A Closer Look at the Sources of Informetric Research

Judit Bar-Ilan

barilaj@mail.biu.ac.il
Department of Information Science, Bar-Ilan University, Ramat Gan, 52900 (Israel)

Abstract
Currently existing data sources for informetric research are far from being perfect. Being aware of the limitations and a closer inspection of the data we work with can improve the validity and interpretation of our findings. In this paper I discuss current limitations of several data sources, emphasize the ever-changing nature of these sources and recommend trying to understand the specific problems and limitations at the time the study is conducted instead of relying on previous studies regarding possible limitations.

Introduction
We are living in the "information age": incredible amounts of information are available to us through the Internet. The Web has existed for twenty years only, yet the large majority of the data sources for informetric research are available through the Web. ISI's Web of Science (now a Thomson Reuters company) was launched in 1997 (Thomson, 2007), before that ISI data were only available through commercial providers (e.g. Dialog and STN), on tapes or CDs (from 1989 and onwards), or in the "ancient times" in print. In November 2004 two additional major citation databases appeared on the Web: Elsevier's Scopus (2004) and Google Scholar (Acharaya, 2004).

Not only the citation databases are online, but all major scientific journals appear now in electronic format beside the traditional printed version. There are already well-established journals that appear in electronic format only. This trend has begun in the late 1990's (Elsevier, 2009), and by now the publishers have digitized many volumes that originally appeared in print only. And of course, one cannot ignore the astronomical amounts of "digitally-born" data on the Web, which also include valuable information for informetric research in general and specifically for webometrics. Thus electronic access to data has become the norm. The computing power and the storage capabilities have also increased by several orders of magnitude over the last two decades, and there are easily accessible and often open-source software tools that enable to collect and analyze large quantities of data even on a personal computer. It has become easy to conduct "desktop or poor-man's bibiliometrics" (Moed, 2009). The data for informetric research have never been perfect, but now that informetric analysis can be conducted with much greater ease than before, it is even more important to understand the limitations and problems of data sources and methods and to assess the validity of the results. In the following sections I discuss some limitations of the existing sources. Often there are no easy solutions to overcome the problems, but by being aware of their existence one can provide better interpretations of the research findings.

The citation databases
The citation databases are major sources of informetric research. Each has specific limitations, and because the indexing and retrieval policies of the databases change from time to time, and the changes are not necessarily retroactive, these changes may cause internal inconsistencies in the databases. In addition, when using multiple databases, either for more

Proceedings of ISSI 2009, pages 1–9.
Edited by B. Larsen and J. Leta.

comprehensive data collection or for comparison, one must be aware of the differences in the applied algorithms and policies. In the following I give a few examples of these problems.

The Web of Science and the Journal Citation Reports

A lot has been written on the ISI Citation Indexes and the way ISI computes the journal impact factor. In this paper we will discuss the web-based product, the Web of Science (WOS). Coverage is one of the main reasons for criticising the ISI Citation Databases: poor coverage of non-English publications, insufficient coverage of the social sciences and poor coverage of the arts and humanities (see Moed, 2005, chapter 7 for an extensive discussion of coverage by discipline).

An additional issue related to coverage is the date the database started to cover the publication or the discipline. In the third quarter of 2008, ISI integrated the Proceedings Indexes into WOS, but proceedings are indexed only from 1990 and onwards, whereas journals in the Science Citation Index are indexed from 1900 and in the Social Science Citation index from 1956 and onwards. Thus the coverage of publications of active researchers is non-uniform – their works before 1990 are only covered if they appeared in journals. This of course is given and cannot be changed, but this must be taken into account, especially in areas where proceedings are an important publication venue, for example in computer science (Bar-Ilan 2006 and 2009).

