issi newsletter #12
international society for scientometrics and informetrics
quarterly e-zine of the international society for scientometrics and informetrics
vol. 03 / nr. 4
december 2007

# EDITORIAL

## ■ The ISSI e-Newsletter in 2007 – in a nutshell

Another year has passed: time to revise what we have done this year.

First of all, right at the beginning of the year ISSI members elected a new president (Ronald Rousseau) and a few new board members. The other outstanding event, of which the newsletter was supposed to report officially, was the biennial ISSI conference held in Madrid, Spain.

As for the e-Newsletter itself, 25 publications (12 articles, 6 conference calls & reports, 3 interviews & short biographies and 4 editorials) from 12 countries were published in the four issues of 2007 (short communications not included).

By mere coincidence, the number of authors and other contributors, who took part in preparing the newsletter in 2007, is also 25.

This year's Hall of Fame, without whom none of the issues of the ISSI e-Newsletter could have been carried out, is the following (in alphabetical order): *Aparna Basu • Balázs Schlemmer • Bettina Berendt, Bihui Jin • Birger Larsen • Eric Zimmerman • Grant Lewison • Henry Small • Hildrun Kretschmer • Isabel Gómez • Isidro F. Aguillo • Judit Bar-Ilan • Katherine W. McCain • Lennart Björneborn • Marek Kosmulski • María Bordons • Mike Thelwall • Peter Ingwersen • Ping Zhou • Raf Guns • Rickard Danell • Ronald Rousseau • Sonia Vasconcelos • Stefanie Haustein* and *Wolfgang Glänzel.*

Thanks for all of you for your committed work! We hope that amongst many new contributors, we can welcome you back next year, too!

On behalf of the editorial board, I also wish a very successful and prosperous 2008 to all our readers!

Balázs Schlemmer
*technical editor*

# CONTENTS



# Editorial Board

# SCIENTIFIC COMMUNICATION OF THE FUTURE

## WISSKOM CONFERENCE AT RESEARCH CENTER JÜLICH, 6-8 NOVEMBER, 2007


Stefanie Haustein
Heinrich-Heine-University Düsseldorf
haustein.stefanie@web.de


It was for the fourth time that the Central Library of the Research Center Jülich in Germany invited to its biennial conference addressing current scientific issues concerning libraries, scientific publication, and science in general. Themed "Scientific Communication of the Future", this year's WissKom conference focused particularly on e-science, scientific indicators, primary data management, communication in research and teaching, and Web 2.0.

The conference was opened by Rafael Ball, head of the Research Center's Central Library, claiming that the future of scientific communication had already begun with the digitalization of information resources and the introduction of electronic communication tools into science. Scientific communication is affected by a tremendous change of communication infrastructure and modalities and has become a holistic approach including information supply, scientific content, processing and publication of scientific expertise, and implementation of long-term availability structures, and challenges information and communication scientists and technologists, as well as librarians and scientists of various disciplines (BALL, 2007).

### ■ E-science

The first day of the conference was devoted to e-science as a vision of a new form of scientific networking. Katrin Weller (Heinrich-Heine-University Duesseldorf) identified "different levels of networking approaches for a future e-science scenario" (WELLER et al., 2007) and introduced the interesting idea of combining those to get a technology that handles scientific problems with networked computing power, obtain data, and run comprehensive and generally accessible archives, where primary data and publications are stored and collaboratively maintained. Interest groups would have the opportunity to comment and debate, existing ontologies and semantic technologies could cater for information integrity. Before this ideal case of e-science can be established, however, a large number of technical and organizational problems have to be solved and, above all, the scientific community has to be prepared to toe the line (WELLER, 2007). Anne-Katharina Weilenmann (Swiss National Library Bern) outlined the development from cyberscience to e-science and described the challenges of information overloads, a subject that overshadowed the whole conference. With DILIGENT (Digital LIbrary Infrastructure on Grid Enabled Technology) and the Parzival-project, she introduced two projects that examined the sustainability and suitability for daily use of e-science (WEILENMANN, 2007). The e-science session was concluded with Thomas Lippert (Jülich) report about Supercomputing at the Research Center in Jülich.

### ■ Communication in Research and Teaching

Elena Semenova and Martin Stricker of Humboldt University Berlin, the Swiss historians Peter Haber and Jan Hodel and Sonja Hierl of HTW Chur contributed to a session referring to communication in research and teaching. Still in its infancy, the Berlin project aims at the development of an ontology for academic disciplines. It shall provide a framework of terms, central layers, and relationships of and between scientific disciplines in order to advance interdisciplinary set-up, research, and data exchange in science communication (SEMENOVA & STRICKER, 2007). Haber and Hodel offered a glance at scientific communication of historians, who have only adapted slowly to the digital change in media environment and are now confronted with the current trends including Web 2.0. (HABER & HODEL, 2007). Sonja Hierl urged the need to provide collaborative competences and academic working skills from the beginning of academical education. New concepts like Web 2.0 and the goal oriented usage of digital libraries and library services have to be embedded in the teaching process. The problems that Haber and Hodel described could thus be avoided in the first place. Hierl reported from DIAMOND (Didactical Approach for Media Competence Development), a project that proved successful in preparing students to take part in modern scientific communication (BAUER et al., 2007).

### ■ Web 2.0

In a conference on scientific communication of the future a session on Web 2.0 cannot be missing. Three speakers reported from practical experience. Christian Haenger introduced the interesting idea of using social tagging in academic libraries as a means to supplement and optimize traditional intellectual indexing. His project wants to examine if non-categorized digital documents can be indexed best either by tags, automatically, or by a combination of both (HAENGER & KRAETZSCH, 2007). Web 2.0 is currently heavily discussed in development research organizations and primarily used in form of RSS feeds, blog software, and social bookmarking services (VON ITTER, 2007). Steffen Leich-Nienhaus delivered insight into scientific information supply at DaimlerChrysler and showed the effects of the infrastructural peculiarities of communication within an organization (LEICH-NIENHAUS, 2007).

### ■ Primary Data Management

Michael Diepenbroek (Bremen), Jan Brase (Hanover), and Harald Krottmaier (Graz) met the challenge of primary data management in a session thus entitled. Krottmaier concentrated on implementing special audio-visual and text-based search into library infrastructure in order to find non-textual digital documents (KROTTMAIER, 2007). Diepenbroek and Brase referred to two projects that archive and register primary biology and geoscience data and make them available for a larger audience (DIEPENBROEK & GROBE, 2007; BRASE & KLUMP, 2007). In a following discussion skeptical opinions were voiced whether raw data should be stored for future generations of if it was better to discard them.

