Web-Based Tools for Citation Data Management

Nancy K Herther
Searcher; May 2008; 16, 5; Platinum Periodicals
pg. 18

Private research corporations and well-endowed private universities vie for funding and the prestige that comes with it. Between 1998 and 2003, for example, the National Institutes of Health (NIH) doubled its science funding. However, about 5 years ago, research funding became stagnant, or even declined, creating an increasingly competitive marketplace for dwindling resources (1). At the same time, due to economic pressures and globalization, research itself has become a primary factor in competitive positioning — for individual researchers, their institutions, and even for national standings in our global community.

Citation analysis has become a critical measure of research value, whether for grant applications, formal promotional processes, or just to enhance public perception of individuals or institutions (2). This has significantly increased the use of Thomson’s Web of Science (WoS) — and other citation sources — to gather data on “citedness,” using this to rank the output of researchers and the relative importance of their institutions or the journals in which they publish.

Increasing Sources of Citations

The increased desire to use citedness to measure value and relative positioning is spurring the development of new databases and software — from Elsevier’s Scopus, CiteSeer, and other citation-based products — to the development of cited references in Google Scholar and, increasingly, the inclusion of cited reference information in scholarly databases from such vendors as CSA and EBSCO (3).

Certainly the increasing number of databases and websites with citation data has become significant. These resources make it possible to go far beyond the 9,000-plus journals indexed in WoS. Although many purists would question whether or not looking comprehensively for all potential citations makes sense, in practice, most clients want whatever they can get.

Using citation information as a gauge of quality research is not without criticism. Even more troubling is the fact that there is no standard, accepted method, or yardstick for analyzing the productivity and importance of researchers, their published works, or their institutions (4).
Research is big business and a KEY to national rankings and international status.

by Nancy K. Herther
Anthropology and Sociology Librarian
University of Minnesota Libraries

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Is More Always Better?

Anyone who has worked on citation analysis — whether for competitive intelligence, marketing, or academic assessment or for more traditional bibliometric analysis — has learned the inherent difficulties with the process. The difficulties can range from making sure that you have everything on your person (common names are notoriously troublesome) to achieving a confidence level in terms of the universe of potential cited references. Good citation analysis depends on the integrity of the reference lists, which, in turn, can depend on such apparently minor issues as the quirks of the various style manuals used by the different journals.

More data adds significant complexity to the task. If the data comes from a single source, you can at least be assured that the field tags will be common. But what about large data sets from various resources? Merging these to create a single database and working with all the potential formulas you might need to reach references is a major problem for researchers.

Last year, two information science faculty did a citation assessment of their departmental colleagues using the WoS, Elsevier’s Scopus, and Google Scholar (GS). The results were sobering: “The WoS data took about 100 hours of collecting and processing time, Scopus consumed 200 hours, and GS a grueling 3,000 hours” (5). This example could be considered as close to ideal as possible: The faculty were in the same field, the same department, and actively cooperated with the research. Imagine the difficulties added when one is doing competitive research without this level of buy-in! Thankfully, there are some good sources of assistance and useful products entering the marketplace.

Citation Analysis Tools in the Databases

Database vendors have increasingly added key analysis tools to their software programs. For example, the Analyze feature on WoS allows users to generate statistical frequency summaries for up to 100,000 records, sorting, ranking, and displaying them by selected fields, including these:

- **Subject Categories** – identified categories of interest
- **Source Titles** – limit results to selected journals
- **Document Titles** – select or eliminate documents by type, such as review papers
- **Authors** – limit to particular authors actively publishing for a particular topic
- **Publication Years** – view profiles based on date of publication

This process allows users to do a more in-depth analysis of their search results while still working with results in the database.

Scopus’ Refine Results allows for similar refinement or visualization of search results. SciFinder Scholar also offers users aids, such as searching name variants, and analyze/refine to limit or refine initial search results.

Features such as these offer users — especially end users — intelligent ways to limit results, eliminate false drops, and to more easily see how their search results fit into a larger perspective of subject categories, document types, etc.

