As we may search – Comparison of major features of the *Web of Science*, *Scopus*, and *Google Scholar* citation-based and citation-enhanced databases

Peter Jacso

Department of Information and Computer Science, University of Hawaii, Honolulu, HI 96822, USA

Keywords: *Google Scholar*, *Scopus*, *Web of Science*.

As They May Have Thought

IT may appear blasphemous to paraphrase the title of the classic article of Vannevar Bush[1] but it may be a mitigating factor that it is done to pay tribute to another legendary scientist, Eugene Garfield. His ideas of citation-based searching, resource discovery and quantitative evaluation of publications serve as the basis for many of the most innovative and powerful online information services these days.

Bush 60 years ago contemplated – among many other things – an information workstation, the Memex. A researcher would use it to annotate, organize, link, store, and retrieve microfilmed documents. He is acknowledged today as the forefather of the hypertext system, which in turn, is the backbone of the Internet.

He outlined his thoughts in an essay published in the *Atlantic Monthly*. Maybe because of using a non-scientific outlet the paper was hardly quoted and cited in scholarly and professional journals for 30 years.

Understandably, the *Atlantic Monthly* was not covered by the few, specialized abstracting and indexing databases of scientific literature. Such general interest magazines are not source journals in either the *Web of Science* (*WoS*)[2], or *Scopus*[3] databases. However, records for items which cite the 'As We May Think' article of Bush (also known as the 'Memex' paper) are listed with appropriate bibliographic information. *Google Scholar* (*G-S*)[4] lists the records for the Memex paper and many of its citing papers. It is a rather confusing list with many dead links or otherwise dysfunctional links, and a hodge-podge of information related to Bush.

It is quite telling that (based on data from the 1945–2005 edition of *WoS*) the article of Bush gathered almost 90% of all its 712 citations in *WoS* between 1975 and 2005, peaking in 1999 with 45 citations in that year alone. Undoubtedly, this proportion is likely to be distorted because far fewer source articles from far fewer journals were processed by the Institute for Scientific Information for 1945–1974 than for 1975–2005. *Scopus* identifies 267 papers citing the Bush article. The main reason for the discrepancy is that *Scopus* includes cited references only from 1995 onward, while *WoS* does so from 1945.

Bush's impatience with the limitations imposed by the traditional classification and indexing tools and practices of the time is palpable. It is worth to quote it as a reminder. Interestingly, he brings up the terms 'web of trails' and 'association of thoughts' which establishes the link between him and Garfield.

> Our ineptitude in getting at the record is largely caused by the artificiality of systems of indexing. When data of any sort are placed in storage, they are filed alphabetically or numerically, and information is found (when it is) by tracing it down from subclass to subclass. It can be in only one place, unless duplicates are used; one has to have rules as to which path will locate it, and the rules are cumbersome. Having found one item, moreover, one has to emerge from the system and re-enter on a new path.
>
> The human mind does not work that way. It operates by association. With one item in its grasp, it snaps instantly to the next that is suggested by the association of thoughts, in accordance with some intricate web of trails carried by the cells of the brain. [….] Selection by association, rather than by indexing, may yet be mechanized.

This is exactly the point where Eugene Garfield enters, with his characteristic fervor for thinking and doing better. He envisioned 50 years ago in his landmark paper[5] how the limitations of descriptor-based look-up of scholarly documents in abstracting/indexing publications could be overcome by using the references cited by the authors in their papers. This is a different approach than the one presented by Bush because cited references have always formed a part of scholarly papers. They represent millions of past associations in a formal and highly structured – if not standard – way. These can be reconstructed

e-mail: jasco@hawaii.edu

from the source documents without the type of intellectual efforts implied by Bush. Garfield got his idea for creating a citation index from the highly successful Shepard's Citations which traced (and still trace) court cases and decisions for each US jurisdiction. Garfield's unpublished master thesis for the MLS degree at Columbia University bore the title 'Shepardizing the scientific literature'[6]. Frank Shepard was not the first (but was certainly the most famous) for creating a citation index. As it was pointed out by Weinberg[7], the earliest Hebrew manuscript citation index dates back to the 12th century.

