Data sorces for performing citation analysis

an overview

Report

Author(s):
Neuhaus, Christoph; Daniel, Hans-Dieter

Publication date:
2006

Permanent link:
https://doi.org/10.3929/ethz-a-005281766

Data sources for performing citation analysis: An overview

Christoph Neuhaus & Hans-Dieter Daniel

Accepted for publication in the Journal of Documentation,
revised version June 30, 2006

Christoph Neuhaus
ETH Zurich
Professorship for Social Psychology and Research on Higher Education
Zaehringerstrasse 24
CH-8092 Zurich
Switzerland

E-mail: neuhaus@gess.ethz.ch

Prof. Dr. Hans-Dieter Daniel
(1) ETH Zurich, Professor for Social Psychology and Research on Higher Education
(2) University of Zurich, Evaluation Office
Muehlegasse 12
CH-8001 Zurich
Switzerland

E-mail: daniel@evaluation.unizh.ch

Abstract

Purpose: To provide an overview of new citation-enhanced databases and to identify issues to be considered when they are used as data source for performing citation analysis.

Design/methodology/approach: Reports the limitations of Thomson Scientific's citation indexes and reviews the characteristics of the citation-enhanced databases Chemical Abstracts, Google Scholar and Scopus.

Findings: Suggests that citation-enhanced databases need to be examined carefully, with regard to both their potentialities and their limitations for citation analysis.

Originality/value: Presents a valuable overview of new citation-enhanced databases in the context of research evaluation.

Keywords: citation analysis, citation index, citation-enhanced database, Chemical Abstracts, Google Scholar, Scopus

Category for the paper: General review

Introduction

The listing of references in publications is a convention among scientists for giving credit or recognition to the value of previous work (Merton, 1988). The application of citation analysis to research evaluation is founded on this tradition. It aims to estimate the varying contributions of scholarly work to the advancement of knowledge. Assuming that scientists cite the work that they have found useful in pursuing their own research, the number of citations received by a publication is seen as a quantitative measure of the resonance and impact that this publication has created in the scientific community.

Most commonly, the main resource for citation analysis are the citation indexes produced by *Thomson Scientific* (formerly *Institute for Scientific Information).* Besides their multidisciplinary nature, citation indexing was the major reason why this service had an unique position among bibliographic databases for many years. Thomson Scientific is, however, no longer the only service offering citation-enhanced databases on the market. In recent years, several database producers have noticed the potential of citation indexing and manually added cited references to a subset of their records. Among others, the discipline-oriented databases *Chemical Abstracts* produced by the *American Chemical Society*, *MathSciNet* by the *American Mathematical Society* and *PsycINFO* by the *American Psychological Association* have introduced citation indexing to their bibliographic databases. This change in practice took place, even though indexing of cited references is still a very laborious and expensive task. In addition to citation indexing of traditional bibliographic databases, new abstracting and indexing services have also emerged. With electronic availability of scholarly documents it became possible to automate data collection from very large resources at relatively low cost. Several bibliographic databases were established which automatically extract bibliographic information and cited references from electronic documents retrieved from digital archives and repositories. Some of these databases offer sophisticated features for citation searching and provide detailed information on download frequencies, which may serve as an additional basis for assessing the resonance and impact of publications. Some remarkable services are *CiteSeer*, which focuses primarily on literature in the fields of computer and information science, *RePEc*, which covers research papers in economics, and *SMEALSearch*, which indexes academic business documents. All these use *autonomous citation indexing* (Lawrence, Giles, & Bollacker, 1999), which results in a cost reduction for citation indexing. Beyond these discipline-oriented databases two multidisciplinary databases have attracted much attention: In 2004, the scientific publisher *Elsevier* launched its abstract and indexing database *Scopus*, which covers about 15,000 peer-reviewed journal titles, and *Google* introduced its free service *Google Scholar*.

In this paper, we provide an overview of citation-enhanced databases (cf. Jacsó, 2004). After describing the Thomson Scientific citation indexes and their outstanding position among bibliographic databases, we outline their limitations in respect to citation analysis, especially in the context of research evaluation. We subsequently introduce new citation-enhanced databases. Considering in more detail the databases Chemical Abstracts, Google Scholar, and Scopus, we review their potentialities and limitations, both as data source and as platform providing analytical tools for citation analysis. It goes, however, beyond the scope of this paper to describe specific features and limits of these citation-enhanced databases. We instead discuss implications for citation analysis in the context of research evaluation, as including more data sources does not necessarily lead to more valid assessment of research performance (Moed, 2005).

The Thomson Scientific citation indexes

The origins of citation analysis as a widespread assessment tool of research performance can be traced to the mid-1950s, when Garfield proposed the groundbreaking concept of citation indexing (Garfield, 1955). With the introduction of the *Science Citation Index* (SCI), the *Social Sciences Citation Index* (SSCI) and the *Arts & Humanities Citation Index* (A&HCI) by the Institute for Scientific Information (now Thomson Scientific), systematic analyses of the impact and influence of scholarly work as well as of trends in science became available. Basically Garfield's citation indexes serve both as a bibliographic and as a citation database. Complete bibliographic information as well as *all* cited references on *all* items published in journals covered are included in the citation indexes. Citation data is one of the main reasons why Garfield's citation indexes have an exceptional position among the bibliographic databases worldwide. Additionally, the multidisciplinary nature of the citation indexes, which provides unique possibilities to study multi- or interdisciplinary research activities, and the consideration of *all* contributing authors as well as *all* their institutional affiliations make them appropriate for performing citation analysis, particularly in the context of research evaluation (Moed, 2005, p. 113f).