### ■ Science Indicators

Day three of WissKom was dedicated to science indicators and showed that new ones emerged almost daily but that a standard has not yet be found. The need for standards was emphasized by Wolfgang Glänzel (Leuven). He pointed out that bibliometric indicators have to meet criteria of scientific methods in order to be reliable and valid. He warned against the misuse of popular indicators like the Impact Factor or the h-index by people who lack the knowledge to interpret the resulting rankings properly. Glänzel provided an insight of the development of bibliometrics and addressed the well-known problem that arose when the application area of bibliometric indicators was extended to research evaluation and politics. Three "hot topics" of bibliometric research were identified including the h-index which still deserves study especially in its possible application on meso level. Glänzel certifies bibliometrics future viability if new approaches of research overcome known weaknesses, structural-based methods are implied, and research and service are connected consistently (GLÄNZEL & DEBACKERE, 2007). Patrick Vanouplines (Brussels) described the value added when available data are linked reasonably. He presented a study that aimed at helping researchers, who want to publish open access to find a journal with the highest impact

factor, by merging information from DOAJ, JSTAGE, and SciELO with data from Thomson Scientific's JCR. A simple table with the IF listed next to the OA journal title was obtained. Vanouplines encouraged the University Library of Lund maintaining the DOAJ and Thomson to collaborate to arrange such a service and called attention to the journal eigenfactor estimating the time library users spend with a specific journal, as an alternative to the IF (VANOUPLINES & BEULLENS, 2007). Dirk Tunger (Jülich) introduced three studies where bibliometrics was used as a part of trend observation (TUNGER, 2007), one of them being the EU-project SMART (Foresight Action for Knowledge-Based Multifunctional Materials Technology) which was taken up again by Show-Ling Lee-Mueller (Jülich). She described how bibliometric analyses were employed to complement expert interviews and discussions and data screening of review articles to create a European map of excellence in materials science (LEE-MUELLER & SCHUMACHER, 2007). The ten rules James Pringle (Philadelphia) offered as a "helpful guideline when planning or interpreting citation-based analysis" (PRINGLE, 2007) should already be known by any information professional working with Web of Knowledge. Pringle advised care in using citation data as a management tool and to never substitute the views of experts, which was also emphasized by Henning Moeller (Karlsruhe), who reported from five years of practical experience of measuring research results of the 15 German Helmholtz Centers, the Research Center Jülich being one of them. He described the problems that arise from the usage of twenty quantitative indicators to measure scientific achievements (MOELLER, 2007).

Apart from the presentations the conference also provided a poster session where research results and various issues of scientific communication were presented. Companies involved in scientific communication were given the opportunity to present their products as well. The conference closed with a panel discussion where experts were confronted with the audience's questions and ideas of future scientific communication. A successful supporting programme including a dinner at the Research Center's lake cafeteria and a tour of the site - including the laboratory of Jülich's this year's Nobel Prize in Physics laureate Peter Gruenberg - topped off the conference. WissKom 2007 took up many aspects of scientific communication and introduced helpful ideas of how to cope with today's and future challenges. Handling the tremendous amount of data remains apparently one of the main challenges of information society and bibliometrics is still in search of adequate indicators and generally accepted standards.

## References


BALL, R. (2007): WissKom 2007. Wissenschaftskommunikation der Zukunft. 4. Konferenz der Zentralbibliothek Forschungszentrum Jülich, 6. - 8. November 2007, Beiträge und Poster.

BAUER, L., BOELLER, N., HERGET, J., HIERL, S. (2007): Konzepte zur Foerderung der Wissenschaftskommunikation. Der Churer Ansatz zur Vermittlung von kollaborativen Kompetenzen, in: Ball, R. (Editor): WissKom 2007, 81-92.

BRASE, J., KLUMP, J. (2007): Zitierfähige Datensätze. Primärdaten-Management durch DOIs, in: Ball, R. (Editor): WissKom 2007, 159-168.

DIEPENBROEK, M., GROBE, H. (2007): PANGEA® als vernetztes Verlags- und Bibliothekssystem für wissenschaftliche Daten, in: Ball, R. (Editor): WissKom 2007, 147-158.

GLÄNZEL, W., DEBACKERE, K. (2007): Bibliometrie zwischen Forschung und Dienstleistung, in: Ball, R. (Editor): WissKom 2007, 209-222.

HABER, P., HODEL, J. (2007): Historische Fachkommunikation im Wandel. Analysen und Trends, in: Ball, R. (Editor): WissKom 2007, 71-80.

HAENGER, C., KRAETZSCH, C. (2007): Collaborative Tagging als neuer Service von Hochschulbibliotheken, in: Ball, R. (Editor): WissKom 2007, 123-134.

KROTTMAIER, H (2007): Die Systemarchitektur von PROBADO: Der allgemeine Zugriff auf Repositorien mit nicht-textuellen Inhalten, in: Ball, R. (Editor): WissKom 2007, 169-176.

LEE-MUELLER, S.-L., SCHUMACHER, G. (2007): Einsatz bibliometrischer Analysen im EU-Projekt zur Technologiefrüherkennung SMART, in: Ball, R. (Editor): WissKom 2007, 263-272.

LEICH-NIENHAUS, S. (2007): Wissenschaftliche Informationsversorgung am modernen digitalen Arbeitsplatz, in: Ball, R. (Editor): WissKom 2007, 107-122.

MOELLER, H. (2007): Messen, Steuern, Regeln – zum Controlling der Helmholtz-Forschung, in: Ball, R. (Editor): WissKom 2007, 273-272.
PRINGLE, J. (2007): The ISI Web of Knowledge as a Management Tool, in: Ball, R. (Editor): WissKom 2007, 253-262.
SEMENOVA, E., STRICKER, M. (2007): Eine Ontologie der Wissenschaftsdiziplinen. Entwicklung eines Instrumentariums für die Wissenschaftskommunikation, in: Ball, R. (Editor): WissKom 2007, 61-70.
TUNGER, D. (2007): Bibliometrie als Teil eines Trenderkennungs-Systems in der Naturwissenschaft, in: Ball, R. (Editor): WissKom 2007, 235-246.
VANOUPLINES, P., BEULLENS, R. (2007): Merging information sources to obtain the impact factor of open access journals, in: Ball, R. (Editor): WissKom 2007, 223-234.
VON ITTER, S. (2007): Wissenschaftskommunikation in der Entwicklungsforschung / Entwicklungszusammenarbeit. Web 2.0 und Communities of Practice – ein Beitrag aus der Praxis, in: Ball, R. (Editor): WissKom 2007, 95-106.
WEILENMANN, A. (2007): Von Cyberscience zu e-Science, in: Ball, R. (Editor): WissKom 2007, 25-32.
WELLER, K., MAINZ, D., MAINZ, I., PAULSEN, I.: Semantisches und vernetztes Wissensmanagement für Forschung und Wissenschaft, in: Ball, R. (Editor): WissKom 2007, 33-46.