As an example, let us consider the ISSI conferences. Proceedings of four conferences are indexed: 1999 (Colima), 2001 (Sydney), 2005 (Stockhom) and 2007(Madrid). The conferences that took place before 1999 and the 2003 conference in Beijing are not indexed. There is no uniform name for the conference series, three of them can be found when looking for ISSI in the publication name, but the 1999 proceedings is not, but a search for "scientometrics and informetrics" when searching in the publication list page works (see Figure 1).

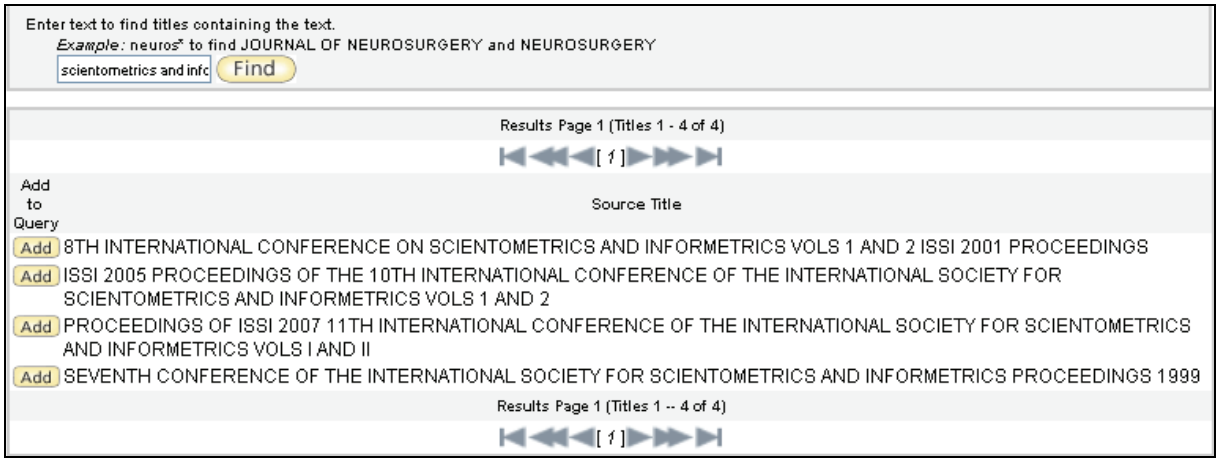

Figure 1. The titles of the different ISSI conferences as indexed by WOS.

Interesting to note that at the time I prepared the paper on the inclusion of the Conference Proceedings Indexes (Bar-Ilan, 2009), when searching for an author all items that were either published by the author or edited by him/her were retrieved. This was not clear on the search form, but was explained in the help files, as retrieved in January 2009 (see Figure 2). This indeed was a somewhat illogical feature, and by April 2009 one can search for author and editor separately (see Figure 3). This change in policy is an excellent example of the point I am trying to make in this paper: one must not rely on what has been said in the past about

database features, but one must check in depth the situation as it is at the time of data collection. Note that at the time of writing, when searching from "All Databases" and not from the "Web of Science" page, "author" still means author and/or editor.

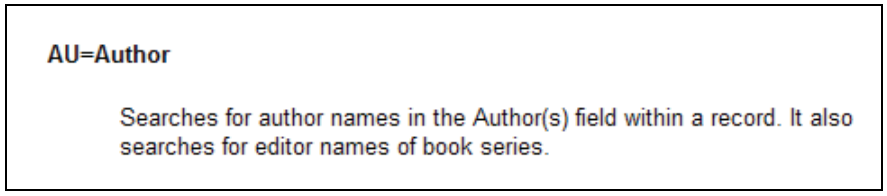

Figure 2. Excerpt from the WOS help file as of January 13, 2009.