The Thomson Scientific citation indexes take into consideration a core set of editorially selected internationally oriented journals. The *Science Citation Index Expanded* (SCIE) edition covers approximately 6,500 peer-reviewed journals, whereas the Social Sciences Citation Index comprises 1,900 journals cover-to-cover and 3,300 journals partially, while the Arts & Humanities Citation Index comprises 1,100 journals cover-to-cover and 7,000 journals partially; the citation indexes overlap in their coverage of the literature to some extent (approximately 310 journals are covered by SCIE and SSCI, 60 journals by SSCI and AHCI, 10 journals by SCIE and AHCI, and 30 journals by all three citation indexes). This selective approach is based on the empirical finding that a majority of influential papers are published in a minority of journals. Approximately 2,000 journals account for around 85% of published articles and 95% of cited articles included in the Science Citation Index (Garfield, 1996).

Limitations of the Thomson Scientific citation indexes

For all of their power and benefits, the Thomson Scientific citation indexes have some limitations that are of crucial relevance for citation analysis as an assessment tool of research performance. Among these constraints is the limited coverage of the citation indexes. As outlined above, Thomson Scientific processes only a selected set of journals for its citation indexes. While these accessed journals tend to be the highest impact peer-reviewed journals, they represent only a fraction of scientific work that is documented. Hence, coverage relates to the extent to which the citation indexes cover the written scholarly literature in a field (Moed, 2005). Two aspects must thereby be differentiated: (a) the importance of journals in a field's written communication system, and (b) the extent to which the citation indexes cover the journal literature in a field.

Thomson Scientific restricts their coverage of sources to the journal literature, with the exception of some highly-cited book series and conference proceedings. Thus, other types of scientific communication such as books and chapters in edited books, conference proceedings, technical reports and patents are not taken into consideration or only to some extent. This, however, means that bibliometric analysis based on Thomson Scientific's citation indexes is less applicable in those fields of science in which the internationally oriented scientific journal is not the main medium for communicating research findings. In mathematics, engineering, economics, and particularly in other social sciences and arts & humanities journals were found to be less central in the scholarly communication system than in other disciplines (Moed, 2005, p. 133). In mathematics scientists often refer to preprints, whereas in engineering and applied sciences, conference proceedings and technical reports play an important role as a primary information source. In economics, other social sciences and arts & humanities books play an important role in the scholarly communication system. Consequently, bibliometric indicators derived from data in the Thomson Scientific citation indexes will be problematic in those fields, as follows: (a) because only journal literature is covered, bibliometric indicators will be based on a small fraction of research output, excluding other types of scientific communication, such as books and chapters in edited books, conference proceedings, technical reports, etc., (b) citations from journals to other publication types are compiled in the Thomson Scientific citation indexes, but such citations can be incorporated only in small bibliometric studies because retrieval is very laborious and time-consuming, and (c) citations from non-journal documents are not processed and are forever excluded from bibliometric analysis (Hicks, 1999). In comparative citation rankings of individual scientists, for instance, focusing solely on the journal literature may lead to wrong conclusions. For the field of sociology, Cronin and Snyder (1997) finds evidence that there may be two distinct populations of highly cited authors, one which is highly cited in monographs and one which is highly cited in the journals. Similarly a study of philosophy, sociology and economics by Lindholm-Romantschuk (1996) shows only a small number of authors whose monographs as well as journal articles are highly cited. Excluding non-journal documents from citation analysis thus may underestimate or even overlook a scientist's individual contributions to knowledge. Hicks (1999) reviews the social science bibliometric literature and summarizes the findings, namely: “Books are very highly cited individually and collectively account for about 40% of citations. Citations to and from books are distributed differently from citations to and from journal articles. The centrality of books in the scholarly communication in the social sciences contrasts with their absence in literature databases, including the SSCI.” (Hicks, 1999, p. 201).

In other fields, the scholarly communication system has changed rapidly over the last decade, providing new avenues for publishing and disseminating research findings such as preprint and postprint servers, and Open Access journals. This movement toward electronic publishing has been commented upon repeatedly in the literature. According to Youngen (1998), who analysed citations to preprints in physics and astronomy, the importance of electronic preprints in the dissemination of primary research information is growing. Thomson Scientific faces these changes in the scholarly communication system by developing a *Web Citation Index*, the launch of which was announced in November 2005. Using technologies developed by NEC Laboratories America including autonomous citation indexing (Lawrence, Giles, & Bollacker, 1999), the multidisciplinary citation index gathers scholarly content from institutional and subject-based repositories and adds cited reference searching to electronic documents such as conference proceedings, technical reports, preprints, dissertations, and other grey literature.

With regard to the extent to which the citation indexes cover the journal literature in a field, the Thomson Scientific citation indexes have been confronted with the steady criticism of alleged journal coverage, both in terms of disciplinarity and nationality: “Emphasis is generally placed on the over-representation in the database of developed, English-speaking countries (notably the USA) and biomedically oriented research fields at the expense of, *inter alia*, Third World countries, nations using a non-Latin alphabet, technology-oriented research fields and mathematics” (Braun, Glänzel, & Schubert, 2000, p. 251). Braun and co-workers analysed the representativeness of the Science Citation Index’s coverage on the basis of science- and technology-related journals listed in the *Ulrich’s International Periodicals Directory*. In the large majority of cases under study, the Science Citation Index journal set proved to be fairly balanced as compared to the much broader journal set by Ulrich’s International Periodicals Directory. Contrary to general belief, no distorting bias in favour of medicine among disciplines and the USA among countries could be observed. Moed (2005) analysed adequacy of coverage on the basis of cited references and draws a similar conclusion: “[...] ISI coverage of the *journal* literature is in most main fields *excellent* to *very good*, except for those parts of social sciences as sociology, education, political sciences and anthropology, and particularly for humanities” (Moed, 2005, p. 135).