Getting Help Beyond the Databases

Although we have many more sources for citation data today, we have yet to see the development of a single product that provides a true replacement to the venerable WoS. However, Elsevier and others are promising serious competition in the future.

Any Thomson/ISI client can use the company’s own research services, headed by Henry Small (certainly one of the foremost experts in this area), to create tailored data sets from WoS databases [http://scientific.thomson.com/products/rgi]. If the stakes are high and WoS data is considered essential, this is a worthwhile, fast, and reliable option.

Often researchers use spreadsheets, relational databases (such as Microsoft’s Access), and statistical packages (such as SPSS) to analyze citation data. Given the dominance of these types of products for other research applications, most researchers find these programs easy to use.
and familiar. Chaomei Chen, Drexel University (Philadelphia) professor of information science and technology, notes that although these general software products may suffice for the occasional user, integrated, specialized tools would be more efficient and probably less error-prone than using multiple (general) tools, "It boils down to whether one should learn to use new tools and whether it will be worth the candle. If all the tools I already know can get me the same results fast enough, then I will need something extra to motivate me to learn to use a new tool."

Collegial Help Aids

In the early days of the microcomputer industry, groups of techies provided a major role in product development and support by creating independent user groups and actively working with vendors on product support and development. These folks were called hackers before this term degraded into a word denoting malicious marauders of software and the Internet. Over the years, as researchers have worked to do their own sophisticated citation analysis, they have similarly developed their own software programs or work-arounds using standard spreadsheets or other programs.

A few of these programs have been made available over the Internet for others, generally at no charge. As with the early microcomputer software industry, these entrepreneurs generally encourage the active involvement of users to help in product support and ongoing development.

The first product offered to help with citation data manipulation was apparently Biblimetric Toolbox. Created by Terrence Alan Brooks (Information School, University of Washington), the product was released in the mid-1980s for $49.95 on CD-ROM. Written in Turbo Prolog under DOS 2.0, it is now available freely over the Internet but is no longer actively supported or updated by Brooks.

Anne-Wil Harzing's Publish or Perish "was originally developed inhouse for the benefit of my own application for promotion to full professor." Intended to empower "individual academics to do the kind of analyses that were typically done by bibliometricians or administrators," Harzing notes her program was developed to help the "casual academic user wanting to assess their impact for e.g. a promotion."

Loet Leydesdorff, professor at the Amsterdam School of Communications Research, has made various programs available over the Web. "I make the programs for my own use, for my Ph.D. students, and for use in class," Leydesdorff notes. "I [make] them available in order to support the community and the research

continued on page 50

KEY Links and Resources for Citation Research

If you want advice or to share your ideas with experts, these resources are some of the best available today in the area of citation analysis:

SIGMETRICS
http://web.utk.edu/~gwhitney/sigmetrics.html

SIGMETRICS is a Virtual Special Interest Group of the American Society for Information Science and Technology that "covers bibliometrics, scientometrics and informetrics, but also metrics as related to the design and operation of Digital Libraries and other information systems interpreted broadly."

ACM SIGMETRICS
http://www.sigmetrics.org/

A listerv of the Association for Computing Machinery concerned with performance evaluation and related topics. The group publishes Performance Evaluation Review, a quarterly newsletter.

Scientometrics
Springer ISSN 0138-9130 (Print) 1589-2861 (Online)
http://www.springerlink.com/content/101080/

Described as "an International Journal for all Quantitative Aspects of the Science of Science, Communication in Science and Science Policy," the journal publishes key articles in the areas of scientometrics, informetrics, bibliometrics, and related areas.

Journal of the American Society for Information Science and Technology (JASIST)
American Society for Information Science and Technology (ISSN 0044-8217)

"This journal serves as a forum for new research in information transfer and communication processes in general, and in the context of recorded knowledge in particular. Concerns include the generation, recording, distribution, storage, representation, retrieval, and dissemination of information, as well as its social impact and management of information agencies. There is a strong emphasis on new information technologies and methodologies in text analysis, computer based retrieval systems, measures of effectiveness, and the search for patterns and regularities in measures of existing communication systems."

