issi newsletter #10

ISSI international society for scientometrics and informetrics

quarterly e-zine of the international society
for scientometrics and informetrics

vol. 03 / nr. 2
june 2007

# EDITORIAL

## ■ ‘Scientometrics’ contents on ISSI website

In June 2007 ISSI launches a new service. Based on an initiative of *Eugene Garfield*, Founder and Chairman Emeritus of the Institute for Scientific Information (ISI, Philadelphia, PA, USA), the editors of the ISSI Newsletters decided to regularly post and archive the contents of the current issues of the journal Scientometrics on the Website of our Society. This service will be public.

The contents list will comprise bibliographic information, names and affiliation of all (co-)authors as well as the abstracts and the complete postal and electronic addresses of the corresponding authors. In addition, we will alert the community to forthcoming issues through the ISSI and SIGMETCRICS listserv.

The idea of providing this service is supported by the editorial team of Scientometrics under the direction of the *Tibor Braun*, founder and Editor-in-Chief of the journal. We also wish to thank *András Schubert*, Editor of Scientometrics, for providing us with the material we need to build up and maintain this new service.

**Wolfgang Glänzel**
Editor-in-Chief
Secretary/Treasurer ISSI

# CONTENTS

# INTRODUCING SONIA VASCONCELOS

Awardee (2007) of the Eugene Garfield Doctoral Scholarship

Sonia Vasconcelos is a Ph.D.student at the Science Education Program (PEGeD) of the Medical Biochemistry Institute (IBqM) of the Federal University of Rio de Janeiro (UFRJ). Her doctoral research on scientometrics has been conducted under the supervision of Professors Jacqueline Leta and Martha Sorenson. She is an English teacher and has designed and taught the first course on scientific writing offered through the Extension Coordination of the Health Sciences Center (CCS) at UFRJ (2004-2006). She is also the language editor of Annals of Magnetic Resonance and member of the Council of Science Editors (CSE). She is a microscopist at the Catalysis Center (NUCAT/COPPE) at UFRJ and member of the Brazilian Society for Microscopy and Materials (SBMM).

Her interdisciplinary background has made it possible for her to look at interactions between language & science. She presented her preliminary doctoral results at the ISSI 2005 Doctoral Forum, in Sweden, and her scientific writing project at the First International iPed Conference 2006: Pedagogic Research and Academic Identities, in the UK. She has recently been accepted to participate in the ESF-ORI First World Conference on Research Integrity, in Lisbon (September, 2007).

One of her interests is to gain a broader insight into research integrity-related publication issues. Her academic contributions in scientific writing include a book chapter entitled "The Scientific Global Village Bridged by a Language" (2003), "Correlating English Proficiency to International Publication Rates for Brazilian Scientists" (2006), a brief report published in the Journal for the European Medical Writers Association, and "Writing up Research in English: Choice or Necessity?" (2007), published in the Journal of the Brazilian College of Surgeons."

As a microscopist, she has co-authored 08 conference posters (1997-2007) and participated in international conferences. Her most recent contributions related to her ongoing PhD research are "English Proficiency: A Potential Science Indicator?" (ISSI 2007) and "Scientist-Friendly Policies for Non-Native English-Speaking Authors: Timely and Welcome", accepted for publication in the Brazilian Journal of Medical and Biological Research (in press).

**Title of her doctoral dissertation**:
Science in Brazil: A Scientometric and Linguistic Approach.

**Abstract**: The current study draws upon the relationship between the scientometric and linguistic scenarios of research in Brazil, whose official language is Portuguese. We have worked on a sample of 51 223 authors registered in the Brazilian National Research Council (CNPq) and evaluated their academic profile, which includes national and international publications output and data on their English proficiency. We initially focused on all academic fields to assess the authors' most developed English skill, which is reading, and then focused on the sciences and the writing skill. We have found that authors with higher rates of publication in English-language international journals are those with better writing skills. This similar trend has been noted for the country's most productive scientific fields in a 4-year period (2001-2004). We are now carrying out quantitative analysis of time leading up to publication in such journals, collecting data on the first-round turnaround time and mean acceptance time of manuscripts submitted by Brazilian authors. Also, qualitative analysis of the peer-review process of these journals and the writing of manuscripts by Brazilian authors are underway. Our preliminary data have pointed to language constraints that could potentially compromise time involved in getting published. So far, our data have shown to be consistent with our hypothesis that English proficiency may be considered a potential indicator of publication output in English-language international journals. We have looked at the Brazilian research scenario as an exemplar. In addition, this doctoral research may be seen as response not only to local, but also global demands to broaden the scope of the approach to S&T indicators in developing countries. Thus, these are the major goals of this study:

- To investigate the relationship between the scientometric and linguistic scenarios of research in Brazil;
- To investigate the possible influence of the linguistic profile of Brazilian scientists on their publication output in English-language international journals;
- To gain some quantitative and qualitative evidence on the correlation between the written English proficiency of Brazilian authors and time leading up to publication
- To broaden the discussion of Science &Technology (S&T) indicators for assessing research output of non-native English-speaking developing countries.

# 12th NORDIC WORKSHOP ON BIBLIOMETRICS AND RESEARCH POLICY

## – CALL FOR PRESENTATIONS –

**13-14 September 2007**
**Royal School of Library and Information Science**
Birketinget 6, DK-2300 Copenhagen S, Denmark

Bibliometric researchers in the Nordic countries have arranged annual Nordic workshops on bibliometrics since 1996:

1996 in Helsinki
1997 in Stockholm
1998 in Oslo
1999 in Copenhagen
2000 in Oulu
2001 in Stockholm
2002 in Oslo
2003 in Aalborg
2004 in Turku
2005 in Stockholm
2006 in Oslo

The general scope of the workshop is to present recent bibliometric research in the Nordic countries and to create better linkages between the bibliometric research groups and their PhD students. The workshop language is English and the workshop is open to participants from any nation.

### Call for presentations

The 12th Nordic Workshop on Bibliometrics and Research Policy will be held in Copenhagen, 13-14 September 2007. The workshop format is interactive and informal: All participants are requested to make a presentation of a research paper or a research idea, but no paper need to be submitted. Please register by email to Birger Larsen (blar@db.dk) if you wish to participate and also submit a max 200 word abstract on what you will present as soon as possible and no later than **August 1st 2007**.

This year's Keynote Speaker will be Dr. Gunnar Sivertsen from NIFU/STEP in Oslo. He will talk on *Publication patterns in complete bibliographic data (all scientific journals and books) at all Norwegian universities.*

Note that there are **no fees for participating** in the Nordic workshops on bibliometrics. However, travel and accommodation have to be arranged by the participants themselves.

### Important dates

Deadline for registration and abstract submission: August 1st, 2007.
Workshop: September 13-14, 2007.

Please visit the workshop website at http://vip.db.dk/blar/nbw2007/ for more information.

### Workshop Organisers

Birger Larsen, Lennart Björneborn and Peter Ingwersen
Royal School of Library and Information Science, Denmark

# THE DOCTORAL FORUM AT THE ISSI2007 CONFERENCE

Birger Larsen[1] Rickard Danell[2]
(Chairs of the Doctoral Forum)

[1]Department of Information Studies, Royal School of Library and Information Science,
Birketinget 6, DK-2300 Copenhagen, Denmark; blar@db.dk

[2]Inforsk, Department of Sociology, Umeå University,
SE-901 87 Umeå, Sweden; rickard.danell@soc.umu.se

While the Vino español informal welcome reception will be the first point on the programme for most ISSI2007 delegates a small group of senior researchers and PhD students will have begun their conference work early that morning at the ISSI2007 Doctoral Forum.

This is the second Doctoral Forum to be held at an ISSI conference – the first was held at ISSI2005 in Stockholm, where 12 students presented their work and received feedback on their projects (See *ISSI Newsletter*, 1(3), Sept. 2005 for more information). The purpose of the Doctoral Forum at ISSI is to provide doctoral students in the field with an environment in which to discuss their research projects with senior researchers and other doctoral students. An important motivation in the establishment of the Forum is to facilitate the interaction between students and experienced researchers in the field, in particular for those students who rarely get a chance to do so. This is especially important in a relatively small field such as ours where the research groups are small and scattered around the world, and doctoral students rarely have a chance to interact with experienced researchers in the field. In addition, the Forum is an opportunity for the doctoral students to set up contacts with other students who are active in the field. Indeed, the Doctoral Forum is deliberately placed *before* the main conference in order to encourage interaction amongst students themselves and between students and senior researchers during the rest of the conference.

The doctoral students applied with four-page papers describing their doctoral research project and the issues that they in particular wish to received feedback on. We received a total of 22 applications (a 32% increase from last time), of which 12 were accepted for presentation at the Forum. Brief summaries of the projects can be seen below. The participating students come from eight different countries, and represent a broad range of perspectives on the field. Six senior researchers who are experienced in different subfields of Scientometrics and

Informetrics will give feedback to the students and discus their projects with them. The participants are split into two groups, allowing each student up to 20 minutes for presentation and 25 minutes for discussion.

We wish to thank the Programme and Conference Chairs for inviting us to organise a doctoral forum at ISSI2007. It is our hope that the ISSI conference will continue to arrange a Doctoral Forum, and that it can contribute to the development of new researchers in the field and thus to the continued growth of the international community of Scientometric and Informetric researchers.