To summarize, when the Thomson Scientific citation indexes are used as an assessment tool of research performance, a function the databases have not primarily been designed for, the selective coverage of the journal literature in a field can pose methodological problems, and the importance of internationally oriented journals in the written communication system in a field becomes crucial.

Another limitation concerns the problem of delimitating fields. In discipline-oriented databases such as Chemical Abstracts, Medline, or Physics Brief documents are assigned to fields and subfields on the basis of a hierarchically structured subject classification scheme. Experts attribute classification codes and index terms, respectively, to each paper in addition to author keywords. In the Thomson Scientific citation indexes, however, publications are not classified through a paper-based subject assignment. To measure and compare national output in fields, journals as a whole are clustered into subject categories, i.e. each paper is attributed to the field to which the journal belongs. This method, however, fails for papers in multidisciplinary journals such as Science or Nature, which are not attributed to any specific field at all (Glänzel, Schubert & Czerwon, 1999).

Emergence of new citation-enhanced databases

The time in which Thomson Scientific was the only service offering citation indexing is gone. Recently, the scientific publisher Elsevier launched its multidisciplinary abstract and indexing database Scopus, which covers approximately 15,000 peer-reviewed journal titles, thus providing a broad coverage of scientific, technical, medical and social sciences literature. Moreover, discipline-oriented databases, such as Chemical Abstracts, MathSciNet and PsycINFO, have noticed the potential of citation searching and have started to enhance a subset of their records with cited references. Although much smaller than the multidisciplinary databases compiled by Thomson Scientific and Elsevier, these discipline-oriented databases may provide broader coverage of the written communication system for specific fields, particularly with regard to non-journal documents. Finally, with electronic availability of scholarly documents through services such as arXiv and Cogprints, several bibliographic databases were established which automatically index scholarly documents from a wide variety of resources. Through technologies, such as autonomous citation indexing developed by NEC Laboratories America, manual information extraction is replaced by parsing algorithms that automatically extract bibliographic information and cited references from electronic documents retrieved from digital archives and repositories, resulting in a reduction of cost for citation indexing.

Electronic collections of academic and professional literature such as *IngentaConnect*, *MetaPress*, or *Highwire Press* and publisher archives such as *BioMed Central*, *ScienceDirect*, and *Wiley Interscience* comprise another category of resources that list cited references for a given publication. Even electronic versions of journals such as *Nature*, *Science*, or *Applied Physics Letters* may contain references. Another resource is *Amazon*'s *Search Inside!* feature, which provides information on cited books and citing books. Although some of the latter services offer sophisticated features for citation searching, their usefulness for performing citation analysis in the context of research evaluation is limited. Electronic collections and publisher archives have a restricted domain defined by their own journals and publications, respectively. In Elsevier's ScienceDirect archive, one cannot find citations from non-Elsevier journals. Therefore, the overview of citation-enhanced services in table 1 is restricted to discipline-oriented and multidisciplinary bibliographic databases. The names of the databases are listed along with information on subject area and publication types covered.

****

take in Table I

****

Multidisciplinary databases are of particular interest because they provide broad subject coverage, thus providing unique possibilities to study multi- or interdisciplinary research activities and to discover hidden subject relationships. In the next section, we will review the multidisciplinary databases Scopus and Google Scholar as well as the discipline-oriented database Chemical Abstracts. The latter provides a remarkable coverage of chemistry *and* related subject areas including biology and life sciences and is the only combined journal and patent citation source. We will address characteristics essential for citation analysis such as coverage and options for browsing and searching. However, it exceeds the scope of this paper to describe all features of these databases in detail. There are several papers which have covered this (see Jacsó, 2005a, for a comparison of Web of Science, Scopus, and Google Scholar).

Chemical Abstracts

The online databases provided by *Chemical Abstracts Service* (CAS), which indexes scientific documents since 1907, represent the world's most important compendia of chemistry and related sciences such as biology and life sciences, engineering sciences, materials sciences, medical sciences, and physics. Chemical Abstracts' extensive coverage includes journals, books, conference proceedings, dissertations, technical reports, preprints and patents from 1907 to the present. CAS processes journal articles from nearly 9,500 scientific journals worldwide for its database; among them, 1,500 key journals are indexed cover-to-cover. Chemical Abstracts also includes over 21,600 records for journal articles dated before 1907. Furthermore, patents of chemical, biochemical, and chemical engineering interest are covered from more than 50 patent-issuing authorities around the world, including the European Patent Office (EPO), the Japanese Patent Office (JPO), the United States Patent & Trademark Office (USPTO), and the World Intellectual Property Organization (WIPO).

Cited references are included for journals, conference proceedings, and basic patents from the USPTO, EPO, WIPO, and German patent offices from 1997 to the present. Patent examiner citations from British and French basic patents are included as of the beginning of 2003. Cited references are available for displaying and linking through the client software *SciFinder* and *SciFinder Scholar*, respectively, as well as through the online host *STN International.* In SciFinder (Scholar) the feature "Get Related" identifies cited references or citing references, respectively, for a single document or for a set of documents. This feature is similar to the "Cited References" and "Times cited" link appearing on the full record for an individual document in *Web of Science*. Furthermore, SciFinder (Scholar) includes tools to analyse search results by author, publication year, journal title, document type, and institutional affiliation, to name a few (see table 2). In turn, STN International offers an unparalleled combination of features for browsing, displaying and searching cited references. Additionally, STN International provides powerful tools for statistical analysis, thus providing numerous possibilities for performing citation analysis. More detailed information on citation searching in Chemical Abstracts using STN International are available directly from CAS (2005) and documented in the literature (Ridley, 2001).