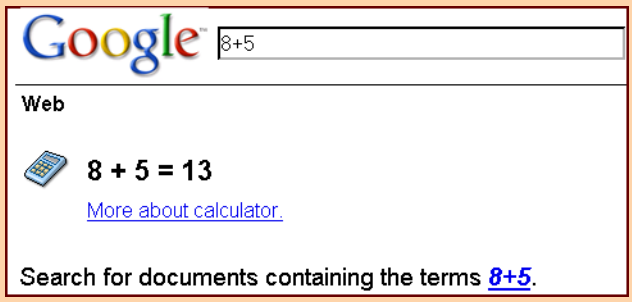
# INFORMATION (RETRIEVAL) + METRICS = INFORMETRICS

## (THE GOOGLE WAY)

Raf Guns
Informatie- en Bibliotheekwetenschap
University of Antwerp
raf.guns@ua.ac.be

*In August 2003, Google unveiled its calculator as a new addition to its popular search engine. The aim of this contribution is to introduce this potentially useful feature to the toolbox of webometricians and informetricians and point out some of its interesting and amusing characteristics.*

*Unlike some other Google services – such as Google Groups, Maps or Scholar –, the Google calculator does not have its own URL or interface. It is completely integrated in and accessed through the main search interface: any query which resembles a calculation is automatically evaluated by the calculator. The actual search engine results are still available behind a link (or sometimes shown below the calculator result):*



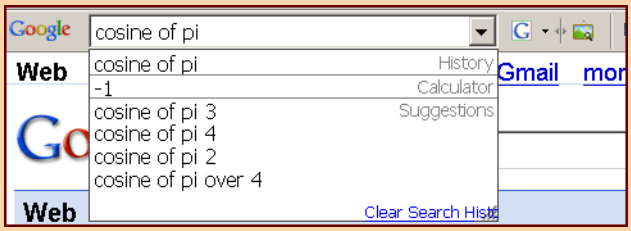

*The calculator can of course do more complex calculations as well, including conversions between many different units. Some examples:*

| Query | Result |
|---|---|
| eight + five | eight + five = thirteen |
| e^(pi*i)+1 | e^(pi*i)+1 = 0 |
| phi | the golden ratio = 1.61803399 |
| cube root of pi | cube root(pi) = 1.46459189 |
| 6 feet to meters | 6 feet to meters = 1.8288 meters |
| 2 gallons in table spoons | 2 US gallons = 512 US tablespoons |
| What is the speed of light in miles per hour? | the speed of light = 670 616 629 miles per hour |

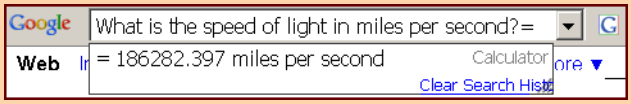
As we can see from these examples, the calculator supports both queries in 'conventional' mathematical notation and queries in (simple) natural language. Several of these examples are difficult or impossible to perform with a pocket calculator. It even recognizes calculations phrased as questions! The latter feature is, however, fairly limited: paraphrases of the same question were not picked up by the calculator.

Each result is presented as '*calculation = solution*', but note that not each calculation is a verbatim repetition of the original query: 'phi' becomes 'the golden ratio', 'table spoons' becomes 'US tablespoons' etc. Presumably, each query is first translated to a standardized notation and subsequently evaluated.

We also point out that the calculator nicely integrates with the suggestion feature in the Google Toolbar (http://toolbar.google.com) and the search field of the Firefox web browser: both immediately show the result in a drop-down box.




This suggestion feature only recognizes fairly straightforward calculations. Especially natural language calculations seem more difficult to discern from 'normal' web queries. To 'force' the calculator to process them, one can simply append an '=' sign to the end of the query.




Given its useful features and easy reachability (considering today's ubiquitous web access), the Google Calculator makes for a very convenient and fast pocket calculator. However, since its launch, avid web users have also been exploring the darker alleys of the calculator's advanced features and have come up with queries like the following:

| Query | Result |
|---|---|
| the answer to life the universe and everything | the answer to life the universe and everything = 42 |
| sqrt(9) +7^2 in roman numerals | sqrt(9) + (7^2) = LII |
| mach 1 in furlongs per fortnight | mach 1 = 2 046 124.55 furlongs per fortnight |
| number of horns on a unicorn acre in tea spoons per light year | number of horns on a unicorn acre = $7.7675003 \times 10^{24}$ US teaspoons per light year |

The first one is a playful reference to Douglas Adams' *The Hitchhiker's Guide to the Galaxy* and illustrates that Google engineers are not devoid of a sense of humour! The last ones simply border on the curious and absurd, and one may wonder whether they are actually intended by Google. Nevertheless, even the last example seems mathematically correct: 1 acre (number of horns on a unicorn = 1) is a measure of area. Tea spoons are a measure of volume ($m^3$), thus tea spoons per light year are a measure of area as well ($m^3/m = m^2$).

As a concluding note, it should not come as a big surprise that since 2003 other search engines – most notably Windows Live Search and Yahoo! Search – have also implemented a calculator. These seem generally less sophisticated than Google's, but Live Search does have one edge over Google: it can handle simple algebra. For instance, the query '5x + 2 = -8' yields the result '5x+2=-8 : x=-2'.

Unfortunately, neither these services nor Google's have extensive official documentation. To find out 'what works and what does not', one is therefore left to trial-and-error and unofficial sources like other websites. Some references:

http://www.google.com/help/calculator.html:
official documentation, fairly limited

http://www.googleguide.com/calculator.html:
unofficial overview

http://www.googleguide.com/help/calculator.html:
unofficial 'cheat sheet'

http://www.soople.com/soople_intcalchome.php:
provides another interface to the calculator

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.