Finally, as chairs, we would like to express our sincere thanks to the senior researchers for their efforts involved in participating in the Forum. They are:

Dr. Bluma **Peritz** (Hebrew University of Jerusalem, Israel) ▪ Dr. Ed **Noyons** (Leiden University, the Netherlands) ▪ Ms. Linda **Butler** (The Australian National University, Australia) ▪ Dr. Mike **Thelwall** (University of Wolverhampton, UK) ▪ Dr. Peter **Ingwersen**, (Royal School of Library and Information Science, Denmark) ▪ Dr. Ronald **Rousseau**, (Catholic School for Higher Education Bruges-Ostend (KHBO), Belgium)

# SUMMARIES OF STUDENT PROJECTS

### An Investigation into Image Tagging

▪ **Emma Angus**

*School of Computing & Information Technology, University of Wolverhampton (United Kingdom)*

In the past several years there have been significant changes in the way people are using internet technology; the main change being the pronounced emphasis on collaboration, user contribution and user community. The term coined for this change is 'Web 2.0'. Web 2.0 applications vary greatly, ranging from websites where you can add, organise and share: bookmarks (e.g., del.icio.us), academic references (e.g., CiteULike.org), videos (e.g., youtube.com) and photographs (e.g., Flickr.com). The one thing that all of these websites have in common is their emphasis on the sharing of resources among users. Resources are generally annotated with 'tags', which are freely chosen keywords assigned by the user and not drawn from any controlled vocabulary. The collaborative and ad hoc nature of tagging systems dictates that they lack the essential properties characterising more traditional classification using controlled vocabularies and there is much discussion as to whether this impacts negatively upon retrieval precision, or whether it positively encourages serendipitous resource discovery which is unattainable using more controlled vocabularies.

Using webometric data collection, classification and informetric analysis my research project aims to investigate the tagging practices of users who upload images to the Flickr website and aims to assess how tagging both resembles and differs from traditional classification and indexing.

### Electronic Resources and Institutional Repositories in Informal Scholarly Communication and Publishing

▪ **Isabel Galina Russell**

*School of Library, Archive and Information Studies, University College London (United Kingdom)*

Academic institutions are developing ways to manage, disseminate, detect and access increasingly copious amount of digital material produced by members of the academic community. In recent years, scholarly repositories have become an important tool for electronic resource management and publication. One area that has rarely been studied is the informal publication of electronic material within these repositories. The main objectives of this research are to shed substantial light on the use, demand and visibility of electronic resources by the academic community outside the framework of formal electronic publishing. A key aspect is to develop appropriate metrics to determine the use and impact of these resources. Transaction server logs and link analysis will be explored, in conjunction with qualitative methods, as possible methodologies for aiding in the development of indicators for measuring the production and impact of new types of electronic resources which differ from formal electronic publications. These results should contribute towards further understanding of these electronic resources and the extent to which they are affecting the scholarly communication and publishing systems.

### Mapping of Research in Information and Library Science: Influences, New Directions and Changes that Occurred During the Years 1985-2003: A Bibliometric Study

■ **Luba Gornstein**
*The Hebrew University of Jerusalem (Israel)*

The main goal of the present research is to establish and define the conceptual ideas of LIS, to examine the nature of the field and especially the changing developments that occurred in the field during the period under study.

The population of the study consists of all research papers published in the 58 core journals listed by JCR. Comparison of the JCR lists for the specific period was made in order to define the journals with the highest impact factor. The theoretical and functional definitions of the research papers which appear in the already selected journals, every three years, will be examined by applying three main research methodologies: faceted classification, content and reference analysis.

A special designed questionnaire for the present study includes: "demographic" details for each paper and a subject classification, based on Peritz's definition of research and classification and on ISA taxonomy for LIS. The two classification schemes were modified according to the great developments that occurred in the field during the period under study.

Other expected contributions from the present research are: the establishment of the conceptual epistemic outline of LIS as an active and developing field of research and for its theoretical foundations the construction of an up-to-date taxonomy for subject matters which will serve additional studies.

### Link Analysis: Methods to Map Cooperation and Geopolitical Relationships

■ **Kim Holmberg**
*Department of Information studies, Åbo Akademi University (Finland)*

This research will use link analysis to study what kind of information can be discovered from local government websites in the region of Finland Proper and how this information can be used. The overwhelming goal is to establish what kind of networks, relationships and cooperation between the municipalities can be read from the links between, to and from the municipal websites. To do this, some new methods and practices for link analyses are developed. Methods used include webometric research methods, hyperlink network analysis, similarity measures, classification of links and link creation motivations and interviews. Researchers from the field of local government administration will be interviewed to confirm the results. The results so far have shown that interlinking between and co-inlinking to local government bodies in Finland follow a strong geographic, or rather a geopolitical pattern and that governmental interlinking is mostly motivated by official cooperation that geographic adjacency has made possible. Such results may give a new and interesting view on the ongoing discussions about municipal merges in the region of Finland Proper and maybe even give hints about future merges.

### Footprints Through Science – Using Citations to Assess the Path Towards Applicability

■ **Miloš Jovanović**
*Fraunhofer INT Euskirchen (Germany)*

The question of how fundamental research and applicable technologies are connected has been discussed for a long time. With finding answers, Policy makers could use indicators to better direct scientific funding and companies could focus on fundamental research that would eventually lead to an application. In my dissertation, I discuss the possibility of finding links from fundamental research towards application through the discovery and analysis of so called "Genesis Articles" (GA). A GA is defined as a paper in which a new technology or scientific discovery is presented to the scientific community for the first time. Through the articles that cited this GA and articles which have a similar topic, the path of the knowledge of the GA is traced through the scientific journals and through time. By analyzing citation levels, times of publication in different journals, the journals themselves, citing and publishing institutions and patents, I try to find a general method, with which GAs can be analyzed. The Web of Science (WoS) is used as a database, along with INSPEC and COMPENDEX to help find journals which might be read by practitioners but which one cannot find on WoS. In addition, other visualization and analysis programmes are used.

### Productivity and Prestige among Brazilian Scientists

■ **Paula Leite da Cunha e Melo**
*Medical Biochemistry Institute, Federal University of Rio de Janeiro (Brazil)*

Written communication, especially in scientific journals, is one of the most important features of modern science. Nevertheless, not all scientists have strong publication records. The motivations for publishing have been widely studied and include some issues such as prestige and productivity. Both issues are addressed in our study, which discusses the academic and scientific outputs of productive and non-productive Brazilian scientists. Our results suggest that the distribution of a particular type of fellowship, an indicator of prestige, is correlated with scientific productivity, which in turn is associated with career stage. However, gender comparisons indicate that productivity is not the only factor underlying fellowship awards. Also, our results show that geographical location and institutional features bring some advantages for scientists to get published and obtain other benefits. According to our preliminary data, the reward system of this community corroborates both the accumulative advantage theory and the Mathew effect principle. Further data are being collected individually from the online available curriculum of Brazilian scientists, in order to analyse variables related to elitism in science. In our view, this may contribute for a more contending proposal about the relationship between publishing rate and prestige in Brazilian science.

### The Maturing of Mexican Science: an Historic and Bibliometric analysis of its Development from 1980-2004

■ **María Elena Luna-Morales**
*Library and Information Studies, Universidad Nacional Autónoma de México (Mexico)*

The objective of the present thesis is to identify the main patterns of growth of Mexican science from an analysis of the scientific literature published by Mexican institutions in mainstream science and engineering journals from 1980 to 2004. Main tools used are the Science Citation Index expanded version, a data mining program developed by the Research Center in Energy of the UNAM, and Pajek. In order to determine the stages of development in Mexican science, use will be made of bibliometric techniques, of mining and visualization of data. The combination of these techniques complemented with the historical aspects that gave rise to the institutionalization and profesionalization of science in Mexico, will help to identify more precise and little known patterns on the development of Mexican science in the 25 years. With this investigation one looks for: (1) to contribute to the study of science in Mexico, (2) to demonstrate that the application of this type of methodologies is within the scope of the Mexican librarian.

### Can Patent's Value be Measured?

■ **Alba Martínez Ruiz**
*Department of Statistics and Operations Research, Technical University of Catalonia (Spain)*

Patent indicators have been traditionally used for studying their economical value, which has been related to an ample variety of indicators such as patent lifetime, family size, forwards and backwards citations, exclusion rights, breadth patent, claims and technological scope. Moreover these indicators have been related to novelty, inventive activity, international scope and the patenting strategy of his owner.

On the other hand, the economic value of the patent has been related with commercial value of a company, the economic growth of a country or region, the technological performance of a country, the research and development results and the value of innovation. However, there is little research that reports the structural relation between the variables that influence the patent value in a multidimensional approach. The purpose of this doctoral thesis is to propose a structural model and a measurement model that allows estimating the influence of the different components of technology information in the patent value. To construct the structural model it has been considered the latent variables that had been found in the state-of- art in a formative sense. The measurement model contemplates indicators from information contained in patent documents, papers, technological information of surrounding and market information.

### Applying Bradford's Law of Scattering in Digital Libraries

■ **Philipp Mayr**
*GESIS / Social Science Information Centre, Bonn (Germany)*

The background of the research project is that distributed search across multiple databases over the WWW will automatically generate large and heterogeneous document sets for scientific topics. One direction in the project will be the application of the bibliometric Bradford Law of Scattering (BLS) for generating core document sets for subject specific questions. BLS is used to re-order result sets and discover interdisciplinary properties of results from distributed searches. The application of BLS in our project has two different perspectives: 1) BLS as a supporting and optimization mechanism for information retrieval. We are focussing on an automatic change between directed descriptor searching into browsing. The subject specific result lists from the different databases will be combined and re-ranked according to Bradford's method (most productive journals for a topic first; compare Bradfordizing). After that, documents/journals in the Bradford nucleus can be delivered for browsing. 2) BLS as a way to analyse the consequence of using automatic descriptor transformations. We postulate that using descriptor cross-concordances will enlarge and complete an interdisciplinary document space which is per se only a plus in document recall. We further believe that BLS can help to analyse and evaluate the effect of automatically transferring controlled terms while subject searching.

