rico" (BAMPR), only Puerto Rico's biomedical journals, were selected for the study. Data of these two journals were obtained from MEDLINE/PubMed. To include older articles not listed in that database a manual hard copy search was also done. Data of cancer-related scientific output of Puerto Rico in international journals were obtained from the Science Citation Index.

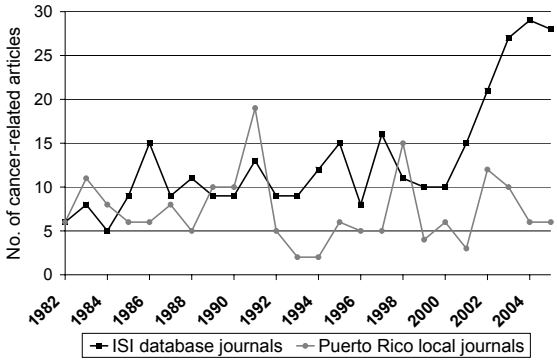
Inclusion criteria's: Terms for the retrieve were identified using the cancer definition from the National Cancer Institute. When searching the SCI database papers authored by at least one scientist working in Puerto Rican institution were identified by retrieving all the records whose affiliation field contained at least one institution located in Puerto Rico.

Statistical analysis: Contingency tables were used to describe the frequency distributions and means of the scientific production. To assess the time trend in the bibliometric indicators the Annual Percent Change (APC) was calculated. The statistical analysis was performed in the computer software Stata/SE (version 9).

Results

As shown in Figure 1, cancer-related articles published in international journals have dramatically increased since 2000 with the highest peak in 2004 (n=29). A comparison between local and international journals (Table 1) showed more articles published in international journals than in local journals with an overall increase of 5.75% per year (APC). A higher collaboration index is observed for international journal (6.52 vs 3.26 authors per article mean). Also, English is the predominant used language and original papers are the most frequent manuscript type used to disseminate the research information in both local and international journals. Table 2 demonstrates that authors affiliated with institutions located in Puerto Rico publish the majority of their research in journals with high visibility (Q=1).

Figure 1. Cancer-related scientific production trends, 1982-2005



Discussion

Our results evidence the growing interest for Puerto Rico's cancer researchers to publish their findings in international journals rather than locals. The growth in cancer publications is consistent with an increase observed in scientific production in Puerto Rico. The co-authorship mean for cancer publications in international journals is higher than the mean previously reported for publications in science and technology by Puerto Rican's co-authors. The fact that original papers are the most frequent manuscript type published in both local and international

journals coincide with trends observed in Puerto Rico in the basic sciences area. Finally, Puerto Rican's researchers published their findings in journals with high visibility demonstrating the good quality of the research performed.

One possible reason to explain the increment of publications in international journals by local researchers and the quality of their research is due to the availability of US grants, local funding and partnerships. As an example, the National Institutes of Health have been funding local researchers through highly competitive grants in the last decades for research, training, education and outreach. Also, in 2002, the National Cancer Institute gave a U-54 grant to the University of Puerto Rico and the University of Texas to establish a Cancer Center in the Island. Locally, in 2004, the government approved legislation to create a Comprehensive Cancer Center with a budget of more than 100 million dollars to spend in the next ten years.

Conclusions

This paper presents a comprehensive approach of the Puerto Rico's cancer-related publication trends, providing valuable information of cancer research for researchers, policy makers and informatics scientists.

Table 1. Descriptive characteristics of cancer-related articles in local and international journals by Puerto Rican researchers

Bibliometric indicators	**Puerto Rico local journals n (%)**	**ISI database journals n (%)**
Scientific production	176 (APC = -0.07)	314 (APC = 5.75)
Co-authorship Index (mean + SD)	3.26 ± 2.67	6.52 ± 5.34
References/article (mean + SD)	17.20 ± 14.81	28.64 ± 22.22
Language		
English	146 (83.0)	313 (99.7)
Spanish	30 (17.0)	0 (0.0)
German	0 (0.0)	1 (0.3)
Manuscript		
Original paper	134 (76.1)	245 (78.0)
Congress minutes	0 (0.0)	45 (14.3)
Notes	5 (2.8)	7 (2.2)
Letters	1 (0.6)	3 (1.0)
Editorial	3 (1.7)	8 (2.5)
Review article	23 (13.1)	6 (1.9)
Special article	10 (5.7)	0 (0.0)

Table 2. Basic indicators related to research visibility of scientific output in international journals

Journals	**n (%)**	**IF**	**Q**
International Journal of Radiation Oncology Biology Physics	41 (13.06)	3.76	1
Cancer	34 (10.83)	4.43	1
Journal of Clinical Oncology	15 (4.78)	11.81	1
Blood	11(3.50)	10.13	1
Cancer Treatment Reports	11(3.50)	Data Unavailable	
American Journal of Clinical Oncology-Cancer Clinical Trials	10 (3.18)	0.93	4
Proceedings of the American Association for Cancer Research	7 (2.23)	Data Unavailable	
Breast Cancer Research and Treatment	5 (1.59)	4.64	1
Journal of Nuclear Medicine	5 (1.59)	4.51	1
Seminars in Oncology	5 (1.59)	Data Unavailable	
American Journal of Epidemiology	4 (1.27)	4.49	1
Investigational New Drugs	4 (1.27)	2.42	2
Journal of the National Cancer Institute	4 (1.27)	11.40	1
New England Journal of Medicine	4 (1.27)	38.57	1

n=numbers of papers, %=percentage
IF=Impact Factor, Q=Quartiles

References

Kamangar F, Dores GM, Anderson WF. Patterns of cancer incidence, mortality, and prevalence across five continents: defining priorities to reduce cancer disparities in different geographic regions of the world. J Clin Oncol 2006 May;24(14):2137-50.

Ortiz AP, Calo WA, Suarez C, Maura M, Suarez E. Bibliometric assessment of cancer related publications in Puerto Rico. Paper submitted to the Journal of the Medical Library Association.

Ortiz-Rivera L, Sanz-Casado E, Suarez-Balseiro CA. Scientific production in Puerto Rico in science and technology during the period 1990 to 1998. Scientometrics 2000;49(3):403-18.

Puerto Rico's Health Department. Puerto Rico's Annual Vital Statistics Report, 2003, San Juan, PR.

Acknowledgments

Study funded by NIH Grant #U54CA96297 from the Puerto Rico Cancer Center / M.D. Anderson Cancer Center Partnership and RCMI Grant #G12RR03051 from the University of Puerto Rico.

Quantification Evaluation of Technical Innovation Capabilities of Chinese Large and Medium Industrial Enterprises

Pang Jing'an, Wang Lian and Cao Yan

pangja@istic.ac.cn, wanglian@istic.ac.cn, caoyan@istic.ac.cn
Institute of Scientific and Technological Information of China, Beijing, 100038(China)

Introduction

Since the beginning of 21st century, the process of industrialization in China has been further accelerated and the general scale of industry has been continuously enlarged. In the process of industrialization, scientific R&D and technical innovation activities carried out by large and medium enterprises have gradually become the fundamental impetus and source to promote the rapid and sustainable development of industry. Based on the up-to-date data released by National Bureau of Statistics, this article makes a full and systematic quantification research on the technical innovation capabilities of Chinese large and medium enterprises, and objectively evaluates the level of technical innovation capability of Chinese large and medium industrial enterprises by using indicators for science and technology in comparison with corresponding figures of foreign developed countries.

General situation of technical innovation activities of large and medium industrial enterprises

In 2004, there were 6,566 large and medium industrial enterprises with technical innovation activities in China, accounting for 23.7% of all the large and medium industrial enterprises.

Manufacturing sector is an industrial cluster in which technical innovation activities are most centralized. In 2004, there were 6,133 large and medium industrial enterprises in Chinese manufacturing sector with such activities, accounting for 93.4% of the total. However, the distribution of technical innovation activities carried out in various industries of manufacturing is uneven.

Quantification analysis on R&D activities carried out by large and medium industrial enterprises

In 2004, there were totally 653,932 R&D personnel in Chinese large and medium industrial enterprises, accounting for 1.86% of all employees. Comparing with the situation of past years, the size of R&D personnel basically shows a gradually rising trend, but the growth speed in recent years has slowed to some extent. In 2004, R&D expenditure of Chinese large and medium industrial enterprises was RMB 95443.11 million. R&D expenditure also shows a rising trend, and the growth rate has accelerated considerably in recent years. The R&D personnel and expenditure of Chinese large and medium industrial enterprises mostly concentrates in the manufacturing sector.

The intensity of R&D expenditure by large and medium enterprises in manufacturing sector was 0.8%, with that by manufacture of medicines being the highest, namely 1.42%.

Technical innovation output of large and medium industrial enterprises

Manufacturing enterprises constitute the main force of technical innovation output. In 2004, large and medium manufacturing enterprises totally realized RMB 2039.2 billion production value of new products and RMB 2026 billion sales income, with RMB 482.4 billion for export thereinto, accounting for 99.2%, 99.2% and 99.4% of all large and medium enterprises respectively.

The number of patent applications by Chinese large and medium industrial enterprises was 42,318, with 30% year-on-year increment. Of the all kinds of patent applications, 13,908 were patents for invention, increasing by 4,513 over that in 2003, with about 50% growth rate.

Technical introduction, assimilation and re-creation of large and medium industrial enterprises

In 2004, Chinese large and medium industrial enterprises expended totally RMB 36.8 billion on introducing foreign technologies, RMB 5.4 billion on assimilating and absorbing foreign technologies, and RMB 7 billion on purchasing domestic technologies. However, the ratio of expenditure on assimilating and absorbing technologies to that on introducing technologies was still relatively low, only 1:0.15. Manufacturing enterprises were still the main body of technical introduction conducted by large and medium industrial enterprises.

Conclusions

(1)Statistics shows that Chinese large and medium industrial enterprises are not actively engaged in carrying out technical innovation activities. Only less than 1/4 of large and medium industrial enterprises have carried out such activities, and this data is obviously low comparing with situation that enterprises in foreign developed countries commonly attach more importance to technical innovation activities. Manufacturing sector is the most centralized industry cluster that carries out technical innovation activities, accounting for 93.4%

of all large and medium industrial enterprises that have done so. The performance of manufacture of medicines is notable, the aforesaid proportion of which even reaches 56.3%. This seems reasonable considering that medicines manufacturing industries of other countries in the world are also generally paying more attention to R&D and technical innovation comparing with other industries.

（2）In China , Enterprises with scientific institutions only account for 32.8% of all large and medium industrial enterprises, which is every unfavorable for building up a stable professional scientific team that carry out long-term key problem tackling research and raising enterprises' key technical R&D capabilities and core competitiveness. It is to be noticed that the although the number of scientific institutions established by large-sized enterprises is relatively small, only accounting for 26.4% of all large-medium industrial enterprises, the personnel for S&T activities and expenditures on S&T activities of them account for 54.5% and 64.4% respectively.

（3）R&D expenditures are the material base and source for enterprise to carry out technical innovation activities. In recent years, both R&D personnel and R&D expenditure of Chinese large and medium industrial enterprises are showing a stably rising trend. However, the proportion of their R&D expenditures in sales income is still rather low, and even for large and medium manufacturing enterprises with higher technological level, the intensity of R&D expenditure is only 0.8%, far behind that of enterprises in foreign developed countries.

（4）The R&D expenditures of Chinese large and medium industrial enterprises are mainly self-owned capital. In 2004, enterprise capital accounted for 94.8% of all R&D expenditures. Currently, 30% of scientific funds in developed countries are used to support enterprises, while in China over 90% of such funds are expended on scientific research institutions and universities and colleges, which accounts for the decreasing proportion of governmental funds in R&D expenditures of enterprises.

As to the goals of R&D activities to be carried out, Chinese large and medium industrial enterprises give more priority to experimental development research, the expenditures for which account for most part of R&D expenditures. This situation is very different from enterprises in foreign developed countries that pay more attentions to basic research. Although enterprises have begun to attach more importance to basic research and application research in recent years, their proportion in R&D expenditures is still low. And at the present time, the technical innovation activities carried out by Chinese large and medium industrial enterprises are basically in-house R&D, and cooperative R&D with other enterprises, especially foreign enterprises and institutions are rather rare. Technical and economic targets of R&D projects are mostly to develop entirely new products and add functions to original products.

（5）Actively carrying out scientific R&D activities has directly promoted continuous growth of technical innovation output of enterprises. In recent five years, for Chinese large and medium industrial enterprises, both the proportion of sales income of new products in that of all products and the proportion of export amount of new products in sales income of new products have realized growth by large amplitude. At the same time, the quantity of patent applications of Chinese large and medium industrial enterprises and the proportion of applications of patent for invention are showing a rising trend.

（6）Technology acquaintance by enterprise is necessary for the development of enterprise. Internally technology acquaintance by enterprise is mainly realized through carrying out R&D activities, while externally technology acquaintance is mainly realized through introducing foreign technologies, assimilating and absorbing them and then re-creating. At present, the expenditures of large and medium industrial enterprises on introducing foreign technologies decrease to some extent, while the expenditures on assimilating and absorbing technologies are gradually increasing. However, the ratio of expenditures on assimilating and absorbing technologies to that on introducing technologies is still rather low, only 1:0.15, far lower than the corresponding indicator of enterprises in South Korea and Japan.

References

National Bureau of Statistics(2005). Statistic data on technical innovation of Chinese large and medium industrial enterprises. Retrieved July 6, 2006 from: http://www.stats.gov.cn/tjsj/qtsj/dzxgyqyzzcxtjzl/index.htm

National Bureau of Statistics (2000-2005). *China Statistical Yearbook*. Beijing: China Statistics Press

Ministry of Science and Technology. *China Science and Technology Indicators*. Beijing: Scientific & Technical Documents Publishing House, 2002, 2004

Ranking Journals in Physics: An Informetric Study of the Citations in the Doctoral Theses of the Indian Institute of Science

K. G. Pillai Sudhier

kgsudhier@gmail.com
Department of Communication & Journalism, University of Kerala, Kariavattom Campus
Thiruvananthapuram – 695 581, Kerala (India)

Introduction and objectives

Citation analysis is one of the important tools among other informetric techniques to find out the core journals in a particular field and is becoming popular in its practical application in libraries and information centres. Studies employing citation analysis methodology are often conducted by evaluating a sample of citations from doctoral theses to develop a core list of journals, and subsequently, to determine what proportion are locally held and the estimated strength of the collection (Bonnie, 1999). Using citation data from a group of theses, it is possible to draw conclusions about scholarly communication at the doctoral level.

Dissertation citation analysis has frequently been proposed as an in-house means to identify journals most important for the research collection.

A number of studies have reported on core journals in various subjects, but no attempt has been made to develop a systematic study in the area of Physics, particularly in the Indian context.

The Indian Institute of Science (IISc) is particularly selected for the study because of its national importance as a centre of excellence in Physics. Since it emerged out of Jamshetji Tata's vision in 1909, it has grown into an advanced research and education centre, indeed, a premier 'brain trust' of the country. IISc is one of India's oldest research institutes approaching its centenary in 2009 and is a premier scientific institution in the world. The Department of Physics established by the Nobel laureate Sir. C.V. Raman in 1933 is the oldest of the departments in the division of Physical and Mathematical Sciences.

The specific objectives of the study are: to prepare ranked list of most cited journals, their coverage in the JCR, to identify Indian journals and their coverage in the JCR, to study the country, language, periodicity and the year of establishment of journals, identify the publishers of journals, to study the application of Bradford's law of scattering in the distribution of journal articles and to study the obsolescence of Physics literature.

Methodology and Source Data

To develop a datay set for this study, a total of 71 doctoral theses awarded during 1999-2003 (5 years) from IISc are selected. The scholars cite 690 journals comprising 11,412 citations in their theses. Informetric analysis of the journal citations appended in the theses are made for the study.

Results and Discussion

Among the different bibliographic forms used by the researchers, journal articles account the maximum constituting 84.48% of the total citations. This is followed by books and others category occupying 10.45% and 1.22% respectively. It indicates that IISc Physicists are heavily dependent on the journal articles.

Top ranked Journals

Frequency of citations or use of a journal is governed by many factors such as availability, language and country of publication, size and frequency of publication, coverage in secondary journals, reprint dissemination, reputation of the authors and so on.

The Physical Review B, a specialised journal in the area of Condensed Matter Physics tops the list with the highest contribution of 1087 (9.53%) citations. Physical Review A, also a publication of APS, is in the second position by accounting 878 (7.69%) articles, while Astrophysical Journal, occupies the third position with 624 (5.47%) citations.

Majority of the most cited journals are published for more than 100 years. Physical Review B (110 Yrs), Physical Review A (110 Yrs), Astrophysical Journal (108 Yrs), Nature (134 Yrs) etc. Among the ranked journals, Physical Review Letters, Applied Physics Letters, Solid State Communication, Physics Review Letters etc. with letters or communication in their titles are in the core list and are the preliminary communication journals. Nature and Science, though they are not letters journals, their main purpose is also reporting preliminary communication and current research. The next remarkable feature of the study is the high status of multi-disciplinary science journals in the core journal list of Physics literature.

The Impact Factor (IF) of journals covered in JCR varied between 0.988 and 32.771. The high IF journal Rev. Mod. Phys. published by American Physical Society with the IF of 32.771 and Immediacy Index 5.826 is ranked 20 by citation count with 125 (1.10%) total articles. The next journal with higher IF 32.182 is Nature, in the 9th rank position followed by Science with IF, 31.853.

Ranked list of Indian journals

Out of the total of 690 journals cited by the IISc researchers only 16 (2.32%) are Indian journals. Pramana-J. of Physics published by the Indian Academy of Science (IAS) is the highly cited Indian journal, with 40 citations (37.04%). Current Science a multi-disciplinary journal brought out jointly by the Current Science Association (CSA) and Indian Academy of Science (IAS) and Bull. Mater. Sci with 18 (16.67%) and 11(10.19%) citations are the second and third most cited journals respectively.

Of the 16 journals cited by the IISc scholars, Jl. of Indian Institute of Science is the oldest one, started in 1914. The highly cited journals from India are either brought out by the government or from a national academy- the IAS.

Coverage of Indian S&T journals in SCI/JCR

Among the 47 Indian journals covered in JCR-2004, Science Edition, two journals, namely Journal of Biosciences and Journal of Genetics have IF of 1.102 and 1.100 respectively. It is the first time, in the history of JCR, that two Indian journals have crossed the IF of 1.000. Among 49 Indian scientific journals covered in JCR- 2005, only Journal of Biosciences has an IF of more than 1.000. The analysis shows that nine out of 16 Indian journals are covered in the JCR 2004 and in which 3 are Physics journals. It is seen that the Journal of Bioscience having IF of 1.102 is the journal of highest IF and its half-life 5.5 is the lowest half-life of all journals cited. Current Science is the journal of highest immediacy index (0.262).

Language-wise distribution of Journals

Majority of journals cited are in English language and it occupies 498 (72.17%) of total journals. Dutch is the second most preferred language (3.04%) followed by Russian (1.16%).

Year of establishment of Journals

153 journals (22.17%) cited by the scholars are published for about or more than 200 years. Majority of the journals (56.52%) are published during 1951-2000. The analysis also shows that 36 (5.22%) journals are in the period of 1801-1900.

Frequency of Journals

Monthly (12/yr) journals are the most cited journals by the scholars and which account for 169 (24.49%) followed by bi-monthly (6/yr) and quarterly (4/yr) journals with a share of 76 (11.01%) and 51 (7.39%) numbers respectively.

Publishers of Journals

It is interesting to see that Elsevier Science is the leading publisher of Physics journals preferred by the researchers and it accounts for more than 10% of the total publishers. North-Holland and Academic Press are the other two leading publishers of journals

Currency of Journals and Obsolescence

The Physicists cite more current journals with time- span of 0-4 (20.94%) and 5-9 (27.02%) in their theses. IISc scholars quote a much lower percentage of older literature and a much larger % (47.96%) of recent literature, i.e., less than 10 years old.

The mean year of Physics journals of IISc theses is found to be 14.19 and the half- life of journals cited by the physicists are calculated as (median year) 10 years.

Application of Bradford's law

Bradford's law of scattering is employed to study the journal literature distribution and also used for identifying the core journals. When the journals are divided into 3 zones, where 1/3 rd of the total citations covered by each zones. But this journal distribution does not follow the Bradford's distribution pattern and therefore it will not fit the law.

When the Leimkuhler and Egghe's formulae are employed for the verification of Bradford's law of scattering the result is different. Since Bradford assumes that there should be minimum 3 zones, here also it is assumed to be 3 and the value of the Bradford's multiplier is calculated as 12. It is seen that the number of journals in the nucleus is 4 and the mean value of the Bradford multiplier is 12.60. Here the percentage of error is most negligible, and the data will fit well the Bradford law.

The study identifies only 4 journals as the core journals, out of the total of 690 journals. The 3 zones are not exactly the 1/3 rd of total citations as proved by Bradford. There is no exact match in the proportion of number of journals, nor an exact match in the number of papers of each group. When the 80/20 rule is applied to citation data, i.e. 20% of the documents contribute 80% of the total citations and 20% of such documents are usually called 'core documents/core collection'. Here 20.14% (239) of the journals accounts for 88.33% (9998) of the total citations. Thus 80/20 rule not exactly follow the distribution of pattern of IISc journal citations.

Summary and Conclusion

Citation pattern of source document indicates that journal citations have been quoted more (84.67%) which shows journals are, in fact, the best vehicle for communication of information. It further shows that Phys. Rev. B is the most cited journal of IISc Physicists and Pramana is the Indian journal. Only 16 Indian journals (2.32%) are cited by the researchers and out of this 9 journals are covered in JCR 2004. English is the prime language and the USA is the leading country of journal publication. Monthlies are the most cited journals and Elsevier Science is the leading publisher of journals. Half-life of journals is 10 years and mean year of journals is 14.19. The study shows that the journal distribution pattern deviate from Bradfords' law and the 80/20 rule.

Reference

Ambia, Golan (1991) Use of periodicals in Physics in Delhi University Science library, IIT, NPL library: a comparative study. Libri. 41, 2, 98-105.

Towards the Construction of a Global Bibliometric Indicator

Marios Poulos*, Nikolaos Korfiatis** and George Bokos*

*mpoulos@ionio.gr, gbokos@ionio.gr,
Dept. of Archives and Library Sciences Ionian University, Palea Anaktora 49100, Corfu (Greece)

** nk.inf@cbs.dk
Dept of Informatics, Copenhagen Business School. Copenhagen (Denmark)

Introduction

It's conventional wisdom that the availability of scholar related information available on the World Wide Web has been dramatically increased. At the same time the Web becomes the main medium of publication and distribution of information. The results of surveys lead us to the understanding, improvement and the appropriate exploitation of the Web for the scientific research. Methods, techniques and tools originated in the application of bibliometric studies in traditional scientific publications, have affected the majority of bibliometric methods (Bar-Ilan et al. 2002).

The most significant bibliometric techniques which are used in the above applications are: Citation Analysis, Scientometrics, Scholarly Communication Studies, Infometrics, Web-metrics, Social Network Analysis (Bar-Ilan et al. 2002, Bjorneborn et al. 2001, Casserly et al. 2003). However, as Snizek indicates, the problems of the above indicators are focused on a multi-parametric variable number of interdependent values (Snizek, 1995). In this paper we introduce a method in which we evaluate each paper with regard to two parameters. First, with respect to the journal in which this paper has been published and second, with respect to the value of the cited references of this paper. These indicators are presented as follows. For the evaluation of the value of a published journal the impact factor of this journal is used with often some negative examples (Smith, 1997). For the evaluation of each of the cited references of the paper in question we take into account the following indicators: (a) the impact factor of the published journal. (b) The degree of the significant relation (SR) of the scientific area, which is calculated between the cited and the published journals (Gomez et al. 1999). (c) The degree of the significance of the social relation between the authors of the published and the cited references as indicated by metadata schemas (Brickley et al. 2006, Korfiatis et al. 2006).

To this end in this research in progress paper we attempt to take into consideration the above factors in order to conceptualize a theoretic evaluation model. We used a technique of the fuzzy set theory in which all the above best indicators should be added in a general mathematic matrix. Finally, the global indicator of each paper is calculated with regard to the best integral model using the cross-correlation test.

Method

The proposed global matrix creates a sophisticated schema, where each indicator is graduated by five fuzzy logic concepts. In particular, we consider a membership relation $\mu_A(u)=1$ in which "u" is an element of the set A, and $\mu_A(u)=0$ in the case "u" is not an element of the set A. The degree of membership is defined by a through a generalized characteristic function called membership function $\mu_A(u):U\rightarrow[0,1]$, where U is called the universe, and A is a fuzzy subset of U. The values of the membership function are real numbers in the interval [0,1], where 0 means that the category of the bibliometric indexes is not a member of the set and 1 means that it belong entirely (N. Kasabov 1996). One way of defining a membership functions is though an analog function.

The representing value for a fuzzy set transformation is given with a Boolean type value yes (1) or no (0). Thus a combined number of values [0,0,0,0,0] may correspond to the value 0, while a combined number of values [1,1,1,1,1] corresponds to the value 1. The other combination, for example [0,1,1,1,1] or [1,1,0,1,1], etc, gives value 0.75. At this stage we constructed an ideal bibliometric indicators set [1,1,1,1,0] which represents the max fuzzy set value for a citation.

The Implementation of the Method

This section is divided in four stages.

In the first stage, we create a theoretic article, which is represented using a matrix (see equation 1), which has 5xM dimensions, where the M is the maximum number of the published articles which belong to the same thematic area with the investigated paper.

$$A=\begin{matrix}\mu_{A_{11}}(u) & \cdots & \mu_{A_{M1}}(u)\\ \mu_{A_{12}}(u) & \cdots & \mu_{A_{M2}}(u)\\ \mu_{A_{13}}(u) & \cdots & \mu_{A_{M3}}(u)\\ \mu_{A_{14}}(u) & \cdots & \mu_{A_{M4}}(u)\\ \mu_{A_{15}}(u) & \cdots & \mu_{A_{M5}}(u)\end{matrix} \quad (1)$$

Thus, a vector *C=[A]* of *1x5xM* dimension is constructed, which is then used in the next stage. Thus, each column of the proposed matrix represents

the publication, which cites the investigated article. In total, the above matrix depicts a degree of representation of all the citations of the investigated paper with regard to the total number (M) of the publications in the same subject area. In any case, the number of columns is completed. In the empty registries we consider that these are represented by the vector (0,0,0,0,0).

In the second Stage, In this stage, we conceptualize an ideal theoretic article, which is represented using a matrix A_1 (see equation 1).

$$A_1 = \begin{matrix} 1 & . & . & . & . & . & . & . & . & 1 \\ 1 & & . & & & & & & & 1 \\ 1 & & & & . & & & & & 1 \\ 1 & & & & & & & & . & 1 \\ 0 & . & . & . & . & . & . & . & . & 0 \end{matrix} \quad (2)$$

Then the coordinate vector of the above matrix is:

Thus, a vector $\hat{C}_x = [A_1]$ of 1x5M dimension is constructed, which is then used in the next stage.

For example we adopted the number M=10^4. Thus, the vector $\hat{C}_x$ has (1x10^4) dimensionality.

In the third stage, we constructed a matrix, labeled A_2 (see equation. 5) which represents an article, which has been cited four times. The citations of this article are represented by the first (4) four columns and the other columns are completed by (M-4) or (10^4-4) vectors which contain zero elements (0, 0, 0, 0, 0).

$$A_2 = \begin{matrix} 0.25 & 0 & 0.75 & 1 & 0 & . & . & . & . & . & . & . & . & 0 \\ 0.50 & 0.75 & 0.50 & 0.25 & 0 & & . & & & & & & & 0 \\ 0 & 0 & 0.25 & 0.50 & 0 & & & & & . & & & & 0 \\ 0.75 & 0.25 & 0.50 & 1 & 0 & & & & & & & & . & 0 \\ 0 & 0.50 & 1 & 0.25 & 0 & . & . & . & . & . & . & . & . & 0 \end{matrix} \quad (3)$$

Then, the coordinate vector of the above matrix is:

Thus, a vector $\hat{C}_x = [A_2]$ of *1x5xM* or for example 1x10^4 dimensionality is constructed.