# THE MISSING LINK BETWEEN GOOGLE SCHOLAR AND PLAGIARISM PREVENTION?

## HOW CITATION ANALYSIS CAN HELP STUDENTS LEARN ABOUT THE NATURE OF KNOWLEDGE


Bettina Berendt
Katholieke Universiteit Leuven, Belgium
bberendt@gmx.net
http://www.cs.kuleuven.be/~berendt


Two sources of chagrin to scientists are their own literature search and their students' papers. In this article, I argue that three often-posed and seemingly different questions related to these two sources are in fact closely related, and I propose a solution strategy to address all three issues: Why is literature search suboptimal (i.e., do we have to make do with Google/Google Scholar searches)? Why are our knowledge sources so poor (i.e., why do the data extracted by Google Scholar and similar services contain so many errors)? Why do students plagiarise (and what can we do about it)?

**Q: Why is literature search still unsatisfactory?**
**A: Because we are not utilizing the full power of scientometric analyses in publicly available tools.**

We can today find literature by using general-purpose search engines, open digital libraries like Google Scholar (http://scholar.google.com), Citeseer (http://citeseer.ist.psu.edu), Citebase (http://www.citebase.org), arXiv (http://www.arxiv.org), fee-based special-purpose services such as Web of Science (http://scientific.thomson.com/products/wos/) or Scopus (http://www.scopus.com), or social-Web tagging services like CiteULike (http://www.citeulike.org) or Bibsonomy (http://www.bibsonomy.org). The basic format for results is a list of documents matching the given keyword, possibly ranked by some (usually non-disclosed) relevance algorithm. Some tools allow the user to navigate from one article to another article that is related by citation (cites or cited-by), textual similarity, co-citation (co-cited with), equality of metadata or tags, etc.

However, available tools operating on live digital libraries rely on the user to read the papers or abstracts in order to answer higher-level questions such as: What topics and groups of papers are there? What are current research fronts? What are the semantic relations between two papers / groups of papers? This is unfortunate because a large number of classical bibliometric methods and visualizations exist for creating such top-down models of literature (see Chen, 2003, for a survey, and Janssens, Leta, Glanzel, & De Moor, 2006, for an extensive study of optimal text-citation information combinations). The challenge is how to make these available for the layperson's analysis of live digital libraries (*challenge #1*).

**Q: Why are our knowledge sources still unsatisfactory?**
**A: Because information extraction of citation metadata remains a hard problem.**

Neither the manual nor the automatic extraction of bibliographic metadata is error-free. An example of the first (different names in WoS for non-English-language institutions) was recently discussed in this Newsletter by Kosmulski (2007); examples of the second (e.g., author

names becoming part of the title and vice versa) are familiar to every user of services that use autonomous citation indexing, for example Google Scholar or CiteSeer. The major reason for the first are different kinds of human error; the major reason for the second is that the underlying computational problem ("bibliographic-metadata information extraction" – extracting from the free text of a reference in a PDF file a structured representation like [author, year, title, …]) has, in spite of its conceptual simplicity, remained a hard problem. A recent survey of different algorithmic approaches by Day et al. (2007) reports error rates averaged over bibliographic fields on a standard dataset between ca. 10 to ca. 30%. Again, non-standard settings like non-English language sources exacerbate the problem. For example, we found an increase in error rates (for relatively easy-to-extract fields only) by 30 percentage points when comparing performance on an English-language sample with a mixed German-English language sample (Berendt, Dingel, & Hanser, 2006). The challenge is to (i) improve these algorithms, (ii) better leverage human effort and (iii) enable lay users to understand and take into account the limitations of algorithms they cannot change, for example, by explaining language-related extraction problems (*challenge #2*).

**Q: Why are academic standards degrading?**

**A: Because the role of citing is ill-understood by students.**

Creatives today (and that includes students and scientists) are faced with a curious dilemma: the limitlessness of available and findable material and the increasing technological ease with which the age-old cultural techniques of copying, transforming, and publishing can be appropriated by everyone is paralleled by increasingly relentless industrial and public persecution of such activities as "violations of intellectual property rights". The use of IPR argumentation may make it impossible (or too costly) to publicly portray something as personal as one's life (in which a music-band poster adorns a teenage bedroom wall) (Lessig, 2001), and it may suppress public criticism (cf. the case of Scientology requesting Google to take Scientology-critical Web pages off the index on grounds of violations of the Digital Millenium Copyright Act, see McHugh 2003). These developments are opposed by many including artists and scientists.

A widespread view is: nearly all content that one may need is already there, copying is easy (and gives better results than recreating the content yourself), and doing it helps the artist become better known. Law and the police are out to help the latter solely to placate BigMoney Music Industry.

And isn't the writing of a term paper essentially the same? Nearly all content that one may need is already there, copying is easy (and hasn't the author said it much nicer than one ever could) and doesn't hurt the author (who wants to be read and wants to influence) or anyone else. University regulations and lecturers are out to detect and sanction "plagiarism" solely to harass students.

The reported evidence shows that such content re-use may be as popular as copying music: In questionnaires, up to 90% of students stated that they used other authors' text without reporting sources, and plagiarism-tested document samples showed high incidences of unmarked text re-use (e.g., Sattler, 2006; Rutgers University, 2003; Weinstein und Dobkin, 2002, see also the list of references at Howard 2007 and http://www.umuc.edu/distance/odell/cip/links_plagiarism.shtml#incidence).

The evidence also suggests that a number of students plagiarise out of helplessness or ignorance of citation norms, that hardly any student conceives of plagiarism as something that in fact harms *them*, and that most students do not employ the information inherent in citations for finding further sources etc. at all, or at most in a very unsystematic way.

In what sense does plagiarism harm the student? It deprives her of practising an essential scientific skill, that of synthesizing and reprocessing existing material in order to truly understand it, of understanding the fabric of science that leads to scientists being able to stand "on the shoulders of giants". Most importantly, it deprives the student of practising and thereby learning a skill that is the foundation of this fabric of science: to differentiate between an *uncritically referring attitude* and a *critically referring attitude*. The former relies on the everyday assumption that what we perceive equals reality;


Figure 1: A screenshot of the grouping/labelling phase of document search and retrieval.

the latter views reality as a text that we have to decipher, which can and must be done using (different) theories (Holzbrecher, 2001).

Luckily, the huge increase in the popularity of search engines with citation-based ranking (like Google) and Weblogs with their heave reliance on different forms of hyperlinks that often link to highly opinionated content now make it easier for teachers to explain, even to beginning students, the notion and importance of a critically referring attitude.

For the purposes of this article, I regard plagiarism as just one extreme form of a more general (and even more common) problem: difficulties and weaknesses in dealing with the re-use and re-processing of scientific publications and other sources.[1]

The challenge is how to go beyond discussing these issues in general, how to create active-learning scenarios that illustrate the advantages (for the student) of being part of a community that references meticulously and produces correct citation metadata (*challenge #3*).