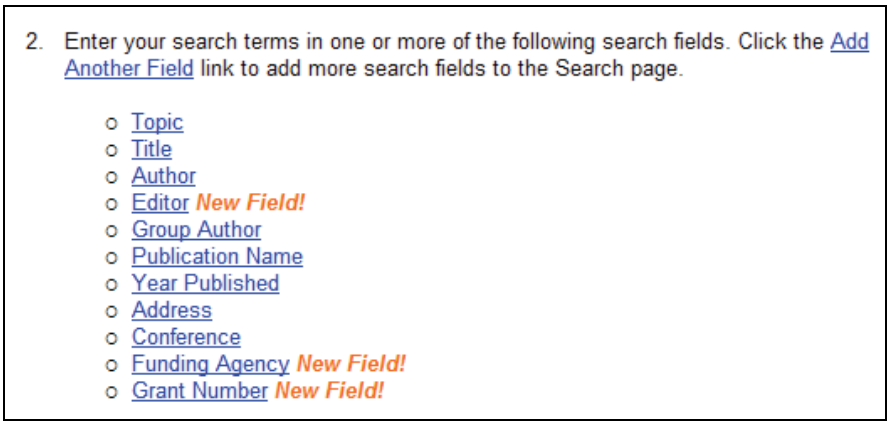

Figure 3. Excerpt from the WOS help file as of April 13, 2009.

Changes over time are an important issue, because usually changes are not retroactive. A few examples for the ISI databases are: inclusion of abstracts since 1991; first author vs. all authors. Note that even now, for non-source items only the first author is indexed. As an example, consider Egghe and Rousseau's "Introduction to Informetrics" (a non-source item) with 298 listed references when searching for cited references of Leo Egghe, but not a single citation is attributed to Ronald Rousseau.

Another recurring issue is the way ISI computes the impact factor. One problem is with ISI's definition of citable documents: when counting the number of publications only citable documents are taken into account, while for citations, citations to "non-citable" items are also counted (Moed & van Leeuwen, 1995). The question is how are "citable" documents defined? Moed and van Leeuwen found strong evidence that in 1995, "citable documents" meant articles, notes and reviews, although they were unable to find an "official definition". There is no clear definition as of now either, but David Pendlebury (2008) from the Thomson Research Services Group writes: "Although all primary research articles and reviews (whether published in front-matter or anywhere else in the journal) are included, a citable item also includes substantive pieces published in the journal that are, bibliographically and bibliometrically, part of the scholarly contribution of the journal to the literature. Research at Thomson has shown that, across all journals, more than 98% of the citations in the numerator of the Impact Factor are to items considered "citable" and counted in the denominator". Thus it is to be assumed that the term "citable" is journal dependent, which is quite reasonable, but it would be nice to know how the decision reached on of what is citable and what is not. Pendelbury's explanations are a reaction is to an article by Rossner, Van Epps and Hill (2007) complaining about the lack of integrity and transparency in the way ISI computes journal impact factors.

Scopus and SCImago

The major complaint against Scopus is that it has systematic coverage, including citation data from 1996 and onwards only. Over time this problem will become less and less serious, because informetric research usually studies recent activities. Scopus (like WOS) has been working hard on author identification, the results are getting better over time, but there is still more work to do. As an example consider an author search on Cronin Blaise in Scopus. The results are displayed in Figure 4. Clearly all six identities are the same and should be grouped together, and the most recent affiliations are wrong. If we extend the search to Cronin B, we get 70 results, which include several additional publications of "our" Blaise Cronin, under Cronin, B (3 publications) and under Cronin Biaise from Indiana University (3 publications).