In fourth stage the extracted set of auto-correlation coefficients $\hat{C}_x$ of an ideal-cited paper case were submitted to the cross-correlation procedure (Morisson et al 1976) along with another cited paper $\hat{C}_y$ as described below:

$$r = \frac{\sum_{i=1}^{5*10^4-1} (\hat{C}x_i - \bar{\hat{C}}x_i)(\hat{C}y_i - \bar{\hat{C}}y_i)}{\sqrt{\sum_{i=1}^{5*10^4-1} (\hat{C}x_i - \bar{\hat{C}}x_i)^2 (\hat{C}y_i - \bar{\hat{C}}y_i)^2}} \quad (4)$$

The extracted cross-correlation coefficient is a number between -1 and 1, which measures the degree to which two variable sets are linearly related.

Conclusions and Further Research

In this paper, we attempted to produce a global bibliometric index for a published article. Thus, we constructed a global ideal article with (5) five high score indicators for each citation. For this implementation, we used the fuzzy logic theory in order to classify in (5) five degrees the bibliometric indicators. Finally, we extracted a bibliometric indicator for a simple article by correlating this with the ideal article. This paper may be considered as the first attempt to construct the proposed IGBI indicator and thus the research presented here should be considered as research in progress. Our major concern is to evaluate the IGBI index against well known comparisons such as journal status (Bollen et al. 2006) using an extensive citation graph. Furthermore, the extension of a semantic vocabulary such as COAP and the construction of a citation index to provide real time IGBI analysis is also an important step.

References

Bar-Ilan, J. & Peritz, B. (2002). Informetric theories and methods for exploring the internet: an analytical survey of recent research literature. *Library Trends*, 50 (3), 371-392.

Bjorneborn, L. & Ingwersen, P. (2001). Perspective of webometrics. *Scientometrics*, 50 (1), 65-82.

Brickley, D. and L. Miller.(2006). FOAF Vocabulary Specification" Retrieved July, 2006 (http://xmlns.com/foaf/0.1/).

Bollen, J., Rodriguez, M.A., Van de Sompel, H., (2006). Journal Status, *Scientometrics*, 69, (3)

Casserly, M. & Bird, J. (2003). Web citation availability: analysis and implications for scholarship. *College & Research Libraries*, 64 (4), 300-317.

Gomez, I. &, Maria Teresa Fernandez and Jesus Sebastian (1999). Analysis of the structure of international scientific cooperation networks through bibliometric. *Scientometrics*, 44(3) 441-457.

Korfiatis, N., Poulos, M., & Bokos, G. (2006). Social Metadata for the Impact Factor. *The Electronic Library (TEL) Journal.* (Forthcoming)

Morrison, N., Donald, F., 1976. Multivariate Statistical Methods, (New York; McGraw-Hill Book Company,)

N. Kasabov (1996). Foundations of Neural Networks, Fuzzy Systems, and Knowledge Engineering, MIT Press,. Fuzzy systems. (sections 3.1.1-3.1.2, pp.167-169).

Smith R. (1997). Journal accused of manipulating impact Factor, *British Medical Journal*, 314, N. 7079 , 463-466.

Snizek W. E. (1995). Some observations on the use of bibliometric indicators in the assignment of university chairs. *Scientometrics* 32(2), 177-120.

A Bibliometric Analysis of Book Based Literature

Hampus Rabow

hampus.rabow@lub.lu.se
Lund University, Centre for Bibliometrics and Research Evaluation, Box 134 221 00 Lund (Sweden)

Background
Some areas of research, including most of the humanities, a large part of the social sciences, and many technological disciplines, have presented considerable challenges for bibliometric analysis. Such an analysis is normally based on citations in journal articles, whereas most of the research results in these areas are published in reports, book chapters and monographs.

It is of course possible to trace citations to non-journal literature in journals, but it seems rather likely that such a skewed sample will introduce various forms of bias. It is, for example, quite conceivable that a work could be highly cited in monographs and yet rarely occur in the reference lists of journal articles.

Aims and methodology
In a project at Lund University, we have developed some algorithms to extract references, citations and significant concepts from a corpus consisting of monograph dissertations in the humanities and social sciences in Sweden. Most of this corpus is Open Access, which has facilitated the project and increased its applicability.

There are several aims that the project hopes to achieve. One is to define various metrics to define proximity between literature. Another is to discover various types of standard texts. The types or categories are based on the context of citation in terms of the standard disposition of scholarly texts. A third aim, which is closely linked to the second, is to define metrics of research impact for fields with complex patterns of progress. In some fields in the humanities, for example, there may be little consensus about the what the main achievements are. Different schools of though may use rather different standards and apply different criteria. A fourth aim is to compare the citation impact in book literature with the impact in journal literature, and specifically, to compare citations to books in journals with citations to books in books.

Results
The project is still not completed at the time of writing, and thus most of the results are fairly preliminary. More definite results will, however, be presented at the Madrid conference. The algorithms are still being improved, but they have already provided some useful results.

The vast majority of references are identified correctly and most of the citations are also correct. The main problem at this point is to avoid a bias in the result because citations to some forms of literature are more difficult to identify. We are doing manual checks of random samples to see how big this problem is. One, not entirely satisfying, possibility is to use these manual checks to calculate bias score and then use these scores to correct the results.

We have discovered several interesting metrics of proximity. One makes use of the main data sources, such as e.g. archives, statistical collections, and bodies of literature. Another tries to identify the main methodological or theoretical background by means of references to standard works or authors. As for citation impact, the results are still too preliminary to be discussed here. One difficulty is that besides a fairly limited set of "core literature", the bibliographic coupling, even between closely related works, can be very weak.

Discussion and next step
During the late 1990s a number of information specialists began to mine web publications for citations. One of the most successful examples was CiteSeer developed by C. Lee Giles, Steve Lawrence and others at NEC Research Institute. The early approaches were fairly simple, often relying on a limited set of simple rules to identify citation elements. The original rules were improved and adapted through some forms of machine learning..

A frequently used approach has been to supply a list of templates with simple rules describing possible elements of a citation. Gobinda Chowdhury and others have argued that sufficiently well structured citation template would be sufficient to automatically create reliable citation databases. This approach has recently been criticised by Min-Yuh Day and others. They argue that rule-based models become too difficult to deploy, since they are dependent on the domain and require frequent tweaking by experts.

Our system is template-based, but only a few generic templates are supplied at the outset. These generic

templates are used mainly to verify the validity of the other templates automatically extracted from the texts.

Our parsing algorithms are not intended to be universally applicable, but to solve the specific problems associated with monograph literature. Our immediate tasks are to solve the bias problem, to find reliable ways to cluster the "core literature", to achieve higher accuracy in identifying references.

We later also intend to expand the body of literature to be parsed, and especially investigate if there are differences between full-length books written by the same author(s), edited volumes where each chapter is written by different authors book reviews in journals and research papers in journals .

Acknowledgments
The project is financed by the Swedish Royal Library, and has benefited from discussion in the OACI group.

References

Archambault, É., Vignola-Gagne, É., Côté, G., Larivière, V., & Gingras, Y. (2006). Benchmarking scientific output in the social sciences and humanities: The limits of existing databases. Scientometrics, 68(3), 329-342.

Day, M.-Y., Tzong-Han Tsai, R., Sung, C.-L. & al. (2007). Reference metadata extraction using a hierarchical knowledge representation framework. Decision Support Systems, 43, 152-167.

Goodrum, AA, McKain, KW, Lawrence S & al. (2001). Scholarly publishing in the Internet age: a citation analysis of computer science literature. Information Processing and Management, 37(5), 661-675

Hicks, D. (1999). The Difficulty of Achieving Full Coverage of International Social Science Literature and the Bibliometric Consequences. Scientometrics, 44(2), 193-215.

Kostoff, R. N., & Martinez, W. L. (2005). Is citation normalization realistic?. Journal of Information Science, 31(1), 57-61.

Lawrence S. Giles C.L., Bollacker K. (1999). Digital libraries and autonomous citation indexing. Computer, 32(6), 67-71.

Lawson M, Kemp N, Lynch MF, & Chowdhury GG. (1996). Automatic extraction of citations from the text of English-language patents - An example of template mining. Journal of Information Science, 22(6), 423-436.

Nederhof, A. J. (2006). Bibliometric monitoring of research performance in the Social Sciences and the Humanities: A review. Scientometrics, 66(1), 81-100.

Van Raan, A. F. J. (7/1996). Advanced bibliometric methods as quantitative core of peer review based evaluation and foresight exercises. Scientomterics, 36(3), 397-420.

Van Raan, A. (1998). In Matters of Quantitative Studies of Science the Fault of Theorists is Offering too Little and Asking too Much. Scientometrics, 43(1), 129-139.

Research in Doñana National Park: Preliminary data

Cristina Ramo

cristina@ebd.csic.es
Estación Biológica de Doñana CSIC, Avda. Mª Luisa s/n, Pabellón de Perú, 41013 Sevilla (Spain)

Introduction

Doñana National Park, created in 1969, is located on the southwestern coast of the Iberian Peninsula and covers over 54,252 ha including four large ecosystems: beach, dunes, scrub-woodland, and marshes. This park provides the scientific community with an exceptionally large and biodiverse study area with 21 different biotopes, 803 species of vascular plants (including 48 endemics to Spain, and four to the Doñana area), seven freshwater fish, 30 estuarine fish, 12 amphibians, 19 reptiles and 29 mammals. The avifauna, with 361 species, represents 70% of all the species present in Europe, of which 119 species breed regularly in Doñana. In winter, up to 700.000 waterbirds concentrate in the marshes, making Doñana one of the most important European wetlands. Amongst the vertebrates there are numerous threatened species such as the Iberian Lynx (*Lynx pardina*) and Spanish Imperial Eagle (*Aquila adalberti)*, species endemic to the southwestern part of the Iberian Peninsula.

Research is one of the priority objectives for the declaration of Doñana National Park as established by the Doñana Law (Law 91/1978, of 28 of December), and the Director of the Doñana Biological Station, a research institute belonging to the Spanish Research Council (CSIC), is responsible for the coordination of the research programmes to be developed in the Park. The scientists interested in carrying out research in the Park must send their proposals to the Research Coordination Office in Doñana Biological Station. The present poster analyzes the research developed in Doñana National Park during the period 1978-2005 in the Spanish context.

Methods

The analyses are based on the databases of research projects and publications managed by the Research Coordination Office (Doñana Biological Station, CSIC). These databases are permanently updated with information provided by researchers and through searches in international (ISI Web of Science and Wildlife & Ecology Studies Worldwide) and national (ICYT and ISOC) databases.

Results

During these 28 years, a total of 410 research activities were carried out (300 research projects, 90 short visits and 20 monitoring projects). Research projects are of broad scope, generally three years long. Short visits of one or several days deal with very specific issues. Finally, monitoring projects are meant for researchers interested in collecting specific physical or biological data (depth of water tables, counts of a specific species, meteorological data, etc.) to obtain long series of data, but outside the framework of a research project.

The main subjects of the research activities were: birds (94 investigations), flora (86), mammals (80), invertebrates (77) and hydrology (46). Each research can cover more than one subject.

Considering the institutions responsible for the investigations, 37% were Spanish universities, 29% Doñana Biological Station, 14% foreign institutions, 11% other CSIC institutes, and 9% other Spanish institutions.

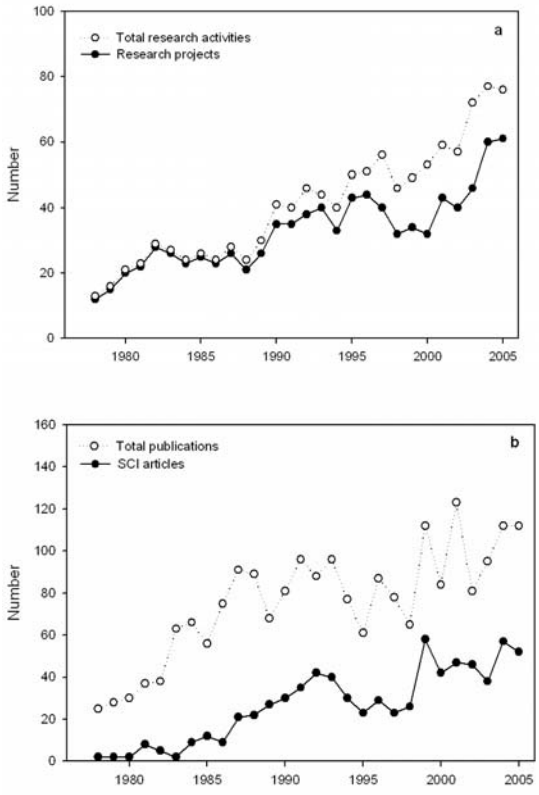


Figure 1. Evolution of the number of research activities (a) and publications (b) over the period analyzed.

The output of these research were 2114 publications: 738 articles in SCI journals, 1 article in a SSCI journal, 435 articles in ICYT journals, 16 articles in ISOC journals, 280 articles in other journals, 161 books and 483 book chapters.

Although the annual number of publications increased from 25 in 1978, to 112 in 2005 (X4.48), not all the different publications had the same performance. While SCI articles increased 28.5 times and book chapters 13.5, ICYT and other articles remained stationary, and books presented a slight increase.

We define the annual productivity as the ratio obtained by dividing the number of publications by the number of projects in the same year. There is an increase in productivity along the years if we consider the SCI articles ($r_s = 0.70$, $n = 28$, $p < 0.001$). On the other hand, the productivity does not increase if we consider the total number of publications ($r_s = 0.005$, $n = 28$, $p = 0.981$).

Considering the ranking of journals by category subject in descending order of impact factor, 41% of SCI articles were in journals positioned in the first quartile, 31% in the second, 18% in the third and 10% in the fourth.

Table 1. Evolution of the percentage of SCI articles (over total publications), average of impact factor and percentage of SCI papers in the first quartile over the period analyzed.

Period	% SCI	Average IF	% Q1
1978-1981	11.7	0.945	14.3
1982-1985	12.6	0.879	28.6
1986-1989	24.5	0.867	17.7
1990-1993	40.7	0.933	34.0
1994-1997	34.3	1.188	32.7
1998-2001	45.1	1.407	39.3
2002-2005	48.3	1.747	53.4

The main categories of the SCI articles were: zoology (excluding birds, 178 articles), ecology (174), environmental sciences (168), ornithology (73), marine and freshwater biology (60), biodiversity conservation (44), plant sciences (43), toxicology (34) and geosciences multidisciplinary. (Each article can be in more than one category subject).

The results will be compared to publication trends observed in other scientific areas in Spain (Jiménez, Moya and Delgado, 2003; Grupo de Bibliometría, 2006).

Conclusions

We can conclude that research conducted in Doñana National Park has experienced a remarkable increase in both activity and productivity (SCI articles) along the last 28 years (figure 1). Besides, a trend towards internationalisation is observed in the outputs (higher percentage publications in SCI), as well as higher visibility (increase in average IF of SCI papers, higher percentage of documents in first quartile) (Table 1).

The results presented here are part of the assessment exercise of Doñana National Park. They will be completed and used as a tool for science policy management.

Acknowledgements

The author thanks Rocío Astasio, Franca Jordá, Héctor Rodríguez and Mara Sempere for updating databases.

References

Grupo de Bibliometría (2006). *La investigación del CSIC a través de sus publicaciones científicas de difusión internacional* (1981-2003). Madrid, CINDOC.

Jiménez, E., Moya F.D. & Delgado E. (2003). The evolution of research activity in Spain: the impact of the National Commission for the Evaluation of Research Activity (CENEAI). *Research Policy*, 32, 123-142.

The Evolution of the Sleep Scientific Literature over 30 Years: A Bibliometric Analysis

Claude Robert[*], Concepción S. Wilson[**], Jean-François Gaudy[*], and Charles-Daniel Arreto[*]

[*]*claude.robert@univ-paris5.fr, charles-daniel.arreto@univ-paris5.fr, jean-francois.gaudy@univ-paris5.fr*
Université Paris Descartes, Laboratoire d'Anatomie Fonctionnelle, 1 rue Maurice Arnoux,
92120 Montrouge, (France)

[**]*c.wilson@unsw.edu.au,*
School of Information Systems, Technology and Management, University of New South Wales,
Sydney, New South Wales, (Australia)

Introduction

During the last few decades, the evolution of our society has placed enormous pressure on the natural human sleep process, resulting in numerous sleep disorders that have important physiological, public health and socio-economic consequences. Recently, two studies (Robert et al. 2006a,b) introduced a one-year snapshot of the sleep scientific literature in medicine and biology, focusing on the European Union productivity. By extending these two studies, this paper provides a more comprehensive and global view of the growth and development of sleep research activity for the past 30 years.

Method

The raw data for this study were obtained in December 2005 from the Excerpta Medica database EMBASE using the Dialog information retrieval search service for the years 1974 to 2004 inclusive. To retrieve all aspects of the sleep-related literature, a broad set of thesaurus terms from EMTREE (http://www.embase.com/emtree) were searched along with their equivalent 'free-text' terms. Data collected from the online searches were parsed and manipulated in Excel spreadsheets to provide the input for the various figures and tables.

Results

Over the four time periods studied, while the total number of publications doubled (2.2), that of the sleep literature quadrupled (4.1) (Figure 1).

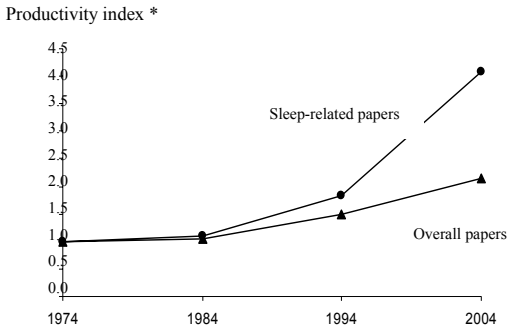


Figure 1. Growth of the Sleep literature in EMBASE over thirty years: 1974-2004. *The productivity index is obtained by dividing the number of papers of the years under consideration by the corresponding number of papers published in 1974

The list of the most prolific countries with publications on sleep in four time periods are displayed in Table 1.

	1974		1984		1994		2004
Country	**Number of papers** *(Rank)*		**Number of papers** *(Rank)*		**Number of papers** *(Rank)*		**Number of papers** *(Rank)*
United States	847 *(1)*		1034 *(1)*		1450 *(1)*		3106 *(1)*
United Kingdom	202 *(3)*		234 *(2)*		379 *(2)*		708 *(2)*
Germany	146 *(4)*		159 *(4)*		284 *(3)*		520 *(3)*
Italy	95 *(6)*		96 *(7)*		179 *(7)*		400 *(4)*
Japan	110 *(5)*		138 *(5)*		257 *(5)*		396 *(5)*
France	210 *(2)*		160 *(3)*		263 *(4)*		362 *(6)*
Canada	85 *(7)*		104 *(6)*		194 *(6)*		343 *(7)*
Spain	16 *(21)*		17 *(20)*		73 *(12)*		282 *(8)*
Australia	41 *(12)*		60 *(9)*		110 *(8)*		234 *(9)*
Brazil	49 *(10)*		52 *(10)*		82 *(10)*		218 *(10)*
The Netherlands	22 *(18)*		52 *(10)*		77 *(11)*		182 *(11)*
China	-		1 *(48)*		9 *(32)*		152 *(12)*
Switzerland	60 *(8)*		70 *(8)*		91 *(9)*		147 *(13)*
Sweden	39 *(13)*		42 *(12)*		72 *(13)*		125 *(14)*
India	30 *(14)*		13 *(23)*		23 *(22)*		120 *(15)*
Belgium	30 *(14)*		38 *(13)*		68 *(14)*		111 *(16)*
Turkey	4 *(35)*		2 *(43)*		15 *(26)*		103 *(17)*
Israel	9 *(28)*		25 *(15)*		48 *(16)*		87 *(18)*
Finland	19 *(19)*		30 *(14)*		62 *(15)*		69 *(19)*
Poland	29 *(16)*		11 *(25)*		21 *(23)*		63 *(20)*
Argentina	26 *(17)*		22 *(18)*		35 *(18)*		62 *(21)*
Taiwan	-		2 *(43)*		14 *(14)*		59 *(22)*
Austria	13 *(23)*		24 *(16)*		33 *(19)*		57 *(23)*
New Zealand	7 *(31)*		8 *(29)*		19 *(24)*		54 *(24)*
Mexico	6 *(32)*		11 *(25)*		16 *(25)*		50 *(25)*

Table 1. Outputs of the main countries involved in Sleep research during the last three decades

While 59 countries published papers about sleep in 1974, in 1984 there were 65 countries, 76 in 1994, and 96 in 2004. In all the four periods, the US ranks first with consistently over one-third of the articles; and for the 2004 ranking, the US is followed by the UK, Germany, Italy, Japan, France and Canada. The production of sleep literature by these countries has increased by nearly 3.5 between 1974 and 2004;

additionally, these seven leading countries remained relatively constant in all the time periods. Throughout all the four periods, about 45% of the contributing countries published 5 or fewer articles.

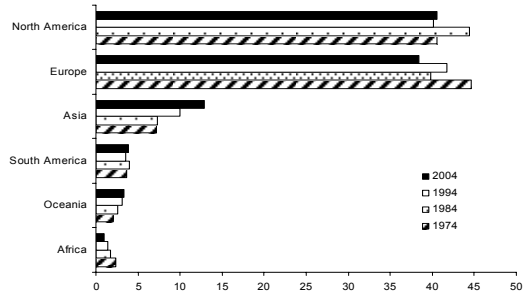


Figure 2. Geographic distribution of the sleep literature for four time periods.

In all periods, publications were concentrated in the two main industrialized areas: North America and Europe (around 40% for each region), with the leadership alternating between the two (Figure 2). A third region, Asia has merged showing continuous growth, from 7.2% in 1974 to 12.8% in 2004). Though significantly lower numbers of articles in sleep research, Oceania has also shown continuous growth, from 2.0% in 1974 to 3.3% in 2004. South America has remained constant between 3.5 to nearly 4%. Conversely, the productivity of Africa has continued to decline from slightly over 2% in 1974 to just under 1% in 2004.

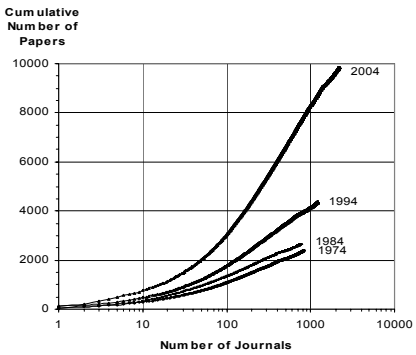


Figure 3. Bradford distributions of the number of sleep-related papers over journals for four time periods

Figure 3 shows that during the first two periods (1974 and 1984), the number of journals remained stable (828 in 1974, and 777 in 1984). In the second two periods (1994 and 2004) the numbers increased dramatically (1231 in 1994, and 2186 in 2004).
The trend lines of the USA and EU have similarly evolved during the last three decades (Figure 4). Both remained at a constant level from 1974 to 1990 (around 1000 publications/year). From 1990 to 2004, both grew dramatically resulting in a more than three-fold increase in productivity (above 3000 publications in 2004).

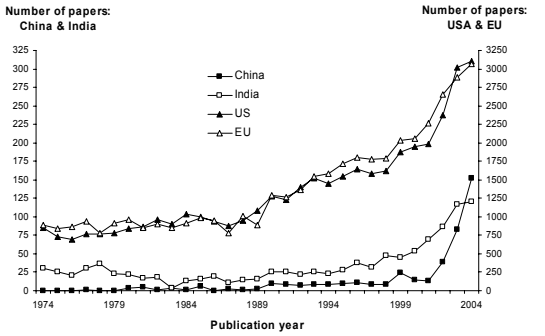


Figure 4. Growth of publications for the two leading research communities, the USA and the EU, and for two emerging countries, China and India.

India has had a constant low level of sleep publications (around 25 publications/year) from 1974 to 1994; however there was a rapid growth from 1995 to 2004 with 120 publications. China has evolved spectacularly: up until 1990, it had only a few publications per year; in the 1990s it had about 10 publications per year; however, it exploded in the 2000s reaching 152 publications in 2004.

Discussion

The sleep literature in the EMBASE database started to take-off in the mid 1980s, mainly due to the widespread introduction of computer technology during this period.
Although the major sleep research countries have not changed substantially during the period examined, the emergence of new countries such as China (anticipated to be one of the five major actors in sleep research in the future), India, Turkey and Spain should be noted.
The investment of the EU in over 30 projects either focusing on sleep or in which a link with sleep is stated (Robert et al. 2006b); the strong tradition of sleep research in North America; the continuous creation of journals dedicated to sleep (most began in the last two decades), the development of national and international professional societies focusing on sleep, and the periodic organization of congresses focusing on sleep are convincing evidences of the importance accorded to sleep by the international scientific community

References

Robert C.,Wilson C.S.,Gaudy J.-F. & Arreto C.-D., (2006a) A year in review: Bibliometric glance at sleep research literature in medicine and Biology. *Sleep & Biological Rhythms*, 4: 160-170.

Robert C., Wilson C.S.,Gaudy J.F. & Arreto CD (2006b), A snapshot of EU publications in sleep research: A scientometric survey. *Scientometrics*, 67(3): 385-405.

Is the European Web British or American?

Gwendolyn Rogge and Ronald Rousseau

gwendolyn.rogge@khbo.be, ronald.rousseau@khbo.be
KHBO (Association K.U.Leuven), Industrial Sciences and Technology
Zeedijk 101, B-8400 Oostende (Belgium)

Goal
To find out whether residents of different European countries prefer using British English (BE) or American English (AE) on the Web.

Methods
We have collected terms, words or phrases which we thought to be typically British, and this together with their typically American synonyms. If necessary, we have added context words in order to delineate the precise meaning of the term.

These terms were entered in Exalead (and Google as a control). Preference was given to Exalead as its reported numbers of sites are more stable than Google's. As a rough criterion we considered terms for which the ratio BE/AE was larger than 2 in .uk sites and at the same time smaller than 0.3 in .edu sites (and .gov as a control) as terms that distinguish between British and American usage. Here BE/AE refers to the ratio of the number of sites using the British variant of the word and the number of sites using the American one. Please note though that we have used the plough-plow pair (Table 1) which strictly speaking does not satisfy this criterion.

We have utilised the following ten pairs of queries, shown in the order BE-AE (Table 1).

Table 1. Pairs of queries

Terms: BE - AE	BE/AE in .uk	BE/AE in .edu
behaviour – behavior	5.31	0.09
colour – color	4.11	0.06
centre – center	3.32	0.04
plough – plow	6.99	0.54
aluminium – aluminum	6.43	0.08
CV AND name - resume AND name	2.34	0.26
aubergine – eggplant	4.66	0.03
"waste bin" - " trash bin"	14.85	0.25
windscreen AND car - windshield AND car	6.69	0.05
"mobile phone" - ("cell phone" OR "cellphone")	13.30	0.15

Next, these terms were used for conjunctive queries (AND) in Exalead with site:"country indicator", e.g. site:de, or site:be. We have calculated the BE/AE ratio for each country and each term.

Hypothesis
Before performing the experiments, we assumed that countries would follow either British style or American style. We further assumed that Western European countries (including the Nordic countries) would follow British style, while Eastern European countries would rather follow American style (being less formally educated in Queen's English).

Results
It turned out that the investigated countries rarely follow one style uniformly. Hence it became rather difficult to draw the line between BE-style countries and AE-style countries. Using different criteria, with the highest weight given to the median BE/AE ratio, we obtained the following list.

The following countries follow British style: the United Kingdom (.uk), Ireland (.ie), France (.fr), Switzerland (.ch) and Finland (.fi). Also 'Europe', i.e. .eu-sites, follows British style.