The meta-challenge is that all these problems are interdependent: The quality of scientific writing (in the sense of how carefully and correctly references are included and formatted) influences the correctness of citation metadata. The quality of scientific writing and the quality of teaching/learning scientific method correlate. The quality of citation metadata influences the results and usefulness of search results. Only good search results, in turn, will motivate scientists and students to make an effort too towards creating high-quality scientific writing. Therefore, each challenge needs to be addressed, and addressing each challenge will help solve the others.

## A solution approach

A large number of academic publications and Web sites describe teaching strategies for preventing plagiarism (see for example the list at http://www.umuc.edu/distance/odell/cip/links_plagiarism.shtml#assignments). We have taken a complementary approach: the development of search engines as part of authoring tools that demonstrate the advantages of having good citation metadata and make them easier to produce. In (Berendt et al., 2006), we presented an interactive software tool that allows its users

[1] It must be pointed out that an exact definition of plagiarism is not straightforward and that therefore the classification of a particular piece of text as plagiarised must be done with great caution. The occurrence of self-plagiarism especially in the work of "mature" scientists presents another conceptual and didactical problem. For reasons of space, the present article cannot discuss these complex issues in detail, let alone further this debate. For a list of references and Web resources, see for example http://www.umuc.edu/distance/odell/cip/links_plagiarism.shtml#theory

to interactively search, group and label documents based on a system proposal derived from a citation-based clustering, and thus to create a conceptual model of the domain. Clustering can be done based on bibliographic coupling or co-citation, and in the more recent versions of the tool, text analysis is used to enhance the grouping's meaningfulness. The tool also supports publication of the results for discussion with peers. A screenshot of the grouping phase of search is shown in Fig. 1. First user studies have shown the tool to be rated as useful and usable, and as supporting literature search effectively.

This tool addresses *challenge #1*. (The current prototype works on the CiteSeer database.) It can also be used in teaching, where it is generally met with a lot of enthusiasm. However, experience shows that this needs to be accompanied by targeted tasks such as those shown in Fig. 2. We propose tool-instruction combinations like this one to address *challenge #3*.

Investigate the citation analyses in the CiteSeer archive:

- Group 1: Starting from a paper of your choice, compare the sets of papers that are "similar" by different criteria. What are the commonalities and differences between these two sets?
- Group 2: Starting from Kleinberg's and/or Brin & Page's papers *(discussed in the previous lecture)*, identify 3 current research topics concerning improvements over the basic algorithms
- Group 3: What feature of CiteSeer is related to the "miserable failure" phenomenon? Find at least one paper that describes algorithmic approaches to exploiting this. Explain commonalities and differences.
- Group 4: Find an archive that is similar to CiteSeer but covers Economics. Find 3 commonalities and 3 differences between the archives.
- (Using the grouping tool) Group 5: Identify different research topics within the field of „link analysis".

Figure 2: An example of exercises for practicing the use of citation-analysis tools

The use of such tools in teaching invites discussions of other aspects of the nature of knowledge, such as what "common knowledge" is, why using it does not require a reference, and how this relates to other notions of knowledge beyond the "property-rights view" that centres on individual and tradable ownership. The results of many clustering runs also produce groupings that invite discussions on a number of well-known problems caused by the current publication-and-citation-focused policies for evaluating scientists. These include the "encouragement of overly large research groups", "repetition", "small and insignificant studies", and the "publication of half-baked ideas", as well as the "creation of publishing pacts" and "clique building" (Parnas, 2007).

In our work starting with that described in Berendt et al. (2006), we have developed algorithms and interfaces for addressing *challenge #2*, in particular (ii): better leveraging human effort. However, while this can reduce error rates, like all other known approaches it does not bring accuracy to 100% – bibliographic-metadata information extraction remains an open problem.

## ■ Conclusion

In this article, I have argued that problems of literature search and plagiarism are in fact related via their common roots in poor citation metadata, insufficient tool support, and insufficient understanding of the role and importance of citing. I have identified three challenges and proposed that a better understanding of (autonomous) citation indexing is essential not only for computer scientists working in that area, but for every scientist. I have described a tool for improving the interaction with such digital libraries and outlined how it can also be used in education.

As an outlook, I return to *challenge #2*: how to (i) improve the algorithms that extract citation and other bibliographic metadata from full texts, (ii) better leverage human effort and (iii) enable lay users to understand and take into account the limitations of existing algorithms.

I propose that *challenge #2* can only be fully addressed if the community relies on, and continues to develop, open-source software (like that of CiteSeer). This will enable an inspection of the code for possible sources of error or bias, easy extension (for example, by lexica of journal names in languages other than English), and extensive testing for gauging quality – and quality increments – of the extraction algorithm on which all further scientometric analysis rests. This in turn will substantially help education (*challenge #3*) and spur the development of more and better tools (*challenge #1*).

## References


Berendt, B., & Dingel, K., & Hanser, Ch. (2006). Intelligent bibliography creation and markup for authors: A step towards interoperable Digital Libraries. In *Proceedings of the 10th European Conference on Research and Advanced Technology for Digital Libraries* (pp. 495-499). Berlin etc.: Springer LNCS 4172.

Chen, C. (2003). *Mapping Scientific Frontiers: The Quest for Knowledge Visualization.* London: Springer.

M.-Y. Day, M.-Y., Tzong-Han Tsai, R., Sung, C.-L., Hsieh, C.-C., Lee, C.-W., Wu, S.H., Wu, K.-P., Ong, C.S., & Hsu, W.-L. (2007). Reference metadata extraction using a hierarchical knowledge representation framework. *Decision Support Systems, 43,*152–167.

Janssens, F., Leta, J., Glanzel, W., & De Moor, B. (2006). Towards mapping library and information science. *Information Processing and Management, 42,* 1614-1642.

Holzbrecher, A. (2001). *Wissenschaftliches Schreiben (Tipps & Tricks).* http://www.ph-freiburg.de/fakultaet-1/ew1/holzbrecher/wissenschaftliches-schreiben.html [14 Dec 2007]

Howard, R.M. (2007). *Plagiarism: Some sources on causes and incidence.* http://wrt-howard.syr.edu/Bibs/PlagIncidence.htm [14 Dec 2007]

Kosmulski, M. (2007). Lack of consequence in English translations of university names ruins their scientific reputation. *ISSI Newsletter, 03/September 2007,* 46-48.