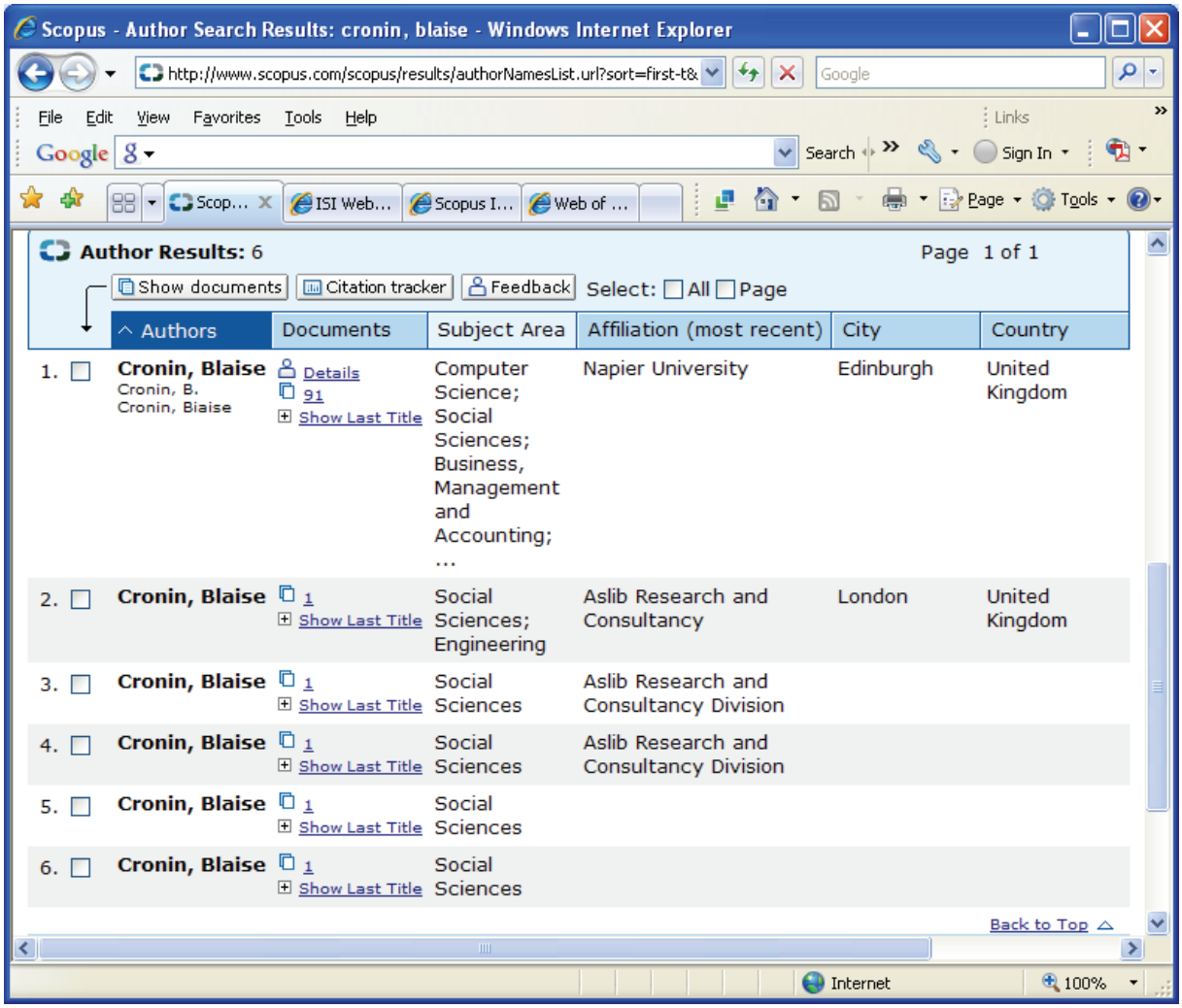

Figure 4. An example of author identification on Scopus as of April 13, 2009.