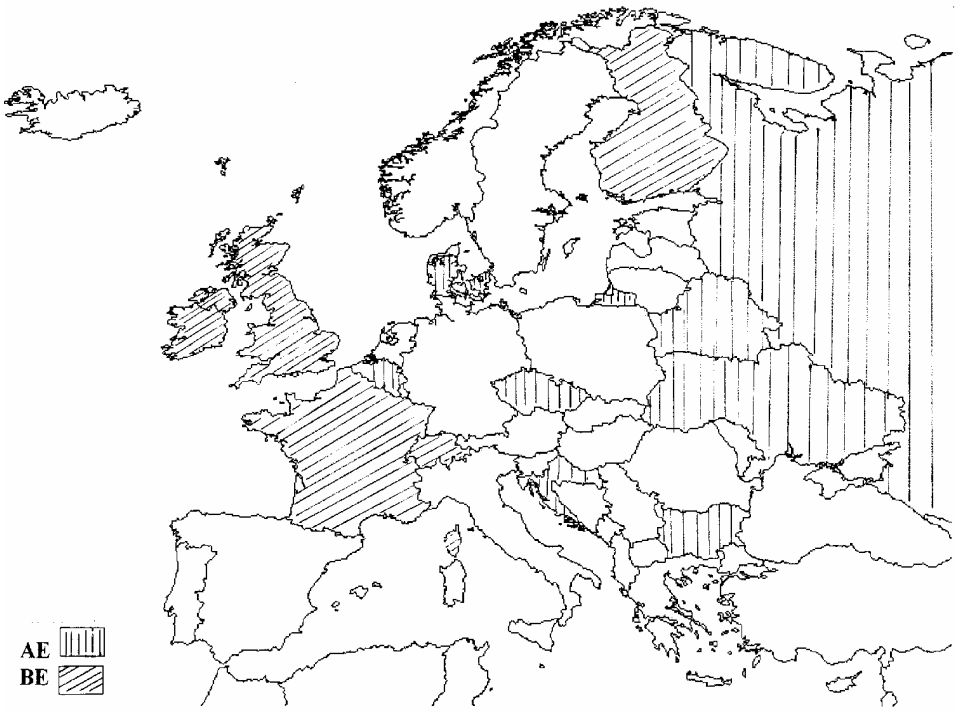
Countries following American style are (besides the United States itself): Russia (.ru), Ukraine (.ua), Belgium (.be), Bulgaria (.bg), Czech republic (.cz), Croatia (.hr), Belarus (.by) and Denmark (.dk).

Surprisingly, Denmark and Belgium belong to the latter group. The other European countries which we investigated have no specific preference. This group consists of: Norway, Sweden, the Baltic republics, Spain, Slovenia, Austria, Germany, Portugal, the Netherlands, Poland, Iceland, Hungary, Romania, Italy, Albania, Greece and Bosnia-Herzegovina.

Among the used terms it is found that Europeans mostly use the term aubergine and rarely eggplant (yet, Australia uses the term eggplant), aluminium and almost never aluminum, plough and almost never plow. Yet, it seems that except for the British all Europeans write color and not colour, and talk about windshields instead of windscreens.

Conclusion
The results of our small test are rather mixed. European English is not British English, but Europe is not Americanised either. Maybe it can be said that Europe uses English as a lingua franca, i.e. a language derived from British English, but with terms borrowed from whatever regional variant the

user is most comfortable with. The main results of this test are shown on below map of Europe .



Quantitative CV-Based Indicators for Research Quality, Validated by Peer Review

Nadine Rons and Arlette De Bruyn

Nadine.Rons@vub.ac.be, Arlette.De.Bruyn@vub.ac.be
Research Coordination Unit, R&D Department, Vrije Universiteit Brussel, Brussels (Belgium)

Introduction

In a university, research assessments are organized at different policy levels (faculties, research council) in different contexts (funding, council membership, personnel evaluations). Each evaluation requires its own focus and methodology. To conduct a coherent research policy however, data on which different assessments are based should be well coordinated. A common set of core indicators for any type of research assessment can provide a supportive and objectivating tool for evaluations at different institutional levels and at the same time promote coherent decision-making. The same indicators can also form the basis for a 'light touch' monitoring instrument, signalling when and where a more thorough evaluation could be considered.

This poster paper shows how peer review results were used to validate a set of quantitative indicators for research quality for a first series of disciplines. The indicators correspond to categories in the university's standard CV-format. Per discipline, specific indicators are identified corresponding to their own publication and funding characteristics. Also more globally valid indicators are identified after normalization for discipline-characteristic performance levels. The method can be applied to any system where peer ratings and quantitative performance measures, both reliable and sufficiently detailed, can be combined for the same entities.

Method

From ex post peer review evaluations of research teams by international expert panels, two types of data sets are available: peer ratings on a series of aspects and quantitative performance measures for a series of standard output categories. Linear correlations between the two are calculated. Before doing so two types of normalizations are performed:

- Normalization for team size: The peer ratings are size-independent. To also create size-independent performance measures, these are calculated per full time equivalent leading staff (potential principal promoters of research projects).
- Normalization for discipline-characteristic performance levels: Both peer ratings and quantitative performance measures are normalized per discipline (aligning mean values and standard deviations), to correct for discipline-dependent evaluation and performance levels.Performance measures that are significantly positively correlated with peer review results are selected as indicators for research quality. In order to avoid accidental occurrences, performance categories that are only present in the output of a minority of the teams are not taken into account.

Material

Data & Research Disciplines

Data were examined for a series of disciplines which were evaluated using the same standard methodology and which included a sufficient number of teams (involving evaluations finalized from 2000 to 2006).

Performance measures investigated in this context until now are different types of publications and project funding. These correspond to categories in the central research database of the university, also used to automatically extract researchers' CV's.

The quantitative peer review results include an overall evaluation as well as scores on scientific merit, planning, innovation, team quality, feasibility, productivity and scientific impact.

Key Figures

- 6 research disciplines and expert panels: Economics, Engineering, Informatics, Law, Philosophy & Letters, Political & Social Sciences
- 57 evaluated teams, 9 to 11 teams per discipline
- 263 full time equivalent postdoctoral level staff
- 63 experts from 11 countries
- 427 returned evaluation forms
- 8 peer review indicators
- 23 scientific publication categories from the university's CV-format + ISI-category
- 21 external project-funding categories

Reliability

To obtain significant correlations between results from different evaluation systems is not evident, as discussed by Moed (2005). Evaluations are designed to support particular decisions (e.g. funding) and do not necessarily consider aspects outside their focus, which however may be important in other evaluations. The peer review method used for this analysis produces peer ratings for a broad series of aspects and contains several precautionary measures to ensure reliable results (confidentiality, panel procedure, site visit, bias verification). It was designed in 1996-1997 taking into account recommendations and known

problems following from earlier experiences as much as possible (Cozzens, 1997; Kostoff, 1997; Martin, 1996).
Reliability of the quantitative performance measures is ensured by data collection (for the files presented to the experts) in close collaboration between the central research administration and the teams.

Findings & Discussion
Generally valid correlations
Some publication categories are significantly positively correlated with peer ratings, globally as well as for almost all disciplines separately (without any significant negative correlation coefficients). This is observed for:
- "Articles in journals with international referee-system" and
- "Publications in journals indexed by the Thomson Scientific SCIE, SSCI or AHCI" (ISI-publications, largely overlapping with the previous category),
- Followed by "Communications at international conferences integrally published in proceedings".
This shows that even in domains where books are a prominent form of output, international, peer reviewed journal publications are a good indicator for research quality at team level.
Figure 1 shows how higher correlation coefficients are obtained after normalization per discipline (all disciplines included except Law for which different publication categories were used).

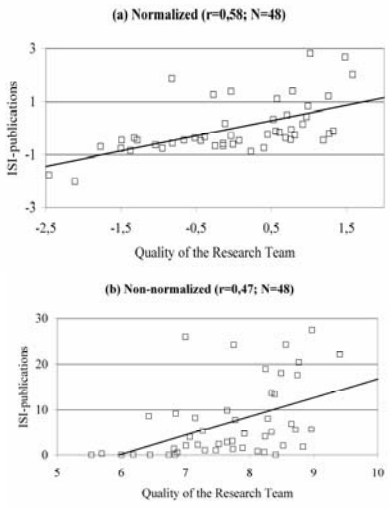


Figure 1. ISI-publications vs. peer ratings for "Quality of the Research Team".

Differences between disciplines
Disciplines differ in the particular categories of publications or project funding which are significantly correlated with peer ratings. These are in line with discipline-dependent typical funding channels (e.g. for applied or policy oriented research).
Differences between performance categories
Correlations with peer ratings differ for related performance categories, such as publications in journals with international, national or without referee system. It is therefore important to be able to distinguish between sufficiently fine categories in order to select appropriate indicators for evaluation. Broad performance categories may merge important performances with less important or even counterproductive ones. Obtaining significant correlations with such "mixed" performance measures is less evident and using them as indicators could be rewarding the wrong performances.

Conclusions & Further Research
This study for a first series of six disciplines shows that correlations between peer ratings and performance measures allow identifying core performance indicators, per research discipline as well as for larger research domains. Such a set of core indicators can be used as a common supportive tool for different kinds of evaluations, or it can be used in a monitoring instrument.
For evaluation purposes, the core performance indicators should be accompanied as much as possible by international reference values per discipline. International reference values however will not be available for locally defined performance categories. If also no national or regional reference values are available, averages within the institution could be constructed, provided a sufficiently large population is available.
Of course, while certain performance indicators may in general be related to quality as seen by peers, this does not necessarily imply these indicators' ability to distinguish between performances of individual researchers or even teams, unless correlations are perfect. Therefore, in the framework of an evaluation, interpretation of indicators by a committee remains necessary.
Future work will include an extension of the set of core performance indicators towards other disciplines (after results of their evaluations become available) and an investigation on reference values.

References
Cozzens S.E. (1997). The Knowledge Pool: Measurement Challenges in Evaluating Fundamental Research Programs. *Education and Program Planning, 20*(1), 77-89.
Kostoff R.N. (1997). *The Handbook of Research Impact Assessment* (7th ed., DTIC Report Number ADA296021). United States.
Martin B.R. (1996). The Use of Multiple Indicators in the Assessment of Basic Research. *Scientometrics, 36*(3), 343-362.
Moed, H.F. (2005). Peer review and the use and validity of citation analysis. In *Citation analysis in research evaluation* (chap. 18). Dordrecht: Springer

Scientific Portfolio Analysis of a Scientific Area by a Competitive Position Approach

Hervé Rostaing, Nicolas Barts, Valérie Léveillé

herve.rostaing@univ-cezanne.fr, nicolas.barts@univ-cezanne.fr, valerie.leveille@univ-cezanne.fr
Université Paul Cézanne Aix Marseille III, LSIS UMR 6168, Campus Saint Jérôme, F-13397 Marseille (France)

Introduction

The bibliometric study was carried out in 2006 for an Advisory Scientific Committee (Comité d'Orientation Scientifique) organized by the 3 universities of Aix-Marseilles. The aim of such a Scientific Committee was not to carry out a research assessment of the three universities or its actors, but much more to analyze the research potential and express recommendations on the scientific orientations to improve their research activities.

To help them, several instruments were put at their disposal: presentation documents of the research organization in the universities and their research groups; gateway on Internet presenting all these research groups and the administrative structures on which they depend... Among these instruments, the steering committee of this event considered that it was necessary to propose a bibliometric study of the three universities' scientific activities.

Methodology

The bibliometric study was intended to be an instrument for helping in the strategic analysis of the three universities' scientific policy. Thus, the methodology employed for this bibliometric study was strongly inspired by strategic analysis methods carried out in companies.

Strategic analysis process

The greatest part of these strategic analyses is founded on the principles of economic competition. The whole of the approaches of these company strategy analyses is usually presented according to *a strategic analysis process* realized in four stages:

-Identification of the Strategic Business Units: these SBU correspond to the unit of analysis.

-Choice of the competitive analysis criteria for each SBU: the chosen criteria are considered as the key factors of economic success for the strategic activity fields.

-The competitive position of each SBU: performance assessment of its company to those of its main competitors according to the criterion or the criteria chosen for each SBU.

-Portfolio management: this last stage seeks to offer a common representation of the strategic position of the company as a whole, judged by its SBU portfolio assessment. This common representation of the SBU portfolio is constructed as a matrix or a "business grid" according to two indicators: competitive position and an indicator showing the position in the life cycle (dynamism notion of the SBU).

The methodology of the bibliometric study follows the strategic analysis process to offer a "competitive" position view of the Aix - Marseilles universities.

Definition of the strategic scientific units

The steering committee considered that the research fields covered by the various Doctoral Schools (*Ecole Doctorale*) represent perfectly the scientific activities of the three universities.

In the Aix-Marseilles area, twelve Doctoral Schools were accredited to the three universities: seven include sciences and technology fields and five social sciences and humanities fields.

Selection of competitors: the set of scientific areas

The steering Committee singled out 7 French areas (Bordeaux, Grenoble, Lyon, Montpellier, Nice, Strasbourg, Toulouse) and 5 European areas (Barcelona, Leeds, Milan, Tübingen, Turin).

This arbitrary choice is mainly based on strategic positioning expectations from the Steering Committee.

Even so, this choice tried to respect two criteria:

-limitation of the size effect by selecting scientific areas having a scale rather similar to the scale of the Aix-Marseilles area

-keeping a scientific coherence by selecting scientific areas having a spectrum of scientific activities characteristic of the Aix–Marseilles area.

Evaluation of the dynamics over the years of strategic scientific units

For evaluating the evolution of the activities of the 13 areas in the scientific fields, the data were collected over a 24-year period (1981-2004) divided into 4-year windows.

Measuring criterion of the scientific activity

The scientific activity of a research center can take many forms. In this study, only the activity of academic research will be taken into account as measuring criterion of the scientific activity. Furthermore, this activity of academic research was restricted to only the production of scientific contributions indexed by the databases produced by Thomson-ISI (SCI, SSCI and A&HCI).

The bibliometric study could not take into account the scientific fields corresponding to the Doctoral

Schools of social sciences and humanities because of the lack of reliable and accurate data from these databases. In the end, only strategic scientific units covered by the 7 Doctoral Schools in sciences and technology fields could be used for the competitive position analysis: Mathematics and computer science; Physics and materials sciences; Physics, Modeling and engineering science; Environmental sciences; Chemical Sciences; Sciences of the Human Movement; Health and Life Sciences.

The portfolio of the strategic scientific units

By analogy with the strategic matrices or the business grids developed by consultancy firms, the portfolio of the strategic scientific units of the Aix-Marseille area is graphically represented according to two indicators (fig 1).

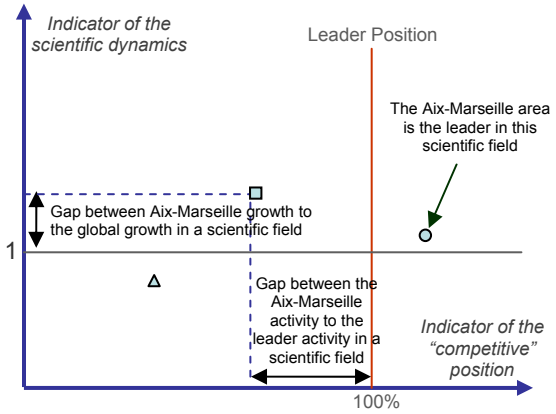


Figure 1. competition position of scientific portfolio

The scientific fields (strategic scientific units) are scattered according to these two axes for a 4-year window:

- indicator of the "competitive" position on the x-axis: the research activity of the Aix-Marseilles area in a strategic scientific unit is positioned on the x-axis in proportion to the research activity of the most active competitor in this field. The position of the leader of each strategic scientific unit is symbolized by a vertical line which has abscissa 100%. Thus the gap between the dot of a scientific field and this vertical line corresponds to the effort that the Aix-Marseilles area should make to catch up with the leader of this scientific field.
- indicator of the scientific dynamics on the y-axis: a strategic scientific unit is positioned on the y-axis according to the growth of the scientific activity of the Aix-Marseilles area in this field. This growth measures the evolution (expressed as a percentage) of the production between this time window and the previous time window. To standardize this measurement for the set of the strategic scientific units, this growth is reduced by the measurement of the growth of the areas set in this scientific field. Thus, the horizontal line having the ordinate 1 symbolizes the position at which the Aix-Marseilles area has a growth equivalent to the global growth for this time window. A strategic scientific unit located above this horizontal line shows that the Aix-Marseilles Area has a dynamics more significant than the dynamics of the areas set for this period of 4-year window.

As the two axes are relative indicators, it is possible to locate the strategic scientific units for all the 4-year windows on the same graph (fig. 2). Such a representation of the scientific portfolio makes it possible to position the scientific fields in relation to the "competition" by taking into account the history of the activity dynamics of the Aix-Marseilles area in each one of these fields.

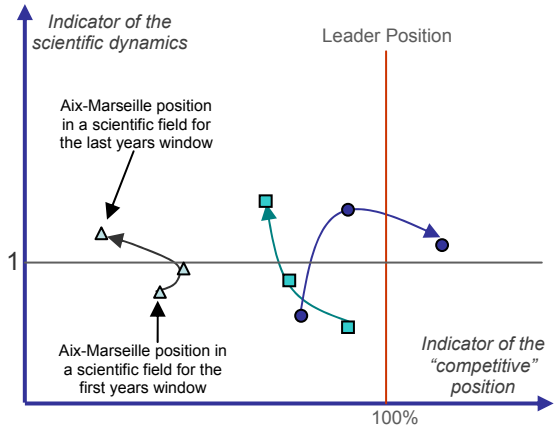


Figure 2. competition position dynamics of a scientific portfolio

Implementation

Data collection

The WoS version of the Citation Index was used for collecting the data. This data collection was carried out by taking particularly into account the boundary identification of the "competitors" areas and the boundary identification of the seven strategic scientific units (selection of a scientific journals panel for each Doctoral Schools).

Results

The collected data were compiled in "scoreboards". A first scoreboard compiles the evolution of the global production of each scientific area for the 6 periods of 4-year windows. A similar scoreboard is built up for each strategic scientific unit. Based on all these scoreboards, the competition position dynamics of a scientific portfolio for Aix-Marseille area was computed and graphically represented.

References

Ernst, H. (2003). Patent information for strategic technology management. *World Patent Information*, 25, 233-242.

Vinck, D. (Ed.). (1991). *Gestion de la recherche. Nouveaux problèmes, nouveaux outils*. Bruxelles: De Boeck-Wesmael.

Analyzing Grey Literature from Postgraduate Programs in Social Communication in Brazil: Network of Influence and Citation Analysis

Samile Andréa de Souza Vanz* and Renato Fabiano Matheus**

*samilevanz@terra.com.br
UFRGS (Universidade Federal do Rio Grande do Sul), Programa de Pós-graduação em Comunicação e Informação, Rua Ramiro Barcelos, n. 2705, CEP 90035 005, Porto Alegre (Brazil)

**renato.fabiano@bcb.gov.br
Banco Central do Brasil, ADBHO / COINF, Avenida Álvares Cabral 1605, 3o andar, CEP 30170-001, Belo Horizonte (Brazil)

Introduction

The importance of the scientific indicator production has been discussed for a long time in many fields of knowledge (BARRE, 1997; VELHO, 1986). More recently, the necessity of the regional indicators began to be evident (SPINAK, 1998). In this sense, this research studies the profile of scientific communication in the Social Communication field in Brazil.

Methodology

The citation data was extracted from 100 dissertations approved between 1998 and 2000 in three different postgraduate programs in Communication in South Brazil: Universidade Federal do Rio Grande do Sul (PPGCOM/UFRGS), Pontifícia Universidade Católica do Rio Grande do Sul (PPGCOM/PUCRS) and the Universidade do Vale do Rio dos Sinos (PPGCOM/UNISINOS). The whole group of dissertations defended during the period was analyzed, except for one text that could not be obtained. Data has been extracted from the reference section in the texts, tabulated and analyzed in a database created in the software Statistics Packet for Social Science (SPSS) version 13.0. Network analysis and visualization were made by the softwares UCINET version 6.142 (BORGATTI & EVERETT & FREEMAN, 2002) and Netdraw, respectively. Citation frequency was analyzed using the computer program LOTKA 1.02 (ROUSSEAU & ROUSSEAU, 2000). After the quantitative analysis, interviews were carried out with advisors involved in the dissertations in study. Seven professors, who were considered influential or related to the results of this study, therefore being able to explain them, were interviewed. The reports were analyzed and they were incorporated to the data discussion.

Data analysis

The data collected resulted in 7,648 citations. The citation analysis shows a high incidence of citations to books and chapters of books (72.5%), what brings into evidence the predominance of this kind of document in the grey literature of Communication field in Brazil. Papers from national and international journals were responsible for only 7.8% of citations, almost the same amount represented by newspaper and magazine articles (7%). Personal communication, interview and lecture took up 2.2% of citations. Portuguese was the predominant language in the citations (76.1%). English was found to be the second most cited language (9.8%), followed by Spanish (9.6%) and French (2.5%). However, this result does not indicate that the majority of texts were from Brazilian authors, but yet that translated texts, mainly from French are used a lot. The distribution of citations by date of publication revealed that the majority of cited documents were published in the 90's (60.4%). However, we found that classic authors are the majority of the core of most cited authors, which reinforces the idea that most of the cited work consists of translations or reedited material, not originals. Five national journals, two Latin-American journals, a Spanish journal and a British journal can be found amongst the most cited journals. A total of 3,435 authors were cited. The results show the lack of a concise group of authors who are strongly cited, since 32% of authors were cited only once. The most cited author, the Italian Umberto Eco was cited 83 times, representing only 1.1% of the total citations. The citation frequency observed follows a power law distribution (R^2=0.9853) and fits Lotka's Law. When data about citation per author was tested using LOTKA program the result obtained was C = 0.7116 and β = 2.3531, and the maximum deviation is 0.0049, leading to the acceptance of the Lotka's Law (1%, critical value=0.0278, max D=0.0049; 5%, critical value=0.0232, max D=0.0049; 10%, critical value=0.0208, max D=0.0049). In order to prove the existence of a relationship between the cited authors and the three postgraduate programs, we proceeded to a correspondence analysis with the core of authors cited 20 times or more. The independence hypothesis was rejected with a sig .000. The social network analysis - SNA - provides an illustration of this relationship (Figure 1). During the interview with advisors they were asked about the possible relationship between the

postgraduate program and specific authors. One of the advisors states the existence of a relationship between the most cited authors in the dissertations and programs, in the scope of visiting professor contracts and also in technical collaboration, co-authorship of books between author being cited and advisors, and so on. Case and Higgins (2000) have already mentioned these trends in Communication, showing that the advisors are prone to have previously worked with cited authors or even have a friendship with them.

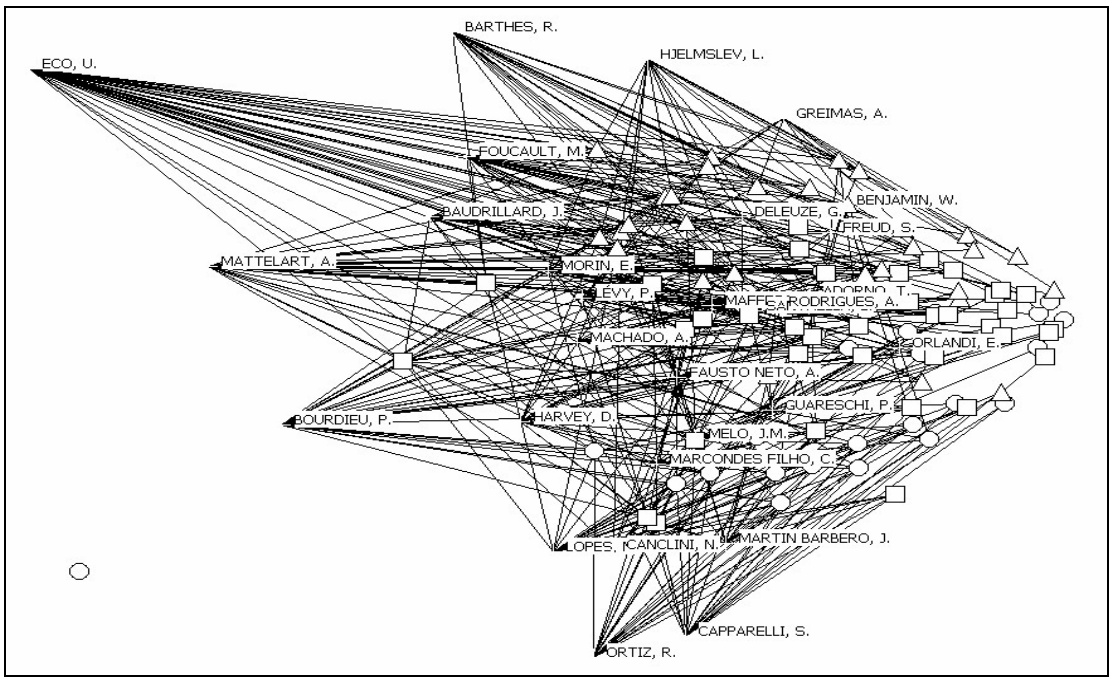


Figure 1. Citation network of the cited authors and the three postgraduate programs: PPGCOM/UFRGS (O), PPGCOM/PUC(□)ePPGCOM/UNISINOS(Δ).

Conclusion

The analyses carried on show a high incidence of citations to books and chapters of books (72.5%). Portuguese was the predominant language in the citations (76.1%), showing the preference for translations into the native language. The distribution of citations by date of publication revealed that the documents were published in the 90's (60.4%). The bibliometrics and network analysis applied indicated the relationship between the cited author and the three postgraduate programs. The frequency of cited authors followed a power low distribution and fitted the Lotka's Law. The advisors interviewed recognize the lack of a theoretic *corpus* of its own. However, the studies about the literature used in Brazilian national publications is still in its beginning and, citation studies allow us to know and monitor the authors who are cited in the period of theoretical formation in which Communication is found to be. It is necessary that the field sheds knowledge on itself in order to make decisions, and, if that is the case, decide to change its course.

References

Barre, R. (1997). The European Perspective on S&T Indicators. *Scientometrics,* 38 (1): 57-70.

Borgatti, S. P. & Everett, M. G. & Freeman, L.C (2002). Ucinet 6 for Windows: software for social network analysis - User's Guide. Analytic Technologies, Retrieved January 31, 2007 from: http://www.analytictech.com/ucinet/ucinet.htm

Case, D. O. & Higgins, G. M. (2000). How Can we Investigate Citation Behavior? a study of reasons for citing literature in Communication. *Journal of the American Society for Information Science*, 51 (7): 635-645.

Rousseau, B. & Rousseau, R. (2000). LOTKA: a program to fit a power law distribution to observed frequency data. *CYBERmetrics,* 4 (1), paper 4. Retrieved January 7, 2007 from: http://www.cybermetrics.cindoccsic.es/pruebas/v4ilp4.htm.

Spinak, E. (1998). Indicadores Cienciométricos. *Ciência da Informação*, 27 (2): 141-148.

Velho, L. (1986). The Meaning of Citation in the Context of a Scientifically Peripheral Country. *Scientometrics*, 9 (1-2): 71-89

.

National and International Dimensions of the Triple Helix in Japan: University-Industry-Government and International Co-Authorship Relations

Yuan Sun,* Masamitsu Negishi* and Loet Leydesdorff **

*yuan@nii.ac.jp, negishi@nii.ac.jp
National Institute of Informatics (NII), 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430 (Japan)

** loet@leydesdorff.net
Amsterdam School of Communications Research (ASCoR), University of Amsterdam, Kloveniersburgwal 48, 1012 CX Amsterdam (Netherlands)

Introduction

International co-authorship relations have grown exponentially during the past two decades (e.g,, Zitt et al., 2000; Glänzel, 2001; Wagner & Leydesdorff, 2005). Coauthored publications across sectors of the economy (university, industry, government) have been used as indicators of the Triple Helix model (Leydesdorff, 2003) and more generally for the study of science-technology relations (Narin et al., 1997) and entrepreneurial science (Tijssen, 2006).

This study builds on ongoing work of a Japanese team (Sun et al., 2006) to categorize all Japanese publications in the international scholarly databases. By combining information on both domestic university-industry-government relations and international coauthorship relations, we are able to make a further elaboration of the measurement of the Triple Helix model in terms of the mutual information.

Data

During the period 1981-2004, 1,277,823 research articles with at least one Japanese address were published in the three ISI-databases (Science Citation Index, the Social Science Citation Index, and the Arts & Humanities Citation Index). All records were standardized in terms of precise addresses and where necessary name changes because of mergers and acquisitions were corrected so that a set of unique address identifiers could be generated. In many cases, this implied also the unification of different English translations of the same Japanese names. The database is not complete for 2004 because of delays in publication years versus tape years.

Method

This data allows us to extend our previous studies of university-industry-government relations with the international dimension. The mutual information in two dimensions (or transmission T) follows directly from the Shannon formulas. T is defined as the difference which it makes when two probability distributions are combined:

$H_i = -\Sigma_i p_i \log_2 (p_i)$; $H_{ij} = -\Sigma_i \Sigma_j p_{ij} \log_2 (p_{ij})$

$H_{ij} = H_i + H_j - T_{ij}$

$$T_{ij} = H_i + H_j - H_{ij} \quad (1)$$

Abramson (1963, at p. 129) derived that the mutual information in three dimensions—let us use "u" for university, "i" for industry, and "g" for government—can be defined as follows:

$$T_{uig} = H_u + H_i + H_g - H_{ui} - H_{ug} - H_{ig} + H_{uig} \quad (2)$$

A negative value means that the uncertainty prevailing at the network level is reduced. Because the information is configurational, this reduction of the uncertainty cannot be attributed to one of the contributors (Jakulin & Bratko, 2004).