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Web-Based Tools for Citation Data Management

continued from page 21

efforts of my colleagues... This provides me with feedback for improving the programs."

The advantage to this approach is not only easy access to free software but participation in a community of researchers doing similar research and willing to actively share ideas and advice. As Leydesdorff notes, "In principle, there is a preference for 'science is gift-giving' [Robert Merton], but sometimes one has to be practical. If I would charge people for it, I would feel obliged to provide services which I cannot provide given the priority of my academic objectives."

Garfield Enters the Marketplace

After years of development, Eugene Garfield officially launched his "algorithmic historiography software" product, HistCite, in October 2007, making it available for free 30-day trials.

First, users do their searches in WoS, then mark records for importing into HistCite. Users can then work with the entire downloaded dataset or define very specific subsets to present in spreadsheet formats or in single two-dimensional (2D) graphics. The program also allows you to do institutional and national rankings. Output options include maps, graphs, or grids/tabs of information. You are able to manually add records to HistCite from other databases or bibliographic management systems (such as RefWorks, ProCite, EndNote, etc.), however, this can be a tedious process.

First demonstrated at the University of Pittsburgh 7 years ago, Garfield worked with about 100 beta testers, refining the software prior to its official launch. While HistCite is an excellent basic software system for citation analysis, the price tag will be an interesting factor in its success.

HistCite isn’t for the inexperienced. The software and its documentation make assumptions about the existing knowledge of the user. For power users of bibliometrics or other types of citation analysis, the price is well worth the access to tools and the growing community that the product seems to be engendering. However, will it be able to attract a broader audience? Will enough users be willing to pay for HistCite to allow for further development? Time will tell (6).

Visualizing the Future

The future would seem to lie more with the sophisticated visualization products now coming to market. As more research institutions work to ramp up their knowledge management systems — including citation analyses, expertise databases, institutional measurements, and other data — the institutions have begun investing in systems that merge data sources and allow for detailed analysis and searching on-the-fly, without having to worry about the underlying metrics.

When it comes to the complicated matrices that create links between ideas, researchers, or institutions, looking at 2D or 3D representations of data rather than long tables or charts helps users better understand the data, increasing the ability to find patterns or otherwise make sense of the information presented. Many programs currently available for citation analysis include some type of visual mapping capability.

Collexis [http://www.collexis.com] is one company to watch. It has created impressive 3D, full-color, interactive databases. Collexis’ Expert Profiling System, in development at Mayo Clinic Libraries, Johns Hopkins University, University of California – San Francisco, and University of South Carolina, allows these institutions to merge data sources and to create a single database for information on existing expertise, comparisons with peers, and other features. Named KMWorld’s Trend-Setting product of 2007 with good reason, Collexis’ graphics are spectacular and the system’s fast response excellent. Using an indexing method called “Fingerprinting,” the software creates document profiles (as opposed to using word matching) to link data. Screen shots don’t do the product justice; it’s something you will want to see in person. In February, Collexis announced a joint product, Thomson Collexis Dashboard, that would tie its data mining tool, Collexis Knowledge Dashboard, to Thomson’s WoS biomedical content. Thomson Collexis Dashboard will expand search and analysis of journals and cited research back to 1985.

Citation Analysis Software on the Web

So, what are some of the options available today for serving the ravenous needs of these “power” citation searchers? Published articles on citation analyses generally include methodology sections that discuss the specific strategies, software, and models used in their assessments. These can often provide invaluable information to anyone embarking on this research. Bibliometric (and scientometric and informetric) researchers have established listservs and other tools to share information as well.

The following software programs represent some of the best on the web today and are generally available as freeware. All

continued on page 32
Beyond WoS
INCLUDING CITED REFERENCES TODAY

The number of scholarly databases, websites, and online vendors adding cited reference features is growing rapidly. Here is just a sampling of databases and vendors offering cited references today.