Another interesting issue is journal categorization. This is especially important when we are interested in journal rankings within categories (e.g. JCR categories). ISI's 2007 JCR category for Information and Library Science contains 56 journals. Scopus defines only 30 subject areas, but by downloading the complete journal list from the Scopus site (http://info.scopus.com/detail/documents/title_list.xls) one can see more refined classifications into multiple categories. Lists of journals in the smaller subject categories can also be retrieved from SCImago (2007). The SCImago portal aims to provide journal and country specific indicators derived from Scopus data. It turns out that the Scopus and SCImago categories for Library and Information Sciences are far from being identical. The

SCImago list contains 92 journals and the Scopus list contains 133 journals. A possible reason could be that the Scopus list represents the journals indexed by Scopus as of March 2009 and the SCImago site presents data as of 2007 (similarly to the current JCR which provides citation data from 2007), but the differences are considerable, so there must be additional reasons for the differences. For example the Journal of Documentation and ARIST are missing from the SCImago list. When comparing the Scopus list with the JCR list, all 56 journals are indexed by Scopus, but seven of them do not belong to the Library and Information Sciences category: three are classified as business/information systems journals, two as health journals and two as communication journals. The Scopus list includes 84 additional journals that do not appear in the JCR list, for example Cybermetrics and D-Lib Magazine, but rather interestingly the Journal of Informetrics is not among the journals in this category; it is primarily classified under Decision Sciences.

The differences between the JCR and the SCImago lists are further emphasized when we consider the ranked lists. We rank the SCImago list according to cites per document (2 years), which is supposed to be the equivalent of the impact factor (see SCImago, 2009). Now JASIST is ranked fifth as opposed to 13th on the JCR list, but the question is fifth or thirteenth out of what? Journals rankings are often used as proxies for journal quality by decision makers and we have to make sure that they are aware of the meaning of such rankings. Of course citation counts are also dependent on the citation database, as an example, JCR reports a 2007 impact factor of 1.436 for JASIST and 1.472 for Scientometrics (ranked 12th) as opposed to 1.77 for JASIST and 1.76 for Scientometrics (ranked 7th). The issue of the sources of data used for computing the h-index was discussed in (Bar-Ilan, 2008).

Google Scholar

A lot has been written on Google Scholar, some write rather negatively about it and emphasize its weaknesses (e.g. Jacsó, 2008a & 2008b), while others praise it (e.g. Harzing & Wal, 2008 & 2009) and some emphasize the great amounts of time needed in cleansing the data (e.g. Meho & Yang, 2007, Bar-Ilan, 2006). Harzing developed "Publish or Perish" (http://www.harzing.com/pop.htm), a very useful tool for retrieving data from Google Scholar.

The major weaknesses of Google Scholar besides the need for extensive data cleansing are that:

1) It is not clear whether Google is committed to continue to maintain and develop Google Scholar – it is still in beta four and a half years after is was launched.
2) It does not disclose its data sources, and there is no clear list of journals and proceedings that are covered.

On the positive side:

1) It is free and quite heavily used by students and academics.
2) Google Scholar is less sensitive to typing/spelling errors than WOS or Scopus and manages to group together some misspelled items.
3) In my experience it indexes new material relatively fast.
4) It also covers areas not well-covered by WOS or Scopus (Walters, 2007).

As an example of a problematic retrieval that could be probably easily corrected by Google, consider the query 'journal citation "impact factor"' (impact factor as a phrase), limited to the year 2000. Google Scholar (GS) reports 4,820 results for this query, whereas WOS retrieves only 21 results and Scopus 25 results for the same query. Is GS's coverage so much greater? Let us take a closer look at the first result page of Google Scholar (see Figure 5):