Lessig, L. (2001). *The Future of Ideas: The Fate of the Commons in a Connected World.* New York: Random House.

McHugh, J. (January 2003). Google vs. Evil. *Wired Magazine 11(01).* http://www.wired.com/wired/archive/11.01/google_pr.html [14 Dec 2007]

Parnas, D.L. (2007). Stop the Numbers Game. Counting papers slows the rate of scientific progress. *Communications of the ACM, November 2007/Vol. 50, No. 11,* 19-21.

Rutgers University (2003). New Study Confirms Internet Plagiarism Is Prevalent. http://ur.rutgers.edu/medrel/viewArticle.html?ArticleID=3408 [14 Dec 2007]

Sattler, S. (2006). *Plagiate in Hausarbeiten. Empirische Prüfung direkter und indirekter Rational Choice Modelle anhand einer Leipziger Studierendenbefragung.* Magisterarbeit, Februar 2006. Institut für Soziologie, Universität Leipzig, Germany. Summary available at: http://www.uni-leipzig.de/~sozio/content/site/wiss_arb/414.pdf [14 Dec 2007]

University of Maryland University College Center for Intellectual Property (undated). *Plagiarism.* http://www.umuc.edu/distance/odell/cip/links_plagiarism.shtml [14 Dec 2007]

Weinstein, J. and Dobkin, C. (2002). *Plagiarism in U.S. Higher Education: Estimating Internet Plagiarism Rates and Testing a Means of Deterrence.* http://webdisk.berkeley.edu/~Weinstein/Weinstein-JobMarketPaper.PDF [3.8.2005]

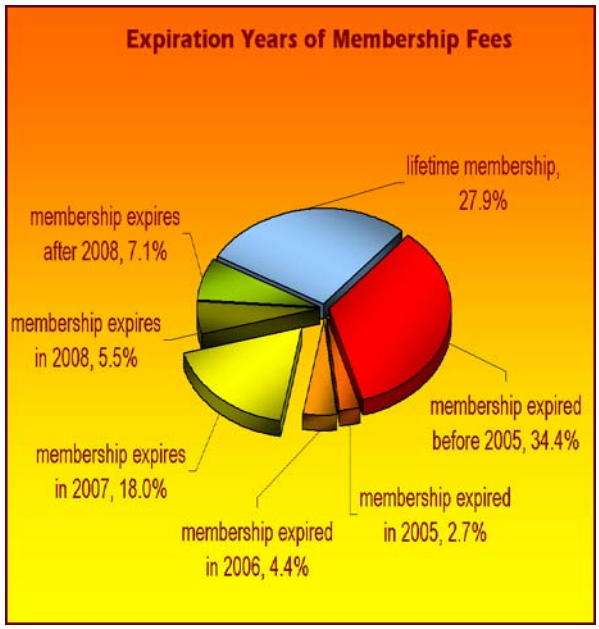

# REMINDER

**Dear Society Members,**

Those members, whose membership will expire this calendar year, are kindly requested to update their membership in time. This is also necessary to guarantee continuous access to the Newsletter. You can find the expiration date on your white or green membership card. Nonetheless, if you have any question concerning your membership status, please, contact me by email. For detailed information consult http://www.issi-society.info/membership.html. Owners of a gold membership card are, of course, subscribed for lifetime.

With very best wishes for 2008 I remain yours sincerely,


Wolfgang Glänzel
Secretary-Treasurer ISSI
Wolfgang.Glanzel@econ.kuleuven.ac.be

# THE MEASUREMENT OF SCIENCE AND TECHNOLOGY IN CHINA

Ping Zhou[a,b]

[a] Katholieke Universiteit Leuven, Steunpunt O&O Indicatoren, Leuven, Belgium
[b] Institute of Scientific and Technical Information of China, Beijing, PR China
ping.zhou@econ.kuleuven.be

**Abstract**

This paper introduced the background about the measurement of science and technology in China and selectively introduced the most recent statistic results released by the Institute of Scientific and Technical Information of China.

## 1. Introduction

At the annual news conference about the statistic results of China's scientific and technical publications held on November 15, 2007, the Institute of Scientific and Technical Information of China (ISTIC) released its new statistic results on China's publications in science and technology (S&T): China has become the second largest country in terms of S&T publications in 2006 (ISTIC, 2007), which was a big jump from the fourth position one year ago in 2005 (ISTIC, 2006).

As an affiliated organization of the Ministry of Science and Technology (MOST), ISTIC started its statistic work in 1987 with the support of the MOST by establishing a database entitled China Scientific and Technical Papers and Citations Database (CSTPCD). The construction of CSTPCD is similar to that of the SCI. Journals are selected based on the standard set by ISTIC (Wu et al., 2004). At the beginning the CSTPCD covered 1,189 journals, but this number was increased to 1723 in 2006 which is around 38% of the total 4,497 S&T journals in 2003 (Ren, 2005). Since its establishment, the CSTPCD has been widely used by research institutions, scientific management organizations, and individual scientists for measuring research output (Wu et al., 2004).

In addition to establishing the CSTPCD, ISTIC conducts statistic analysis based on the CSTPCD for domestic publications and on international databases for China's international publications. The international databases included are both the Science Citation Index (SCI) on CD-ROM and the Science Citation Index Expanded (SCIE), Engineering Index (EI), and Index to Scientific & Technical Proceedings (ISTP). Both the SCIE and the ISTP are products of the Institute of Scientific Information (Thomson Scientific, Pliladelphia, PA, USA), while the EI is a product of Elsevier. Since 2005, ISTIC has expanded its scope by covering the MEDLINE which is compiled by the U.S. National Library of Medicine (NLM) and published

on the Web by Community of Science. MEDLINE is the world's most comprehensive source of life sciences and biomedical bibliographic information. It covers over 4800 journals from over more than 70 countries. In addition, ISTIC does some general statistic analysis for Chinese patent based on the United States Patent and Trade Office (USPTO), Japanese Patent Office (JPO), and European Patent Office (EPO).

Statistic results are officially released annually in the form of a news conference; in 2007 already the 15th official release took place. The release of ISTIC's results is a great event for Chinese science community, and attracts attention from Chinese media as well. Major results published at the conference include: Chinese international publications and citations, domestic publications and citations, disciplinary distributions, regional distributions, publication distributions of important Chinese institutions, international collaboration, high impact publications, and statistic results about Chinese journals as well. In the following I will summarize some results (ISTIC, 2007) that might be of interest to the international community.