### Mapping the Intellectual Structure of Nanotechnology Using Bibliometrics, Social Network Analysis and Ethnography

■ **Staša Milojević**
*Department of Information Studies, University of California, Los Angeles (United States)*

My project focuses on the study of nanotechnology as a scientific field, using a combination of bibliometrics, social network analysis, and qualitative methods such as interviews and document analysis. My research questions are: What is the structure and evolution of nanotechnology as a collective scientific endeavor? Does nanotechnology correspond in any sense to what various researchers in the past have modeled as a discipline or a specialty? Can the combination of the above methods offer a fruitful way of studying whether the fields such as nanotechnology are disciplines? Bibliometrics offers a powerful set of methods and measures for studying the intellectual structure of a field as represented in citations. Nanotechnology has an extensive and detailed literature, which permits the application of bibliometric techniques. I will supplement it by applying the methods of social network analysis, by studying the relationships among nanotechnology researchers, the chains of teachers and students, as well as contemporary associates and rivals. The mapping of the structure of the field will be completed with the views of the development of the field obtained through interviews with nanotechnology researchers themselves, as well as the analysis of the classification schemes used by databases holding nanotechnology material, handbooks and encyclopedias.

### Modelling Scholarly Publishing Patterns in Social Sciences and Humanities

■ **Hanna-Mari Pasanen**
*Research Unit for Science, Technology and Innovation Studies (TaSTI), University of Tampere (Finland)*

In the sciences, articles in refereed international journals predominate publishing. In social sciences and humanities, books, book chapters, and non-scholarly publications play a significant role as well. The objective of the study is to model publishing behaviour in social sciences and humanities at micro-level. Using bibliographic data, I will model the scholars' publishing patterns as their output is composed of several types of publications. The data consists of 3,500 scholars and 37,000 publications in two Finnish universities.

The individual and departmental factors affecting on 1) the orientation towards different publication types and 2) publishing productivity will be examined. The main focus of the study will be on the application of statistical techniques on modelling publishing patterns. For example, classic multivariate techniques and analysis of compositional data will be applied.

### Webometric analysis of international connectivity within science

■ **Tina Ruschenburg**
*Institute for Science and Technology Studies, Bielefeld University (Germany)*

International communication and collaboration is essential to modern science. As its name suggests, the World Wide Web appears to be a natural starting point for international studies on the global scientific community. However, after the first decade of webometrics and an enormous progress in the development of methodology, there is still comparably little empirical web research on an international level. Presumably, one reason for this diffident occupation with international questions is the considerable number of methodological hurdles and uncertainties connected to multi-national studies on the web. This dissertation project pursues the aim to evaluate the use and the potential constraints of webometrics for the empirical analysis of international connectivity within science. A case study on a small number of research institutes in the field of oceanography is being conducted in order to answer a set of ten research questions. Among other issues, it will be examined whether top-level-domains are suitable indicators of nationality, whether there are fundamental differences regarding the functions of international and domestic links, and whether the extensive country- and language-specific search options offered by Microsoft Live deliver reliable results. In addition, the international profiles given by basic webometric indicators and by standard publication measures will be compared.

# 8th COLLNET MEETING 2007 REPORT

Hildrun Kretschmer
COLLNET Center,
Borgsdorfer Str. 5, D-16540 Hohen Neuendorf, Germany
The School of Humanities and Social Sciences,
Dalian University of Technology, Dalian, China
kretschmer.h@t-online.de

The 8th COLLNET Meeting took place in conjunction with the Third International Conference on Webometrics, Informetrics, Scientometrics and Science and Society from 6-9 March, 2007, in New Delhi, India (www.collnet-delhi.de). This conference was hosted by the Society for Information Science (SIS), India, the Damodar Academy of Scientific & Educational Research (DASER), India, COLLNET, Germany, and co-hosted by the National Institute of Science,

Technology & Development Studies (NISTADS), New Delhi, as part of their 'Silver Jubilee Celebrations'.

- DASER was established in 2002 is a registered non-profit making society. The society has been working in the area of education and awareness creation for capacity building of masses (especially rural), the area of health, engineering, science and agriculture. The other area of priority is promotion and undertaking of R&D for rural upliftment. To meet the objectives the academy extends arms to collaborate with various national and international institutions engaged in similar activities.
- SIS was formed during 1975 with the aims to promote interchange of information in the discipline of information science and its subdivision amongst the specialists and between specialists and the public, to encourage and assist the professionals to maintain the integrity and competence of the profession and to foster a sense of partnership amongst the professionals engaged in these fields.
- NISTADS is one of the 38 institutes/laboratories of the Government of India's Council of Scientific and Industrial Research (CSIR), New Delhi.
- COLLNET, founded in 2000, is a global interdisciplinary research network under the title "Collaboration in Science and in Technology" (www.collnet.de). Eight COLLNET meetings were held in conjunction with international conferences or workshops on science studies (Berlin, Germany, in 2000, New Delhi, India, in 2001, Sydney, Australia, in 2001, Beijing, China, in 2003, Roorkee, India, in 2004, Stockholm, Sweden, in 2005, Nancy, France, in 2006 and New Delhi, India, in 2007). The focus of COLLNET is to examine the phenomena of collaboration in science, its effect on productivity, innovation and quality, and the benefits and outcomes accruing to individuals, institutions and nations of collaborative work and co-authorship in science.

Selected papers of earlier COLLNET Meetings were published in the *Journal of Information Management and Scientometrics (Incorporating the COLLNET Journal)*. In 2006 the international editorial board of this journal has decided to launch a new half-yearly journal from 2007 to be published in June and December each year. The journal aims to publish the papers presented at the COLLNET Meetings/Conferences and other articles of repute for the benefit of researchers and professionals. It will be available in print as well as in online form, fulfilling a long standing demand of the colleagues. The *COLLNET Journal of Scientometrics and Information Management,* published by TARU Publications, New Delhi (www.tarupublications.com), was officially launched during the 8th COLLNET Meeting in New Delhi.

The COLLNET Meeting included a pre-conference tutorial, three plenary talks by Donald deB. Beaver (USA), Wolfgang Glänzel (Belgium) and **G. J. Samadhanam** (India) highlighting the basic knowledge on the theme collaboration in science and in technology as well as about 30 papers presented by scientists from about 15 countries. In total, about 70 scientists from Asia, Europe and America have participated in the conference.

The conference was focussing on the following main topics:

- collaboration in science and technology
- measurement and evaluation of research performance
- webometrics and knowledge management
- medicine and life sciences related studies
- combination and integration of qualitative and quantitative approaches

The conference was preceded and concluded by roundtable discussions organized by NISTADS, SIS, DASER and COLLNET about upcoming projects. The time between these satellite events was an important opportunity of fruitful exchange among the participants. In particular, possible joint projects and the creation of training modules for practitioners/students were discussed.

We have just launched the COLLNET JOURNAL OF SCIENTOMETRICS AND INFORMATION MANAGEMENT, half-yearly, English. The inaugural issue vol.1#1 (June 2007) was released at the Third International Conference on Webometrics, Informetrics, Scientometrics and Science and Society & Eighth COLLNET Meeting held at New Delhi, during 6-9th March 2007.

For instructions of authors, please have a view at www.collnet-delhi.de under "COLLNET Journal". The subscription rate is $120.00 for Print+online subscription. The publishers can be contacted at info@tarupublications.com.

This news would be of interest to the professionals world over and if possible, kindly include this as a NEWS ITEM in your institution or journal.

Sincerely yours,

**Hildrun Kretschmer**
Associate Editor

# WEB SCIENCE: WHAT CAN INFORMATION SCIENCE CONTRIBUTE?

Mike Thelwall
School of Computing and Information Technology,
University of Wolverhampton,
Wulfruna Street, Wolverhampton WV1 1SB, UK.
m.thelwall@wlv.ac.uk

In November 2006 a new academic field was announced: 'Web science', supported by the Web Science Research Initiative of the University of Southampton and Massachusetts Institute of Technology (MIT), and a new journal: 'Foundations and Trends in Web Science'. The first issue of this journal was devoted to an exposition of Web science (Berners-Lee et al., 2006). The abstract begins with "This text sets out a series of approaches to the analysis and synthesis of the World Wide Web, and other Web-like information structures. A comprehensive set of research questions is outlined, together with a sub-disciplinary breakdown, emphasising the multi-faceted nature of the Web, and the multi-disciplinary nature of its study and development. These questions and approaches together set out an agenda for Web Science, the science of decentralised information systems" (Berners-Lee et al., 2006). The backing of Web inventor Tim Berners-Lee ensured wide coverage.

The initial reaction of some people (myself included!) was outrage: How can anyone claim to have invented Web science when so many people have been conducting Web research already? For example (and conflating the Web and the Internet for simplicity) there are already articles and a thesis defining Internet research (Rall, 2007; Silver, 2004), there are review chapters on information science Webometrics (Thelwall, Vaughan, & Björneborn, 2005) and search engine use for research (Bar-Ilan, 2004) as well as other review articles surveying Web research in many different areas (Arasu, Cho, Garcia-Molina, Paepcke, & Raghavan, 2001; Weare & Lin, 2000) and even a Workshop on Web Science Research Methods in 2004 (http://cybermetrics.wlv.ac.uk/AoIRASIST/).