****
take in Table II
****

As a free service CAS provides *Science Spotlight* (available at http://www.cas.org/spotlight), which identifies the publications and patents most frequently cited in documents covered in the Chemical Abstracts. As a supplement, the publications and patent families most frequently requested by researchers using CAS search services are highlighted as well as the most intriguing documents from each quarter as selected by CAS scientists.

Google Scholar

Google Scholar, released in November 2004 in beta version, is a freely available service that crawls the content of scholarly documents from a wide variety of sources. Google Scholar covers journals, books, conference proceedings, dissertations, technical reports, preprints and postprints, and other scholarly documents from all areas of science. Documents are located from various academic publishers, preprint and postprint servers, bibliographic databases and from digital repositories of several universities, research organizations and government agencies. Some prominent collections include the Association for Computing Machinery, arXiv, BioMed Central, Blackwell, HighWire Press, IEEE, IngenaConnect, NASA Astrophysics Data System, PubMed, Nature Publishing Group, RePEc, Springer and Wiley Interscience. These sources are, however, not all indexed entirely and some major publishers, including Elsevier and the American Chemical Society, have declined to cooperate with Google Scholar, thus significantly limiting its coverage of peer-reviewed journal literature. Generally, the extent to which the documents picked up by Google Scholar cover the written scholarly literature and especially the journal literature in a field is unknown, as Google does not disclose any information about the sources processed, nor the document types included, nor the time span covered. Evidence suggests, however, that at the moment Google Scholar's content is a modest subset of the content retrieved directly from publishers' archives, Thomson Scientific's citation indexes, Scopus and traditional bibliographic databases such as PsycINFO (Jacsó, 2005c). But Google Scholar has the potential to become more comprehensive than any single bibliographic database as it collects documents from a wide variety of sources.

The search interface of Google Scholar is simple and easy to use. Search options include some limiting criteria such as author, article title, journal title, publication year and subject area (see table 2). Results are returned in a relevance-ranked order, which relies primarily on the full text of each document and its citation count. Thus, results emphasise documents that are cited more often, creating a bias toward older literature. In this regard, some sort options would be helpful.

In practical terms, Jacsó (2005c) has explored the precision and recall performance of Google Scholar. He exposes significant shortcomings in the extent and the quality of the information retrieved. In particular, he uncovers unreliable search options, which lead to inaccurate and misleading results, duplicate records due to erroneous or incomplete bibliographic information, problems in automatically extracting bibliographic information from electronic documents such as author and publication year, and problems in matching cited and citing references. Consequently, citation counts should be treated with reservation.

Certainly, Google Scholar is an important service for those who do not have access to expensive multidisciplinary databases such as the Thomson Scientific citation indexes or Scopus. However, Google Scholar currently processes its sources in an unsystematic, unpredictable and fragmentary manner. For lack of adequate options for browsing, searching and saving results in structured output formats it is difficult to make even elementary bibliometric analyses efficiently. At least in its beta version, Google Scholar is not yet a useful choice for citation analysis, but it may develop into a sophisticated tool.

Scopus

In 2004, Elsevier released its ambitious Scopus abstract and indexing database covering over 15,000 peer-reviewed journal titles, including coverage of approximately 500 Open Access journals, 700 conference proceedings, and 125 book series. Altogether, Scopus indexes more journals than Thomson Scientific's citation indexes, and offers greater coverage of Open Access journals, but lacks the depth of coverage in years of journals, going back as far as 1966 selectively. As of the date of this paper, the majority of journal titles were found in the physical sciences (5,500 journal titles, including chemistry, engineering, mathematics, physics, etc.) followed by the health sciences (5,300 journal titles, including the entire *Medline*), the life sciences (3,400 journal titles) and the social sciences (2,850 journal titles, including arts & humanities).

As the Thomson Scientific citation indexes, Scopus considers all contributing authors as well as all of their institutional affiliations wherever applicable. Thanks to indices, it is even possible to assign the appropriate institutional affiliation to each author. Cited references are currently included from 1996 onward. Approximately 1,250 unique journal titles, however, are fed directly from Medline into Scopus and do not include cited references. Cited references are available for displaying, backward and forward linking and searching. Scopus provides plenty of searchable fields such as author, publication year, journal title and institutional affiliation as well as cited reference searching. Beyond the generic cited reference index, Scopus has separate indexes for cited author, cited year, cited title, cited source and cited pages. These are comprehensive options for citation searching, facilitating the performance of citation analysis using different approaches. Similar to Web of Science, Scopus offers the feature "Related Documents", which returns a list of documents that share cited references with the currently selected document. Through its refinement option, Scopus provides an overview of search results according to author, publication year, journal title, document type and subject area.

Released in January 2006 the *Scopus Citation Tracker* further enhances citation analysis by enabling citations to be viewed year on year, providing users a powerful way to explore citation data over time. The Scopus Citation Tracker tabulates citation data, showing how often the individual documents have been cited in individual years as well as in total. If necessary, the tabular representation can be exported as a text file.