## 2. An overview on China's publications in 2006[1]

### 2.1 Publications in international databases

Based on data from the SCIE, EI and ISTP, China ranked second after the USA in terms of international journal and conference publications in 2006, while China still held the fourth position in 2005. Figure 1 shows the evolution of Chinese publications in international databases in the past ten years.

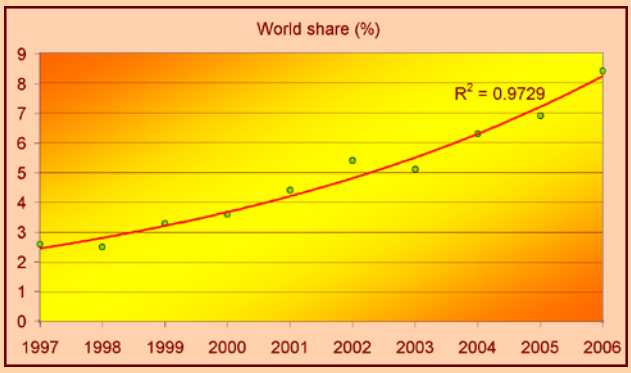



Figure 1. Evolution of Chinese international publications.

[1] ISTIC used SCIE to calculate publication counts. Publications include all types of documents. However, for citation counts the SCI CD-ROM edition was used.

In order to obtain a more detailed picture of Chinese publication output in international journals and is citation impact, I will break down publication counts by data sources.

(i) BASED ON THE SCIE: Chinese publications took 5.9% share in the SCIE, which keeps China standing at the fifth position for the third year. The former four countries were the USA, the UK, Germany and Japan. Among Chinese publications in the SCIE, 59.3% had received at least one citation each in the period 1997-2006, and 225 Chinese publications were "highly cited" in the sense that they had received more than 100 citations each. Impact of Chinese publications in terms of citations does not assort quite well with the amount of publications, in particular, China ranked only 13th in terms of citation counts in the past ten years (1997-2006).

(ii) BASED ON THE EI: Chinese publications took 14.6% in the EI, making China the world's 2nd largest country in terms of publications. The first rank is hold by the USA.

(iii) BASED ON THE ISTP: In addition to publishing in academic journals, Chinese researchers become more active in attending international conferences. In 2006, Chinese researchers had attended 2,139 international conferences. Chinese publication share in the ISTP was 9.0% in 2006, making China the second largest country in terms of ISTP publications. The first five countries are the USA, China, Japan, Germany and the UK (see also Glänzel et al., 2006).

(iv) BASED ON THE MEDLINE: There were 87 Chinese journals covered by the MEDLINE in 2006. The number of Chinese publications increased to 31,118, which is 13.3% growth compared to the corresponding number (i.e., 27,460) in 2005.

### 2.2 Publications in the domestic database (CSTPCD)

Data for domestic publications are based on the CSTPCD which covered 1,723 Chinese journals in 2006. There were 404,858 publications with Chinese researchers as the first authors. Compared to 2005, China's domestic first-author publications had increased by 14.0%. There were 2,970 publications which were not first-authored by Chinese researchers.

In addition, ISTIC started to collect data of domestic publications in cross-disciplines and social sciences in 2005. Based on 381 such kind of journals, Chinese researchers published 96,348 papers in 2006.

## 3. China's leading disciplines

### 3.1 Relative disciplinary strength in 2006

#### *3.1.1 In the international community*

Disciplines that contributed the most to China's international publications or that received the most citations among China's international publications are listed in Table 1.

Table 1. Top 10 disciplines in terms of international publications or citations

| Ranks by Publication Counts | | Rank by Citation Counts | |
|---|---|---|---|
| Ranks | Disciplines | Disciplines | Ranks |
| 1 | Chemistry | Chemistry | 1 |
| 2 | Computer Science and Technology | Physics | 2 |
| 3 | Physics | Biology | 3 |
| 4 | Electrics, Communication and Automatic Control Technology | Materials Science | 4 |
| 5 | Materials Science | Mathematics | 5 |
| 6 | Biology | Basic Medicine | 6 |
| 7 | Mathematics | Mechanics | 7 |
| 8 | Dynamic and Electronic Engineering | Clinical Medicine | 8 |
| 9 | Geoscience | Geoscience | 9 |
| 10 | Civil Architecture | Dynamic and Electronic Engineering | 10 |

Note: The citation counts in 2006 were based on the citations to Chinese publications according to the 2001-2005 volumes of the CD-ROM edition of the SCI.

#### *3.1.2 In the domestic community*

China's domestic publication activity is somewhat different from that in the international community. Table 2 presents the list of the top ten disciplines in which China was most active or received most citations in the domestic community.

Table 2. Top 10 disciplines in terms of domestic publications or citations.

| Ranks by Publication Counts | | Rank by Citation Counts | |
|---|---|---|---|
| Ranks | Disciplines | Disciplines | Ranks |
| 1 | Clinical Medicine | Clinical Medicine | 1 |
| 2 | Electrics, Communication and Automatic Control Technology | Geoscience | 2 |
| 3 | Computer Science and Technology | Agriculture | 3 |
| 4 | Agriculture | Biology | 4 |
| 5 | Basic Medicine | Basic Medicine | 5 |
| 6 | Biology | Chemistry | 6 |
| 7 | Chemistry | Electrics, Communication and Automatic Control Technology | 7 |
| 8 | Pharmacy | Dynamic and Electronic Engineering | 8 |
| 9 | Chemical Engineering | Chemical Engineering | 9 |
| 10 | Metallurgy and metallography | Computer Science and Technology | 10 |

Note: The citation counts in 2006 were based on the citations to domestic publications included in the CSTPCD in the period 1988-2006.

China's most active fields in terms of international conference presentations were *computer science and technology, electronics, communications and automatic control, physics, civil engineering, kinetic and electrical science, materials science, geology, chemistry, basic medicine, information science*, and *systems science*.

### 3.2 China's leading fields in the past ten years (January 1997–August 2007)

In the past ten years (i.e., 1997-2006), China was more active in materials science and chemistry in terms of publications. Either field took over 10% world shares within its own field. Above all in materials science, China had the second highest world share in terms of publication, while the USA took the highest. China's citation impact in this field was strong as well: China ranked fourth in terms of citation counts received in the past ten years.

In terms of publication counts, China had entered world top ten in *materials science, chemistry, mathematics, computer science, engineering technology, physics* and *multi-disciplines*.

In terms of citation counts, China was among the top ten countries in *materials science, mathematics, chemistry, engineering technology, multi-disciplines, physics, computer science*, and *geology*.