The extension with a fourth dimension ("f" for foreign) is straightforward:

$$T_{uigf} = H_u + H_i + H_g + H_f - H_{ui} - H_{ug} - H_{uf} - H_{ig} - H_{if} - H_{gf} + H_{uig} + H_{uif} + H_{ugf} + H_{igf} - H_{uigf} \quad (3)$$

Because the number of foreign papers without a Japanese address which is needed for the computation of H_f is not contained in the data, we use the sum total of Japanese papers with an international coauthor as a proxy for the relevant international environment in this study.

Results

National system of innovations

First turning to the mutual information in three dimensions for the Japanese national system of innovations, operationalized as a Triple Helix of university-industry-government relations, we see a trend breach around 1995. Before 1995, industry was the second largest producer of academic papers in Japan. However, the loss of synergy within the system is not caused by the decline in the number of papers with only an industrial address, but by the decline in university-industry co-authorship relations. The negative contributions to the mutual information are an effect of the relations in the network. This latter development began already before 1995, but is reflected in the national system with some delay. The mutual information between universities and

government agencies is approximately stable during this same period (1993-2004).

Addition of the international dimension

The addition of the international dimension changes the normalization of all distributions because the total set is expanded,

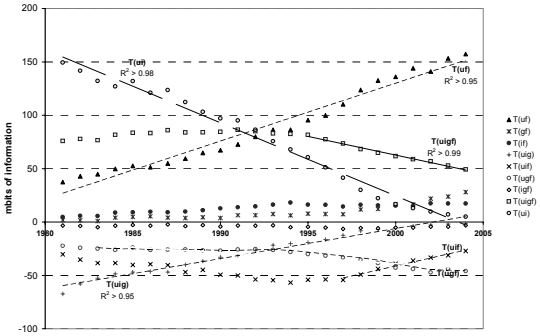


Figure 1. The mutual information in three and four dimensions among Japanese papers with a university, industrial or governmental address, and international co-authorships.

The upper half of Figure 1 shows two major trends in the system which are negatively correlated ($r = -0.99$; $p < 0.01$). First, there is the ongoing expansion of collaboration between university researchers in Japan with international colleagues. (This upward trend line (for T_{uf}) is indicated in the figure with ▲.) This collaboration is to the detriment of university-industry relations (indicated in this figure with ○ and as a negative trend). Both trends are robust and taking place over the full period of time (1981-2004). These long-term, i.e. cultural, trends have a negative influence on the integration of the national system as is exhibited in the lower part of the figure.

The various mutual informations in three dimensions are pictured in the lower part of this graph because they are all negative. However, the figure shows that the triple-helix synergy (T_{uig}) is continuously weakening during this period and even approaching zero in the most recent years. Among the trilateral systems university-government-foreign co-authorship relations have gained momentum (that is, T_{ugf} has become more negative).

Since 1995, the uncertainty in the total system of university-industry-government-foreign authorship relations (T_{uigf}; □) has been steadily reduced. In other words, the international co-authorship relations of the national partners have increasingly been integrated in the national system. The synergy in this configuration is no longer to be found in the national system without taking this international dimension of the publication system into account.

.

Conclusions

The Triple Helix can be considered as a indicator of dynamics within a knowledge-based economy (Leydesdorff, 2006). Using the operationalization of the Triple Helix in terms of mutual information transmissions between and among the helices, the international dimension can be endogenized into the Triple Helix model as a fourth dimension. The measure suggests that long term trends of internationalization among Japanese academic scholars led around 1995 to the opening of the national system of innovations to the international dimension. As a case, Japan may be atypical because it was weakly internationalized during the early 1990s. However, we expect similar reactions to have taken place in other parts of the world because of the globalization of the knowledge-based economy after the demise of the Soviet-Union and the opening of China.

References

Abramson, N. (1963). Information Theory and Coding. New York, etc.: McGraw-Hill.

Glänzel, W. (2001). National characteristics in international scientific co-authorship relations. *Scientometrics*, 51, 69-115.

Jakulin, A., & Bratko, I. (2004). Quantifying and Visualizing Attribute Interactions: An Approach Based on Entropy. from http://arxiv.org/abs/cs.AI/0308002

Leydesdorff, L. (2003). The Mutual Information of University-Industry-Government Relations: An Indicator of the Triple Helix Dynamics. Scientometrics, 58(2), 445-467.

Leydesdorff, L. (2006). T*he Knowledge-Based Economy: Modeled, Measured, Simulated*. Boca Rota, FL: Universal Publishers.

Narin, F., Hamilton, K. S., & Olivastro, D. (1997). The increasing link between U.S. technology and public science. *Research Policy*, 26(3), 317-330.

Sun, Y., Negishi, M., & Nisizawa, M. (2006). Coauthorship Linkages between Universities and Industry in Japan. Paper presented at the 9th International Conference on Science and Technology Indicators, Leuven, Belgium.

Tijssen, R. J. W. (2006). University-industry interactions and university entrepreneurial science: towards measurement models and indicators. Research Policy, 35(10), forthcoming.

Wagner, C. S., & Leydesdorff, L. (2005). Mapping the Network of Global Science: Comparing International Co-authorships from 1990 to 2000. International *Journal of Technology and Globalization*, 1(2), 185-208.

Zitt, M., Bassecoulard, E., & Okubo, Y. (2000). Shadows of the past in international cooperation: Collaboration profiles of the top five producers of science. *Scientometrics*, 47(3), 627-657

International Communication Patterns in an Emerging Interdisciplinary Field - Applications of Geographic Information Systems in Public Health

Dimitar T. Tomov* and AbdulKader A. Murad**

*dimtomov@yahoo.com
Prof. Paraskev Stoyanov Medical University of Varna, Faculty of Public Health, Department of Health Economics and Management, 55 Marin Drinov St., BG-9002 Varna (Bulgaria)

**amurad22@yahoo.com
King Abdulaziz University, Faculty of Environmental Design, Department of Urban & Regional Planning, P. O. Box 80210, Jeddah, 21589 (Kingdom of Saudi Arabia)

Introduction

About 20 years ago, geographic information systems (GIS) were launched for use in public health (Gatrell & Loytonen, 1998; Gatrell, 2002; Murad 2004) and education (Murad, 1998). Modern computer-aided information retrieval in scientific databases through Internet enables hot-topic-oriented bibliometric research not only by experienced scientometricians but also by beginners and research policy makers at different level. We demonstrate a complex evaluation method for dynamic science stratification on the example of GIS applied in public health.

Material and methods

Our simple method uses a set of parameters for comparative analysis of the international scientific communications. Secondary documents on GIS applications in health abstracted in both *EMBASE* and *MEDLINE* through OVID in 1986-2005 were examined. We analyzed the number of items per year, the types of primary publications (Tomov & Mutafov, 2004), the journal titles (number, thematic profile, country and disciplinary belonging), the authors' names (number of unique names, authors per paper per year and country belonging), the affiliations (thematic profiles and types), and the rate of appearance of new specialized journals. The titles of relevant papers and journals containing them, the abstracts, author's names and affiliations were semantically analyzed to illustrate the international-ization, institutionalization and conceptual diversification of this interdisciplinary domain (Tomov, 2001a) and to compare the information-retrieval capacities of these databases (Tomov, 2006) thus suggesting the role of applied scientometrics for GIS research promotion.

Results and discussion

Table No 1 demonstrates the different coverage of *EMBASE* and *MEDLINE* concerning both indexing and abstracting the relevant papers.

Usage of both databases is required to warrant a truly comprehensive retrieval of publications by GIS researchers and information scientists. The editorial policy reflected in these indexing systems differs when single countries of journals, journal titles and journal issues are concerned. The number and content of key words/descriptors assigned to one and the same paper in both databases varies to a different extent in single years and journals.

Table 1. Total number of publications abstracted in *MEDLINE* and *EMBASE*

Time span	*MEDLINE*	*EMBASE*
1986-1990	3	9
1991-1995	40	86
1996-2000	167	227
2001-2005	1050	689
Total	1260	1011

There is a permanent increase of the world publication output in terms of absolute number of journal articles, journals containing them, countries of journals and authors as a whole and of single papers. The true amount of authors which papers have been abstracted in *EMBASE* is greater to certain extent as in 1982-1986 only the first three authors per paper with more than four authors have been listed in this database. There is a great variety of journals' disciplinary thematic belonging and thus of nominations of journals and authors' institutions proving the active involvement and practical accomplishment of ideas and methods from many technological, biomedical and clinical disciplines and other sciences. The most common author's institutions are the following: department, institute, centre, clinic, unit, laboratory, and hospital.

The number of the countries of authors engaged in this specific research sharply increases in 2001-2005 (Table 2).

Table 2. Countries of authors with publications abstracted in *MEDLINE* and *EMBASE*

Time span	*MEDLINE*	*EMBASE*	total
1986-1990	3	2	4
1991-1995	5	14	16
1996-2000	19	37	38
2001-2005	66	58	72
Total	67	65	78

The dramatic augmentation of the amount of publications in narrow interdisciplinary fields such as memory and learning, memory and cognition, adipobiology, AIDS/HIV and nanotechnologies (Tomov, 2006) is an essential peculiarity of modern science. Similarly to most publication-output distributions, the USA science occupies a dominant position in any aspects of the publication output figures in this field followed by the UK, Germany, France, China and Japan. The gradually rising international presentation of the authors from Africa, Asia, Latin America and Eastern Europe stresses and proves the role of theoretical and practical experience of specialists trained in the USA and Western Europe. The science stratification in this particular field has emerged in the last decade of the 20th century and continues until presence.

The most popular publication type is the original journal article although there are numerous review articles along with some editorials, letters to the editor, single conference papers, meta-analyses and multicentre studies (Table 3). This variety indicates a rising maturity of the research field.

The enormous growth of the world publication output in this filed testifies to the awareness of its social significance and to the properly gained public recognition of these mutually related issues in the era of information society and knowledge-based economics.

The substantial differences between *EMBASE* and *MEDLINE* concerning the degree of relevance of the primary publications assigned to this specific topic raises the question of the sufficiently qualified indexing work of the teams involved and enhances the responsibility of the authors of the publications for the information value of the titles and abstracts of their own papers.

Our results convincingly demonstrate the enormous capacities of the computer-aided scientometric approach to modern interdisciplinary science.

Conclusion

Contemporary applied scientometric methodology (Tomov & Mutafov, 1996; Tomov, 2001a) should become more popular in university curricula. Information retrieval on this and other similar topics in these databases can successfully be combined by using other miscellaneous, technological and biomedical secondary electronic information sources such as *Science Citation Index CDE*, *Social Sciences Citation Index CDE*, *INSPEC*, and *BIOSIS* for the definition and subsequent implementation of an effective science policy management in small countries.

For instance, the most up-dated version of the Internet-accessible ISI Web of Knowledge launched recently by Thomson Scientific (Philadelphia, PA, USA) where a lot of disadvantages of the older ISI editions (Tomov, 2001) were eliminated fairly serves in this respect. In this way, many young scientists from these countries could aspire for a better international visibility through effective engagement in fruitful international collaboration.

References

Gatrell, A. & Loytonen, M. (1998). GIS and Health. London: Taylor and Francis.

Gatrell, A.C. (2002). Geographies of Health: an Introduction. Oxford: Blackwell.

Murad, A.A. (1998). Application of Geographical Information Systems for Educational Facilities Planning. Ph.D. thesis. Newcastle Upon Tyne: University of Newcastle Upon Tyne.

Murad, A.A. (2004). Creating a GIS application for local health care planning in Saudi Arabia. International Journal of Environmental Health Research, 14, 185-199.

Tomov, D. T. (2001). Some critical remarks on the stop word lists of ISI publications. Journal of Documentation, 57, 798-808.

Tomov, D. (2001a). The unity of interdisciplinarity, institutionalization and internationalization of science: reflections from/on cell biology. Biomedical Reviews, 12, 41-55.

Tomov, D.T. (2006). Adipobiology: a scientometric view to its increasing significance. Biomedical Reviews, 17, 113-118.

Tomov, D.T. & Mutafov, H.G. (1996). Comparative indicators of interdisciplinarity in modern science. Scientometrics, 37, 267-278.

Tomov, D.T. & Mutafov, H.G. (2004). Dynamic stratification of document-type distributions in interdisciplinary biomedical research. Scripta Periodica, 7 (Suppl.), 12-18

Table 3. Document-type distributions of publications abstracted in *MEDLINE* and *EMBASE*

Document type	*EMBASE*			*MEDLINE*		
	time span	n		time span	n	
		documents	journals		documents	journals
editorial	2001-2005	9	9	2003-2005	5	4
letter	2003-2005	5	5	2004	1	1
review	2000-2005	55	41	1997,2002-2005	69	42
multicentre study	-	-	-	2003-2005	5	4
meta-analysis	-	-	-	2005	1	1
conference	2001-2005	44	21	-	-	-

An Analysis and Comparison of Citation Data between Journals of Physics, Chemistry and Engineering

Ming-yueh Tsay

mytsay@nccu.edu.tw
Graduate Institute of Library, Information and Archival Studies
National Chengchi University, Taipei (Taiwan)

Introduction

A journal citation analysis may be an important tool for the academic assessment of an individual researcher or a research institute. More importantly, a journal citation analysis can also show how knowledge is changing, or becoming obsolescent, and help optimize an information retrieval system. The primary objective of the present study is to investigate the relationship and difference between each pair of five kinds of citation data, i.e. citation count, impact factor, immediacy index, citing half-life and cited half-life for journals of physics, chemistry and engineering. Physics and chemistry are two of the most important subjects for science, while engineering can be considered as the application of basic sciences.

Methodology

In the present study, the citation data of essential journals for physics, chemistry and engineering were drawn from SCI Journal Citation Reports (JCR) on the Web 2002. Pearson correlations from each pair of citation data were calculated to assess the significant relationship between each of the 15 pairs of these citation data. The Fisher's Z-transform was used to test the significant difference between the strength of Pearson correlation coefficient for each pair of citation data of the three subject areas. Moreover, the T-test was employed to examine the mean difference of citation data among essential journals of physics, chemistry and engineering. The essential journals for the three subjects were selected from the most current edition of *Guide to Information Sources in the Physical Sciences* (Stern, 2000), *Information Sources in Chemistry* (Bottle & Rowland, 1993), and *Guide to Information Sources in Engineering* (Lord, 2000). Journals common to both Guide to the Literature tools and JCR Science edition web 2002 are selected as the samples for the statistical test of journal citation data. Consequently, the essential journal samples of the present study for the subjects of physics, chemistry, and engineering are 236, 95 and 94, respectively.

Results and Discussion

Comparison among the significant correlations listed in Table 1 demonstrates the similarity in the existence of significant correlations among source item, citation account, impact factor, and immediacy index and between cited half-life and citing half-life for the three subjects under study except that there is no correlation between source item and impact factor and immediacy index, respectively, for the physics subject. The similarity is also demonstrated for the existence of significantly negative correlations between citing half-life and other citation indicators, except positive for cited half-life, for the subjects of physics and chemistry, but not for the engineering subject. However, difference rises for the correlation between the cited half-life and other citation data among the three subjects. In general, there is no significant correlation between the cited half-life and other citation indicators. To substantiate this general observation, the Fisher tests are conducted in the next section further clarify the strength of the significant correlation between each pair for journals in these three subject fields, the Fisher tests were conducted. It was revealed that 14 out of 22 pairs show no significant difference. In particular, the three subjects present the same strength of the correlation between source item and citation count, between citation count and impact factors and between cited half-life and citing half-life. For the correlations between citation count and immediacy index and between impact factor and citing half-life, no significant difference is observed between essential journals of physics and engineering. For the correlations between citing half-life and source item and citation count, respectively, there is no significant difference between physics and chemistry journals.

Table 2 illustrates the mean values of citation indicators for the essential journals of the three subjects under study. Superficially, except for the cited half-life and citing half-life, the mean values of various citation indicators for the essential journals of engineering deviate significantly from those of the physics and chemistry. The last two subjects are relatively close to each other. On average, a chemistry journal published slightly more papers than a physics journal, and a physics journal publishes about three times of papers as an engineering journal does. Because there are more articles to be cited, a chemistry journal gets more citations than a physics journal does, and a physics journal possesses citations more than four times of an engineering journal. The average impact factor for a physics or chemistry journal is around 2.10, while the average impact factor for an engineering journal is 0.921 only. The T-test of

mean difference for citation data among the three subject journals demonstrates that for the 18 pairs among the three subjects, there are 8 pairs without significant difference in citation data. In particular, for all the citation indicators of this study except the cited half-life, there is no significant difference between the essential journals of physics and chemistry. On the other hand, significant difference is generally observed between engineering and physics or between engineering and chemistry for most of the citation indicators.

Conclusions

For all the citation indicators of this study except the cited half-life, there is no significant difference between the essential journals of physics and chemistry indicating similar citation behavior among the scientists. However, significant mean difference is generally observed for most of the citation indicators between engineering and physics or between engineering and chemistry demonstrating the difference in citation behavior among engineering researchers and scientists in physics or chemistry. There is no significant mean difference in the citing half-life among the journals in physics, chemistry and engineering.

Comparison among the significant correlations from the Pearson tests demonstrates the similarity in the existence of significant correlations among source item, citation account, impact factor, and immediacy index and between cited half-life and citing half-life for the essential journals of physics, chemistry and engineering. Difference arises for the correlation between the cited half-life and other citation data among the three subjects of physics, chemistry and engineering. In general, there is no significant correlation between the cited half-life and other citation indicators.

The three subjects of this study present the same strength of the correlation between source item and citation count, between citation count and impact factor, and between cited half-life and citing half-life. Significant difference in strength exists for the three pairs among the three subjects between impact factor and immediacy index.

Acknowledgements

This study was supported by the National Science Council, Taiwan, NSC93-2413-H-004-017. Data collection by I-ling Chen, a graduate student of Department of Library and Information Science, Tamkang University, is very much appreciated.

References

Bottle, R.T. & Rowland, J.F.B. (Ed.). (1993). *Information Sources in Chemistry*, 4th ed. London: Bowker-Saur, chapter 2.

Lord, C. R. (Ed.). (2000). *Guide to Information Sources in Engineering*. Englewood, Colo: Libraries Unlimited, chapter 4.

Stern, D.(Ed.). (2000). *Guide to Information Sources in the Physical Sciences*. Englewood, Colo: Libraries Unlimited, chapter 3

.

Table 1. Pearson Correlation Coefficient (α=0.05) between Citation data of I: Physics; II: Chemistry; III: Engineering

		Source Item		Citation Count		Impact Factor		Immediacy Index		Cited Half-Life	
Citation Count	I	.835(.000,233)									
	II	.882(.000,93)									
	III	.871(.000,94)									
Impact Factor	I	.088(.182,233)	I	.237(.000,236)							
	II	.315(.002,92)	II	.412(.000,94)							
	III	.260(.011,94)	III	.444(.000,94)							
Immediacy Index	I	.118(.075,230)	I	.271(.000,230)	I	.898(.000,230)					
	II	.526(.000,90)	II	.650(.000,90)	II	.835(.000,90)					
	III	.267(.009,94)	III	.418(.000,94)	III	.596(.000,94)					
Cited Half-Life	I	-.138(.036,232)	I	.032(.627,235)	I	.130(.047,235)	I	.106(.111,229)			
	II	-.170(.104,92)	II	-.030(.774,94)	II	-.161(.123,93)	II	-.214(.044,89)			
	III	.021(.849,88)	III	.228(.033,88)	III	-.046(.672,88)	III	.153(.155,88)			
Citing Half-Life	I	-.298(.000,232)	I	-.223(.001,233)	I	-.143(.029,233)	I	-.127(.054,230)	I	.464(.000,232	
	II	-.290(.006,88)	II	-.312(.003,89)	II	-.549(.000,89)	II	-.585(.000,88)	II	.326(.002,88)	
	III	.022(.839,90)	III	.053(.618,90)	III	-.228(.031,90)	III	.036(.735,90)	III	.335(.002,86)	

Table 2. Mean values of citation indicators for three subject journals

Subject	Source item	Citation count	Impact factor	Immediacy index	Cited half-life	Citing half-life
Physics	407.24	10568.03	2.108	0.419	7.028	8.115
Chemistry	451.46	14866.73	2.076	0.326	7.962	8.283
Engineering	132.21	2546.32	0.921	0.142	7.828	8.001

Verification of Increasing Trend Detection

Yuen-Hsien Tseng[*], Yu-I Lin[**], Chun-Hsien Kuo[***] and Yi-Yang Lee[***]

[*] *samtseng@ntnu.edu.tw*
Digital Media Center, National Taiwan Normal University (Taiwan)

[**] *jg141@mail.jges.tpc.edu.tw*
Taipei Municipal Univ. of Education (Taiwan)

[***] *chkuo@mail.stpi.org.tw, yylee@mail.stpi.org.tw*
Science & Technology Policy Research and Information Center, National Applied Research Lab. (Taiwan)

Introduction

Monitoring research trends has always been a concern of decision makers, since it helps resource allocation and technology forecast. There have been studies on detecting these trends in scientometrics and other fields alike. However, most encountered the lack of expert feedback to the results (Noyons, 1998). In addition, effectiveness of such prediction is rarely known quantitatively. In a project to analyze a large set of scientific publications in the agricultural areas, we have the opportunity to detect upward trends for a group of experts and have their feedback in trend type labelling. We then analyze the effectiveness of our detection and prediction methods based on this feedback. This work is to report our preliminary findings about this evaluation.

Method

Given a collection of publications, our method to identify upward trends involves two steps. First, the knowledge structures of the collection are detected based on a hierarchical clustering approach. Ideally, terms are grouped into concepts. These concepts can be further clustered into topics, which in turn can be clustered into categories or domains. We generate the structures based on the co-occurrence of controlled terms at this stage, as they yield high-precision and sufficient results for expert feedback.

Second, the time series of each structure (i.e., hierarchy of concepts, topics, or categories) is created and an index to predict its tendency is calculated. Specifically, the publication dates of the documents are grouped by a specified interval. Number of documents in the same intervals is counted to yield the frequency sequence over time. From the time series, three trend indices are compared to know which one is the best predictor. The first is average increase per year (*aipy*):

$$aipy = \frac{1}{N}\sum_{i=1}^{N-1}\frac{y_{i+1}-y_i}{y_i}\times 100 \quad (1)$$

where y_i denotes the number of documents in the ith interval and N is the number of intervals in the time series. This index has been used in a foresight survey in Japan (STFC, 2003). The second index, denoted as *slp*, is the slope of the regression line that best fits the data in the time series. This is the most commonly used prediction method in statistics. However, it has a large variation among sequences of different frequency scales, which may make it difficult for humans to predict its trend type based on its value. The remedy is to convert all the numbers in the sequence into their z scores and calculate the slope of the regression line based on the new sequence. This is our third index and is denoted as *slp_z*.

To identify upward trends for experts, these structures are sorted by the trend index in decreasing order. If the trend index is a good predictor, the inspection in this order should be efficient in finding upward trends. There is a tool to evaluate this sort of effectiveness, i.e., trec_eval (Buckley) used in information retrieval evaluation, such as TREC (Text REtrieval Conference). Two measures can be read from the output of trec_eval, i.e., NAP (Non-Interpolated Average Precision Rate), defined as:

$$NAP = \frac{1}{R}\sum_{i=1}^{R}\frac{i}{Rank_i} \quad (2)$$

where R is the number of all relevant items (true upward trends) and $Rank_i$ is the position of the ith relevant item in the ordering; the other is Pre@R, which denotes the precision rate by the R position in the ordering, i.e., n/R, where n is the number of relevant items in the top R items. The Pre@R measure is simpler to interpret, while the NAP measure can distinguish slightly different orderings.

Data

A set of six domains regarding safe agriculture were enumerated by a group of experts. For each domain a query was formulated to search the ISI's Web of Science database. By limiting the year range from 1996 to 2005, 72500 documents were downloaded. We use controlled terms for structure detection, i.e., those terms from the field of SC (Source Category) and DE (Descriptor). For each selected field, terms co-occurred in the same records were counted. This pair-wise count was normalized by the individual occurrence of each term. Similarity based on this calculation was used in a complete link clustering algorithm. From the SC field, a total of 80 clusters

were found and from the DE field, there were 1617 clusters. We sampled 50% of clusters from SC and 10% from DE and removed those clusters or sub-clusters having less than 30 documents. The resultant clusters with their time series were then presented to 9 experts for trend type labelling as a benchmark for evaluating the proposed trend indices. When labelling the trend type, the experts were advised to judge the type based on their knowledge. If this did not help, the time series of the cluster can be consulted. If this did not help either, the documents in the cluster can be examined. If all these efforts failed, the cluster was labelled 'inconclusive', denoted as '?'. Five types of trend can be labelled if it is decidable, i.e., 'sharp increase', 'increase', 'fluctuation', 'decrease', and 'sharp decrease', denoted as '++', '+,' '=', '-', and '--', respectively.

Results

Table 1 shows the statistics of the experts' judgment. As can be seen, 43 term groups (43/155=27.74%) from SC were inconclusive in their trend type. This percentage for DE is 22.89%. Since the overall tendency is upward, there is no sharp decreasing case for SC and DE. The decreasing cases are also rare. It is our concern to distinguish the sharp increasing or increasing cases from the others. Based on the NAP and Pre@R measures, the prediction effectiveness (see Figure 1 for NAP) of *slp* and *slp_z* perform equally well, while *aipy* performs poorly when the year span is 1. However, when the year span varies from 1, 2, to 5 (a year span of 5 corresponds to dividing the 10-years collection into two periods), the performance of *apiy* increases, while that of *slp_z* decreases.

The next evaluation is to know how performance would be affected if we only use the first *n* years of data for prediction. As shown in Figure 2, for the *slp* index the NAP performance may drop as low as 25% for the sharp increase cases when only the first 8 years of data were used. In other words, when we are at the 8th year to predict the trends in the 10th year, the performance is 75% of that when we have the full 10 years of data.

Implications and Conclusions

We have learned from the scientometrics literature that different year spans may be used to create the time sequence and different indices were chosen for trend observation. However, the effectiveness of these parameters was hardly known, quantitatively and comparatively. Our work provides clues to better interpret the results when a certain choice was made. For example, when *aipy* was chosen for trend prediction, it should be noted that it is only effective for large year span or short time series. Also when the collection is split into only two periods, both *slp* and *apiy* can be used. In any cases, *slp* is recommended since it performs well under the various conditions that we evaluated.

Our goal is to explore the best way to predict upward trends. The current results show that using the slope of linear regression on the time series yields the best result. Our work is important to know which index is the best under a certain condition. If a good trend index is used, the inspection in the order sorted by the index should be efficient in finding upward trends, which could relieve the burden of analysts on monitoring any upward trends from a large stream of scientific publications.

Acknowledgments

This work is support in part by NSC 95-2221-E-003 -016- and NSC 95-2524-S-003-012-.

References

Buckley, C. trec_eval IR evaluation package. http://trec.nist.gov/trec_eval/.

Noyons, E.C.M. & van Raan, A.F.J. (1998). Mapping Scientometrics, Informetrics, and Bibliometrics. Retrieved November 23, 2006 from http://www.cwts.nl/ed/sib/home.html.

STFC (Science and Technology Foresight Center) (2004). *The 8th science and technology foresight survey*. National Institute of Science & Technology Policy, Japan.

Table 1. Statistics of experts' judgment.

Field	(sub-)clusters	++	+	=	-	--	?
SC	155	18	57	37	0	0	43
DE	249	20	61	97	14	0	57

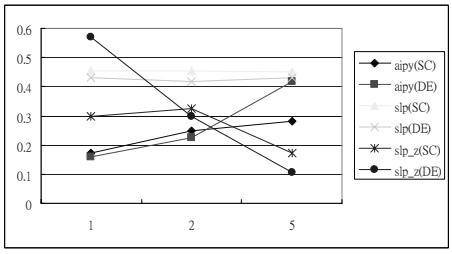


Figure 1. Prediction effectiveness when year span varies from 1, 2, to 5. (Only NAP for ++ is shown)

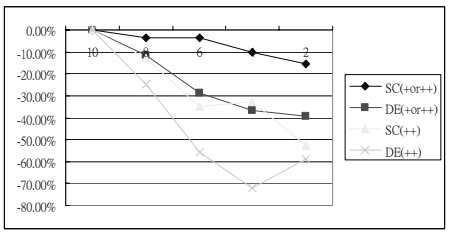


Figure 2. Performance drop for *slp* using only the first n years of data (n varies from 10, 8, 6, 4, to 2).