## ■ What is Web science?

The content of the Web science framework (Berners-Lee et al., 2006) is revealing. From a bibliographic perspective, its 299 references are dominated by computer science but with contributions from a range of social sciences, including input from the social science-based Oxford Internet Institute. A large section of the article deals with the semantic Web, and it is a key goal of Web science to boost this project. No Webometricians were cited, and neither were the most high profile internet researchers of the Association of Internet Researchers (AoIR), such as Rafelei and Wellman. Both Information Science and Communication Science seem to have been ignored; surely a cause for concern if Web science is to be an effective multidisciplinary initiative. Nevertheless there are overlaps: I counted 19 references to articles that I have also cited or attended as talks.

Overall, I think that Web science actually means Web computer science and does not seriously seek to address issues of social science other than those which support the computer science objectives associated with system engineering. More specifically, and as stated in the text of the introductory article, Web science is decentralised information systems research, or the science of the Web as an Information System (Berners-Lee et al., 2006). Some information science topics are deemed relevant to Web science, however. There are sections on folksonomy (p30), metadata (p35) and ontology (p27), with this kind of information seen as essential to a well-functioning Web. Given that MySpace has overtaken Google as the most visited Web site for U.S. users (Prescott, 2007), the social side of the Web and studies of the uses to which it is put seem to be of vital importance, even from the computer science perspective.

If the Web science initiative helps to unify and direct computer science and other research towards a set of clear and useful goals, then it will be a positive step. The Web science initiative is almost certain to have some impact within computer science because of its high-profile backers, but whether it becomes established in the longer term, and whether other disciplines are also affected, is less sure. Probably these depend upon the ability of Web science to deliver practical and theoretical results that can justify the original vision, and upon whether the Web evolves in new unexpected directions. The arguably mixed fortunes of the Semantic Web may also have an influence, as will the increasing importance of the social side of the Web (e.g., MySpace) and the apparent unpredictability of new developments.

In summary, Web science does not intend to usurp Webometrics (as I initially feared) or social science Web research. Rather, it is an initiative aimed at encouraging focused computer-science led research into making the Web into an even more effective decentralised information system. If successful, all Web users may gain, and, as a by-product, Web researchers will probably have access to more powerful tools and theories for measuring and analysing the Web.

## ■ Web science degrees: Can information science contribute?

In addition to the call for research, the Web science initiative is likely to lead to the creation of several Web science courses. For example a Web science degree has been announced in a joint initiative between Southampton University and MIT. Should information scientists argue that techniques from our field could usefully be included? The following quote helps the case for the information science: “because humans are the creators of Web pages and links between them, their interactions form emergent patterns in the Web at a macroscopic scale. These human interactions are, in turn, governed by social conventions and laws. Web science, therefore, must be inherently interdisciplinary; its goal is to both understand the growth of the Web and to create approaches that allow new powerful and more beneficial patterns to occur" (Berners-Lee, Hall, Hendler, Shadbolt, & Weitzner, 2006).

I think that two Webometric research areas have a particularly strong case for inclusion: search engine evaluation and link analysis. Search engine evaluation (Bar-Ilan, 2004; Rousseau, 1999; Vaughan & Thelwall, 2004) analyses the reliability of search engine results, but not from the computer science perspective of assessing how effective the search engine is at delivering useful results to the end user. Instead the focus is document-centred, investigating issues such as the coverage of search engines, biases in coverage, difference between engines and changes in results over time. This kind of research perspective is important because of the key role that search engines play as gateways to the Web, something that can easily be overlooked by Web site designers (Van Couvering, 2007).

Link analysis is another useful information science contribution to Web science. In fact, there is already both a significant body of link analysis research in computer science (called Web structure mining) and an acknowledgement of the importance of bibliometrics in motivating some of this research, including Google's PageRank algorithm (Brin & Page, 1998). The unique information science contributions are the development and formalisation of the alternative document model approach to link counting (Thelwall, 2004) and the application of link analysis to mining information about the impact of Web sites and other Web content (Aguillo, Granadino, Ortega, & Prieto, 2006), as well as to highlight the relationships between collections of Web sites, as manifested by links (Heimeriks & van den Besselaar, 2006). Knowledge of this large-scale multiple Web site analysis perspective can help Web science courses by building intuition and understanding about how the Web works. This would be particularly useful to support mathematical Web growth modelling (Pennock, Flake, Lawrence, Glover, & Giles, 2002) and topological modelling (Björneborn, 2006; Broder et al., 2000) which are already part of Web Science, but operate at a high level of abstraction.

## ■ References

Aguillo, I. F., Granadino, B., Ortega, J. L., & Prieto, J. A. (2006). Scientific research activity and communication measured with cybermetrics indicators. *Journal of the American Society for Information Science and Technology, 57*(10), 1296-1302.

Arasu, A., Cho, J., Garcia-Molina, H., Paepcke, A., & Raghavan, S. (2001). Searching the Web. *ACM Transactions on Internet Technology, 1*(1), 2-43.

Bar-Ilan, J. (2004). The use of Web search engines in information science research. *Annual Review of Information Science and Technology, 38*, 231-288.

Berners-Lee, T., Hall, W., Hendler, J., Shadbolt, N., & Weitzner, D. J. (2006). Creating a science of the Web. *Science, 313*(5788), 769-771.

Berners-Lee, T., Hall, W., Hendler, J. A., O'Hara, K., Shadbolt, N., & Weitzner, D. J. (2006). A framework for Web science. *Foundations and Trends in Web Science, 1*(1), 1-130.

Björneborn, L. (2006). 'Mini small worlds' of shortest link paths crossing domain boundaries in an academic Web space. *Scientometrics, 68*(3), 395-414.

Brin, S., & Page, L. (1998). The anatomy of a large scale hypertextual Web search engine. *Computer Networks and ISDN Systems, 30*(1-7), 107-117.

Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., et al. (2000). Graph structure in the Web. *Journal of Computer Networks, 33*(1-6), 309-320.

Heimeriks, G., & van den Besselaar, P. (2006). Analyzing hyperlink networks: The meaning of hyperlink-based indicators of knowledge. *Cybermetrics, 10*(1), Retrieved August 1, 2006 from: http://www.cindoc.csic.es/cybermetrics/articles/v2010i2001p2001.html.

Pennock, D., 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, 99*, 5207-5211.

Prescott, L. (2007). Hitwise US consumer generated media report.Retrieved March 19, 2007 from: http://www.hitwise.com/.

Rall, D. N. (2007). *Considering internet scholarship - in theory and practice.* Southern Cross University, Lismore.

Rousseau, R. (1999). Daily time series of common single word searches in AltaVista and NorthernLight. *Cybermetrics, 2/3*, Retrieved July 25, 2006 from: http://www.cindoc.csic.es/cybermetrics/articles/v2002i2001p2002.html.

Silver, D. (2004). Internet/cyberculture/digital culture/new media/fill-in-the-blank studies. *New Media & Society, 6*(1), 55-64.

Thelwall, M. (2004). *Link analysis: An information science approach.*San Diego: Academic Press.

Thelwall, M., Vaughan, L., & Björneborn, L. (2005). Webometrics. *Annual Review of Information Science and Technology, 39*, 81-135.

Van Couvering, E. (2007). Is relevance relevant? Market, science, and war: Discourses of search engine quality. *Journal of Computer-Mediated Communication, 12*(3), Retrieved May 14, 2007 from: http://jcmc.indiana.edu/vol2012/issue2003/vancouvering.html.

Vaughan, L., & Thelwall, M. (2004). Search engine coverage bias: evidence and possible causes. *Information Processing & Management, 40*(4), 693-707.

Weare, C., & Lin, W. Y. (2000). Content analysis of the World Wide Web-Opportunities and challenges. *Social Science Computer Review, 18*(3), 272-292.

# A NOTE ON THE CONNECTION BETWEEN THE HIRSCH INDEX AND THE RANDOM HIERARCHICAL MODEL

Aparna Basu

G.N. Ramachandran Knowledge Centre for Genome Informatics,
Institute of Genomics and Integrative Biology,
Mall Road, Delhi 110007, INDIA
aparnabasu.dr@gmail.com

***In this note I explore the possible connections between the h-index and the Random Hierarchical model originally formulated as a model for Bradford's Law in bibliometrics.***

## Introduction

The *h* index is a new measure of scientific achievement suggested by Hirsch[1] in 2005. The h index of a scientist is defined as the number of his papers that have at least *h* citations each. The meaning of *h,* as a relative measure of achievement, is intuitively clear; a scientist with a higher *h* has a larger number of papers that are each cited a larger number of times than another scientist with lower *h*. The implication is that the work of the former has had a wider impact, and he can be ranked higher than the latter on a scale of scientific achievement.

Hirsch's work has been followed by a number of papers that analyse the properties[2,3,4,5], modify or refine the index so as to better bring out an acceptable ranking order of eminence corresponding to perceptions in the scientific community[6,7]

In the present note, my intention is to point out that an interim equation used by Hirsch is structurally similar to the distribution obtained in the Random Hierarchical (RH) model[8] used earlier as a model of Bradford's law. This has motivated us to examine the mathematical underpinnings of the Hirsch index and compare with the RH model. If the RH model is regarded as a model of the distribution of citations to papers, then the *h* index is that particular point in the distribution where the citation level of a set of papers is equal to the number of papers with that citation level. I argue that the RH model may be regarded as general distribution of which the *h* index is a special case.