To conclude, China's continuous progress in science and technology is impressive indeed (Zhou & Leydesdorff, 2006; Kostoff, 2007). However, it is worth mentioning that such progress is mainly reflected by its gross output especially the total number of publications. Citation impact of Chinese publications does not yet keep pace with the dynamic growth of its publication output. Chinese publications attract still less citations than expected on the basis of the world standard. According to ISTIC's statistics for Chinese publications in relevant fields during the period of January 1997 to August 2007, there is no single discipline in China whose mean citation rate (c/p) was above world average. China's deviation from the world standard is the least in mathematics and the largest in molecular biology and genetics. In other words, China's negative deviation from the world average in mathematics in this period actually amounts to -0.74 (=1.97-2.71), while that in molecular biology and genetics is -15.22(=8.90-24.12).

**Acknowledgement**

I am grateful for Prof. Wolfgang Glänzel from the Steunpunt O&O Statistieken, Katholieke Universiteit Leuven, for his comments and revisions. I would like also thank Mr. Zheng Ma from the statistic team of the ISTIC for providing some detail information relevant to their statistic results.

**References**

GLÄNZEL, W., SCHLEMMER, B., SCHUBERT A., THIJS B. (2006), Proceedings literature as additional data source for bibliometric analysis, *Scientometrics*, 68 (3), 457–473.

ISTIC (2006), 2005 年度中国科技论文统计结果 (Statistics of Chinese publications in 2005). Beijing: Institute of Scientific and Technical Information of China.

ISTIC (2007), 2006 年度中国科技论文统计结果 (Statistics of Chinese publications in 2006). Beijing: Institute of Scientific and Technical Information of China.

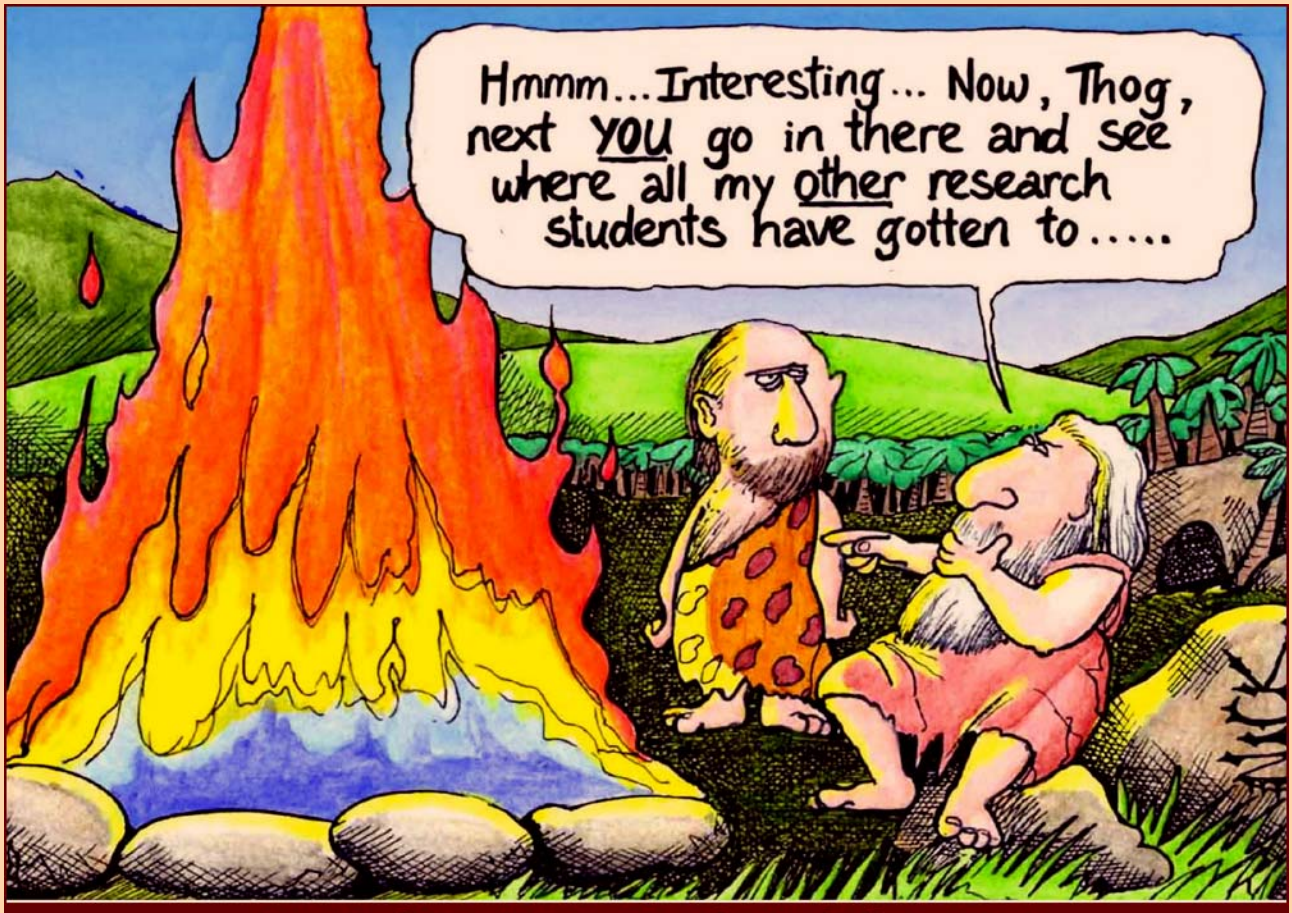
Kostoff, N. R., Briggs, M. B., Rushenberg, R. L., Bowgles, C. A. , Icenhour, A. S.,. Nikodym, K. F, Barth, R. B., Pecht M. (2007), Chinese science and technology % Structure and infrastructure. *Technological Forecasting & Social Change*. 74. 1539-1573.

Ren, S.L. 2005. Editing scientific journals in Mainland China, *European Science Editing*, 31 (1), 8–9.

Wu, Y.S., Pan, Y.T., Zhang, Y.H., Ma, Z., Pang, J.A., Guo, H., Xu, B., & Yang Z.Q. (2004). China scientific and technical papers and citations (CSTPC): history, impact and outlook. *Scientometrics*, 60 (3), 385–397.

Zhou, P. & L. Leydesdorff. 2006. The emergence of China as a leading nation in science, *Research Policy*, 35 (1), 2006, 83–104.

## CARTOON




**Proto-Professor Algarth Zag, pioneer in fire research.**