Recent Trends in Renewable Energy Research: A Bibliometric Perspective

Ali Uzun

azun@metu.edu.tr
Department of Statistics, Middle East Technical University 06531 Ankara (Turkey)

Introduction

Renewable energy is energy which can be replenished at the same rate it is used. Currently renewable energy sources contribute only about 15% of the human energy use worldwide but "Sayigh (1998) estimated that 60% of all our energy will come from renewable energy sources by the year 2070". Mankind's traditional use of wind, water, and solar power are widespread in developed and developing countries. The motivation for writing this paper is grounded in identifying those countries contributing the most to the field of renewable energy by finding answers to the following questions:

1. How the countries are ranked in renewable energy research in terms of total output of articles from 2002 to 2006?
2. What are the annual growth rates of the number of articles in countries?
3. What are the most productive countries as measured by the Activity Index (AI) which has been used widely in bibliometric evaluations?

Data and Method

The study is based on a total number of 8505 articles published by 33 countries in 13 major renewable energy journals over the period 2002-2006. These journals are indexed in SCI-Expanded database of Thomson ISI. The articles are credited to countries in accordance with author addresses.

Moving averages of the number of articles are calculated to reduce the random fluctuations from year to year. Mean Rate of Growth (MRG) of the number of articles per year from 2002 to 2006 is calculated using 3-year moving averages. Activity Index (AI) for each country is calculated to eliminate the biases introduced by country sizes so that cross-national comparisons can be made. AI was proposed for the first time by Frame (1977) and used among others by (Schubert and Braun, 1986), Carpenter et al. (1988).

AI is defined as follows:

$$AI = \frac{p_i}{p_0} \text{ x100}$$ where p_i and p_0 are the proportion of renewable energy articles in the total output of articles of country i in the SCI database and the proportion of renewable energy articles in the total output of articles of all countries in the SCI database, respectively. MRG is defined as follows:

$$\text{MRG} = \frac{MA_3 - MA_1}{MA_1} \text{ x100}$$

Findings

Table 1 shows the distribution of the 8505 articles by country, moving averages for 3-year periods denoted by MA_1 for 2002-2004, MA_2 for 2003-2005, MA_3 for 2004-2006, the MRG values as well as the AI values for the countries. We briefly note a few of the details that may be found through inspection of Table 1. (1) When the countries are ranked by the number of renewable energy articles, the USA, India, Japan, China, Germany, Turkey, Spain, England, Sweden, and France emerged as the top ten countries. These countries account for 65% of the total output. (2) MRG values given in the sixth column indicate that the top ten countries are Finland (175%), Iran (133%), New Zealand (85.7%), Taiwan (74%), Austria (70%), China (59%), Singapore (50%), Portugal (44.4%), India (42%), and Netherlands (40%). The three countries ranked at the bottom of the list are Hungary, Russia, and Israel having growth rates -25%, -20%, and -11.8%, respectively. (3) AI values given in the last column indicate that the top ten countries include Turkey (AI = 425), India (350), Mexico (316), Greece (315), Denmark (238), Sweden (227), Spain (157), Iran (147), Taiwan (143), and Norway (135). The USA, Germany, France, and England are well below the mean AI = 100 for the 33 countries.

Another interesting observation which is not shown in Table 1 is that India, Turkey, and China are the most active countries in biomass and bioenergy resources. China is also one of the top producers of hydrogen energy research. The USA and Germany having low profiles in renewable energy research turn out to be the leading countries in solar energy, solar energy materials/ solar cells, and photovoltaic technology in general. Equally interesting is the fact that Denmark accounts for 28.7% of the articles in wind energy which is the fastest growing of the renewable energy sector. The USA, New Zealand, Mexico, and Italy are the

countries having maximum number of articles in geothermal energy.

Table 1. National outputs and growth rate of the number of articles in renewable energy for 2002 – 2006

Rank[a]	Country	TRE[b]	MA_1	MA_2	MA_3	MRG (%)	TSCI[c]	AI
1	USA	1312	233	259	285	22.3	1,202,517	58
2	INDIA	722	116	116	165	42.2	111,546	350
3	JAPAN	660	123	114	130	5.7	370,990	96
4	CHINA	614	93	104	148	59.1	292,251	114
5	GERMANY	487	93	82	95	2.1	334,775	79
6	TURKEY	464	83	96	105	26.5	59,055	425
7	SPAIN	388	63	74	88	39.7	133,672	157
8	ENGLAND	325	57	59	72	26.3	275,914	64
9	SWEDEN	320	58	58	73	25.8	76,050	227
10	FRANCE	270	46	43	58	26.0	243,190	60
11	CANADA	259	48	48	57	18.7	174,171	80
12	ITALY	229	43	47	45	4.6	180,755	68
13	MEXICO	224	43	45	52	20.9	38,351	316
14	AUSTRALIA	219	44	45	47	6.8	110,602	107
15	NETHERLANDS	212	35	39	49	40.0	98,428	115
16	S. KOREA	211	34	38	46	35.3	113,609	114
17	GREECE	195	35	38	47	34.3	33,437	315
18	TAIWAN	185	27	36	47	74.0	70,046	143
19	DENMARK	181	30	31	40	33.3	41,107	238
20	BRAZIL	178	35	38	43	22.8	75,005	128
21	SWITZERLAND	142	25	30	31	24.0	73,418	105
22	ISRAEL	83	19	17	17	-11.8	52844	85
23	FINLAND	80	8	13	22	175	38,135	113
24	RUSSIA	70	15	8	12	-20.0	119,148	32
25	NORWAY	67	11	11	13	18.2	26,926	135
26	AUSTRIA	66	10	13	17	70.0	39,477	90
27	POLAND	62	11	10	12	9.1	64,303	52
28	ARGENTINA	55	11	10	11	0.0	23,989	124
29	PORTUGAL	55	9	10	13	44.4	24,335	122
30	NEW ZEALAND	52	7	12	13	85.7	21,561	130
31	IRAN	51	6	9	14	133.3	18,719	147
32	SINGAPORE	51	8	11	12	50.0	26,684	103
33	HUNGARY	19	4	4	3	-25.0	22,124	46
Sum		8505					4,587,195	

Key:
[a] The ranking is made in descending order of the number of renewable energy articles.
[b] TRE stands for the total number of renewable energy articles
[c] TSCI denotes the total number of articles published in all journals included in the SCI database.

Conclusion

Based on an overview of the data presented here one may conclude that India and Turkey maintained a high profile in the mainstream renewable energy literature in terms of the number of articles and Activity Index. This is not unexpected when one realizes that both countries are poor in oil but rich particularly in bioenergy resources. This high priority is consistent with energy policies toward sustainable development of these countries. The research activities in terms of the AI values for the rest of the countries except Hungary, Russia, and Israel are found to be somewhat close to the 33 countries' mean of 100.

References

Carpenter, M.P., Gibbs, F. Harris, M. Irvine, J. Narin, P. (1988). Bibliometric profiles of academic institutions: An experiment to develop output indicators. *Scientometrics*, 14, 213-233.

Frame, J.D. (1977). Mainstream research in Latin America and the Caribbean. *Interciencia*, 2, 143-148.

Sayigh, A.A.M. (1998). Renewable energy efficiency, policy and environment, *Proceedings of the World Renewable Energy Congress V, Part I- Forward.*

Schubert, A. & Braun, T. (1986). Relative indicators and relational charts for comparative assessment of publication output and citation impact. *Scientometrics*, 9, 281-291.

From Web Citation to Web Co-citation: Discovering Relatedness on the Web

Thanh-Trung Van, Michel Beigbeder

{van, mbeig}@emse.fr
Centre G2I, Ecole Nationale Supérieure des Mines de Saint Etienne, 158 cours Fauriel,
42023 Saint Etienne (France)

Introduction

For a long time, citation-based methods have been used to find relatedness between scientific papers beside content-based methods. Kessler (1963) proposed the bibliographic coupling method. In this method, the similarity between two papers is computed by the number of their co-references. He supposed that if two papers have common references in their bibliographies, they may focus (entirely or partially) on the same topic. Marshakova (1973) and Small (1973) independently proposed another method called "co-citation". In this method, the relatedness between two papers is computed by their co-citation frequency. The co-citation frequency is the frequency that two papers appear together in the bibliography section of a third paper.

These two methods have been used widely for about 40 years. However, both of them have their limits. In the bibliographic coupling method, the relatedness between two papers is fixed since their publication date because they are computed by the number of their co-references which is unchanged. In the co-citation methods, with the time two related papers may receive more and more citations and their co-citation frequency can increase. However, if we want to know this citation information, we have to extract from the citation graph of the actual paper collection or read from a citation database which is usually limited. That is why in this paper we propose a new approach to compute co-citation relatedness between scientific papers which can overcome this limit. We call this method *Web co-citation*.

Methodology

Traditional co-citation method with Web of Science

We decided to choose Web of Science as a citation database in our experiments. The Web of Science of Thomson ISI is an important database which is used widely for citation studies. The Web of Science provides access to current and retrospective multidisciplinary information from approximately 8,700 of the most prestigious, high impact research journals in the world. Besides, it also provides an API which facilitates the access to its database without using a Web browser[1]. Another important reason to use Web of Science (WoS) is that it contains most of journals and transactions used in our experiments.

Web co-citation method using the Google search engine

Recently, a new method for citation analysis called *Web citation analysis* begins attracting the research community. Web citation analysis finds citations to a scientific paper on the Web by sending the query containing the title of this paper (as phrase search using quotation marks) to a Web search engine and analyze returned pages (Vaughan & Shaw, 2003). In our Web co-citation method, we compute the co-citation similarity of two scientific papers by the frequency that they are mentioned together on the Web. If two papers are mentioned together in the same Web document, we can assume that they have a (strong or weak) relation. The search engine used in our experiment is the Google search engine. To find the number of time that a paper is "cited" by Google we need only to send the title of this paper (as phrase search using quotation marks) to Google and note the number of hits returned. Similarly, to find the number of times that two papers are "co-cited", we send the titles of these two papers (in the same query) to Google.

Experiments

In this work we conduct experiments for evaluating performance of different methods: bibliographic coupling and co-citation with Web of Science and Google. The experiments described here are simulations of personalized searching in a digital library using user profiles.

Test Collection

The test collection that we use in our experiments is the collection used in INEX 2005[2]. In the first step we remove all elements that are not scientific papers. After this process, the collection contains 14237 documents. There are also many topics with relevance assessments distributed with the collection. Each topic represents an information need. In our experiments we use only CO topics which do not contain structure of documents to create user queries. INEX uses a two-dimensional, multi-valued scale for relevance assessments of each topic. However in our experiments we use precision/recall metrics with

[1] http://scientific.thomson.com/support/faq/webservices/

[2] http://inex.is.informatik.uni-duisburg.de/2005/

binary scale (relevant/non-relevant). Therefore we did a transformation on the relevance assessments of INEX: if a document has at least one element which is judged relevant (entirely or partially), this document will be considered as relevant; otherwise it will be considered as non-relevant. There are 29 original CO topics but only 20 topics that have more than 30 relevant documents will be used for experiments.

As mentioned above, our experiments are simulations of personalized searching using user profiles. In this case, 20 topics represent different information needs of 20 different people. For each topic, we choose some relevant papers as "pseudo user profile" of this person (5 in average in our experiments). (Generally, a user profile is a set of information that represent interests and preferences of a user.) The selected papers are chosen among the highly relevant papers to the correspondence topic and those that receive many citations from other documents. After the preparation step, we use the **zettair**[3] search engine to index the INEX collection, then we send 20 queries (which are formed from above topics) to **zettair**; with each query we take the first 300 documents for re-ranking using "user profiles" of correspondence topic. The similarity between a document d and a user profile p is computed as:

$$similarity(p,d) = \sum_{d' \in p} similarity(d',d)$$

In the above formula, *similarity(d',d)* is the similarity (bibliographic coupling and co-citation) between a document d' in profile p and document d. The co-citation similarity between two papers d and d' is defined as:

$$sim_cocitation = \ln\left(\frac{cocitation(d,d')}{citation(d).citation(d')}\right)$$

The bibliographic coupling similarity is computed by a similar formula. The final score of a document is obtained by combining its original score computed by **zettair** and the similarity document-profile by a product function.

Results

The experimental results are presented in Figure 1 and Table 1. In the Figure 1, the metric used is precision/recall. The Table 1 shows precision at 5, 10, 20, 30 documents.

From the experimental results, we can see that the co-citation method using the Web of Science database do not bring any improvement, it even causes a slight performance decrease. The bibliographic coupling method performs better but not very clearly. The Web co-citation method using the Google search engine is the best, it brings 15.06% improvement for the precision at top 30 documents.

[3] http://www.seg.rmit.edu.au/zettair

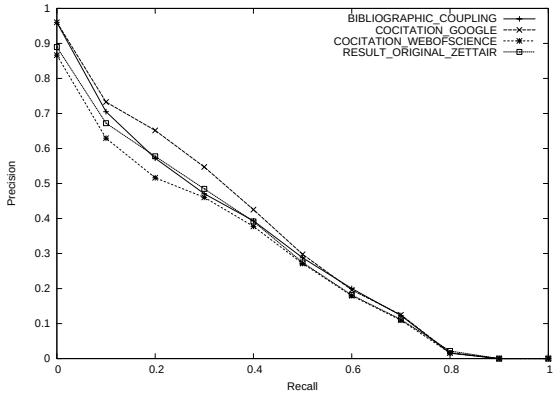


Figure 1. Re-ranking search results of **zettair** with different citation-based methods

Table 1. Precision at 5, 10, 20, 30 documents

Prec.	Original Result	Bib. Coupling	Co-C. WoS	Co-C. Google
@5	0.6600	0.7300	0.6300	0.7100
@10	0.6150	0.6050	0.5800	0.6800
@20	0.5375	0.5600	0.5075	0.6025
@30	0.4867	0.4883	0.4533	0.5600

Conclusions

In this paper we consider two famous citation-based methods: bibliographic coupling and co-citation. We propose a new approach to find co-citation relatedness between scientific papers using the Google search engine. The experiments show that such approach could be more efficient than the traditional approaches.

References

Kessler, M.M (1963). Bibliographic coupling between scientific papers. *American Documentation, 14, 10-25*

Marshakova, I. (1973). System of document connections based on references. *Nauchno-Tekhnicheskaya Informatsiya Seriya 2* (pp. 3-8)

Small, H.G. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *JASIS*, 24(4), 265-269

Vaughan, L. & Shaw, D. (2003). Bibliographic and Web citations: what is the difference*? JASIST,* 54(4), 1532-2882

English Proficiency: A Potential Science Indicator?

Sonia Vasconcelos, Martha Sorenson and Jacqueline Leta

sonia@peq.coppe.ufrj.br, sorenson@bioqmed.ufrj.br, jleta@bioqmed.ufrj.br
Science Education Program (PEGeD), Medical Biochemistry Institute (IBqM), Federal University of Rio de Janeiro (UFRJ) (Brazil)

Introduction

English has widely been acknowledged as the *lingua franca* of science (Montgomery, 2004). Although publication is not solely a linguistic enterprise, writing science in this "English-only" research world might be a hurdle for non-native-English speaking (NNES) researchers, especially. Recently, a relationship among research funding, English proficiency (estimated by TOEFL scores) and publication output has been reported (Man et al, 2004). The authors suggest that there is "a significant relationship of national spending on research and TOEFL scores to publication output of developed countries". Their results have a particular focus on publication in highly-ranked medical journals but challenge the assumption that traditional indicators (citations, funding, and collaboration, among others) alone may explain research output. Concerning the productivity of NNES developing countries in Latin America, *e.g.*, Brazil and Argentina, publication output in ISI-indexed journals has steadily increased (Glänzel, Leta & Thijs, 2006). This growth notwithstanding, publications from these countries still represent a tiny fraction of those journals. This fact may be attributed to many factors, including a limited number of publishing outlets in English and the language barrier itself (Zhoua, P. & Leydesdorff, L., 2006). Indeed, good command of English cannot be taken for granted in academia. A significant number of researchers from NNES countries, *e.g.*, those from South America and Asia, do not speak English even as a second language. Therefore, considering the possible influence of this factor on their productivity may be reasonable. Besides, developing writing skills takes time. Where no academic writing support is provided, what is the English of such authors like? What if it plays a considerable role in their getting published? The literature on scientometrics has not provided the answers. We believe that the role of writing skills in getting published may be misrepresented in the assessment of publication output. Our hypothesis is that English proficiency may be a potential indicator of research performance of NNES countries, and we look at Brazil as an exemplar. We have investigated the academic profile of a sampling of Brazilian researchers, which includes their English proficiency. Our aim is to contribute to a broader discussion of a possible relationship between English proficiency and the productivity of NNES researchers in today's "English-only" scenario.

Data set

This work draws upon a quantitative analysis of research productivity and its relationship with English proficiency This analysis is based on data from a database provided by the Brazilian Research Council (CNPq). It contains the academic profile of researchers, including their publication output (national/international) from 1997 to 2005 and their English proficiency. It covers information of 51,223 researchers, all having a PhD degree. As for English proficiency, data on the authors' self-evaluation of the four English skills, *i.e.*, speaking, reading, listening, and writing, are included.

Results

According to the database, *reading* and *comprehension* [listening] rather than *writing* and *speaking* are the strongest English skills among the 51,223 researchers. Around 78% of the authors in the sample state to have "good reading skills", whereas 44.4% state to have "good writing skills". As our main interest is written English proficiency, our data focuses on *writing*.

Writing skills among fields

The analysis of the different fields of knowledge revealed some remarkable differences in terms of *writing* (Figure 1). More researchers from internationally-oriented research fields, such as hard (chemistry & physics) and earth sciences, engineering, health and life sciences, state to have "good writing skills" as compared to those from more nationally-oriented research fields (54.8% in the hard and earth sciences versus 25.8% in the humanities, for example).

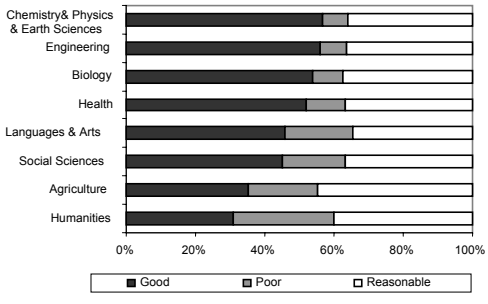


Figure 1. Researchers with different English writingskills among academic fields (%).

Conversely, a larger number of authors from these fields tend to have "poor writing skills. This factor may be associated with their research focus, mostly on national themes. Thus, developing the writing skill in English to write for an international audience may not be an issue for them (Ortiz, R., 2004). The analysis of this group is rather particular and will have a different approach.

Writing skills, productivity and fields

The 51,223 researchers were split into 9 groups according to the number of publications in international journals (2001-2004). The distribution of researchers' groups is based on their written English *proficiency* (Table 1). The fraction of authors with "good writing skills" increases as the number of papers (published in the 4-year time frame) increases, *i.e.*, and there is a trend towards the most productive researchers having "good writing skills". On the other hand, a very tiny fraction of authors with "poor writing skills" are among the most productive. Indeed, the largest number of authors from this group has no international publications or published only a range of 1-2 publications in the period. Whether productive authors have developed their *writing* skills during their academic career or are productive exactly because they had good *writing* skills at the beginning of their academic life is an open question.

Table 1. Distribution of researchers according to their number of publications and English writing skills. (a) = all researchers; (b) = authors with"good writing skills"; (c) = authors with "poor writing skills".

Papers	(a)	(b)	(c)	b/a	c /a
0	31,157	10,191	5,620	32.7%	18.0%
1-2	9,725	5,197	757	53.4%	7.8%
3-5	5,091	3,293	201	64.7%	3.9%
6-10	3,068	2,288	64	74.6%	2.1%
11-15	1,115	889	16	79.7%	1.4%
16-20	489	402	7	82.2%	1.4%
21-30	375	320	1	85.3%	0.3%
31-50	154	132	3	85.7%	1.9%
>50	49	45	0	91.8%	0.0%
Total	51,223	22,757	6,669	44.4%	13.0%

Discussion and Conclusion

Our preliminary data indicates that English proficiency, especially in writing, seems to play a role in the publication output of these Brazilian researchers. Currently, traditional indicators used to assess NNES developing countries' productivity are considered effective. Nevertheless, they fall short of providing an accurate representation of the factors influencing publication in such countries. Our preliminary results have pointed to English proficiency as a potential science indicator in Brazil. We argue that a broader discussion on this language issue should be highly encouraged in academia.

References

CNPq, National Research Council Conselho Nacional de Desenvolvimento Científico e Tecnológico – Diretório de Grupos de Pesquisa, available at http://dgp.cnpq.br/ , 2006.

Glänzel, W., Leta J., Thijs, B. (2006). Science in Brazil. Part 1: A macro-level comparative study. *Scientometrics*, Vol. 67, No. 1 67–86.

Man, J.P.., Weinkauf, J.G., Tsang, M., Sin, D.D., (2004). Why do some countries publish more than others? An international comparison of research funding, English proficiency and publication output in highly ranked general medical journals. *European Journal of Epidemiology,* 19, 811–817.

Montgomery, S. (2004). Of towers, walls, and fields: perspectives on language in science. *Science*, 303, 1333-1335.

Ortiz, R. (2004) English and the social sciences. Revista Brasileira de Ciências Sociais, 19-54, 5-22

Zhoua, P., Leydesdorff, L., (2006). The emergence of China as a leading nation in science, Research Policy, 35, 83–104.

Common Statistics of Heterogeneous Tagging Systems

Jakob Voss

jakob.voss@gbv.de
Common Library Network (GBV), Head Office (VZG)
Platz der Göttinger Sieben 1, 37073 Göttingen (Germany)

Introduction

Authors, institutions, publications, and citations are widely used in informetric studies. Abstract subjects as primary entities of analysis are far more difficult to deal with because it is less obvious what the subject of a given document is. To describe a document's subject several methods and types of knowledge organization systems like abstracting, keywords, and classification codes have been developed in information retrieval. The dominance of search engines that build on automatic indexing has made manually created subject indexing to appear out-of-date. However it is still used with success. Since around two years a revival of web based indexing is going on and the new term *tagging* is used to describe this new systems. In this work several new and old tagging systems are compared and analyzed with methods of informetrics and information science.

Methods

Several tagging systems were selected, namely del.icio.us, Dewey Decimal Classification (DDC), Wikipedia's category system, the JITA classification for library and information science, the AGROVOC thesaurus, and the German social bookmarking service LinkARENA. Tagging data was collected with scripts from RSS-feeds, available database dumps, and library catalogues. To clarify their differences and commonalities the tagging systems were analyzed by a revised typology (Table 1) that was developed in (Voss, 2007). The data was cleaned and histograms were calculated to determine distributions of tags, resources, and users. The statistical software R was used to test, calculate, and visualize distribution patterns and their coefficients.

Results

Tagging systems may differ a lot in user interface, usage, and variety of typology dimensions, but the general structure is comparable. The usual model is a tripartite graph, with resources, users, and tags (Lambiotte & Ausloos, 2005). It was discovered that the distribution of these nodes can be described by simple laws in all analyzed tagging systems. A combined visualization is presented at the poster. However it shows research-in-progress; additional tagging systems and larger data sets are needed to find out if or under what conditions the laws are generally valid and which dimensions of the tagging typology affect their parameters in which way.

Popularity of tags

The frequency distribution of tags is often described by power laws (Halpin, Robu & Shepherd, 2007). Power laws are common in informetrics, bibliometrics, and scientometrics. White (2005) suggested naming this distributions *bibliograms*. In practise the "core" of relatively few top-ranked terms is of most interest, for instance for recommender systems. In modern tagging systems the popularity of tags is frequently visualized with a *tag cloud*. A tag cloud is a visually weighted list of the most popular tags. The power law coefficient seems to depend on the tagging system's vocabulary connectivity.

Number of tags per document

By far most documents are indexed with only one tag. If more tags are assigned to a document then the number of tags per document follows an exponential distribution with mean parameter around $\lambda = 0.5$. The parameter seems to be influenced most by tagging feedback and external rules.

Table 1. Tagging typology dimensions and their variety in the analyzed tagging systems.

Dimension	Found variety
Tagging Rights	Tags by authors, experts, users
Source of Resources	Newly created, collected for own usage and other users
Resource Representation	Online (websites, papers), physical (books)
Tagging Feedback	no feedback, support by lists, automatic suggestions
Tag Aggregation	set-model vs. bag-model
Vocabulary control	Free keywords, managed categories, fixed lists
Vocabulary connectivity	Unconnected, network (thesaurus), tree (hierarchy)
Resource connectivity	Ignored in this study
Automatic tagging	Ignored in this study

Tag levels
Some tagging systems (classifications and thesauri) provide hierarchical relations between tags. All tags without link to a broader term (also 'top terms') build the root you can reach all other descriptors from. The level of a tag is defined as the length of a minimal path to a root. Small classifications have a limited tag level and the number of tags per level grows exponentially. It is shown that in larger classifications and thesauri the distribution of tag levels is more or less normal with little tags at the root and at the deeper levels but most tags with average levels.

Number of broader tags in multihierarchies
In multihierarchical tagging systems (thesauri) the number of broader tags per tag is distributed by a power law. The details seem to be affected by several aspects of the tagging system.

Conclusions
Tagging systems have only little been studied with methods from information science and informetrics. Most research on tagging on the web comes from disciplines that are unaware of the connections to classical subject indexing as known in library and information science. Collaborative tagging has only been compared superficially with professional subject indexing but not with statistics. At this point informetrics is helpful. First comparisons have shown that common structures can be described by simple laws and distributions. Detailed results and visualizations are shown at the poster. Further result should try to get data of more and more diverse tagging systems to find general structures of tagging systems and the properties that affect them.

References
Halpin, H., Robu, V., Shepherd, H. (2007). The Complex Dynamics of Collaborative Tagging. In *Proceedings of the WWW 2007 Conference.*

Lambiotte, R & Ausloos, M. (2005). Collaborative tagging as a tripartite network. In *Lecture Notes in Computer Science*, 3993 (pp. 1114-1117). http://arxiv.org/abs/cs/0512090

Marlow, C., Naaman, M., Boyd, D. & Davis, M. (2006). Position paper, tagging, taxonomy, flickr, article, toread. In *Proceedings of the WWW 2006 Collaborative Web Tagging Workshop.* http://www.rawsugar.com/www2006/29.pdf

Voss, J. (2007). Tagging, Folksonomy & Co – renaissance of manual indexing? In *Proceedings of the 10th International Symposium for Information Science*. Cologne. http://arxiv.org/abs/cs/0701072

White, H. D. (2005). On Extending Informetrics: An Opinion Paper. In P. Ingwersen & B. Larsen (Eds.), *Proceedings of ISSI 2005* (pp. 442-449)

Innovative Capacity Evaluation of Main Countries Based on Patent Analysis

Chun-Chieh Wang [*], Mu-Hsuan Huang [*] and Dar-Zen Chen [**]

[*] *d95126002@ntu.edu.tw, mhhuang@ntu.edu.tw*
Department of Library and Information Science, National Taiwan University, Taipei (Taiwan)

[**] *dzchen@ntu.edu.tw*
Department of Mechanical Engineering, National Taiwan University, Taipei (Taiwan)

Introduction

In the 21th century, competitions between countries become more violent and countries have taken the development of innovative capacity as the major targets. In this article, we use patent indicators generated from utility patents granted in the U.S. patent system. We chose top ten countries of utility patent counts in period of 2000-2005 and assigned patents to five technology areas. We used patent data to compute several indicators: Relative Growth Rate (RGR) (Albert, Yoshida, & Opstal, 1998), Activity Index (AI) (Albert, Yoshida, & Opstal, 1998), Current Impact Index (CII) (Albert, Avery, Narin & McAllister, 1991), Essential Patent Index (EPI) (Chen, Lin & Huang, 2007) and Essential Technology Strength (ETS) (Chen, Lin & Huang, 2007) for each country. The main object of this article is to explore, compare and contrast the innovative capacity difference between countries. The result of this article could be useful to aid the understanding of the critical role that innovative capacity ensures the economic competitiveness of these countries.

Results

In this article we analyzed the innovative capacity of ten countries in five technology areas in period of 2000-2005 through patent indicators. The results we found include the follows.