## ■ The Hirsch Index: Mathematical Analysis

Hirsch first considers a simple linear model of growth of citations. A scientist writes *p* papers every year, and each paper gets a fixed number of citations *c* every year thereafter. According to this model, the paper written the earliest gets the largest number of citations, while the most recent paper gets the least number of citations. The total number of citations received at the end of n years is, **[1]**

$$N_{c,tot} = \sum_{j=1}^{n} pcj = \frac{pcn(n+1)}{2}$$

The number of citations received by the $y^{th}$ paper in a ranked list of papers arranged in decreasing order of citedness is **[2]**

$$N_c(y) = N_0 - (\frac{N_0}{h} - 1)y$$

The total number of papers, $y_m$ is given by **[3]**

$$y_m = \frac{N_0 h}{N_0 - h}$$

This is a very simplistic model in time of the accrual of papers and citations, which is necessarily a complex random process. Hirsch then adopts a stretched exponential which he says is more realistic, **[4]**

$$N_c(y) = N_0 e^{-\left(\frac{y}{y_0}\right)^{\beta}}$$

and derives another relation for the total number of papers **[5]**

$$y_m = h\left[1 + \alpha^{\beta} \ln(h)\right]^{1/\beta}$$

The relationship expressed by Eqn. 5 is strikingly similar to the main result of the Random Hierarchical model that I shall examine in the next section, and compare the two approaches.

## ■ The Random Hierarchical Model

The 'Random Hierarchical' distribution was proposed by Basu[9], as a model for the distribution of articles in journals usually known in information science as Bradford's law[10]. The 'law' is an observed relationship between journals ranked by decreasing 'productivity' (defined in terms of articles published by the journal in some particular knowledge domain), and the total (cumulative) number of articles published in journals down to some rank r. The regularity observed by Bradford was that the cumulative articles produced by journals when plotted against the logarithm of the journal rank, almost always followed a J shaped curve, regardless of the size of the bibliography, time span, etc. Subsequently it was observed that the curve could fall away from the linear end of the J-shaped curve and assume an S-shape, which was referred to as the Groos droop.[11]

The RH model is similar to the Whitworth model of random fragmentation of an entity into N parts [Whitworth, 1901][12]. The RH model was fitted (*see Basu,* 1992) to a sample data set of journals and articles in agriculture given by Lawani[13] and found to give a good fit in all three regions, namely, the 'core' region of highly productive journals, the almost linear portion in the middle of the Bradford curve and the falling Groos droop. Although formulated as a model for the Bradford law of distribution of articles in journals, the RH model is, in fact, context independent, being based on general probabilistic principles without referring to specific details of the time dependent and causal processes by which a distribution is generated. This makes it likely to have wider applicability to situations where similar processes are in operation.

The two main assumptions of Random Hierarchical model were *a*) that the distribution was the result of a random process, and, *b*) that the resulting distribution should be 'hierarchical', i.e., the fragments could be 'ranked'. This last assumption is in the nature of a constraint. The objective was to derive the *most probable* distribution under these conditions. (It may be noted that the *most probable* or *least biased* distribution in the absence of any constraint would be a uniform distribution). Finally, it is assumed that the *most probable* distribution obtained is the same as the distribution generated by the actual time dependent random process.

Applied to citations of an individual scientist, the RH model can be expressed by the relation, **[6]**

$$p = [q(1 - \ln q)]^{\alpha}$$

Where '*p*' is the proportion of the scientist's total citations earned by a proportion '*q*' of his papers from the top of a ranked list of papers ranked in order of decreasing citations.

The argument on which the above relation is derived is briefly outlined below.

Let us assume that the unit line fragments into 2 parts. Clearly the probability of breaking at the two end points is zero. The probability of breaking would be higher at some intermediate point. The most probable (unbiased) point at which it can break is at the mid-point. However this violates the constraint requiring unequal sizes. If the probability is taken to be 0 at the mid-point, then the probability distribution across the unit line is bimodal, peaking at the positions ¼ and ¾, particularly. The resulting sizes for 2 pieces may be written as **[7]**

$$\left(\tfrac{1}{2}\cdot\tfrac{1}{2}\right) \text{ and } \left(\tfrac{1}{2}\cdot\tfrac{1}{2}+\tfrac{1}{2}\cdot 1\right)$$

For N fragments, the resulting sizes will be **[8]**

| Size | Rank |
|---|---|
| $\frac{1}{N}\cdot\frac{1}{N}$ | $N$ |
| $\left(\frac{1}{N}\cdot\frac{1}{N}\right)+\left(\frac{1}{N}\cdot\frac{1}{N-1}\right)$ | $N-1$ |
| $\left(\frac{1}{N}\cdot\frac{1}{N}\right)+\left(\frac{1}{N}\cdot\frac{1}{N-1}\right)+\ldots+\left(\frac{1}{N}\cdot\frac{1}{N-x}\right)$ | $N-x$ |
| $\left(\frac{1}{N}\cdot\frac{1}{N}\right)+\left(\frac{1}{N}\cdot\frac{1}{N-1}\right)+\ldots+\left(\frac{1}{N}\right)$ | 1 |

where ranks have been assigned in decreasing order of size.

The expected size for rank *i* is, **[9]**

$$E_i = \frac{1}{N}\sum_{x=1}^{N-i}\frac{1}{N-x} = \frac{1}{N}\sum_{j=1}^{N}\frac{1}{j}$$

The cumulative sizes of fragments up to rank *i* is **[10]**

$$CE_i = 1 + \sum_{j=1}^{N}\frac{1}{j} - \sum_{j=1}^{i}\frac{1}{j}$$

This result is based on a discrete model, which can be further simplified when the number of pieces *N* becomes large, and **[11]**

$$\lim_{N\to\infty}\left(\sum_{j=1}^{N}\frac{1}{j} - \ln N\right) = \gamma$$

where *g* » 0.57 is Euler's constant. Applying this to Eqn [9], for papers and citations gives,**[12]**

$$\frac{B(r)}{\mu} = r[1 + \ln N] - r\ln r$$

where *B(r)* is the total number of citations to the top *r* cited papers, and *m* is the average citedness of the author *B/N*, *N* being the total number of papers.

Eqn. [12] can be re-written as a relation of the proportion of citations *p* obtained by any given proportion *q* from the top of a ranked list of papers, as **[13]**

$$p = [q(1 - \ln q)]$$

An additional free parameter a may be introduced taking care that the boundary conditions (*p=0, q=0* and *p=1,q=1)* are preserved, giving the relation stated in Eqn.[6],

### ■ Comparison between the Hirsch and RH models

Note that the relation given by Eqn.[6] is strikingly similar in form to that obtained by Hirsch in Eqn.[5], although they have been obtained in entirely different ways. For example, the stretched exponential distribution was not assumed in the RH model, which was obtained from a random fragmentation process, additionally constraining the fragments to be unequal. Moreover, the entities in the equations are also different.

The Hirsch relationship in Eqn. [5] connects the number of papers with at least one citation to the *h* index, which connects the citations and papers of a scientist at a single point in the distribution. The RH relation also connects papers and cumulative citations, but is more general and applies anywhere across the distribution of citations and papers.

This implies that the Hirsch relation should be a special case of the RH distribution, valid when one considers *h* papers as given by Hirsch's definition.

## References

[1] Hirsch, J.E. (2005) An index to quantify an individual's scientific achievement, Proceedings of the National Academy of Sciences, Nov 15, Vol. 102, No. 46, 16569-16572

[2] Glänzel, W. (2006) On the Opportunities and Limitations of the H-index, (in Chinese), Science Focus, 1 (1), 10-11 (English version available at http://eprints.rclis.org/view/people/Gl=nzel,_Wolfgang.html)

[3] Glänzel, W. (2006) On the H-Index – A mathematical approach to a new measure of publication activity and citation impact, Scientometrics, 67 (2), 315-321

[4] Egghe L., Rousseau R. (2006) An informetric model for the Hirsch-index, Scientometrics 69 (1) 121-129

[5] Burrell, Q.L. (2006) Hirsch's h-index: A stochastic model, Journal of Informetrics, 1 (1) 16-25

[6] Egghe L. (2006) Theory and practice of the g-index, Scientometrics 69 (1) 131-152

[7] Jin, BH, Liang, LM., Rousseau., R, Egghe, L. (2007) The R- and AR-indices: Complementing the h-index, Chinese Science Bulletin, 52 (6) 855-863.