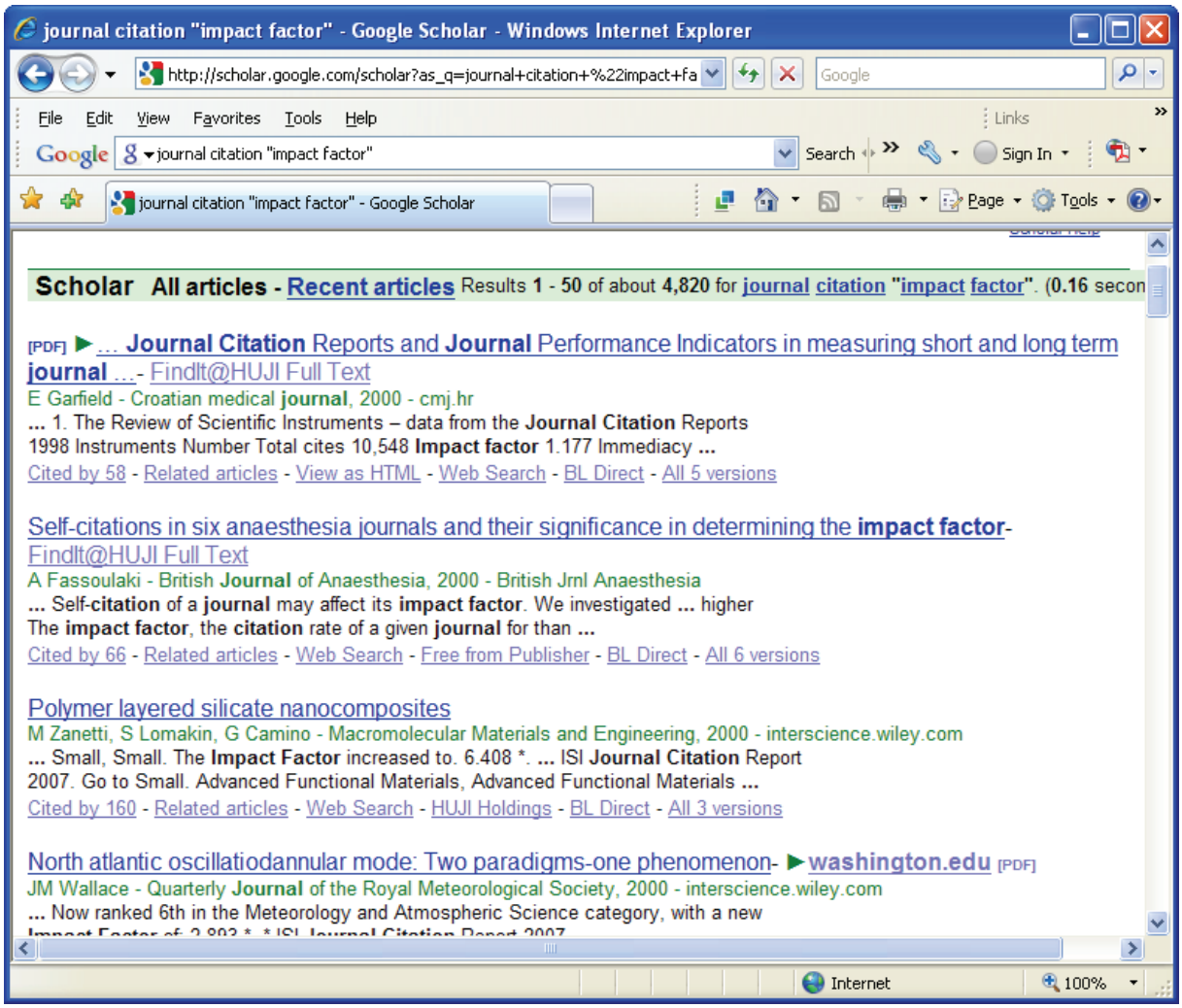

Figure 5. Top results of 'journal citation "impact factor"' from GS on April 13, 2009.

The third result looks very interesting, it has been cited 160 times – it must be highly relevant to the topic! It is not easy to see how 'polymer layered silicate nanocomposites' are related to impact factors, but it is worth to try. After clicking on the result and searching for "impact factor", the mystery is solved: the search terms appear on the bottom right of the page on the side bar (see Figure 6) of every journal on the Wiley Interscience Platform announcing the impact factor of some of the Wiley journals in the current year, and has nothing to do with the specific article or the year it was published.