Finally, Scopus includes an integrated Web search via *Scirus* that is similar to Google Scholar. The search engine provides access to scholarly documents from digital archives and repositories (e.g. arXiv, BioMed Central, Cogprints and RePEc) and to patents, including those from the EPO, JPO, USPTO, and WIPO. Results from the Web search are separate and do not include citation data.

Evaluation of citation-enhanced databases

When evaluating citation-enhanced services for bibliometric purposes, one must consider that bibliographic databases may contribute in two distinct ways to bibliometric analysis: (a) as a data source, and (b) as a platform providing the analytical tools for bibliometric analysis (Hood, & Wilson, 2003). Both contributions are beset with several methodological and technical difficulties, including limited coverage of the scholarly literature, inconsistent and inaccurate data, and limited facilities for browsing, searching and analysing data. Most of these difficulties arise because bibliographic databases are primarily designed for information retrieval and bibliometric analysis represents only a secondary use of the systems. In some cases, the only viable solution to overcome these problems is to download the data of interest and to perform offline data processing and analysis. In order to evaluate the usefulness of a given database for citation analysis in context of research evaluation, some characteristics of the database must be carefully considered, including:

Coverage: An understanding of the sources covered is central to the validity of any bibliometric analysis. As discussed above, coverage relates to the extent to which the sources processed for the database cover the written scholarly literature and in particular the journal literature in a field, since citation analysis in this context of research evaluation primarily focuses on papers published in peer-reviewed journals (cf. Daniel, 2005). Most notably, it must be ensured that coverage is not biased towards particular countries, languages or publishers (e.g. when comparing research performance of different nations). Moreover, the time period of a database may be limited, which makes it impossible to analyse the long-term impact of scientific work.

Important coverage issues to be considered include: Are the sources processed for the database known? Is there a known set of journals covered in the database? Does the database producer fully or partially cover the journals? Does the database contain peer-reviewed as well as non-peer-reviewed documents? Does the database also comprise Open Access journals? Which publication types (e.g. journal articles, books, conference proceedings, technical reports) and document types (e.g. research articles, letters, notes, reviews) are included in the database? How does coverage change over time?

Consistency and accuracy of data: Even in high-quality databases there are many instances of inconsistent and erroneous spellings of author names and a lack of journal title standardisation. In most bibliographic databases information about the institutional affiliation of the contributing authors is taken directly from the journals without any standardisation or is abbreviated in an inconsistent manner. Thus when gathering the publications to analyse, all variations of author names, journal titles and institutional affiliations, including linguistic variations, must be considered. The interpretation of citation data may lead to erroneous conclusions if such factors are not taken into consideration. Especially for individual scientists or research groups, the neglect of a single but prolifically cited publication may produce a large error. As pinpointed by Jacsó (2005c), serious problems arise in databases using autonomous citation indexing. Extracting bibliographic information such as author, publication year and institutional affiliation from electronic documents, detecting duplicate records and matching cited and citing references is still an error-prone task, although sophisticated algorithms were developed in recent years.

Data fields: Each database has a different set of fields, many of which are useful for citation analysis in the context of research evaluation. The basic unit of analysis is a collection of publications that must be selected in the database. Eventually, it depends on the research question addressed and on the approach chosen for data collection whether data fields such as institutional affiliation, document type, or subject area are absolutely necessary or dispensable for data selection. The author field is important for data selection to analyse the contributions of a single scientist to the advancement of knowledge. In some databases, however, not all contributing authors are included, thus complicating data collection (e.g. Chemical Abstracts lists up to ten author names only). Constructing indicators of national or institutional research performance is hardly suitable, when the institutional affiliation is only provided for the first author or the reprint author, respectively. Furthermore, problems may arise, when attribution of authors to their corresponding institutional affiliation is not possible. In order to analyse the impact of a given subject area, standardised information, such as classification codes, index terms, or keywords, may be helpful to select the publications to be analysed. Another important decision to be made in the process of data collection is the determination of which publication types (e.g. journal articles, books, conference proceedings) and document types (e.g. research articles, letters, notes, reviews) to include. Actually, bibliometric analysis is predominantly interested in the primary literature represented by journals. In doing so, only research articles, letters, notes and reviews are incorporated, excluding document types which do not generally constitute an original piece of research or a synthesis of work by others. It is also essential to know the year from which a particular data field is available, especially the year from which records have been enhanced with cited references. The Science Citation Index Expanded format available through the Web of Science, for instance, includes cited references from 1900 onward, while Scopus and Chemical Abstracts do so from 1996. Others include cited references only for a defined set of journals. Completeness of cited references is another crucial issue. In some implementations of PsycINFO, for example, only a fraction of the references could be included in the records due to technical restrictions (Jacsó, 2004).

Browsing options: Given inconsistent and inaccurate bibliographic information, browsing options are essential to look up variants, inconsistent or erroneous spellings, punctuations and abbreviations of author names and journal titles. Browsing is even more important when searching for cited authors and journal titles, as cited references show far more inconsistencies and errors than other data elements in bibliographic records. The process of reconciling individually cited references from different papers to the same target publication is error-prone, because the format of cited references varies widely across different fields and journals. Many authors use ad-hoc abbreviations for journal titles, confuse volume, issue and page numbers, misspell author names, or omit the middle initial (Jacsó, 2005b). In some cases database producers even aggravate the situation by adding their own inconsistencies and errors. Some implementations of databases offer a chance to look up variations by browsing data field-specific indexes. Chemical Abstracts on STN International, for instance, have a comprehensive set of browsable indexes for each data element in cited references as opposed to SciFinder (Scholar), which does not allow browsing the cited reference index at all.