[8] Basu, A. (1992) Hierarchical Distributions and Bradford's law, JASIST, Vol. 43, No.7, 494-500

[9] Basu, A. (1991) in 3rd International Conference of the Society of Scientometrics and Informetrics, Informetrics 91, Bangalore, India

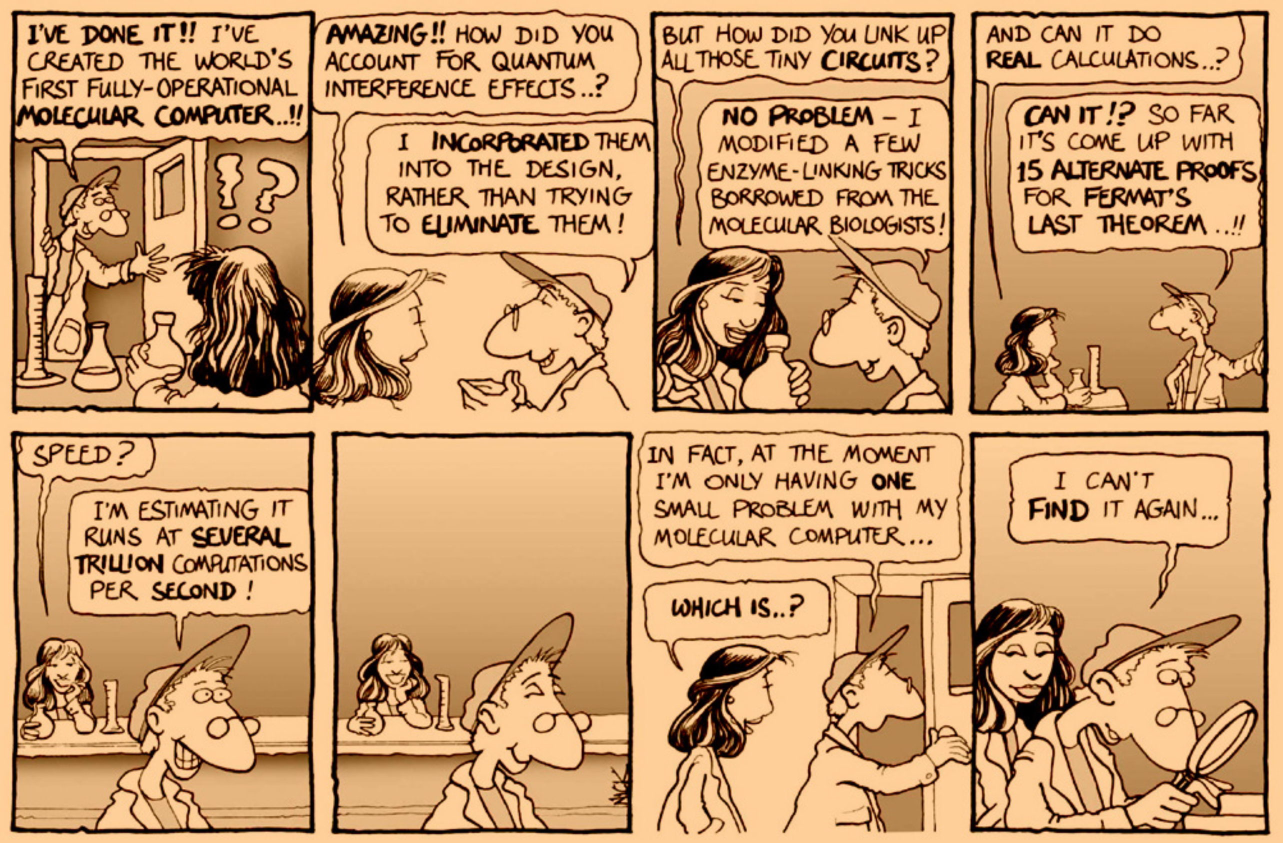
[10] Bradford, S.C. (1948) Documentation, London: Crosby Lockwood; Bradford, SC (1934) Sources of information on specific subjects, Engineering, 137, 85-86

[11] Groos, O.V. (1988) Bradford Law and the Keenan Atherton data, American Documentation, 18, 46

[12] Whitworth, W. A. (1901). Choice and Chance. 5th ed. New York: Hafner.

[13] Lawani, S.M. (1973) Bradford's law and the literature of agriculture, International Library Review, 5, 341-350

## CARTOON


I'VE DONE IT!! I'VE CREATED THE WORLD'S FIRST FULLY-OPERATIONAL MOLECULAR COMPUTER..!!
AMAZING!! HOW DID YOU ACCOUNT FOR QUANTUM INTERFERENCE EFFECTS..?
I INCORPORATED THEM INTO THE DESIGN, RATHER THAN TRYING TO ELIMINATE THEM!
BUT HOW DID YOU LINK UP ALL THOSE TINY CIRCUITS?
NO PROBLEM – I MODIFIED A FEW ENZYME-LINKING TRICKS BORROWED FROM THE MOLECULAR BIOLOGISTS!
AND CAN IT DO REAL CALCULATIONS..?
CAN IT!? SO FAR IT'S COME UP WITH 15 ALTERNATE PROOFS FOR FERMAT'S LAST THEOREM..!!
SPEED?
I'M ESTIMATING IT RUNS AT SEVERAL TRILLION COMPUTATIONS PER SECOND!
IN FACT, AT THE MOMENT I'M ONLY HAVING ONE SMALL PROBLEM WITH MY MOLECULAR COMPUTER...
WHICH IS..?
I CAN'T FIND IT AGAIN...

© Nick Kim (Nearing Zero). Reproduced with the permission of the author.

# SOME NEW APPLICATIONS OF THE H-INDEX*

Wolfgang Glänzel[1,2]
[1]Steunpunt O&O Indicatoren
& K.U.Leuven, Faculty ETEW, MSI, Leuven (Belgium)
[2]Hungarian Academy of Sciences, IRPS, Budapest (Hungary)
wolfgang.glanzel@econ.kuleuven.be

*In this note some new fields of application of Hirsch-related statistics are presented. Furthermore, so far unrevealed properties of the h-index are analysed in the context of rank-frequency and extreme-value statistics.*

## ■ Introduction

Since its introduction in 2005, the h-index has mainly been used as a measure to quantify the research output of individual scientists. This is in line with Jorge E. Hirsch's intensions (Hirsch, 2005). Recent attempts to fine-tune or improve the indicator (e.g., Egghe, 2006 and Jin et al., 2007) or to extend its use to higher levels of aggregation (e.g., Braun et al., 2005) follow the original design. In what follows, we will show some new application possibilities of the h-index in the context of rank statistics. In particular, the properties of the characteristic extreme values of Pareto-type distributions provide the basis of the new statistics. The first application is actually found in the form of a composite indicator strongly related to the h-index. The second application relates the h-index with a generalised version of the Zipf-Mandelbrot law. While the first indicator can only be applied to distributions with finite expectation, that is $\alpha > 1$, the second application even works if $\alpha \leq 1$. Both applications are useful supplements in evaluative studies of research performance at the micro and meso level.

## ■ Theoretical background

In recent papers (Glänzel, 2006, Egghe and Rousseau, 2006, Burrell, 2007), attempts were made to interpret theoretically some properties of the *h*-index and to connect the results with traditional indicators of publication activity and citation impact (Schubert and Glänzel, 2007). The underlying citation distribution was assumed to be Paretian and on the basis of extreme-value statistics, important properties and regularities could be derived from the distribution. Specifically, the dependence of the *h*-index on the basic parameters of the distribution and on the sample size was discussed using Gumbel's characteristic extreme values. In order to further elaborate these new approaches, we briefly summarise the mathematical rudiments.

Let $X$ be a random variable. In the present case $X$ represents the citation rate of a paper. The probability distribution of $X$ is denoted by $p_k = P(X = k)$ for every $k \geq 0$ and the cumulative distribution function is denoted by $F_k = P(X < k)$. Put $G_k := 1 - F_k = P(X \geq k)$. Assume we have a sample $(\{X_i\}_{i=1,\ldots,n})$ of size $n$ where all elements are independent and have the same distribution $F$. Gumbel's $r$-th characteristic extreme value $(u_r)$ is then defined as **[1]**

$$u_r := G^{-1}(r/n) = \max\{k: G_k \geq r/n\},$$

where $n$ is a given sample with distribution $F$ (see *Gumbel*, 1958). The actual rank statistics $R(r) = X_r^*$ (where $X_1^* \geq X_2^* \geq \ldots \geq X_i^* \geq \ldots \geq X_n^*$

are the ordered/ranked elements of the sample $\{X_i\}_{i=1,\dots,n}$ can be considered an estimator of the corresponding *r*-th characteristic extreme value $u_r$.

According to Glänzel (2006), the theoretical *h*-index (*h*) can be defined as **[2]**

$$h := \max\{r:\ u_r \ge r\} = \max\{r:\ \max\{k:\ G_k \ge r/n\} \ge r\}\,.$$

If there exists such index *r* so that $u_r = r$ then we have obviously $h := r$ and we can write $h := u_h$.

## ■ Methods and Results

For simplicity's sake we assume that the citation distribution under study can be approximated by a non-negative continuous distribution. In the case of continuous distributions we will write $F(x)$ and $G(x)$ instead of $F_x$ and $G_x$, respectively. Furthermore, we assume that the underlying citation rates follow a Pareto distribution of the second kind. This general form of the Pareto distribution, also referred to as *Lomax* distribution, can be obtained from the infinite beta distribution if one of the parameters is chosen 1 (e.g., Johnson, Kotz, & Balakrishnan, 1994). In particular, we say that the non-negative random variable *X* has a Pareto distribution (of the second kind) if **[3]**

$$G(x) = P(X \ge x) = N^{\alpha}/(N+x)^{\alpha}\ ,\ \text{for all } x \ge 0$$

Clearly, if *x* is large ($x >> N$) we can neglect the parameter in the denominator and we have **[4]**

$$G(x) \sim N^{\alpha}/x^{\alpha}\ ,\ \text{for } x >> N.$$

Assuming a statistical sample with Lomax distribution and size *n* we obtain **[5]**

$$G(u_r) \sim N^{\alpha}/u_r^{\alpha} = r/n,\ \text{if}\ \ n >> r.$$

Consequently, we have $r \cdot u_r^{\alpha} = N^{\alpha} \cdot n$ and **[6]**

$$\zeta(r) := r^{1/(\alpha+1)} \cdot u_r^{\alpha/(\alpha+1)} = N^{\alpha/(\alpha+1)} \cdot n^{1/(\alpha+1)}.$$

Since the right-hand side does not depend on the particular rank *r*, the left-hand side must be a constant. Furthermore, we have $\zeta(h) = h$ by definition (namely $\zeta(h) = h^{1/(\alpha+1)} \cdot h^{\alpha/(\alpha+1)} = h$). Consequently, $\zeta(r) \equiv h$ for all $r << n$. This property yields the first important result, particularly, **[7]**

$$h = c(\alpha)^{*} \cdot \mathrm{E}(X)^{\alpha/(\alpha+1)} \cdot n^{1/(\alpha+1)},\ \ \text{if } \alpha > 1,$$

where $\mathrm{E}(X) = N/(\alpha-1)$ is the expected value of the underlying Lomax distribution and $c(\alpha)^{*} = (\alpha-1)^{\alpha/(\alpha+1)}$ is a positive real value which only depends on the parameter α. Taking into account that the continuous *Lomax distribution* model often rather poorly fits the empirical discrete, integer-valued distributions, one cannot expect a perfect correlation. Nonetheless, we have found a strong correlation between *h* and $m^{\alpha/(\alpha+1)} \cdot n^{1/(\alpha+1)}$ with *m* being the mean citation rate of scientific journals (Schubert and Glänzel, 2007). Solely the empirical $c(\alpha)^{*}$ value was usually somewhat lower that the theoretical one.