- CiteSeer [http://citeseer.ist.psu.edu], a free service that focuses on computer/information science areas, with a layout for citations very similar to Google Scholar.


- RePEc (Research Papers in Economics) [http://repec.org] is described as a “decentralized database of working papers, journal articles and software components,” including many records from EconLit. The records include field reports and working papers — far more diverse information than found in other databases on each author — all linked, when possible, to full text online. Click on Access Statistics for a fascinating statistical summary of the accesses and downloads to documents in the database.

- MathSciNet [http://www.ams.org/mathscinet], produced by the American Mathematical Society, allows for searching by either “Ref Author” or “Ref Anywhere.” You can also click on the linked PDF for individual records to see a list of references to this record from other SciFinder records.


- CiteBase Search [http://citebase.eprints.org] includes more than 300,000 records in arXiv.org, cernpreprints.org, and BioMed Central. The results can be ranked in the database by most cited author/paper, how often authors/papers are formally cited or accessed through links, etc.

- Elsevier journals [http://www.sciencedirect.com], available through the Science Direct service, link to citations from other journal articles in the database.

- Faculty of 1000 [http://www.facultyof1000.com] for Biology and Medicine includes about 5,000 “leading researchers and clinicians [who] share their expert opinions by highlighting and evaluating the most important articles in biology and medicine.” This approach relies on the evaluations given by experts in the fields.

- Google Scholar [http://scholar.google.com], noted for its innovation, attempts to cover anything available on the web. The results are widely scattered, lacking authority control for basic data elements, and include no explanations on how citation rates are calculated. Still, the numbers and resources are fascinating to track.


- IEEE Xplore database [http://ieeexplore.ieee.org/xpl/运作?SP], a collection of standards and papers from IEEE and IEE, allows you to see other references to some papers in the database by clicking on the Abstracts Plus view for records.

- SciFinder Scholar [http://www.cas.org/SCIFINDER/SCHOLAR] covers more than 9,500 currently published journals (1999 to the present) and patent information from more than 50 active patent issuing authorities. It focuses on chemistry and medicine.

- Scopus [http://www.scopus.com/scopus/home.url] is the Elsevier product released in 2004 that covers approximately 15,000 journals back to 1996. The database is still under construction (especially in the Humanities and Social Sciences areas, where current coverage is very poor). However, it is a good complement to WoS in the sciences.
Bibexcel — Quick Start Guide to Bibliometrics and Citation Analysis
http://www.rhbnc.ac.uk/~uhtm001/bibexcel-primer.pdf

Alan Pilkington, Royal Holloway, University of London, Egham, Surrey

Quick and Dirty Guide to Bibexcel
http://www2.dboj.jws/Files/Quick%20and%20Dirty%20Guide%20to%20Bibexcel.doc

Downloads a Word file with tips on using Bibexcel for bibliometric research.

Bibliometrics Toolbox
http://projects.ischool.washington.edu/tabrooks/hyper/biblio.html

"The Bibliometrics Toolbox will perform a variety of bibliometric measures on a sample of a literature. For example, it will do a rank listing of the productive journals, create a bibliography, and calculate a wide series of measures as defined by Goffman and Warren, Eggle, Leimkuhler, Lotka, Zipf, and so on. The Bibliometrics Toolbox was written in Turbo Prolog under DOS 2.0." As mentioned, the software is now free, comes as a zip file, and includes a user manual. A computer software program that ranks entries by productivity, calculates minimum Bradford zones, and graphs the bibliography. Bibliometrics Toolbox allows division of data sets into groups for analysis.

BibTechMon
http://systemsresearch.ac.at/FTP/schiebel/FTP

BibTechMon is a powerful software for co-word analysis, visualization, and navigation into Co-Word Maps. It offers automated indexing, a feature for standardizing word variants (e.g., endings, plurals, etc.), cluster analysis, and other tools. Previously, it was available free of charge for educational purposes and thesis work on the web. In order to get current details on the program, you should send an email to the developer, Edgar Schiebel, head, technology management, ARC Systems Research GmbH, Tech Gate Vienna [Edgar.Schiebel@arcos.at.ac].