Searching options: Most databases are available in different formats (e.g. through online hosts, CD-ROM, or web-based interfaces). Although these delivery mechanisms are based mostly on the same data, they provide significantly different features for browsing, searching and analysing data. Some implementations of databases offer sophisticated features for cited reference searching, providing separate fields for cited authors, cited publication year, cited journal title and so on. Others only offer a single cited reference index for searching, and still others do not make cited references separately searchable even though they appear as distinct parts in a record (Jacsó, 2005b).

Multifile capabilities are of great value for the incorporation of multiple bibliographic databases into a data collection. Particularly online hosts take advantage of searching multiple databases simultaneously. As databases always overlap to some degree, online hosts also provide commands to remove duplicate records from the search results.

Analytical tools: Like browsing and searching options, the availability of tools to perform statistical analysis also depends on the implementation of the database. Simply because one implementation offers good analysing features with a particular database, this does not necessarily hold true for another implementation of one and the same database. Accessing the Science Citation Index through Web of Science, for example, up to 2,000 search results can be ranked by a particular field, while on STN International the same feature is available for up to 50,000 search results. Thus, not only the data required must be considered, but also the features available for analysing the data. Online hosts such as DIALOG and STN International provide powerful functionalities for statistical analysis, including commands to determine the top authors and journals in a given search result, to extract terms from specific data fields and rank them in decreasing order or to cross-tabulate search results by author and publication year, among others.

Saving and exporting options: Some bibliometric studies require standardisation of the data (e.g. unification of institutional affiliation) before calculating bibliometric indicators. Others aim to visualise data in the form of bibliometric maps, in order to discover the cognitive landscape of a scientific field. In such cases, data must be downloaded for offline processing and analysis. Some databases offer different formats for saving bibliographic records and/or exporting bibliographic records to reference software such as *EndNote* or *Reference Manager*. In contrast, others such as *Google Scholar* do not provide any option for saving or exporting bibliographic information.

Contribution of new citation-enhanced databases to research evaluation

Citation analysis has proved to be an important assessment tool for research evaluation. Conducting citation analysis using the citation indexes produced by Thomson Scientific provides an obvious starting point in assessing research performance, but is bibliometrically restricted to a small fraction of the journal literature. The availability of citation data in other bibliographic databases opens up the possibility to extend the data source for performing citation analysis, particularly to include other publication types of written scholarly communication such as books, chapters in edited books and conference proceedings. The inclusion of other publication types will contribute to the validity of bibliometric analysis when evaluating fields in which the internationally oriented scientific journal is not the main medium for communicating research findings.

Overall, the number of cited references in the Thomson Scientific citation indexes greatly exceeds the number of cited references in other citation-enhanced databases. Especially back in time, the Thomson Scientific citation indexes offer the most comprehensive coverage. The Science Citation Index Expanded format available through the Web of Science goes back to 1900[1], while coverage in the Social Sciences Citation Index starts from 1956, and the Arts & Humanities Citation Index starts from 1975. In contrast, databases enhanced with cited references in recent years cover cited references mostly from the mid-1990's. For some fields, however, discipline-oriented databases may offer the best coverage, both in terms of the journal literature covered and – given the changes in scholarly communication – for other publication types such as preprints and postprints. Bibliometric studies undertaken to determine the usefulness of the new citation-enhanced databases show that searches in those resources may retrieve a number of unique citations. Analyzing citing references for works of chemistry researchers for the years 1999-2001, Whitley (2002) found that 23% of the total number of citing references are unique to the Chemical Abstracts, 17% are unique to the Science Citation Index, and the remaining 60% are duplicated in the two indexes (see also Ridley, 2001). A preliminary study by Bauer and Bakkalbasi (2005) reveals that Google Scholar on average yields higher citation counts for papers published in 2000 in the *Journal of the American Society for Information Science and Technology* than the Thomson Scientific citation indexes, while citation counts retrieved using Scopus were similar to those reported by Thomson Scientific citation indexes.

[1] The *Century of Science* initiative makes available approximately 850,000 publications from 262 scientific journals published from 1900 to 1944. For further information see http://www.thomsonscientific.com/centuryofscience/

In conjunction with the Thomson Scientific citation indexes, new citation-enhanced databases enable examination of the role of the international, peer-reviewed journal literature in the scholarly communication system within a field. Jacsó (2004) found that in the psychology subsections of the Social Sciences Citation Index dating back to 1972 there are more than twice as many records enhanced by cited references than in PsycINFO. Nevertheless, the Social Sciences Citation Index has lower average citation rate per record than PsycINFO because Thomson Scientific does not process books for its citation indexes. Books, however, are more frequently cited than journal articles (see also Hicks 1999). Abt (2004) finds that the NASA Astrophysics Data System reports 15% more citations than the Thomson Scientific citation indexes, mainly from conference proceedings. New citation-enhanced databases even allow tracking of the impact of non-traditional publication types such as preprints and postprints. Thus, the changes in how scientists publish and communicate their research findings may be examined and the implications for research evaluation may be considered.

Finally, the electronic availability of scholarly documents permits the study of emerging research questions, e.g. the analysis of acknowledgements in publications. According to Cronin, McKenzle, Rubio and Weaver-Wozniak (1993) acknowledgments like citations reflect influential contributions to scientific work. An analysis of acknowledgements in five leading information science journals suggests that highly cited authors are also relatively highly acknowledged (Cronin, 2001). Until recently, however, acknowledgments have been accorded relatively little attention in the investigation of scholarly communication, because they are currently not included in major bibliographic databases. Using CiteSeer digital library as data source and applying parsing algorithms to automatically extract acknowledgements from electronic documents, Giles and Councill (2004) show that analysis of acknowledgements uncovers interesting trends, not only in reference to individual scientists, but also regarding the funding agencies and companies that invest in research.