United States: United States performed the strongest ETS in every technology area and leaded all other countries by a wide margin.

Japan: Japan's ETS was ranked second in every technology area and still leaded other countries by a margin, because it had a large number of patent counts.

Germany: Germany's ETS was ranked third in Chemical, Biotech & Drugs and Mechanical.

Taiwan: Taiwan's best ETS rank order was third in Electrical & Electronic but its patenting trend was declaim and weak in this technology area.

South Korea: South Korea got higher ETS in Chemical, Computers & Communications and Electrical & Electronic.

Canada: In Computers & Communications, Canada's ETS was ranked third, the best of five technology areas.

Other Countries: French, United Kingdom, Netherlands and Switzerland's ETS were all ranked after top third countries in each technology area.

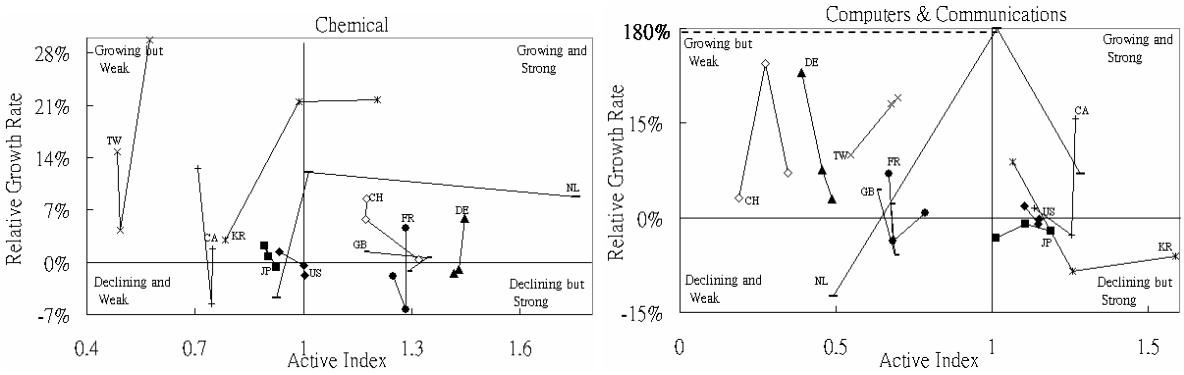
For a further improvement of this innovative capacity evaluation, it is recommended that the analyzed period be longer and technology areas be tuned to the detail classified requirements for specific industrial research demands.

References

Albert, M.B., Avery, D., Narin, F., and McAllister P. (1991). Direct validation of citation counts as indicators of industrially important patents. *Research Policy*, 20: 251-259.

Albert, M.B., Yoshida, P.G., and Opstal, D. (1998). The New Innovators: Global patenting trends in five sectors. U.S. Department of Commerce, Office of Technology Policy. Retrieved November 06, 2006 from: http://www.tech.gov/Reports/09111998.pdf.

Chen, Dar-Zen, Lin, Wen-Yau and Huang, Mu-Hsuan, (2007). Using Essential Patent Index and Essential Technological Strength to Evaluate Industrial Technological Innovation Competitiveness. *Scientometrics,* 71(1): 101-116.



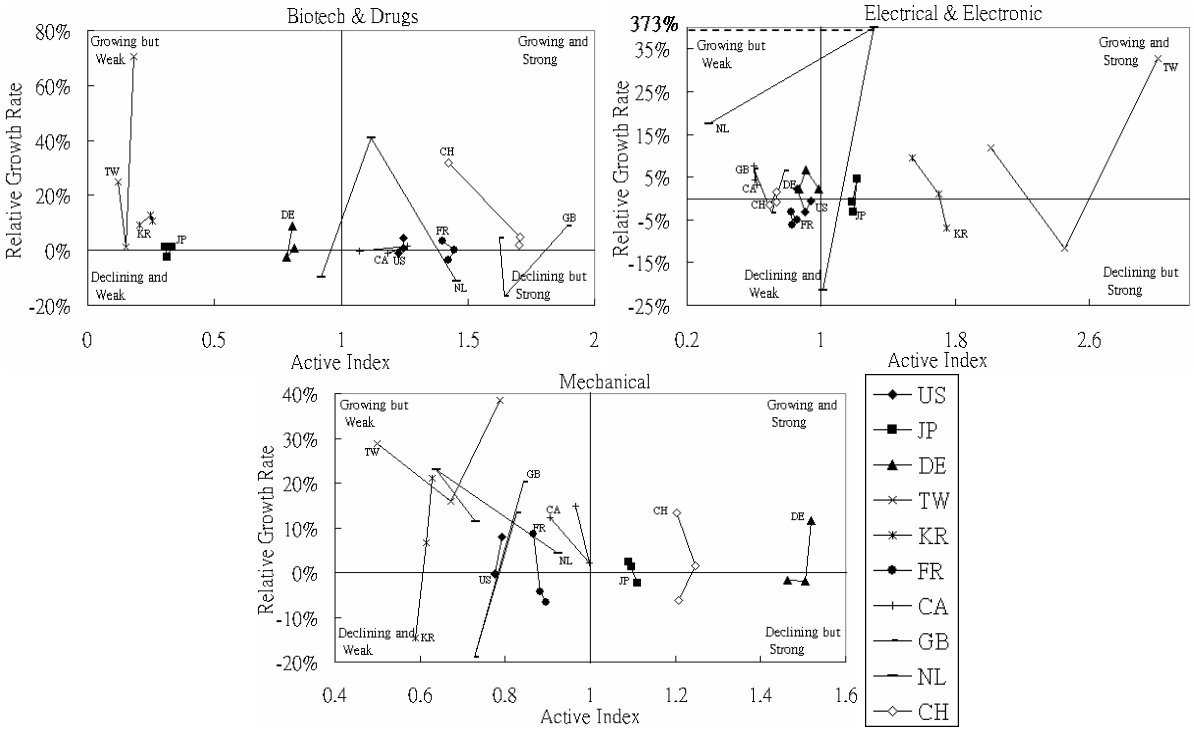


Figure 1. RGR vs. AI of ten countries in five technology areas in three periods.

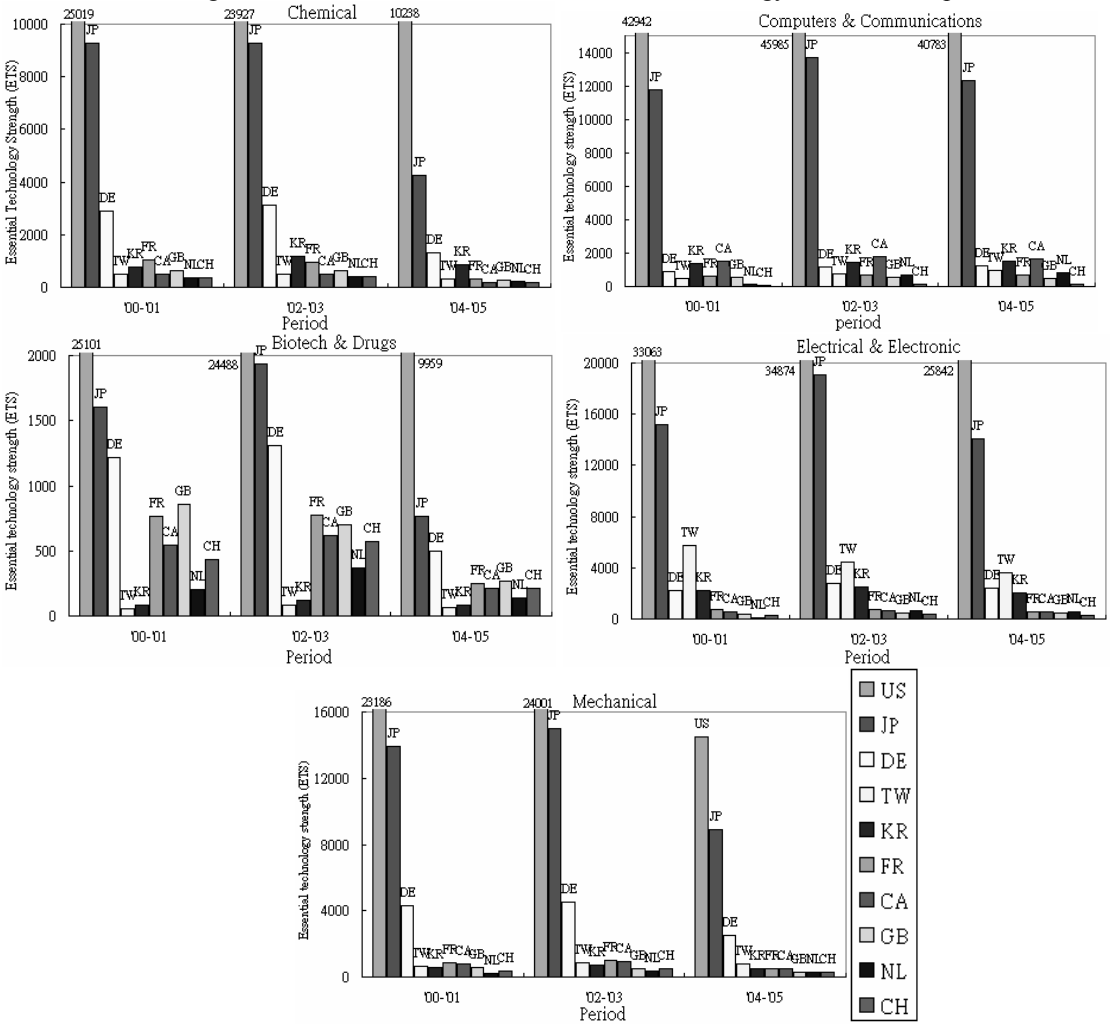


Figure 2. ETS of ten countries in five technology areas in three periods.

Search Characteristics in Different Types of Internet-Based IR Environments: Are They the Same?

Dietmar Wolfram

dwolfram@uwm.edu
School of Information Studies, U. of Wisconsin-Milwaukee, P.O. Box 413, Milwaukee, WI 53201 (USA)

Introduction
Studies of mathematical modelling author and journal productivity frequencies, word usage, and resource usage have examined many of the social aspects of information production and use. With the growth of the Internet, and particularly the World Wide Web, these studies have more recently focused on empirical regularities of Internet users and usage. The broader availability of electronic datasets and more recently Internet technologies have made it possible to study significantly larger samples consisting of hundreds of thousands to many millions of observations.

Although search facilities for different information retrieval (IR) environments on the Internet (Online Public Access Catalogues [OPACs], bibliographic databases, Internet search engines) may appear similar, the content to which they provide access can be fundamentally different. "Popular" search environments such as public search engines provide access to full text content on a variety of topics. Bibliographic databases provide full text or surrogate access to formally published document sources. OPACs primarily deal with access to monographic and serial-level access to document surrogates. Do these differences in content result in differences in search behaviour by users, or are the same behaviours observed across systems? In other words, do Internet users treat search facilities (i.e., search boxes) the same way, or does the content and nature of each environment result in different search regularities? Previous studies that have examined these differences have relied on separate earlier research that might not be comparable (Jansen & Pooch, 2001) by focusing on specific audiences or topic areas that were not necessarily Internet-based.

The purpose of this study is to compare the characteristics of end user searching characteristics using several different public search environments. Datasets consisting of tens to hundreds of thousands of records are analysed for differences in distributions of search characteristics and descriptive statistics of these distributions

Previous Studies
To date, there appear to have been no direct comparisons of Web-based IR transaction log outcomes across different environments, although a number of public search engine logs have been compared, for example, Spink & Jansen (2004), who compare findings of several search services in their monograph, and more recently they perform a comparison of nine datasets representing five search engines (Jansen & Spink, 2006). Jansen & Pooch (2001) do provide a side-by-side comparison of search data from different IR environments, but the comparison is made based on Web and non-Web systems that represent closed, pre-Web IR environments, that are not necessarily based on unobtrusive transaction logs. The current study attempts to perform a direct comparison of these different Web-based environments.

Method
Transaction log datasets were collected from four different types of Web-based IR environments. The amount of data from each system was limited to what was made available to the author by the data providers (Table 1). Datasets included representation from the following environments:

Bibliographic Databank – The University of Wisconsin-Madison libraries provide access to dozens of general and specialized bibliographic reference and full-text databases through one interface. Size: 66,895 queries, 37,499 sessions collected over nine months.

Online Public Access Catalogue – Query data are from the University of Wisconsin-Milwaukee's OPAC. The OPAC indexes millions of monographic, periodical titles and other media content. Records are primarily surrogates, but the OPAC also provides access to a growing amount of full-text material. Size: 18,565 queries, 8,250 sessions collected over one month.

Search Engine – Publicly available query data from the Excite search engine were used. Excite indexes many millions of websites on any topic. Size: 587,145 queries, 327,051 sessions (estimated) collected over one day.

Specialized Search Service – The Medical College of Wisconsin HealthLink service is a full text consumer health information resource that is available to any Web user. The service indexes and provides access to thousands of articles on health-related topics. Size: 377,701 queries 161,277 sessions (estimated) collected over one year.A

method for identifying session boundaries was needed for the Excite and special topic database, as these were not directly recorded by the Web servers. A novel method was used for this purpose by relying on the distribution of time intervals between queries submitted by each identifier. Methods proposed by others either were not applicable or were too labour-intensive given the size of the data sets. The method used for session boundaries was to use the time interval representing the 80th percentile of all inter-query time intervals. When graphed against the average session length for different percentile cut-offs, the 80th percentile corresponds closely to an inflection point representing a slowdown in the increase in session lengths as the query interval increases. To determine the effect of different cut-off points on the average session length, query intervals representing the 10th through 90th percentiles of values were selected with the resulting average session length calculated for the Excite and HealthLink datasets. When graphed using a semi-log plot, there is a notable increase for average session lengths over short query interval increases, but these increases become much smaller with larger cut-off times. The average session values take on an S-shape, with the inflection point occurring high in the distribution. These inflection points signify a marked slowdown in the increase of the average session size, and occur around or just before the 80th percentile. Although not an ideal solution, considering an 80% cut-off point takes into account the characteristics of each dataset and deals with the obvious inter-query time outliers that extend across many hours, which skew the session sizes and average inter-query times, and overestimate the number of longer sessions.

Findings & Conclusions

Data were analysed at the term, query and session level. Only selected findings for the query and session level analysis are reported here due to space limitations (Figures 1 & 2). Based on the mean values and distributions the following observations of each environment can be made:

Bibliographic Databank – Queries are short and of academic interest. Sessions are short and brief in duration, indicating little effort to evaluate results.
Online Public Access Catalogue – Queries are longer and reflect more academic interest. Sessions are longer with longer inter-query times, indicating more review and evaluation of retrieved content.
Search Engine – Queries are of mid-length and of general or popular interest. Sessions are short but reflect the longest time period between queries.
Specialized Search Service – Queries are short and focused. Sessions are long in queries, but reflect very short durations, indicating quick assessment of the returned results for each query.

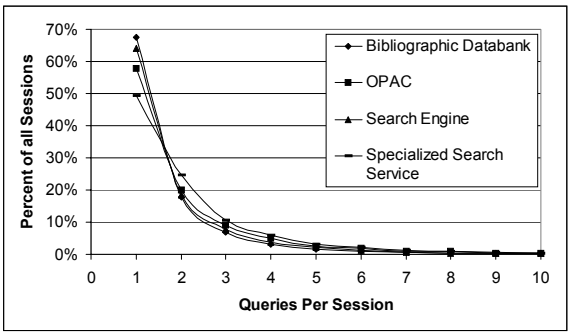


Figure 1. Distribution of Queries per Session

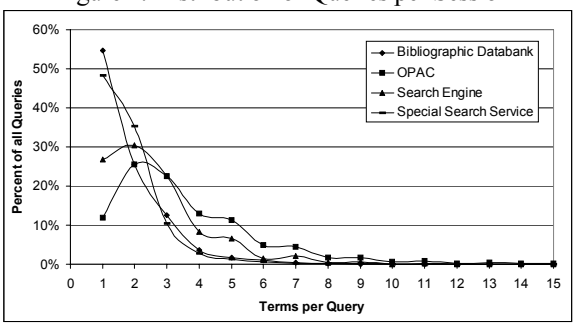


Figure 2. Distribution of Terms Used per Query

Given the differences in the search patterns, it appears that users do engage in different search behaviours and that Web search facilities are not viewed as the same. The two surprising finding are the short query lengths and search sessions for the bibliographic databank. As an example of a more traditional IR environment providing access to formally published resources with more structured searching, one would expect user search behaviours to be more in line with what is observed for the OPAC, where queries are purposeful, longer in length, with more queries per session and longer inter-query times reflecting the perusal of results.

Acknowledgments

The author would like to thank Excite@home, the Medical College of Wisconsin, the University of Wisconsin-Madison Library System, and University of Wisconsin-Milwaukee Golda Meir Library for providing access to the datasets used in this study:

References

Spink, A, & Jansen, B. J. (2004). *Web search: Public searching of the Web*. Dordrecht: Kluwer.

Jansen, B. J., & Pooch, U. (2001). A review of Web searching studies and a framework for future research. *Journal of the American Society for Information Science & Technology*, 52(3), 235-246.

Jansen, B. J., & Spink, A. (2006). How are we searching the World Wide Web? A comparison of nine search engine transaction logs. *Information Processing & Management*, 42(1), 248-263.

Influence of the International Migration of Researchers on National Publications in Three Fields of Engineering

Yasuhiro Yamashita[*] , Sen Ueno[**], Hiroyuki Tomizawa[**] and Masayuki Kondo[***]

[*] *yaya@jm.kj.yamagata-u.ac.jp*
Evaluation and Analysis Office, Yamagata University, 1-4-12 Kojirakawa-machi, Yamagata-shi, 990-8560 (Japan)

[**] *uenosen@nistep.go.jp, tomizawa@nistep.go.jp,*
National Institute of Science and Technology Policy, 2-5-1 Marunouchi 100-0005 (Japan)

[***] *mkondo@ynu.ac.jp,*
Yokohama National University, 79-7 Tokiwadai, Hodogaya-ku, Yokohama 240-8501 (Japan)

Introduction

The purpose of this study is to analyze the relationship between the international migration of researchers and the production of research articles, and to demonstrate methods for analysis based on biographies of the researchers that accompany research articles on three engineering disciplines.

Although few previous quantitative studies have been based on researcher biographies, there are some exceptionally interesting studies that were made in China (Chen & Liu, 2005, Jin, Li & Rousseau, 2004 and Liang et. al., 2001). These studies mainly focused on the age structure. In addition, Ioannidis investigated the international migration of researchers who published high-impact articles, using the ISIHighlyCited.com database compiled by the Thomson Corporation (Ioannidis 2004). We attempted to find out the influence of the international migration of researchers by analyzing data from high-impact international journals.

Methods

Three journals published by the Institute of Electrical and Electronics Engineers, Inc. (IEEE) during two periods - 1991 and 2001 - were used for the purpose of this study. The impact factors (IF) of these journals in their respective fields were the highest or second highest in 2002. The three journals are listed herewith:

a. IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI, IF = 2.923; ranked the second highest in the category of "Computer Science, Artificial Intelligence.")

b. IEEE Transactions on Robotics and Automation (T-RA, IF = 1.048; ranked the highest in the category of "Robotics.")

c. IEEE Transactions on Semiconductor Manufacturing (T-SM, IF = 1.031; ranked the second highest in the category of "Engineering, Manufacturing.")

We regarded the countries in which the universities the researchers graduated from were located as their country of origin. If the university from which a researcher acquired his or her Bachelor's Degree or an equivalent degree was not disclosed, the country of birth was considered as the home country. Information on those researchers who disclosed neither the university from which they graduated nor the country of their birth was gathered from the Internet. Approximately 90% of all the articles were attributed to the authors' places of origin. To avoid overrating the articles written by multi-authors, each author's contribution to an article was equally counted as one divided by number of authors. The articles were counted with regard to the two following factors:

No(A): Number of articles written by authors originating from country/region A,

Ni(A): Number of articles produced by institutes located in country/region A.

Results

Changes in the article production structures in major countries

Figure 1 illustrates the shares of publication of major countries. The originating and terminating points of the arrows designate the shares of publication in 1991 and 2001, respectively. Unknown types were excluded from the calculation of No(A).

The values for the United States are located below the diagonal in all the figures; thus, the United States relatively utilizes researchers of foreign origins for research activities. China (T-PAMI and T-RA), Korea (T-RA) and Taiwan (T-SM) changed their research style during these ten years from that mainly conducted by expatriate researchers to that by native researchers residing them. On the other hand, India, with regard to T-PAMI and T-RA, native researchers produced many articles, but Indian institutes produced virtually no articles in both 1991 and 2001.

Effect of "researchers' migration across regions"

To present more reliable data by avoiding small samples, we categorized each country into one of No(A) of Asia and Oceania is equivalent to or more than those contributed by North America and Western and Northern Europe. However, a significant number of the articles written by researchers of Asian origin were written through the institutes outside Asia and Oceania. Meanwhile, North America and Western and Northern Europe relatively utilized the expertise of nonnative researchers.

The origin of researchers who published through Japanese/US institutes

The United States accounted for 40% to 50% of the global share of Ni(A) (see the pie graphs in Figure 2). However, the breakdown of Ni(A) according to the researchers' places of origin revealed that a large number of articles were produced by nonnative researchers, such as Indian and Chinese.

In contrast, Japan is placed at around the third to seventh positions in terms of the share of publication in the three journals. The breakdown of publication shares revealed that most of the Japanese articles were written by Japanese researchers. This corroborated the statement that outstanding researchers from Asia prefer to migrate to the United States as opposed to Japan.

Conclusion

There are very few international journals that publish author biographies along with articles that they wrote or helped write; hence, the arbitrariness of journal selection and the low levels of confidentiality were inevitable due to the small sample size. For a more reliable analysis, it appears that it would be necessary to analyze these phenomena for a period of three years or more for a larger sample size.

Despite the abovementioned limitations, researchers' biographies may provide valuable information on the invisible structures of research activities that might complement ordinary bibliometric data.

Acknowledgment

We thank Dr. Tadao Saito for giving us the principal idea of our work.

References

Chen, Y. & Liu, Z.(2005). An analysis of co-authorship of management science in China, *Proceedings of ISSI 2005* (ISSI2005) (pp 733–739), Stockholm, ISSI.

Ioannidis, J. P. A.(2004). Global estimates of high-level brain drain and deficit, *FASEB Journal*, 18, 936–939.

Jin, B. H.,Li, L. & Rousseau, R. (2004). Long-term influences of interventions in the normal development of science: China and the cultural revolution, *Journal of the American Society for Information Science and Technology*, 55, 544–550.

Liang, L., Kretschmer, H., Guo, Y., deBeaver, D. (2001). Age structures of scientific collaboration in Chinese computer science, *Scientometrics*, 52, 471–486.

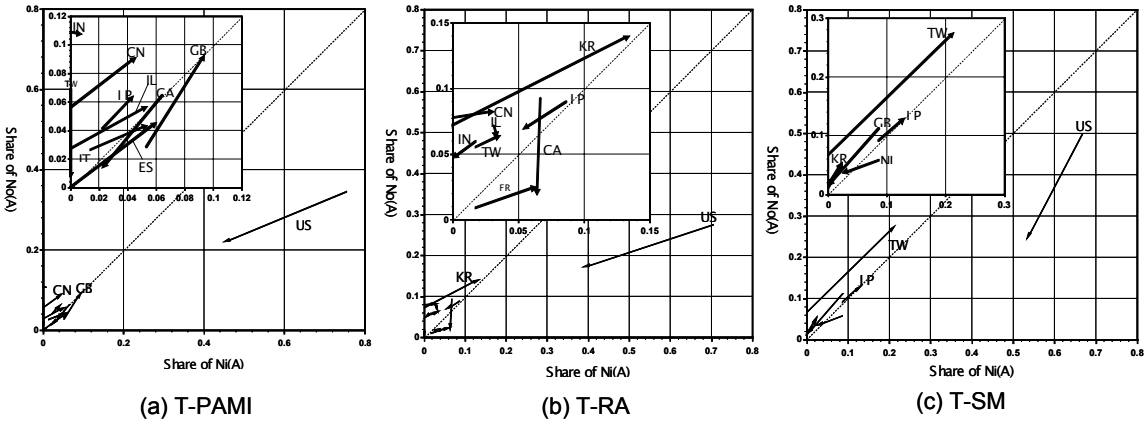


Figure 1. Share of Ni(A) and No(A) of major countries in 1991 and 2001

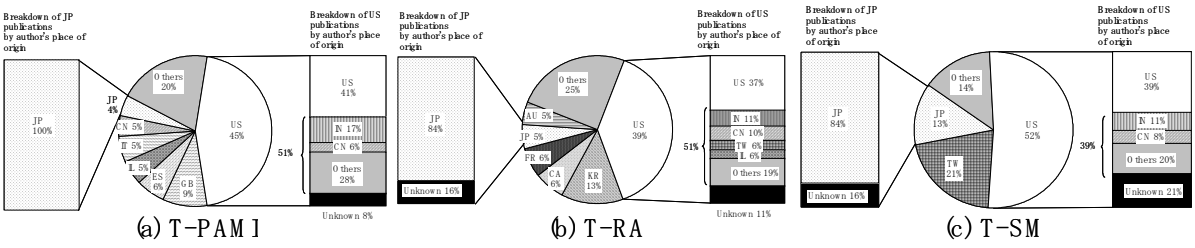


Figure 2. Breakdown of Ni(A) of Japan and US in 2001

Citation Analysis of the Chinese Journal of Medical Science Management

Li-li Yang, Dong-hong Fu, Xiao-ping Bai, Wei-gang Fang, Shu-yin Yao

kgzz@bjmu.edu.cn
Peking University Health Science Center, Scientific Resaerch Division, 38 Xueyan road, Haidian Region ,Beijing 100083 (P.R.China)

Introduction

The Chinese Journal of Medical Science Research Management is a academic journal that focuses mainly on medical research management discipline. The journal makes its best efforts to publicize governmental policies and guidelines for scientific research and report research achievements and progress in the field of medical research management in China. The journal is a platform of intercommunion and showing academic achievement of medical research management for individuals and groups in scientific community in China. Since the publication in 1988 when the journal started publication its quality has been increasing. However, there are some undesirable aspects. We should carefully summarize experience and think over the idea about publishing this journal, which can help to push the development of this discipline and field, to advance its academic level and influence, to improve its quality and make it a leading journal in this subject.

Material and Methods

46 issues of the journal, from the eighth volume 1995 to the eighteenth volume2005 have analysed the traits in respects of quantity, languages and types on the basis of published essays, summaries, lectures and cases. It is used 'Chinese Journal of Medical Science Research Management' as retrieve word and analyzed cited situation in the papers published in it on CNKI-Chinese Citation Database located on National Knowledge Infrastructure, retrieve time was up to 1 March 2006.

Results

These 46 issues published 1181 papers, among them, 765 papers contained references, mean citing rate is 64.78 per cent. The number of cited documents was 2906 pieces, averaging number of citation was 2.46.

The categories of the citation were mainly journal articles and books,1913(65.83%)of total share was published as journal articles, 679(23.36%)as books,63 (2.17 %) as news features and others (government standard, guidelines, dissertation and so on) 251(8.64%), respectively, and the citations were main to Chinese language documents, which accounted for 91.43 per cent .The number of papers in English amounted to 241, while that in Japanese was 8 constituting 8.29 per cent and 0.28 per cent respectively.

1751 Chinese citations located in 315 the Chinese language journals. The top five journals were Chin J Med Sci Res Manage, Science Research Management, Chin J Hosp Admin, R & D Management, Chinese Hospital Management. The references in the top 5 journals ranked by the amount of citation accounted for 54.88% of the total journal citations. 679 Chinese citations come from 490 the Chinese language books, thereinto Medicine Scientific Research Management published by People Health Press in 1991 was cited 123 times by the journal during 11 years. This book has become famous in the field of medicine scientific research management in China. Science, Nature and JAMA are the top three English journals ranked by the amount of citation accounting for 29.63%. The Prices's Index of the journal was 72.00%. The data shows that citing rate of recent references by the journal and thus the ability of the journal to absorb and utilize new information is high .A higher Prices's Index usually reflects more updated content in the publication. Higher Prices's Index also indicates that the researcher is in front of his research field." Gao, et al(2004) indicated that Usually Prices's Index of a newly forming discipline is higher than that of a long-standing discipline,and the average Prices's Index for all disciplines is about 50%". Self-citing rate of the journal was 22.57%.