CiteSpace
http://cluster.cis.drexel.edu/~cchen/citespace

Created and maintained by Chaomei Chen, CiteSpace is a freely available Java application for analyzing and visualizing scientific literature, allowing you to develop maps linking authors, institutions, or concepts.
Download citation data from WoS and start CiteSpace. The visualizations are very clear and the software updates often. The site includes a good user help guide as well. Chen has written about applications of the software in the research press (8).

**FullText.exe**

http://users.fmg.uva.nl/ileydesdorff/software/fulltext/index.htm

A freely available program for academic/research applications, maintained by Loet Leydesdorff of the Amsterdam School of Communications Research, FullText.exe generates a word-occurrence matrix, a co-occurrence matrix, and a normalized co-occurrence matrix from a set of text files and a word list. The output files can be read into standard software (SPSS, Ucinet/Pajek, etc.) for statistical analysis and visualization. A version adapted for the Korean character set is available at http://www.leydesdorff.net/krkwc. A similar program using lines (e.g., titles) instead of long texts can be retrieved at http://www.leydesdorff.net/software/ti/index.htm.

**HistCite**

http://www.histcite.com

In its initial release, HistCite works only on the Microsoft Windows platform and uses the web browser as the user interface. HistCite was optimized for Internet Explorer versions 6 and 7, however, some testing has also been done with both Firefox and Opera.

The software is free for the first 30 days, giving users a chance to evaluate its usefulness before committing to the purchase price. Prices begin at $199 for licensing a single copy and rise to a scale depending on the number of copies needed. Buying from five to nine copies brings a 15% discount, 10 to 24 copies carries a 25% discount, etc. The price includes updates and revisions for that year. Campus or organization-wide contracts are also available. After the first year, you can renew for future upgrades, priced at $89 for a single user.

**Slexe**

http://users.fmg.uva.nl/ileydesdorff/software/slexe/index.htm

Another program freely available for academic usage from Loet Leydesdorff, Slexe.exe takes WoS data as input and produces four output files in dBase III format. These files can be moved into either Excel, Microsoft Access, or SPSS for further processing.

The program is based on DOS-legacy software. It runs in a MS-DOS Command Box under Windows. If you use this software, the creator requests that you reference it as follows:

Publish or Perish
http://www.harzing.com/resources.html#pop.htm

Freeware for personal nonprofit use, the software is available in both Windows and Linux versions. Publish or Perish uses Google Scholar as the core source for raw citations, analyzing and presenting the following statistics:

- Total number of papers
- Total number of citations
- Average number of citations per paper
- Average number of citations per author
- Average number of papers per author
- Average number of citations per year
- Hirsch’s h-index and related parameters
- Egghe’s g-index
- The contemporary h-index
- The age-weighted citation rate
- Two variations of individual h-indices
- An analysis of the number of authors per paper.

TL.exe
http://www.leydesdorff.net/software/ti

TL.exe was created and maintained by Loet Leydesdorff, Amsterdam School of Communications Research, to generate wordlists and a variety of co-occurrence matrices (exportable to popular statistical and network analysis packages). This DOS-era legacy software runs using an MS-DOS command box under Windows. Files can be output in dBaseIV, Excel, or SPSS. Check for other derivative programs at http://users.ing.uva.nl/leydesdorff/indicators/index.htm.

Nancy K. Herther [n-her@umn.edu] is the anthropologist and sociology librarian at the University of Minnesota Libraries in Minneapolis. She is a former columnist for ONLINE and Database (ECcontent) magazines and former editor of CD-ROM Professionals.

References
3. For a good background on the development and scope of cited reference resources, see Dana Roth’s 2005 article, “The emergence of competitors to the Science Citation Index and the Web of Science,” *Current Science*, vol. 89, no. 9, pp. 1538-1539, November 2005. Available freely online at http://www.ias.ac.in/journal/ostc/vol89/nov2005/1538.pdf.