In a previous paper (Bar-Ilan, 2008) I mentioned that Almind & Ingwersen's highly cited paper on 'Informetric analyses on the World Wide Web', was incorrectly attributed to D. Copenhagen. This problem has been corrected since, showing again that the data sources are dynamic and are changing over time.

A major limitation of Google Scholar for informetric data collection is that it does not retrieve more than 1,000 results even if it reports to have found say 4,820 results like in the above case. In informetrics research we often use large datasets, thus if we want to consider using Google Scholar as a data source, this problem has to be solved.

Figure 6. The sidebar of the page displaying the abstract of 'Polymer layered silicate nanocomposites'

Web Search Engines and the Web

In the above section I looked at the three major citation databases and showed that their coverage, features and capabilities change over time. The Web is much more dynamic than bibliographic records. Overall the Web continues to grow, but at the same time Web pages undergo content and design changes, are moved to a different host or directory and disappear from the Web. In an eight year study we (Bar-Ilan & Peritz, 2009) followed the growth, disappearance and changes that occurred to web pages containing the term 'informetrics'. Thus data collection for webometric studies is probably even complex than collecting data for other types of informetric research. Mike Thelwall, in his book on link analysis (2004) explains in details the precautions one must take (data cleansing, validations, understanding and reporting limitations) when conducting webometric studies.

When using search engines for data collection, one must take even greater care. The major search engines are not geared towards webometric research, their aim is to provide "search experience" for the general user. They will add/remove features to please the general user and not the informetrician. Users want fast answers and are usually not interested in the comprehensiveness of the results.

In 2005, I presented a wish-list (Bar-Ilan, 2005) for the "ideal" search engine for webometric research. Browsing that paper, it is easy to spot changes that occurred since. A few examples: the paper mentions the then newcomer search engine Exalead. Although it still exists it has not become one of the major search engines. That paper mentions MSN beta, a search engine that has become Live Search (http://www.live.com/). Simple link queries on Yahoo (e.g. link:http://www.issi2009.org) now automatically transfer the user to Yahoo's site explorer with rather interesting features. MSN used to have some ranking options; these are not existent on Live Search. On the other hand, currently Google allows some personalization of search results for logged in users. Personalization is not widespread, but Google has local search interfaces in a large number of countries, and it presents 'localized' results, results that in Google's opinion better fit the local users' expectations.

In 2005 there was no mentioning of Web 2.0 applications, and their possible use for webometric research. By 2009, there are a number of works in this direction (e.g. Thelwall, 2007 & 2008; Angus, Thelwall & Stuart, 2008), and probably we will see more in the future. Some serious shortcomings of Web search engines for informetric research still exist, for example the maximum number of results retrieved for a query remains 1,000 (for Google and Yahoo) and Google still does not enable to combine link searches with other search terms.

Conclusions

In this paper I tried to demonstrate some limitations and shortcomings of frequently used informetric data sources. The data and the data collection tools change all the time, and the examples in this paper might not be valid in the future. The examples are not important, the major point is that when conducting an informetric study, we should thoroughly check whether the data collection process works as planned and whether the collected data are valid for the purposes of the research.

References

Acharaya, A. (2004). *Scholarly Pursuits*. Retrieved April 12, 2009, from http://googleblog.blogspot.com/2004/10/scholarly-pursuits.html

Angus, E., Thelwall, M., & Stuart D. (2008). General patterns of tag usage among university groups in Flickr*, Online Information Review*, 32(1), 89-101.

Bar-Ilan, J. (2005). Expectations versus reality – Search engine features needed for Web research at mid 2005. *Cybermetrics*, 9, issue 1, paper 2. Retrieved April 12, 2009, from http://www.cindoc.csic.es/cybermetrics/articles/v9i1p2.html

Bar-Ilan, J. (2006). An ego-centric citation analysis of the works of Michael O. Rabin based on multiple citation indexes. *Information Processing and Management*, 42, 1553-1566.