Conclusions

As highlighted by Moed (2005, p. 316), including a greater number of data sources to perform citation analysis does not necessarily lead to more valid assessments of scientific advancement and of scientists' productivity. Given the methodological and technical difficulties in citation analysis, citation-enhanced databases need to be examined carefully, both in regard to their potentialities and their limitations for citation analysis (Moed, 2005). Particularly, they should be explored to determine whether they provide more complete citation data for publication types not covered in the Thomson Scientific citation indexes. Decisive for further bibliometric studies will be which databases perform best as data source for particular fields and time periods. As seen in this paper, each bibliographic database covers unique content, but none is comprehensive. In this respect, new citation-enhanced databases must be viewed more as a supplement than as a substitute to the Thomson Scientific citation indexes. Certainly, the usefulness of citation-enhanced databases will grow as the amount of content increases, e.g. when analysing the long-term impact of scientific work. In the future, citation-enhanced databases could potentially also be used to calculate reference standards to allow for field normalization or metrics similar to the highly controversial journal impact factor calculated on an annual basis by Thomson Scientific. Definitively, coverage is not the only criteria determining the usefulness of bibliographic databases for performing citation analysis. The quality of data must be considered as well as the database implementation's facilities for browsing, searching and analysing data.

In any case, the central assumption of bibliometric assessment of research performance remains the same: scientists refer in their work to the earlier work of other scientists, which they have found useful in pursuing their own research. Obviously, the process of citation is a complex one and assessing research performance by citation analysis is a vulnerable method. Problems such as different motives for giving or not giving a reference to a particular publication, self-citations, or differences in publication and citation practices among fields and subfields have all been clearly outlined (e.g. MacRoberts & MacRoberts, 1996). Despite these limitations, many studies have demonstrated that citation analysis provides useful information for research evaluation and that "*ex ante* peer review should be supplemented *ex post* with bibliometrics and other metrics of science to give a broader and powerful methodology with which to assess scientific advancement" (Daniel, 2005, p.147).

References

Abt, H. A. (2005), “A comparison of citation counts in the Science Citation Index and the NASA Astronomical Data System”, in Heck, A. (Ed.), Organizations and Strategies in Astronomy, Vol. 6, Kluwer Academic Publishers, Dordrecht, pp. 169-174.

Bauer, K. and Bakkalbasi, N. (2005), "An examination of citation counts in a new scholarly communication environment", D-Lib Magazine, Vol. 11 No. 9.

Braun, T., Glänzel, W. and Schubert, A. (2000), “How balanced is the Science Citation Index's journal coverage? A preliminary overview of macrolevel statistical data”, in Cronin, B. and Atkins, H. B. (Eds.), The Web of Knowledge: A Festschrift in Honor of Eugene Garfield, Information Today, Medford, pp. 251-277.

Chemical Abstracts Service (2005), “Cited references in CAplus and CA”, STNotes, No. 24, available at: http://www.cas.org/ONLINE/STN/STNOTES/stnotes24.pdf

Cronin, B. (2001), "Acknowledgement trends in the research literature of information science", Journal of Documentation, Vol. 57 No. 3, pp. 427-433.

Cronin, B., McKenzie, G., Rubio, L. and Weaver-Wozniak, S. (1993), "Accounting for influence: Acknowledgments in contemporary sociology", Journal of the American Society for Information Science, Vol. 44 No. 7, pp. 406-412.

Cronin, B., Snyder, H. and Atkins, H. (1997), "Comparative citation rankings of authors in monographic and journal literature: A study of sociology", Journal of Documentation, Vol. 53 No. 3, pp. 263-273.

Daniel, H. D. (2005), "Publications as a measure of scientific advancement and of scientists' productivity", Learned Publishing, Vol. 18 No. 2, pp. 143-148.

Garfield, E. (1955), "Citation indexes for science. New dimension in documentation through association of ideas", Science, Vol. 122 No. 3159, pp. 108-111.

Garfield, E. (1996), “The significant scientific literature appears in a small core of journals”, The Scientist, Vol. 10 No. 17, p. 13.

Giles, C. L. and Councill, I. G. (2004), "Who gets acknowledged: Measuring scientific contributions through automatic acknowledgment indexing", Proceedings of the National Academy of Sciences of the United States of America, Vol. 101 No. 51, pp. 17599-17604.

Glänzel, W., Schubert, A. and Czerwon, H. J. (1999), "An item-by-item subject classification of papers published in multidisciplinary and general journals using reference analysis", Scientometrics, Vol. 44 No. 3, pp. 427-439.

Hicks, D. (1999), "The difficulty of achieving full coverage of international social science literature and the bibliometric consequences", Scientometrics, Vol. 44 No. 2, pp. 193-215.

Hood, W. W. and Wilson, C. S. (2003), "Informetric studies using databases: Opportunities and challenges", Scientometrics, Vol. 58 No. 3, pp. 587-608.

Jacsó, P. (2004), "Citation-enhanced indexing/abstracting databases", Online Information Review, Vol. 28 No. 3, pp. 235-238.

Jacsó, P. (2005a), "As we may search - Comparison of major features of the Web of Science, Scopus, and Google Scholar citation-based and citation-enhanced databases", Current Science, Vol. 89 No. 9, pp. 1537-1547.