547 out of 1181 papers were cited 1509 times, cited rate accounted for 46.32%. Among these 1509 times, 177 times were cited by dissertations produced by 37 universities and 1332 times were cited by 260 journals. The journal's self-cited was 471 times, self-cited rate was 31.21%. According to the version 2003, 2004, and 2005 of Chinese Academic Journal Comprehensive Citation Annual Report, the journal's cited half-life is 2.9, 2.9 and 3.4 respectively which mean its trend is stable.

Discussion

The Amount of Citation

Citations are the documents quoted by an author when he or she generates his or her academic paper.

Generally, they are listed on the end of the paper and is the important component of an paper. They are not only papers' scientific evidence, authors' ability to absorb information, show continuity of the research and succeed to the previous scientific research so as to trace given subjects' development but also can be regarded as one of the most important indicators to evaluate the quality of the journals. In Chinese Journal of Medical Science Research Management, 1181 source articles with 2906 cited items averaged 2.46 cited references per article. The figure is lower than 3.74 that "Jiang (2001) found that the journals of Disciplines of Science and S&T Management averaged 3.74 cited references per article.'' This situation means that authors' consciousness and ability of drawing on intelligence of this journal-related is weak.

The Types of Citation

The types of citations in the Journal include many special literatures including of standard document, patent document, dissertation and so on, and newspapers amounting to 8.64 percent and 2.17 percent respectively except journals and books, which relates to its discipline characteristics. Because it is used as a management discipline, its idea of releasing journal is to publicize and carry out national health policies, guidelines and criterions and so on .

Citation Used Languages

By analyzing citations used languages we can find out authors' ability to employ home and abroad literatures' information and the journal's ability to draw on information. We know that researchers in this field express and intercourse their research findings in which language, which can reflect they master which language and the ability to read and use it. The mainly used language, Chinese accounts for 91.43 percent followed by English and Japanese amounting to 8.29 percent and 0.28 per cent respectively. This situation suggests that, on the one hand, reference resources applied by medical scientific research administrators come mainly from Chinese literatures. On the other hand, the authors are short of ability to master and employ literatures, as in the future this ability should be improved.

Self-citing and Self-cited Rate

Our research findings show that the self-citing rate of the journal was 22.57 percent and its self-cited rate was 31.21 percent that it is obvious smaller than those of Chin J Hosp Admin, Hosp Admin J Chin PLA, and Journal of Nursing Administration. "Li & Hou(2006) found that their self-cited rates were 0.79, 0.92 and 0.571 respectively". It suggests that medical administration discipline has its independent property compared with other disciplines, on the other hand this discipline is lack of use for the methods, knowledge and research achievements of correlation disciplines.

Conclusion

The most remarkable characteristic of citation analysis approach lies in its objective. The reason is that academic papers should be interacted one another, which shows that knowledge and intelligence can be come in for and used and indicates the scientific development. "Carr & Britton(2003) indicated that such data are amenable to longitudinal analysis and can be used as a measure of a journal's impact on research within a discipline". The citation is that authors use citing literatures when they produce their articles. Generally speaking, citing literatures are listed on the last of papers. Citations are an important part of a paper. On the one hand, they reflect cited paper's visibility, the ability of an author drawing on and using intelligence, showing the continuity and inherited traits on scientific researches so that people can track a special subject. On the other hand, it is an important indicator for assessing the level of journals.

References

Gao, J. Q., Zhang, R.A., Chen, J.P., Shi , Y. (2004). Periodical citation analysis and its information value. *Library and Information Service,* 48, 58-6.

National Knowledge Infrastructur (CNKI). (2003-2005).2002-2004 Chinese Academic Journal Comprehensive Citation Annual Report. *Tsinghua university pres*s.

Jiang,C.L. (2001).Analysis and evaluation on quotations in 5 categories of magazine about disciplines of science and S&T management .*R &D Management*, 13(1), pp.70-72.

Li,Y.J.&Hou H.Q.(2006). The"adjusting"of journal self-citation toimpact factor. *Journal of the China Society for Scientific and Technical Information,* 25,172-178.

Carr,J.,E.& Britton, N.L. (2003). Citation trends of applied journals in behavioral psychology 1981-2000. *J Appl Behav Anal.*, 36(1),pp. 113–117

Assessment of Traditional Chinese Medicine for Substance-Related Disorders Based on Bibliometric Techniques

Lin Yang[*] , Lili Yang[**] and Chengzheng Zhao[***]

[*]*ynl0524@sina.com*
Library of Peoples' Hospital, Peking University, Beijing 100044 (PR China).

[**]*kgzz@bjmu.edu.cn*
Peking University Health Science Center, Scientific Research Division, 38 Xueyan Road, Haidian District,Beijing 100083 (P.R.China)

[***] *zhaocz@bjmu.edu.cn*
National Institute on Drug Dependence, Peking University, Beijing 100083 (PR China).

Introduction

Traditional Chinese medicine (TCM) includes Chinese medicine and acupuncture. TCM herbs consists of natural products including plants, animals and minerals. TCM herbs has been practised in China for more than 2000 years, and for the past 200 years it has been used in the treatment of drug addiction. "Zhao, et al.(2004) indicated that the drug problem in China reappeared in the late 1980s. Beginning in the early 1990s, drug abuse spread quickly. The number of registered drug addicts increased from 70,000 in 1990 to one million by the end of 2002". "Tang, et al.(2006) indicated that Opiate abuse contributes substantially to the spread of HIV/AIDS in China, with intravenous drug use the most prevalent route of transmission (51.2%)." "SHI, et al.(2006) found that there are some advantages in using TCM for opiate detoxification, including less harmful side effects, high safety and ideal effects in the inhibition of protracted withdrawal symptoms and relapse. Co-administration of TCMs with modern medicine shows some synergistic effects in detoxification. Many TCM for detoxification also have efficacy on the rehabilitation of abnormal body functions induced by chronic drug use, including improving immune function, increasing working memory and preventing neurological disorder. Given that TCM is effective in the prevention of relapse and causes fewer side effects, it may be used widely in the treatment of opiate addiction." From bibliometric perspective, this paper analyses the characteristics about the literatures of TCM for the management of substance-related disorders in Chinese biomedicine literature database (CBMnet) during the period 1978-2005, which will be helpful for researchers in the fields of drug dependence, clinic, and teaching to realise the mainstream and trend of TCM for drug detoxification, to determine its weak and strong points, confirm an accurate research direction, and ensure scientific research and clinic activities going on wheels.

Methods

This paper used the data extracting from database CBMnet, developed by Institute of Medical Information, Chinese Academy of Medical Sciences & Peking Union Medical College. We adopted a thesaurus approach, and the search strategy is Substance-Related Disorders and 8 subheadings: diet therapy(DT), massage therapy (AL), qi gong (QL), acupoint therapy (XL), acupuncture therapy (ZL), TCM & western medicine therapy(ZJ), TCM & Chinese herbal drugs (ZH), Chinese herbal drugs therapy (ZD), TCM (ZY). Totally 861 papers about TCM for substance-related disorders were retrieved from CBMnet from 1978 to 2005 and an analysis was conducted on these papers with bibliometric techniques and indicators.

Results

Altogether 861 publications were searched out within the period of 1978-2005. Most of them published after 1990s with 2001 seeing the most publications. No related paper was published in 1990, while 1978, 1980, 1982, 1985, 1987, and 1991 only saw one paper each year. Two papers were published in 1986, three in 1992, seven in 1993, and the trend of publications during the period of 1994-2005 is displayed on figure 1. The number of publications from 1994 to 2001 had increased and reached its peak in 2001. However from then the trend reversed. The reasons that result in this situation need further research.

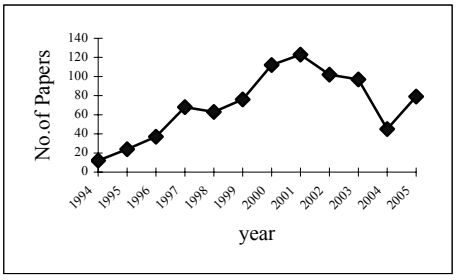


Figure 1 The numbers of the publication from 1994 to 2005

The Journals Carrying These Papers

These 861 papers were released on 273 journals, with 22 journals carrying more than 10 papers, the total number of which was 330 papers amounting to 38.33 %. These 22 journals accounted for 8.09 % of all journals, among which Chinese Journal of Drug Abuse Prevention and Treatment, Chinese Journal of Drug Dependence, Chinese Journal of Integrated Traditional and Western Medicine and Chinese Acupuncture & Moxibustion carried the most, so they could be regarded as core journals on traditional Chinese medicine for substance-related disorders.

Documental Types

According to research attribute, 861 publications were classified into reviews, clinic research papers with specific patient cases, experimental research papers in which there were clear animal experiment or pharmic extraction research findings etc., theoretical research papers were involved in Chinese medical pathologic mechanism, science of prescription etc., and other papers including abstract, translation, meeting papers, and basic popular science(table 1).

Table 1 Documental Types

DT	R	CR	ER	TR	Other	Total
Items	69	512	158	83	39	861
%	8.01	59.47	18.35	9.64	4.53	100

*DT= Documental Type, R= Reviews, CR= Clinic Rresearch //ER= Experimental Research, TR= Theoretical Rresearch

Statistics of Regions

806 out of 861 papers with authors' addresses. They were from 32 provinces, metropolis, and autonomies.

Most papers were produced in Guangdong, Beijing and Shandong where located institutes because of their fast economic development.

Statistics of Authors

There were authors' names on 857 out of 861 papers. Thereinto, the number of author per paper as follows: 250 had one author, 161 two authors, 153 three authors, 93 four authors, 176 five or more and there was a paper created by 14 authors. Total number was 2575 persons/times and mean number of authors per paper was 3 persons

Statistics of Thesaurus

391 papers were about drug detoxification. The thesaurus were opioid, opioid-related disorders, cannabis, morphine, substance-related disorders, substance withdrawal syndrome, heroin, and heroin dependence, which totals 669.Twenty-nine papers were on hormone and other medicines. Thesaurus were substance-related disorders, substance withdrawal syndrome, nephritic syndrome, uveitis, prednisone, hormones, dermatitis, facial dermatoses, corneal ulcer, bronchitis etc. 75 words.

436 literatures were about alcohol. 9 thesaurus were "fatty liver, alcoholic", "liver Cirrhosis, alcoholic", "liver diseases, alcoholic", "hepatitis, alcoholi", "alcohol-related disorders", "alcoholism", "substance-related disorders", "substance withdrawal syndrome", "cardiomyopathy, alcoholic", which totals 559 times.

Five papers were on tobacco. Thesaurus were tobacco use disorder, smoking, tobacco use cessation, and substance withdrawal syndrome two times, respectively. Substance-related disorders was only once.

Conclusion

From the above-mentioned analysis, the following conclusions can be understood. The research activities about TCM curing substance-related disorders are correspondingly active. But they have not become a hot point. The journals carrying papers are relatively concentrated and formed core periodical on using Chinese traditional medicine to abstain from drug. During these 27 years, Chinese biomedicine database embodied 861 papers out of 3995 papers on TCM curing substance-related disorders amounting to 21.55 percent. Acupuncture literatures account for 11.73 percent of Chinese traditional medicine research papers, which is 2.53 percent of total publications on therapy of substance-related disorders. Theoretical research papers and
experimental research papers are relatively fewer amounting to 18.35 % and 9.64 % of the total papers, respectively.

Among publications on TCM treating substance-related disorders, papers on curing alcohol dependence and toxicosis are 436 while curing opioid-related disorders are 391, the rest is on curing hormone dependence and smoking cease etc. The result showed that the theoretical and experimental research activities on using TCM to cure substance-related disorders are still weak and should be strengthened. Acupuncture to cure drug dependence in clinic research activities should be further improved. Researchers should fully take advantages of using TCM to cure substance-related disorders in the further.

References

Shi, J. Liu, Y.L., Fang, Y.X.,et al. (2006). Traditional Chinese medicine in treatment of opiate addiction. *Acta Pharmacologica Sinica,* 27,1303-1308.

Tang, Y.L., Zhao, D., Zhao, C.Z.,er al.(2006). Opiate addiction in China: current situation and treatments. *Addiction*, 101, 657-665.

Zhao, C. ,Liu ,Z. ,Zhao, D. ,et al.(2004). Drug abuse in China. *Annals of the New York Academy of Sciences*, 1025,439-45.

Complex Data Modelling for Informetric Research

Hairong Yu, Concepción S. Wilson, Mari Davis and Fletcher Cole

{hairong.yu ,c.wilson, m.davis, f.cole}@unsw.edu.au
School of Infor. System, Technology and Management, University of New South Wales, Sydney, (Australia)

Introduction
Electronic bibliographic databases provide raw data for research in informetrics, bibliometrics, citation analysis, librametrics, scientometrics, scholarly communication studies, and studies of the Web, webmetrics or cybermetrics (described by Wilson 1999). Hood & Wilson (2003), in discussing well-known problems and limitations to using these resources, acknowledge an unmet need for appropriate data management systems suitable for informetric research and analyses. Several groups have been working on this challenge of providing a usable platform for creating purpose-built databases to gather, format, and tabulate complex and extendable data. But there is still much work to be done. This paper used object-relational modelling for building an object relational database management system that provides great flexibility in analyses of scientific production and activity.

Objective and Rationale of this Research
Using data modelling techniques and adapting Object Relational Database (ORDB) to solve data complexity issues (Ullman 1998), our objective is to create a full-scale Object-Relational Bibliographic and Citation Database to serve informetric research purposes. Our database includes integration of data from many sources to enable analysis on research quality, scientific visibility, collaboration, inter-disciplinarity, or growth of specialties. The emphasis is on data modeling, that is, on defining the database objects and tables. Informetric research uses complex data such as data with hierarchic relationships (categories or classifications), non-hierarchic transitive relationships (bibliographic references), and dynamic facts (impact factor, author career), and often in combination (Järvelin et al 2000). Complex data are characterized as:

1. *Not being primitive, straightforward, first normal form (1NF) data*. For example, authors or journals having more than one name or title due to variations or changes.
2. *Having large numbers of many-to-many relationships*. For example, authors who write many articles and serve many organizations during their careers.
3. *Using self-defined data types*. For example, adoption of different discipline classifications.

None of these issues can be solved by a Relational Database (RDB) because it only allows user-defined data types based on its built-in standard data types rather than object type data structures. With ORDB, object storage and operation capability are added to a full-fledged RDB. Three things can be encapsulated and operated in a well-constructed ORDB: a) traditional field data which are similar to attributes in relational databases; b) simple objects which have lists of attributes in their class plus their methods; and c) complex objects such as time-series, geo-spatial series, or binary media data such as audio and video. Users can still touch and feel familiar tabular structures by using traditional open standard data definition or data manipulation language, such as SQL. Thus ORDB was chosen for this project for the following reasons:

- The capability of conceptualizing and processing complex data and objects allowing composite fields, e.g. arrays, lists, sub-sets or nested tables, hierarchical or Object-Oriented (OO) data, without limits.
- The ability to accommodate various data source formats and standards in a straightforward relational approach for searching on multiple databases or comparing different information retrieval systems.
- Possibility of integrating statistical calculating tools as object functions in user-defined data types to capture the dynamic nature of informetric measures and data.

Although RDB is preferred for keeping large volumes of persistent data because of its stability and capacity, we were driven to look closely at ORDB because RDB has difficulty in representing data involving inheritance hierarchies, and complex relationships (frequent many-to-many relationships). ORDB treats objects as data entities as well as representing complex data, and it reconciles the advantages of both RDB and OO modelling (OOM) for complex data manipulations. Now well established, ORDB has proved to be a viable technology in application.

Key Features of Informetric ORDB
From the complex data modelling aspect, features worth mentioning are:
Object Type – Object types are user-defined data types which allow database system developers to model complex real-world entities. New object types can be created from any built-in database types and any previously created object types, object references, and collection types (e.g. varrays

and nested tables in Oracle). For example, articles in a journal could be well represented by an object.

Type Inheritance – Type inheritance enables developers to create type hierarchies by defining successive levels of increasingly specialized subtypes that derive from a common ancestor object type. Derived subtypes inherit the features of the parent object type but extend the parent type definition. This hierarchy building provides better manipulation of the complex application model. For example, monograph or conference proceedings could be a child type of a parent journal type. The conference proceedings have all the generic attributes of a journal and other specific features, e.g. location, and host organization.

Type Evolution – ORDB data types are easily created, modified, or "evolved" in attribute, method, type definition or metadata level to capture the changes automatically. This is particularly useful for incorporating various citation or bibliographic database formats and standards on field structures, attribute names, data types or naming standardization and still use their existing calculating functions or analysis tools without re-defining and recoding the underlying programs. Journal class can be associated by different subject classifications or even ad hoc ones, in this case, the Library of Congress Classification System (LC System), Journal Citation Report (JCR) and a local category system such as the University of New South Wales' research field, discipline and subject codes.

The main task in modelling is to map the data needed for informetric research into the objects, attributes, and tables of an ORDB, with scope and balance in mind to answer questions on, for example:

- Publication productivity, for example of authors, groups, collaborations, institutions or organizations, and journals.
- Publication, collaboration or visibility profiles over time and by region (state, national or international).
- References-citations made by/from neighboring disciplines.
- Volume and direction of knowledge transferred by journals.

Conclusion and Future Work

The ORDB methodology has been implemented for informetrics with a system enabling researchers to model dynamic and complex data, to interoperate with diverse data formats and attribute standards, to perform sophisticated queries or to embed ad hoc calculations as essential analysis methods in user-defined objects. Such tasks are not feasible when RDBs are adopted. Our implementation has confirmed that the directions we have taken with ORDB methodology are very promising. Benefits to informetric research derive from the querying capability when complex data is modelled expressly for informetric analytic purposes together with the building of purpose-built databases. Some of the previous informetric questions mentioned above can be answered straightforwardly. Query results will be useful also for policy analysts, publishers, editors, library collection developers, subscribers, authors and researchers. The system, which follows industry standards, would be adaptable for many extensible uses, whenever complex and flexible data is required for analysis.Further work on the issues we face and our specific adaptations to data collection and processing are ongoing. Specifically we focus on the database design, implementation, evaluation and its leverage over Web and interoperability with other services. A number of new directions and interesting dimensions include:

- Establishing a more efficient database platform for extended informetrics research.
- Accommodating extended data created by other languages.
- Leveraging the standardized application environment to Web applications.
- Combining database content with text mining for bibliographic analysis (Stix 2005).
- Integrating ORD with Information Retrieval (IR) for documentation-related analyses (inspired by Fisk et al 2003).

Acknowledgment

This research is funded by the John Metcalfe Memorial Fund. The encouragement and support of the School of Information Systems, Technology and Management (SISTM) and members of SISTM's Information Management Research Group at the University of New South Wales are acknowledged.

References

Fisk, J.M.; Mutalik, P.; Levin, F.W.; Erdos, J.; Taylor, C. & Nadkarni, P. (Jan/Feb. 2003), Integrating query of relational and textual data in clinical databases: A case study, Journal of the American Medical Informatics Association, 10(1), pp21-38.

Hood, W.W. & Wilson, C.S. (2003), Informetric studies using databases: Opportunities and challenges, Scientometrics, 58 (3), pp587-608.

Järvelin, K.; Ingwersen, P. & Niemi, T. (2000), A user-oriented interface for generalized informetrics analysis based on applying advanced data modelling techniques, Journal of Documentation, 56(3), pp250-278.

Stix, G. (May, 2005), A software tool elicits previously undiscovered gene or protein pathways by combing through hundreds of thousands of journal articles, Scientific American, 292(5), pp70-73.

Ullman, J. (1998), Object-Relational Features of Oracle. Retrieved August 2006 from http: //www-db.stanford.edu/~ullman/fcdb/oracle/or-objects.html.

Wilson, C.S. (1999), Informetrics, Annual Review of Information Science and Technology, vol. 34, pp107-247.

Analyzing the Impact of Sloan Digital Sky Survey on Astronomical Literature: A Multiple Perspective Approach

Jian Zhang[*], Weizhong Zhu[*], Yunan Chen[*], Michael Vogeley[**] and Chaomei Chen[*]

[*]*jz85@drexel.edu, wz32@drexel.edu, yc337@drexel.edu, chaomei.chen@cis.drexel.edu*
College of Information Science and Technology, Drexel University, Philadelphia (USA)

[**]*msv23*@drexel.edu
Department of Physics, Drexel University, Philadelphia (USA)

Introduction

As the first step of an ambitious project to support scientific discoveries in astronomy, we focus on the impact of large-scale astronomical surveys on scientific discoveries. Trends and patterns in the Sloan Digital Sky Survey (SDSS) literature are analyzed in particular.

Background

There are many studies of astronomical literature. Abt (1981, 1990, 2000) conduct a series of research on astronomical literatures from 1910 to current in terms of annual publication rate, authorship, and internationalization. Davoust & Schmadel (1991) identify 14 "super-productive" astronomers who published over 150 papers between 1969 and 1987. Schulman et al. (1997) confirm that multi-authorship papers formed the mainstream of astronomical publication.

Studies mentioned above focus on astronomical literature in general. In this study, we focus on the largest astronomical survey to date, the Sloan Digital Sky Survey (SDSS), and in particular, scientific outputs resulted from the SDSS data. We uses three perspectives in our analysis, namely statistical analysis, citation analysis, and text analysis.

Methods

In Sept. 2006, publications related to the SDSS survey were retrieved from the Web of Science with search terms: "SDSS" AND "Sloan Digita*" between 1990 and 2006. The dataset contains a total 1,350 bibliographic records from 159 journals. We checked the abstract and title of each paper and removed irrelevant ones from the dataset. Note that the Web of Science only indexes a relatively small but important portion of journals published on a subject; therefore, the subsequent analysis is within the scope implicitly imposed by the selection criteria of Web of Science.

In this study, we use the statistical package SPSS version 11.5 for statistical analysis, CiteSpace (Chen, 2006) for citation and co-citation analysis, and the visual analytic tool Storylines (Zhu, 2007) for text analysis and identifying major latent concepts in the semantic space of the text corpus.

Results

A total of 1,116 research papers relevant to SDSS were published between 1993 and 2006 in 46 journals. A total of 2,636 authors, from 48 countries and 595 organizations worldwide, contributed to these papers. The average number of authors per paper is 2.3. The first SDSS paper appeared in 1993, and then the annual number of papers increased slowly until 1999. Since then, the annual number of publications increased dramatically, close to 50 new papers per year. Table 1 shows the top 5 most productive authors and citations they received.

Table 1. Top 5 most productive SDSS authors.

Author	Papers	Cites
Brinkmann, J	227	7758
Schneider, DP	185	6478
Ivezic, Z	115	5670
York, DG	115	5549
Strauss, MA	106	5167

Figure 1 shows a timeline of burst terms found in citing papers. For example, brown dwarf was a hot topic in 2000, active galactic nuclei in 2004, and star forming galaxies in 2005. The timing of these terms gave us an idea of research focuses and the change of such focuses over time.

In text analysis, Storylines identified 530 dimensions that explain more than 80% of the latent semantic space. For example, the most predominant dimension includes terms such as quasar, agn, luminos function, and power spectrum. Quasars are one of the high-priority research topics that motivated the SDSS survey. In fact, as shown in Figure 2, the term quasar is predominant in several major dimensions of the SDSS latent semantic space. The appearances of a term in multiple latent dimensions indicate that the term has distinct roles in various contexts in the SDSS literature.

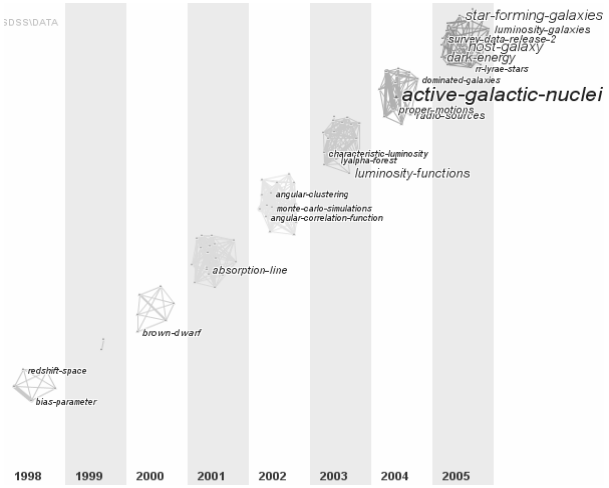


Figure 1. Burst terms in the SDSS papers.

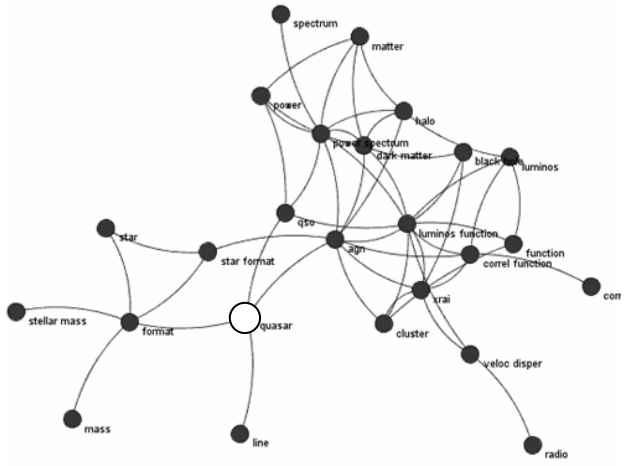


Figure 2. Major concepts in the most predominant latent dimension of the SDSS data.

Discussion and Conclusions

The SDSS Survey has provided numerous data (York, et al, 2000) to the astronomy community. When the survey data became available after 1999, publications on SDSS increased radically. The SDSS data become the real goldmine for astronomers. The SDSS also produced "super star," the productive authors. For example, Brinkmann, J. at Apache Point Observatory, Schneider, D. P. at Pennsylvania State University, Ivezic, Z. at Princeton University, York, D. G. at University of Chicago, and Strauss, M. A. at Princeton University published hundred research papers in 10 years. In terms of citation counts, two papers by Becker R. H. at University of California Davis obtained highest citations. In future studies, we will study the dynamics of the SDSS literature with direct references to specific astronomical discoveries made with the SDSS data.

Citation map generated by CiteSpace offered alternative views to the timelines of hot topics and landmark papers. Storylines is useful in clarifying the nature of each dimension. Each dimension defines a context of astronomical research topics.

In conclusion, large-scale sky surveys greatly affect scholarly communications of the astronomical community. Astronomers have made many discoveries that would not be possible without the SDSS data. We will continue to investigate the dynamic system of scientific activities in this domain with various tools to improve our understanding of the interrelationship between observational data and astronomical discoveries.

References

Abt, H. A. (1981). Some trends in American astronomical publications. *Publications of the Astronomical Society of the Pacific*, 553, 269-272.

Abt, H. A. (1990). Trends toward internationalization in astronomical literature. *Publications of the Astronomical Society of the Pacific*, 102, 368-372.

Abt, H. A. (2000). Astronomical publication in the near future. *Publications of the Astronomical Society of the Pacific*, 112, 1417-1420.

Chen, C. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology*. 57, 359-377.

Davoust, E., & Schmadel, L. D. (1991). A study of the publishing activity of astronomers since 1969. *Scientometrics,* 22*,* 9-39.

Fernandez, J. A. (1998). The transition from an individual science to a collective one: The case of astronomer. *Scientometrics,* 42, 61-74.

Schulman, E., French, J. C., Powell, A. L., Eichhorn, G., & Kurtz, M. J. (1997). Trends in astronomical publication between 1975 and 1996. *Publications of the Astronomical Society of the Pacific*, 109, 1278-1284.

York, D. G., Adelman, J., Anderson, J. E., Anderson, S. F., Annis, J., Bahcall, N. A., et al. (2000). The Sloan Digital Sky Survey: Technical summary. *The Astronomical Journal*, 120, 1579-1587.

Zhu, W., & Chen, C. (2007). Storylines: Visual exploration and analysis in latent semantic spaces. *Computers & Graphics*.

Factor Rotation Methods in Author Co-citation Analysis: A Comparison

Dangzhi Zhao

dzhao@ualberta.ca
University of Alberta, School of Library and Information Studies, Edmonton, Alberta (Canada)

Introduction

Since its introduction by White & Griffith (1981), author co-citation analysis (ACA) has gained great popularity in the study of intellectual structures of scholarly fields and of the implied social structures of the corresponding communities, and as a result has become an important area of information science (e.g., White & McCain, 1998).