Bar-Ilan, J. (2008). Which h-index? - A comparison of WoS, Scopus and Google Scholar. *Scientometrics*, 74(2), 257-271

Bar-Ilan, J. (2009). Web of Science with the Conference Proceedings Citation Indexes – The case of computer science. In B. Larsen and J. Leta (Eds.) *Proceedings of the 12th International Conference on Scientometrics and Informetrics.*

Bar-Ilan, J. & Peritz, B. C. (2009). The lifespan of "informetrics" on the Web: An eight year study (1998-2006). *Scientometrics*, 79(1), 7-25.

Elsevier (2009). *E-journals at Elsevier*. Retrieved April 12, 2009, from http://www.elsevier.com/wps/find/authored_newsitem.cws_home/companynews05_00021

Harzing, A. W. K. & Wal, R. van der (2008). Google Scholar as a new source for citation analysis. *Ethics in Science and Environmental Politics*, 6, 61-73.

Harzing, A.W.K. & Wal, R. van der (2009). A Google Scholar h-Index for journals: An alternative metric to measure journal impact in economics & business? *Journal of the American Society for Information Science and Technology,* 60(1), 41-46.

Jacsó, P. (2008a). Google Scholar revisited. *Online Information Review*, 32(1), 102-114.

Jacsó, P. (2008b). The pros and cons of computing the h-index using Google Scholar. *Online Information Review*, 32(3), 437-452.

Meho, L. I. & Yang, K. (2007), Impact of data sources on citation counts and rankings of LIS faculty: Web of Science versus Scopus and Google Scholar. *Journal of the American Society for Information Science and Technology,* 58(13), 2105-2125.

Moed, H. F. (2005). *Citation Analysis in Research Evaluation*. Dortrecht, Netherlands: Springer.

Moed, H. F. (2009). Presentation at the SSH Bibliometric Database Stakeholder Workshop, Brighton, March 18, 2009. Retrieved April 12, 2009, from http://www.sussex.ac.uk/Units/spru/esf/stakeholder.php

Moed, H.F. & van Leeuwen, Th. N. (1995). Improving the accuracy of Institute for Scientific Information's Journal Impact Factors. *Journal of the Society for Information Science*, 46(6), 461-467.

Pendelbury, D. (2008). *Thomson Scientific Corrects Inaccuracies in Editorial*. Retrieved April 12, 2009, from http://forums.thomsonscientific.com/t5/blogs/blogprintpage/blog-id/citation/article-id/4

Rossner, M., Van Epps, H. & Hill, E. (2007). Show me the data! *Journal of Cell Biology*, 179 (6), 1091-1092,

SCImago (2007). *SJR — SCImago Journal & Country Rank*. Retrieved April 13, 2009, from http://www.scimagojr.com

SCImago (2009). *SJR – Help*. Retrieved April 13, 2009, from http://www.scimagojr.com/help.php

Scopus (2005). *Elsevier is Proud to Announce Today's Launch of the World's Largest Scientific Abstract Database*. Retrieved April 12, 2009, from http://www.info.scopus.com/news/press/pr_041103.asp

Thelwall, M. (2004). *Link analysis: An Information Science Approach*. Amsterdam: Elsevier.

Thelwall, M. (2007). Blog searching: The first general-purpose source of retrospective public opinion in the social sciences? *Online Information Review,* 31(3), 277-289.

Thelwall, M. (2008). Social networks, gender and friending: An analysis of MySpace member profiles, *Journal of the American Society for Information Science and Technology*, 59(8), 1321-1330.

Thomson Scientific (2007). *Company Timeline*. Retrieved April 12, 2009, from http://web.archive.org/web/20071012002902rn_2/scientific.thomson.com/isi/timeline/

Walters, W. H. (2007). Google Scholar's coverage of a multidisciplinary field. *Information Processing and Management*, 43, 1121-1132.