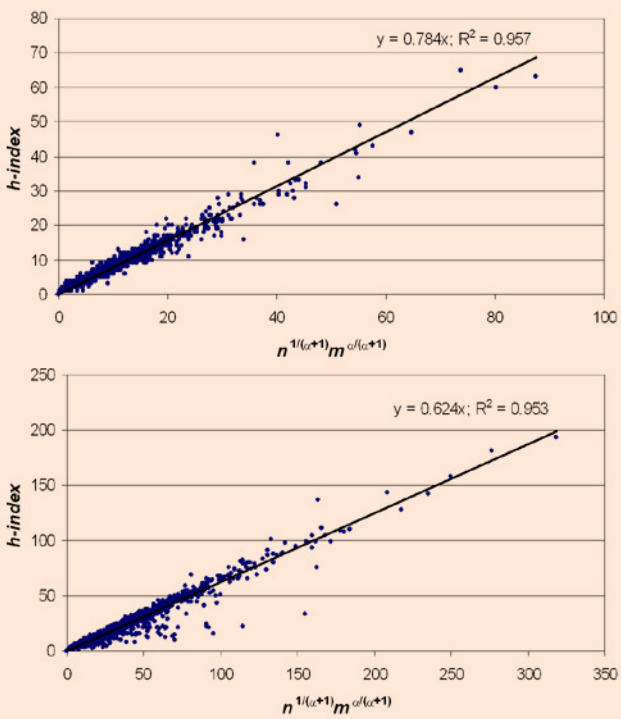


*Figure 1 Correlation of the journal h-index with $n^{1/(\alpha+1)} m^{\alpha/(\alpha+1)}$ in all science fields combined top: citation window: 1980-1982 (α=2); bottom: citation window: 1980–2000 (α=1.5)*

For largely different citation windows we have found solutions with different α values. For small windows comprising an initial period of about three years after publication, an α value around 2 has been found appropriate. For larger windows lower values yield an optimum solution. This change of exponent α with growing time intervals is in line with observations by Vlachý (1976) and Pao (1986).

Figure 1 shows the dependence of *h* on *n* and journal impact measures *m* for papers published in 1980 and indexed in the *Science Citation Index* of Thomson Scientific (Philadelphia, PA, USA). The impact measures have been calculated for a 3-year (top) and 21-year (bottom) citation window, respectively. In the first case α = 2, for the longer citation period α = 1.5 has been chosen.

Both theoretical considerations and empirical analysis lead to the conclusion that the h-index strongly correlates with $m^{\alpha/(\alpha+1)} \cdot n^{1/(\alpha+1)}$ which can be considered a composite indicator combining publication output and mean citation rate. Although this indicator has interesting properties, it is not designed to substitute the h-index. We just mention is passing that a similar indicator for journal impact was already suggested by *Lindsey* (1978) independently from the Hirsch index. For his *Corrected Quality Ratio* (CQR) we actually have $CQR^{0.4} = n^{0.4} \cdot m^{0.6}$, i.e., in this case we have $\alpha = 1.5$.A second important property can be observed when replacing the theoretical values in the left-hand side of Eq. (6) by the corresponding statistics. In particular, we obtain $z(r) = r^A \cdot R(r)^{(1-A)}$ with $A = 1/(\alpha+1)$, where $z(r)$ is expected to be an estimator of $h$ for each $r << n$. In practice the deviation from $\zeta(r) \equiv h$ is quite large for the individual $r$ values but the median $M$ of the empirical $z(r) = r^A \cdot R(r)^{(1-A)}$ values proofed a strikingly robust estimator of $h$. Table 1 presents the corresponding statistics for selected *Price Awardees*. Both publications and citations have been counted from 1972 till May 2007 and no selection has been made for relevant literature. Thus papers in chemistry have been included for Schubert; the same applies for physics publication by van Raan and mathematical papers by Egghe, Rousseau and Glänzel. For the *Hirsch core* we obtained $M \sim h$ (with $\alpha = 2$). Except for Henry Small, $z$ statistics provide robust approximation and the corresponding medians good estimators of the h-index. The reason for the poor fit for Henry Small lies in several highly cited papers of the 1970s on co-citation analysis and an extremely skewed citation distribution. For Henry Small an $\alpha$ value of about 1.0 resulting in constant $z$ values around 20 would be more appropriate.

This new method for analysing the tail properties of Pareto-type distributions based on the $z$ statistics works much better than the model described by Glänzel and Schubert (1988). The latter one was based on transformations of ordered statistics, namely on $r \cdot \ln(X_r^*/X_{r+1}^*) = r \cdot \ln[R(r)/R(r+1)]$ with $r << n$, which were extremely sensitive to ties. These statistics have an exponential distribution with parameter $\alpha$, provided the underlying common distribution of $X_r^*$ is Paretian with the same parameter (Glänzel et al., 1984). In practice, rank statistics of integer-valued discrete distributions often include ties (i.e. $R(r) = R(r+1)$ for some $r = 1, 2, \ldots$) resulting in $r \cdot \ln[R(r)/R(r+1)] = 0$. These ties can heavily distort the fit of the exponential distribution. By contrast, the new $z$ statistics are more robust and much less sensitive to ties (see Table 1). One further important property is worth mentioning, namely that the $z$ statistics can be considered a version of the Zipf-Mandelbrot law (cf. Yablonski, 1980, Egghe and Rousseau, 1990), where the constant value equals the h-index to the power ($\alpha$+1), that is, $r \cdot R(r)^\alpha = \{z(r)\}^{\alpha+1} = h^{\alpha+1}$. Consequently, the case $\alpha = 1$ results in the following version of the classical Zipf's Law: $z(r) = \{r \cdot R(r)\}^{½} = h =$ constant, or equivalently, $r \cdot R(r) = h^2$ .

*(See table Table 1 for Hirsch-type indexes for selected bibliometricians who have been awarded the Price Medal (with $\alpha = 2$) on the next page!)*

## ■ Conclusions

In this paper we have described two new applications of Hirsch-related indexes. The composite indicator, which expresses a multiplicative connection between derivatives of publication output and citation impact, proved surprisingly robust and works at both the meso and the micro level. Its strong correlation with the h-index is independent of the subject area (cf. Schubert and Glänzel, 2007). The $z$ statistics, representing the second application, can be used to analyse the tail of citation distributions in the light of the h-index. At the same time, the h-index proved useful as truncation point for rank frequency analysis, for instance, by applying $z$ and related statistics to the *Hirsch core* (e.g. Burrell, 2007) publications.

## ■ References

Braun, T., Glänzel, W., Schubert, A. (2005), A Hirsch-type index for journals. *The Scientist*, 19 (22) 8.