A fairly consistent set of techniques is available for ACA studies (McCain, 1990a; White & McCain, 1998). ACA takes the author as the unit of its analysis, uses co-citation counts to measure the relationship between oeuvres of representative authors, and applies multivariate analysis techniques (e.g., Cluster Analysis, Multidimensional Scaling, or Factor Analysis) to reveal the underlying structure of a research field represented by these oeuvres.

The goal of Factor Analysis, in particular, is to identify meaningful factors that explain relationships among items, and is used in ACA to reveal specialty structures of a research field and membership of authors in one or more such specialties. Factor Analysis applied in ACA has been shown to provide clear and revealing results as to the nature of a discipline (White and McCain, 1998).

Most ACA studies apply the general techniques of classic ACA to different research fields with little or no modification. Recently, however, interest in exploring alternatives has resurged.

Some studies propose new techniques for mapping author clusters (White, 2003); some explore how to go beyond first-author counting in ACA given newly available citation data sources (Zhao, 2006); others challenge classic ACA techniques, such as Ahlgren, Jarneving, & Rousseau (2003) who propose using cosines instead of Pearson's r as similarity measures.

The present study examines one of the commonly used statistical procedures in author co-citation analysis (ACA) – Factor Analysis. Confusion exists in ACA studies regarding which techniques to choose in designing a factor analysis study. Here, we compare two types of rotation methods in Factor Analysis that have both been used in ACA studies but have not been studied explicitly: oblique rotation vs. orthogonal rotation.

The problem and research questions

Factor Analysis has been used in ACA from the very beginning (White & Griffith, 1981). In addition to the specialty structure of a research field that other multivariate analysis techniques such as the combination of Cluster Analysis and MDS can also identify, Factor Analysis allows individual authors to appear in more than one group (specialty), indicates the level of associations between authors and the specialties to which they belong, and provides measures on the relationship between specialties (White & Griffith, 1982).

The design of a factor analysis study in ACA involves a number of decisions regarding the number of authors, the type of input data, the method for extracting factors and the number of factors, and which factor rotation method to use, as well as which factor loadings to consider meaningful for interpreting factors. Different decisions may well produce different results. It is therefore important to understand the options that are available, the different results that may be obtained using different techniques, and when and why certain techniques may be more appropriate than others.

An examination of ACA studies reveals some commonly used factor analysis techniques. Principle Component Analysis has been the main method used for extracting factors, Kaiser's rule of eigenvalue greater than 1 the main technique for deciding the number of factors to extract (McCain, 1990a); both raw co-citation frequency matrices and Pearson's r correlation coefficient matrices have been used as input to Factor Analysis; both orthogonal and oblique rotations have been used in ACA to facilitate interpretation; various significance levels of factor loadings for interpretation have been reported, ranging from 0.3 to 0.5. And yet, studies are largely silent on why a particular technique is chosen, nor do they compare ACA results obtained using different techniques to help researchers choose among them. We attempt to rectify this situation for the case of rotation methods.

Theoretically, oblique rotations are more appropriate for factor analysis in ACA for two reasons:

(a) As an orthogonal rotation assumes that factors are not correlated, an oblique rotation is often more appropriate to use when it can be expected theoretically that resulting factors would in reality be correlated (Hair et al., 1998). In the case of ACA, factors are interpreted as specialties within a research field, and the correlation between factors can therefore be expected to be high. In practice, high correlations between factors up to between 0.39 and 0.65 were reported in different fields (McCain, 1990a; White & Griffith, 1982; Zhao, 2006).

(b) The degree of correlation between factors is measured by the component correlation matrix produced by oblique rotation. This information is

useful in the analysis of the intellectual structure of a research field.

In practice, however, both oblique and orthogonal rotation methods have been used in ACA studies, even by the same authors, and in both cases with success. For example, McCain (1990) and White & Griffith (1982) apply oblique rotation and discuss its advantages in providing measures on the inter-relationships between factors. Later, as in White & McCain (1998), they use orthogonal rotation instead, without any indication why. Still, the orthogonal rotation in that landmark study produces clear and revealing results as to the nature of the LIS field (White & McCain, 1998).

These are curious discrepancies between theory and practice. It is important to know whether different methods would produce different specialty structures in ACA, and, if so, whether the differences are sufficient to justify one over the other.

Oblique rotation produces two distinct matrices, the structure and the pattern matrix. The loadings in the pattern matrix produced by oblique rotation can be interpreted as the unique contribution of individual authors to specialties, whereas loadings in the structure matrix, which are "simple correlations between variables and factors," contain both this unique contribution and the correlation among factors. As the correlation among factors is expected to be quite high in ACA, it can be difficult to distinguish which authors load uniquely on each specialty in the structure matrix (Hair, et. al., 1998, p. 113). Theoretically, therefore, the pattern matrix is more appropriate for ACA interpretation purposes, whereas in ACA practice, the structure matrix has traditionally been used to discuss results (White & Griffith, 1982; McCain, 1990a). Again, the discrepancies between theory and ACA practice are curious.In summary, the research questions explored in the present study are as follows. (a) What are the differences in factor analysis results in ACA between the use of an oblique rotation and an orthogonal rotation? (b) Is there any evidence whether the structure matrix or the pattern matrix is preferred for reporting factor analysis results for interpretation in ACA if an oblique rotation is used?

Methodology

The research field we analyzed in the present study was Library and Information Science (LIS) defined exactly as in White & McCain (1998), i.e., by its 12 main journals as listed in White & McCain (1998, p. 330), but for 1996-2005. We collected 4,422 records of source papers with 110,785 references altogether.

We followed the techniques of ACA mentioned earlier. A core of 120 authors was selected to represent the LIS research field based on "citedness" (White & McCain, 1998, p. 332) to participate in the final factor analyses. A Java program determined co-citation frequencies of these highly-cited authors and recorded them in a co-citation matrix, which was then input to the factor analysis routine in SPSS.

Factors were extracted by Principal Component Analysis (PCA). Kaiser's rule of eigenvalues greater than 1 resulted in a 12-factor model which explained 84% of the total variance, and the differences between observed and implied correlations were smaller than 0.05 for the most part (99%). The communalities ranged from 0.62 to 0.94, only 4 (or 3%) of which were below 0.7 and 18 (or 15%) of which below 0.8. The model fit was clearly adequate (Hair, et al., 1998). We applied orthogonal rotation (SPSS Varimax) and oblique rotation (SPSS Direct OBLIMIN).

Results and conclusions

Results confirm that a major difference between ACA results from these two types of rotation lies in that the oblique rotation tends to recognize more of the authors' unique research contributions whereas the orthogonal rotation tends to show more of their general research areas. Results from an oblique rotation appear to be clearer and more theoretically meaningful. When an oblique rotation is used, the pattern matrix tends to have a simpler and clearer structure than the structure matrix in that it has fewer authors by far who have high loadings on a large number of specialties, which makes interpretation of results considerably easier.

References

Hair, J.F. Anderson, R.E., Tatham, R.L., & Black, W.C. (1998). Multivariate data analysis (5th edition). Upper Saddle River, NJ: Prentice Hall.

Leydesdorff, L., & Vaughan, L. (2006). Co-occurrence matrices and their applications in information science: Extending ACA to the Web environment. JASIS&T, 57(12), 1616-1628.

McCain, K. W. (1990a). Mapping authors in intellectual space: a technical overview. JASIS, 41, 433-443.

McCain, K. W. (1990b). Mapping authors in intellectual space: population genetics in the 1980s. In C. L. Borgman (ed.), Scholarly Communication and Bibliometrics (pp.194-216). Newbury Park, CA: Sage.

White, H. D. & Griffith, B.C. (1981). Author cocitation: A literature measure of intellectual structure. JASIS, 32, 163-171.

White, H. D. & Griffith, B.C. (1982). Authors as markers of intellectual space: Co-citation in studies of science, technology and society. Journal of Documentation, 38, 255-272.

White, H. D. & McCain, K.W. (1998). Visualizing a discipline: An author co-citation analysis of information science, 1972-1995. JASIS, 49, 327-355.

Zhao, D. (2006). Towards all-author co-citation analysis. Information Processing & Management, 42, 1578-1591.

Bibliometry Analysis of the Research on Women and Health

M.A. Zulueta[*], G. García-Gómez[**], I. Doménech[**], M. Izquierdo[**],P. Moscoso[**]

[*]*ma.zulueta@uah.es*
Faculty of Documentation. University of Alcalá, c/San Cirilo s/n, 28801 Alcalá de Henares, (Spain)

[**] *gema.garciag@uah.es; irene.domencech@uah.es;* monica.izquierdo@uah.es; p.moscoso@uah.es
Faculty of Documentation. University of Alcalá, c/San Cirilo s/n, 28801 Alcalá de Henares, (Spain)

Introduction

A recurring topic in bibliometry is the analysis of the scientific research output generated by different disciplines (Gil-Montoya, J.A. et al 2006.) The Medline database includes descriptors from the Mesh Thesaurus, allowing a more accurate definition of the intellectual content of the documents. Scientific research focusing on the study of specific aspects related to Woman, has strongly increased along the past decade (García García, P. et al. 2005) (Ding, W. W., et al. 2006). The interest in this particular research lies on the need to better understand specific aspects related to women's position in social, economic and health care contexts. Despite the emergence of this subject, to our knowledge there are no bibliometric studies specifically focusing on the scientific literature related to women and health. On that basis we have developed an approach to assess, the relevance from a scientific standpoint, of published studies related to women health as well as other social and health-care related aspects.

Objectives

To analyze from a bibliometric standpoint the results of scientific research specifically focused in the study of Health and Women, and to assess its evolution along time.

Methods and Materials

The Medline database (National Library of Medicine, 2006) was considered the most adequate system for detecting and identifying medical aspects specifically related to Woman. The indexing system of this database is based on the use of a controlled vocabulary, the Medical Subject Heading (Mesh, National Library of Medicine, 2006), which confers it a high level of accuracy and exhaustivity. Employing this database permits to address the relationship between Woman and health, as well as other social, cultural, political, economic and psychological aspects.

A methodology based on the visualization of the relationship between the main descriptors related to the concept Woman was used in order to delimit the initial search terms. (Vargas Ouesada, B., et. al. 2006) To this end, a first approach was carried out with the nine Mesh descriptors obtained by searching for the word Women in the thesaurus, including: Women, Women's Health, Women's Health Services, Women's Rights, Women's Working, Dentist Women, Battered women, Physician's Women and Pregnant Women.

To achieve a compilation of the most relevant documents, the descriptors were searched for in the MajorMesh field, where the tags Human and Females were used in order to exclude documents related to man and/or to animals. This search resulted in 19,524 documents. Subsequently, the 60 most frequent descriptors related to these documents were identified and used in a second search using the same database and methodology. All the recovered records are being examined in order to identify the distribution of the following indicators: Temporal Evolution, Publication Language, Document Type, Journal Publication and Specialties. Moreover, the level of overlapping in comparison to the ISI databases has also been evaluated

Results

The total number of documents corresponding to the 1963-2005 period was 365,789. Along this time the number of documents progressively increased, particularly after 1980 (Figure 1). These documents accounted for 0,17% of the Medline database between 1963 and 1974, whereas their proportion in the last decade was 1,24%.

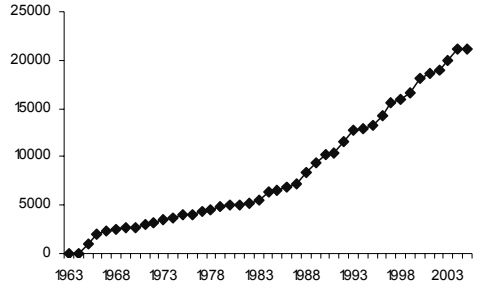


Figure 1. Annual scientific production related to Woman

Publication Language

Most of the documents found were published in English (83%). Other languages included German (3,5%), French (2,94%), Russian (1,57%), Japanese (1,41%), Spanish (1,40%) and Italian (1,26).

Document Type

Medline included in the Publication Type field, documents corresponding to traditional document classification i.e. journal articles, letters or revisions. In addition, it also included in this classification other types of documents distinguished by methodological aspects, such as Clinical Trial, Multicenter Study, Twin Studies; yet, only a small proportion of the documents (0,02%) found did not correspond to any of the traditional types. As expected, the most common document type was Journal Article (n=313.355, 85.7%).

Journals

The total number of journals was 7.734; the distribution of journals containing the higher number of documents is shown in Table I.

Table 1. Most Productive Journals for the field Woman.

Period 1964-1984		**Period 1985-2005**	
Journals	**Docs**	**Journals**	**Docs**
Am J Obstet Gynecol	1444	Lancet	3419
Lancet	1439	AIDS	2666
Cancer	1269	Cancer	2365
Obstet Gynecol	999	Br Med J	2215
Br Med J	939	Soc Sci Med	2096
JAMA	741	Breast Canc R Tr	2005
Cesk Gynekol	672	JAMA	1896
Minerva Ginecol	617	Am J Obstet Gyn	1798
Akush Ginek-Moskow	575	J Clin Oncol	1724
N Engl J Med	523	Br J Cancer	1724
Med J Aust	504	Am J Publ Health	1621
Zentralbl Gynakol	501	Psychol Rep	1589

The analysis of these journals indicates that scientific literature in relation to women health is published in both general and specialized journals. A first analysis reveals a dominant research activity related to clinical, social, psychological and nursing-related aspects. To better understand the evolution of the investigations on this topic we have conducted the study in two separate periods (1963-1984 and 1985-2005) Comparing both periods it is noteworthy the disappearance of journals published in countries under soviet influence and the emergence of the journal AIDS. Another remarkable aspect stems from the comparative analysis of this subject in the ISI database. The number of journals detected in both databases was 1965, containing 170.798 documents. These numbers represent respectively 25% of the journals and 46% of all the documents we found in Medline. Despite the inclusion of the SSCI databases these data further support the broader coverage of Medline to perform studies with a strong social and health-care component.

SpecialtiesIn order to understand which disciplines are the most related to women's health, we have carried out a study of the subject categories pertaining to each of the journals, as indexed by Medline. The most actives disciplines were Medicine, Gynecology, Obstetrics and Neoplasm These stand as the most productive in both periods though with a lesser impact in the second one, where a higher subject dispersion was also noticed. It is of note the importance of Nursing, Acquired Immunodeficiency Syndrome (AIDS), Sexually Transmitted Diseases, and Substance-Related-Disorders, as well as of Social and Psychological aspects and Public Health.

Table 2. Most productive categories for the field Woman.

Period 1963-1984		**Period 1985-2005**	
Category	**%**	**Category**	**%**
Gynecology	9,37	Medicine	5,35
Medicine	8,24	Neoplasms	4,26
Obstetrics	6,53	Gynecology	3,13
Neoplasms	3,37	Obstetrics	2,97
Family-Planning	0,91	AIDS	2,42
Reproduct Med	0,67	Nursing	1,71
Psychology	0,13	Public-Health	1,36
Psychiatry	0,92	Psychology	1,27
Public-Health	0,90	Pediatrics	1,01
Subst-Rel-Disord	0,49	Reproductive-Medicine	0,93
Nursing	0,22	Health-Services-Research	0,83
Geriatrics	0,33	Substance-Related-Disord.	0,71
Biology	0,22	Social-Sciences	0,68
		Geriatrics	0,67
		Psychiatry	0,57
		Radiology	0,51
		Communicable-Diseases	0,51
		Epidemiology	0,49

Conclusions

Our preliminary analysis supports a growth trend of studies on women and health, changing throughout the timeframe studied. As compared to the initial lineal growth, a strong increment in the rate of the research and consequently in the related publications was noticed after 1984. As the documents were distributed by journal and categories, a change in the content and behavior of the researchers relative to their publication habits was observed. Albeit the multidisciplinary nature of the research, there are indications that these studies tend to acquire their own identity.

Acknowledgments

This work has been supported by a grant of the Spanish Ministry of Health (Ref. UAH 142/2005).

References

Ding, W. W., Murray, F y Stuart, T.E. (2006). Gender differences in patenting in the academic life sciences. Science, 313, 665-7.

García García, P., López Muñoz, F., Callejo, J., Martín Águeda, B., Álamo, C. (2005). Evolution of Spanish scientific production in international obstetrics and gynecology journals during the period 1986-2002. Eur J Obstet and Gynecol and Reprod Biol, 123, 150-156.

Gil-Montoya, J. A., Navarrete-Cortes, J., Pulgar, R., Santa, S., Moya-Anegón, F.(2006). World dental research production: an ISI database approach (1999–2003). Eur J Oral Sci, 114, 102–108.

National Library of Medicine. Publication Characteristics (Publication Types) Scope Notes. http://www.nlm.nih.gov/mesh/pubtypes2006.htm

A
Aguillo, Isidro 608, 816, 912
Alcain, Mª-Dolores 672
Aleixandre-Benavent, Rafael 37, 818, 898
Alonso-Arroyo, Adolfo 898
Ambre, Sumeet 457
Aminpour, Farzaneh 820, 822, 868
Andújar, Inés 838
Angus, Emma 824
Archambault, Éric 45, 449
Arreto, Charles-Daniel 926
Ausloss, Marcel 365
Azagra-Caro, Joaquín M. 677

B

Bai, Xiao-ping 958
Bailón-Moreno, Rafael 421
Barden, Elizabeth M. 766
Bar-Ilan, Judit 52, 826
Barjak, Franz 63
Barrios, Maite 828
Barts, Nicolas 932
Bassecoulard, Elise 74, 797
Batista, Pablo D. 184
Beigbeder, Michel 946
Besselaar, Peter van den 85, 803
Bhattacharya, Sujit 95
Böhmer, Susan 397
Bokos, George 920
Bonilla, Ana Isabel 830
Bonitz, Manfred 832
Bordons, Maria 226, 335, 552
Börner, Katy 112, 457
Bornmann, Lutz 149
Borrego, Angel 828
Bosire Onyancha, Omwoyo 100
Boyack, Kevin W. 112, 124, 437
Bozeman, Barry 305
Braam, Robert 136
Brodin Danell, Jenny-Ann 154
Bruyn, Arlette De 930
Burgoon, John 457
Burrell, Quentin L. 162
Buter, R.K. 834
Butler, Linda 170

C

Calero-Medina, Clara 179, 599, 761
Calo, William A. 914
Camí, Jordi 896
Cammarano, Rosaria Rosanna 836
Campanario, Juan Miguel 840
Campiteli, Mônica G. 184
Cañibano, Carolina 838
Cao, Cong 427
Cao, Yan 916
Castellano-Gómez, Miguel 818

Chaichio-Moreno, Juan Antonio 300
Chen, Chaomei 189, 964
Chen, Dar-Zen 952
Chen, Li 201
Chen, Lixin 844
Chen, Yunan 964
Cheng, Yunwe 776
Chinchilla-Rodríguez, Zaida 582
Chu, Heting 207
Coccia, Mario 216
Cole, Fletcher 962
Collazo-Reyes, Francisco 523
Corera-Álvarez, Elena 582
Coslado, Mª Angeles 840
Costas, Rodrigo 226
Cothey, Viv 231

D

Danell, Rickard 154
Dang, Yaru 842
Daniel, Hans-Dieter 149
Davis, Mari 962
Debackere, Koenraad 659
Delgado López-Cózar, Emilio 421
Dhawan, S.M. 241, 856
Diwakar, Sandhya 846
Doménech, I. 968

E

Egghe, Leo 245
Escudero, Cristina 300

F

Fang, Shu 254, 776
Fang, Wei-gang 958
Fernández, M. Teresa 335
Fernández-Izquierdo, Francisco 266
Fernández-López, Juan Antonio 300
Filippo, Daniela de 848, 900
Filliatreau, Ghislaine 797
Foland, Patricia 708
Forti, Enrico 272
Franzoni, Chiara 272, 285
Frenken, Koen 85
Frietsch, Rainer 295
Fu, Dong-hong 958

G

Gao, Feng 842
García, Preiddy-Efraín 266
García-Carpintero, E. 850
García-Gómez, G. 968
García-Romero, Antonio 300
García-Zorita, Carlos 266, 902
Garfield, Eugene 21
Gaudy, Jean-François 926
Gauffriau, Marianne 852
Gaughan, Monica 305
Giménez, Elea 672
Gingras, Yves 314, 449

Glänzel, Wolfgang 321, 408, 736
Gläser, Jochen 860
Gómez, Isabel 335, 830, 848
Gómez, Yuri Jack 346
Gómez-Crisóstomo, Rocío 582
González-Alcaide, Gregorio 37, 818, 898
González-Barahona, Jesús M. 910
Gorelskyy, Roman 708
Gorraiz, Juan 854
Granadino, Begoña 816, 850
Gupta, B.M. 241, 856
Gupta, R. P. 856
Gutiérrez-Carrasco, A. 894

H

Hällsten, Martin 685
Harnad, Stevan 27
Hassan-Montero, Yusef 858
Havemann, Frank 860
He, Fenglan 842
Heimeriks, Gaston 85
Heinz, Michael 860
Hellsten, Iina 365
Henadeera, Kumara 170
Hernández, Tony 830
Herrero-Solana, Víctor 858
Hérubel, Jean-Pierre V. M. 368
Hibino, Aiko 908
Hinze, Sybille 295
Holmberg, Kim 378
Hook, Peter A. 387
Hörlesberger, Marianne 691
Hornbostel, Stefan 397
Hu, Zhengyi 776
Huang, Mu-Hsuan 952
Huang, Yue 842

I

Iglesias, Juan E. 862
Ingwersen, Peter 696, 864
Ins, Markus von 852
Iribarren-Maestro, Isabel 401, 914
Izquierdo, M. 968

J

Janssens, Frizo 321, 408, 659
Jansz, M. 542
Jiménez-Contreras, Evaristo 421
Jin, Bihui 427

K

Kaaouachi, Abdelali 866
Kabiri, Payam 820, 822, 868
Ke, Weimao 457
Klavans, Richard 112, 437
Kondo, Masayuki 956
Korfiatis, Nikolaos 920
Kotar, Mojca 872
Kretschmer, Hildrun 870
Kretschmer, Theo 870
Krichel, Thomas 207
Kuo, Chun-Hsien 942

L

Lambiotte, Renaud 365
Larivière, Vincent 45, 449
LaRowe, Gavin 457
Larsen, Birger 696, 864
Larsen, Katarina 463
Lascurain-Sánchez, Maria Luisa 266, 401
Lee, Yi-Yang 942
Leeuwen, Thed N. van 469, 740
Leite, Paula 874
Lelu, Alain 74
Leta, Jacqueline 480, 874, 948
Léveillé, Valérie 932
Levitt, Jonathan 485, 876
Lewison, Grant 489, 878, 880
Leydesdorff, Loet 499, 936
Li, Baoli 285
Li, Kan 842
Liang, Liming 508, 844
Libkind, I. 542
Lin, Xia 882
Lin, Yu-I 942
Liu, Yuxian 514
Liu, Zeyuan 844
Lo, Szu-chia Scarlett 884
Lucio-Arias, Diana 886
Luna-Morales, Ma. Elena 523

M

Ma, Zheng 201, 792
Macías-Chapula, C.A 894
Markusova, V. 542
Martinez, Alexandre S. 184
Martín-Moreno, Carmen 266
Matheus, Renato Fabiano 888, 934
Mauleón, Elba 552
Maye, Isabelle 852
McCain, Katherine W. 558, 890, 892
Meho, Lokman 568
Méndez-Vásquez, Raúl I. 896
Mendoza-Guerrero, J.A. 894
Meneghini, Rogério 904
Meyer, Martin 631
Miguel-Dasit, Alberto 37, 818, 898
Moed, Henk F. 761
Moneda-Corrochano, Mercedes 421
Moor, Bart De 408, 659
Moreno-Díaz-del-Campo, Francisco-Javier 266
Morillo, Fernanda 335, 900
Morris, Steven A. 766
Moscoso, P. 968
Moya-Anegón, Félix 582
Mugnaini, Rogério 902, 904
Muñoz-Fernández, Francisco José 582
Muñoz-Molina, Antonio 582
Murad, AbdulKader A. 938
Must, Ülle 906

N

Nabiullin, Ahat A. 594
Navarrete-Cortés, José 300
Navarro-Molina, Carolina 37
Negishi, Masamitsu 936
Noyons, Ed C.M. 179, 599, 834
Nozawa, Takayuki 783

O

Ocholla, Dennis N. 100
Ohniw, Ryosuke L. 908
Olesen Larsen, Peder 852
Ollé, Candela 828
Ortega, José Felipe 910
Ortega, Jose Luis 608, 816, 912
Ortiz, Ana Patricia 914
Otamendi, Javier 838

P

Packer, Abel 904
Pan, Yuntao 201, 792
Pang, Jing'an 916
Patel, Pari 295
Pecharromán, Carlos 862
Peritz, Bluma C. 52
Pillai Sudhier, K.G. 918
Plaza, L.M. 850
Polanco, Xavier 613
Ponomariov, Branco 305
Poulos, Marios 920
Pouris, Anastassios 619
Pouris, Anthipi 619
Povedano, Elisa 266

R

Rabow, Hampus 922
Rafols, Ismael 631
Ram, Ashwin 285
Ramanana-Rahary, Suzy 643, 797
Ramo, Cristina 924
Rao, I.K.Ravichandra 655
Robert, Claude 926
Robinson, Simon 63
Robles, Gregorio 910
Rodriguez, Victor 659
Rogge, Gwendolyn 928
Román-Román, Adelaida 266, 672
Romero-de-Pablos, Ana 677
Rons, Nadine 930
Rostaing, Hervé 932
Roulin-Perriard, Anne 852
Rousseau, Ronald 427, 514, 643, 928
Rubio-Liniers, Cruz 266
Ruiz-Baños, Rosario 421
Ruiz-Pérez, Rafael 421
Russell, Jane M. 523

S

Sandström, Ulf 685
Sanjuan, Eric 613
Sanz, Ginés 896
Sanz-Casado, Elías 266, 401, 848, 902, 914
Sastre, N 850
Scharff, Morten 864
Scharnhorst, Andrea 365, 912
Schiebel, Edgar 691
Schlögl, Christian 854
Schmidt, Marion 860
Schneider, Jesper W. 696, 864
Seta, Maurella Della 836
Shelton, R. D. 708
Shibui, Susumu 783
Silverstein, Scot 892
Simpkins, Christopher L. 285
Singh, Dr. KK 846
Small, Henry 718
Smith, Alastair 726
Sobrero, Maurizio 272
Somoza, Marta 828
Song, Il-Yeol 189
Sorenson, Martha 948
Souza Vanz, Samile Andréa de 934
Stuart, David 731, 824
Su, Cheng 201
Suárez, Carlos A. 914
Suárez, Erik 914
Suñén-Pinyol, Eduard 896
Suttmeier, Richard P. 427
Suzuki, Takafumi 783

T

Takeyasu, Kunio 908
Teixeira, Flávio Martins 480
Thelwall, Mike 378, 485, 731, 824, 876
Thijs, Bart 321, 736
Tijssen, Robert J.W. 469, 740
Tomizawa, Hiroyuki 956
Tomov, Dimitar T. 938
Torres-Salinas, Daniel 421
Tsay, Ming-yueh 940
Tseng, Yuen-Hsien 942

U

Ueno, Sen 956
Upham, Phineas 718
Uzun, Ali 944

V

Valderrama-Zurián, Juan Carlos 37, 818, 898
Van, Thanh-Trung 946
Vargas-Quesada, Benjamín 582
Varshavsky, A. 542
Vasconcelos, Sonia 948
Vaughan, Liwen 745
Vilagines, Andreu 828

Vinkler, Peter 753
Visser, Martijn S. 761
Vogeley, Michael 964
Voss, Jakob 950

W

Wagner, Caroline 499
Wang, Chun 776
Wang, Chun-Chieh 952
Wang, Lian 916
Wang, Liya 842
Wen, Yi 776
Wilson, Concepción S. 926, 962
Wolfram, Dietmar 954
Wu, Yishan 201,792

X

Xiao, Guo-hua 254

Y

Yamashita, Yasuhiro 956
Yang, Kiduk 568
Yang, Lili 958, 960
Yang, Lin 960
Yang, Liying 766
Yang, Zhiping 776
Yao, Shu-yin 958
Yoshikane, Fuyuki 783
You, Justin 745
Yu, Hairong 962
Yuan, Sun 936
Yue, Weiping 792

Z

Zhang, Jian 964
Zhang, Xian 254
Zhao, Chengzheng 960
Zhao, Dangzhi 966
Zheng, Yi 776
Zhu, Weizhong 189, 964
Zitt, Michel 74, 643, 797
Zuccala, Alesia 803
Zulueta, M.A……..……968