Jacsó, P. (2005b), "Browsing indexes of cited references", Online Information Review, Vol. 29 No. 1, pp. 107-112.

Jacsó, P. (2005c), “Google Scholar (Redux)”, available at: http://www.galegroup.com/reference/archive/200506/google.html

Lawrence, S., Giles, C. L. and Bollacker, K. (1999), "Digital Libraries and autonomous citation indexing", IEEE Computer, Vol. 32 No. 6, pp. 67-71.

Lindholm-Romantschuk, Y. and Warner, J. (1996), "The role of monographs in scholarly communication: An empirical study of philosophy, sociology and economics", Journal of Documentation, Vol. 52 No. 4, pp. 389-404.

MacRoberts, M. H. and MacRoberts, B. R. (1996), "Problems of citation analysis", Scientometrics, Vol. 36 No. 3, pp. 435-444.

Merton, R. K. (1988), “The Matthew Effect in science, II. Cumulative advantage and the symbolism of intellectual property”, ISIS, Vol. 79 No. 299, pp. 606-623.

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

Ridley, D. D. (2001), "Citation searches in on-line databases: possibilities and pitfalls", Trends in analytical chemistry, Vol. 20 No. 1, pp. 1-10. [F386]

Whitley, K. M. (2002), "Analysis of SciFinder Scholar and Web of Science citation searches", Journal of the American Society for Information Science and Technology, Vol. 53 No. 14, pp. 1210-1215.

Youngen, G. K. (1998), "Citation patterns to traditional and electronic preprints in the published literature", College & Research Libraries, Vol. 59 No. 5, pp. 448-449.

Table I. Overview of new citation-enhanced databases.

Database name	Field	Publication types
ACM Guide	computing	journal articles books conference proceedings dissertations reports etc.
CERN Document Server	particle physics and related fields	journal articles books conference proceedings dissertations reports preprints etc.
Chemical Abstracts	chemistry and related fields	journal articles books conference proceedings dissertations reports preprints patents etc.
Citebase	multidisciplinary	preprints postprints etc.
CiteSeer	computer and information science	journal articles conference proceedings reports preprints etc.
CSA Social Sciences Databases	social sciences	journal articles books conference proceedings dissertations etc.
Current Index to Nursing and Allied Health Literature (CINAHL)	nursing and allied health	journal articles books dissertations etc.
Digital Bibliography & Library Project (DBLP)	computer science, particularly databases and logic programming	journal articles books conference proceedings etc.
IEEE Xplore	technology (computer engineering, biomedical technology, aerospace, telecommunications etc.)	journal articles conference proceedings transactions and standards etc.
MathSciNet	mathematics	journal articles books conference proceedings etc.

NASA Astrophysics Data System	physics, geophysics, astronomy, astrophysics, and instrumentation	journal articles conference proceedings reports etc.
PsycINFO	psychology	journal articles books dissertations reports etc.
RePEc	economics	journal articles books working papers etc.
Scopus	multidisciplinary	journal articles conference proceedings etc.
SMEALSearch	business administration	journal articles working papers white papers consulting reports etc.
SPIRES-HEP	high energy physics	journal articles conference proceedings reports preprints etc.
Web Citation Index	multidisciplinary	preprints postprints etc.

Table II. Comparison of Thomson Scientific citation indexes, Chemical Abstracts, Google Scholar and Scopus.

	Thomson Scientific citation indexes		Google Scholar	Scopus	Chemical Abstracts	
	via Web of Science	via STN International			via SciFinder (Scholar)	via STN International
Coverage: breadth	37 million records 9,100 journal titles (SCIE, SSCI and A&HCI)	30 million records 6,500 journal titles (SCIE)	unknown	27 million records 15,000 journal titles	25 million records 9,500 journal titles patents from more than 50 active patent-issuing authorities	25 million records 9,500 journal titles patents from more than 50 active patent-issuing authorities
Coverage: time period	back-years to 1900 (SCIE), 1956 (SSCI) and 1975 (A&HCI)	back-years to 1974 (SCIE)	unknown	back-years to 1966 cited references back to 1996	back-years to 1907 cited references back to 1996	back-years to 1907 cited references back to 1996
Classification codes and index terms	no	no	no	yes	yes	yes
Browsing options	author cited author cited source title	author source title address cited reference cited author cited source title etc.	not available	not available	not available	author source title address cited reference cited author cited source title etc.
Searching options	author publication year article title topic source title address document type	author publication year article title topic source title address document type	author publication year article title topic source title subject category	author publication year article title topic source title address document type	author publication year article title topic source title address etc.	author publication year article title topic source title address document type

	language	language subject category ISSN etc.		language keywords ISSN etc.		language classification code ISSN etc.
Cited reference searching options	author publication year source title	author publication year article title source title volume page number patent number etc.	not available	author publication year article title source title page number	not available	author publication year article title source title volume page number patent number etc.
Analytical tools	ranking up to 2,000 records by author, publication year, source title, country, institution name, subject category, language, or document type	ranking and cross-tabulation up to 50,000 records by author, publication year, source title, country, institutional affiliation, subject category, language, document type etc.	not available	ranking by author, publication year, source title, subject category, and document type (via refinement option) and analysis of citations over time (via Citation Tracker)	ranking by author, publication year, source title, address, language, document type etc.	ranking and cross-tabulation up to 50,000 records by author, publication year, source title, country, institutional affiliation, language, document type etc.
Saving and exporting options	yes	yes	no	yes	yes	yes