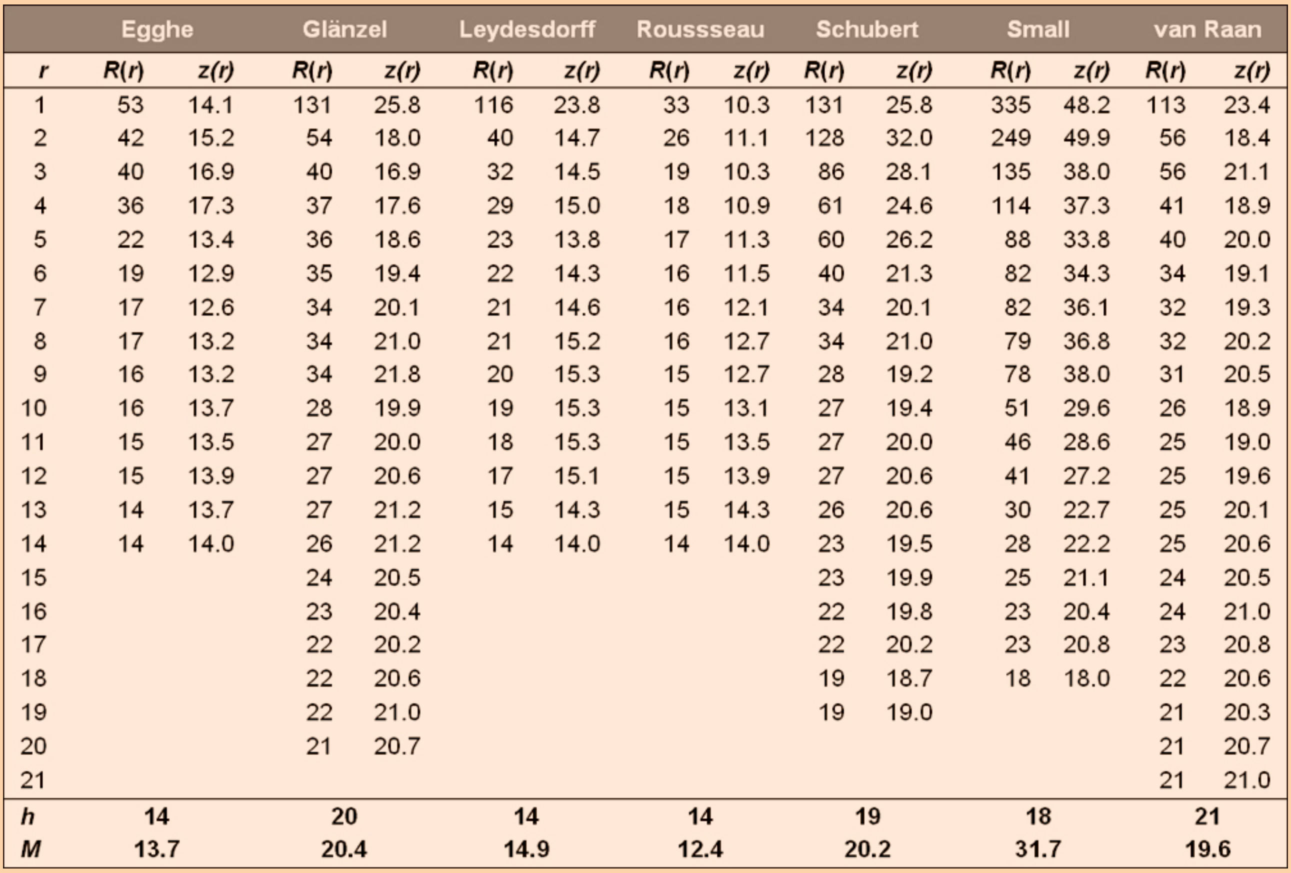
Burrell, Q. L. (2006), Hirsch's h-index: a stochastic model. *Journal of Informetrics*, 1 (1) 16-25

Burrell, Q. L. (2007), On the h-index, the size of the Hirsch core and Jin's A-index, *Journal of Informetrics*, 1 (2) 170-177.

Egghe, L., Rousseau, R. (1990). *Introduction to informetrics. Quantitative methods in library, documentation and information science*. Elsevier Science Publisher. Amsterdam.

Egghe, L., Rousseau, R. (2006). An informetric model for the h-index. Scientometrics, 69(1), 121-129.

Egghe, L. (2006). Theory and practice of the g-index. *Scientometrics*, 69(1), 131-152.



| | Egghe | | Glänzel | | Leydesdorff | | Roussseau | | Schubert | | Small | | van Raan | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *r* | *R*(*r*) | *z*(*r*) | *R*(*r*) | *z*(*r*) | *R*(*r*) | *z*(*r*) | *R*(*r*) | *z*(*r*) | *R*(*r*) | *z*(*r*) | *R*(*r*) | *z*(*r*) | *R*(*r*) | *z*(*r*) |
| 1 | 53 | 14.1 | 131 | 25.8 | 116 | 23.8 | 33 | 10.3 | 131 | 25.8 | 335 | 48.2 | 113 | 23.4 |
| 2 | 42 | 15.2 | 54 | 18.0 | 40 | 14.7 | 26 | 11.1 | 128 | 32.0 | 249 | 49.9 | 56 | 18.4 |
| 3 | 40 | 16.9 | 40 | 16.9 | 32 | 14.5 | 19 | 10.3 | 86 | 28.1 | 135 | 38.0 | 56 | 21.1 |
| 4 | 36 | 17.3 | 37 | 17.6 | 29 | 15.0 | 18 | 10.9 | 61 | 24.6 | 114 | 37.3 | 41 | 18.9 |
| 5 | 22 | 13.4 | 36 | 18.6 | 23 | 13.8 | 17 | 11.3 | 60 | 26.2 | 88 | 33.8 | 40 | 20.0 |
| 6 | 19 | 12.9 | 35 | 19.4 | 22 | 14.3 | 16 | 11.5 | 40 | 21.3 | 82 | 34.3 | 34 | 19.1 |
| 7 | 17 | 12.6 | 34 | 20.1 | 21 | 14.6 | 16 | 12.1 | 34 | 20.1 | 82 | 36.1 | 32 | 19.3 |
| 8 | 17 | 13.2 | 34 | 21.0 | 21 | 15.2 | 16 | 12.7 | 34 | 21.0 | 79 | 36.8 | 32 | 20.2 |
| 9 | 16 | 13.2 | 34 | 21.8 | 20 | 15.3 | 15 | 12.7 | 28 | 19.2 | 78 | 38.0 | 31 | 20.5 |
| 10 | 16 | 13.7 | 28 | 19.9 | 19 | 15.3 | 15 | 13.1 | 27 | 19.4 | 51 | 29.6 | 26 | 18.9 |
| 11 | 15 | 13.5 | 27 | 20.0 | 18 | 15.3 | 15 | 13.5 | 27 | 20.0 | 46 | 28.6 | 25 | 19.0 |
| 12 | 15 | 13.9 | 27 | 20.6 | 17 | 15.1 | 15 | 13.9 | 27 | 20.6 | 41 | 27.2 | 25 | 19.6 |
| 13 | 14 | 13.7 | 27 | 21.2 | 15 | 14.3 | 15 | 14.3 | 26 | 20.6 | 30 | 22.7 | 25 | 20.1 |
| 14 | 14 | 14.0 | 26 | 21.2 | 14 | 14.0 | 14 | 14.0 | 23 | 19.5 | 28 | 22.2 | 25 | 20.6 |
| 15 | | | 24 | 20.5 | | | | | 23 | 19.9 | 25 | 21.1 | 24 | 20.5 |
| 16 | | | 23 | 20.4 | | | | | 22 | 19.8 | 23 | 20.4 | 24 | 21.0 |
| 17 | | | 22 | 20.2 | | | | | 22 | 20.2 | 23 | 20.8 | 23 | 20.8 |
| 18 | | | 22 | 20.6 | | | | | 19 | 18.7 | 18 | 18.0 | 22 | 20.6 |
| 19 | | | 22 | 21.0 | | | | | 19 | 19.0 | | | 21 | 20.3 |
| 20 | | | 21 | 20.7 | | | | | | | | | 21 | 20.7 |
| 21 | | | | | | | | | | | | | 21 | 21.0 |
| ***h*** | **14** | | **20** | | **14** | | **14** | | **19** | | **18** | | **21** | |
| ***M*** | **13.7** | | **20.4** | | **14.9** | | **12.4** | | **20.2** | | **31.7** | | **19.6** | |

*Table 1 Hirsch-type indexes for selected bibliometricians who have been awarded the Price Medal (with α = 2)*

Glänzel, W., Schubert, A., Telcs, A. (1984), *Goodness of Fit Test for the Tail of Distributions*. Bolyai Colloquium on Goodness of fit (Debrecen, Hungary, June 25-28, 1984).

Glänzel, W. (2006), On the h-index – A mathematical approach to a new measure of publication activity and citation impact. *Scientometrics*, 67 (2) 315–321.

Glänzel, W., Schubert, A. (1988), *Theoretical and Empirical Studies of the Tail of Scientometric Distributions*. In: L. Egghe, R. Rousseau (Eds.), Informetrics 87/88, Elsevier Science Publisher, 75-83.

Gumbel, E. J. (1958). *Statistics of extremes*. New York: Columbia University Press.

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. (also available at: arXiv:physics/0508025, accessible via http://arxiv.org/abs/physics/0508025).

Jin, B.H., Liang, L.M., Rousseau., R., Egghe, L. (2007), The R- and AR-indices: Complementing the hindex. *Chinese Science Bulletin*, 52 (6) 855-863.

Johnson, N. L., Kotz, S., Balakrishnan, N. (1994), *Continuous univariate distributions*. Volume 1, 2nd Edition, John Wiley & Sons, Ney York.

Schubert, A. Glänzel, W. (2007), A systematic analysis of Hirsch-type indices for journals. *Journal of Informetrics*, 1 (3), in press. (doi:10.1016/j.joi.2006.12.002)

Pao, M. L. (1986), An empirical examination of Lotka's law. *Journal of the American Society for Information Science*, 37 (1) 26–33.

Vlachý, J. (1976), Time factor in Lotka's law. *Probleme de Informare si Documentare*, 10 (2) 44–87.

Yablonski, A. I. (1980), On fundamental regularities of the distribution of scientific productivity. *Scientometrics*, 2 (1) 3-34.

ISSI Newsletter is published by ISSI (http://www.issi-society.info/).
Contributors to the newsletter should contact the editorial board **by email**.
Wolfgang Glänzel: wolfgang.glanzel@econ.kuleuven.be
Ronald Rousseau: ronald.rousseau@khbo.be
Liwen Vaughan: lvaughan@uwo.ca
Aparna Basu: basu.aparna@rediffmail.com
Balázs Schlemmer: balazs.schlemmer@econ.kuleuven.be
Accepted contributions are **moderated** by the board. **Guidelines** for contributors can be found at http://www.issi-society.info/editorial.html
Opinions expressed by contributors to the Newsletter do not necessarily reflect the official position of ISSI. Although all published material is expected to conform to ethical standards, **no responsibility** is assumed by ISSI and the Editorial Board for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions or ideas contained in the material therein.