While Bush emphasized the associative thinking of the individual researcher, Garfield built his idea on the constantly growing, and already recorded collective associate thinking of the invisible college of researchers through the network of references documented in their published papers. This is an important distinction.

Garfield's vision was also right and much ahead of his time in recognizing how the legal citation index can be enhanced, adapted and applied to scientific literature in order to produce a unified index without disciplinary boundaries. Most indexing/abstracting databases started out as (and remained) discipline-oriented reference sources, such as *Biological Abstracts*, *Psychological Abstracts*, or *Sociological Abstracts*. When Garfield penned his original article and a decade later when he launched *Science Citation Index*[8], there were no multidisciplinary indexing/abstracting databases. It was a breakthrough and remained a unique resource for 40 years. In Fall 2004, Elsevier launched its ambitious *Scopus* service. It was followed by the release of *Google Scholar* in beta version.

Garfield's contributions cover a wide range. They were acknowledged on his 75th birthday by a Festchrift[9]. In it many of the best practitioners, theoreticians and educators of the information profession paid homage to his achievements and personality. I am graced by being offered the opportunity to pay homage in this special issue of *Current Science* on the 50th anniversary of publishing his seminal article in *Science*. I do so by reviewing some of the essential characteristics of contemporary citation-based and citation-enhanced information services to illustrate how they implemented Garfield's vision.

The findings of the test searches tailored for this paper may not apply universally to the databases or to all the disciplinary areas, but a more extensive battery of tests for a variety of other topics and sources corroborated that the database characteristics presented here are rather typical even though the test samples formally do not meet the requirements for a statistically representative sample. The point is to demonstrate the profiles, and illustrate the pros and cons of the three major multidisciplinary citation systems through examples pertinent to the occasion.

It is to be noted that ISI has a companion of open access sources and can pass on the query and can run it in many open access scholarly archives and databases. This is true also for *Scopus* which runs the query automatically and simultaneously in Scirus[10] against even more open access sources. These are important extras, but were not tested for this review. There was no choice with *Google Scholar* which runs the query against the index created from an undisclosed mix of journal article archives of publishers, repositories of preprints and reprints, as well as educational sites and home pages of presumably scholarly people.

As We May Search

There are several papers which mention *Web of Science*, *Scopus* and *Google Scholar*, including a few substantial reviews[11–13]. I re-tested the three major systems for this review in April and May of 2005, but I also relied on the earlier in-depth reviews of *WoS*[14], *Scopus*[15] *Google Scholar*[16] and its updated version[17], as well as on a series of commentaries about citation enhanced indexing/abstracting services[18–20], link-enabled cited references[21], using citation scores for filtering and sorting results[22], software approaches to citation searching[23], and citation browsing[24].

ISI and Elsevier provide substantial factual information and help files about *WoS*[25] and *Scopus*[26], which act as an information hub. From there one may look up the list of journals processed[27,28], or – in case of *Scopus* – also the list of publishers[29], and the list of open access journals[30]. The ISI web-site also includes essays by Eugene Garfield about the concept[31], and history[32] of citation indexing. *Google* provides minimal information about the content of *Google Scholar*. Its list of Frequently Asked Questions[33] provides some information about the software features.

The three databases represent different approaches to citation search services. *WoS* and *Scopus* are commercial databases (at the expensive end of the spectrum – for good reasons). *Google Scholar* is currently an open access database, still in beta version after its launch in November 2004. The expectations are different for fee-based and free databases, but open access should not provide excuse for ill-conceived and poorly implemented search options, and for convoluted, and potentially misleading presentation of information.

The family of ISI citation indexes which makes up the core of *WoS* was created from the get-go by the inclusion of all references cited by papers in the primary (source) documents. Creating traditional bibliographic records is an error-prone process. Creating entries for cited references is even more so.

Considering the vagaries of identifying the quintessential bibliographic elements of references from the very different reference styles used across thousands of journals which form the source base of the ISI citation indexes has been and remains a daunting task. This is even more true for the millions of unique citations given to hundreds of thousands of cited sources: books, articles,

conference papers, government documents, notes, transcripts, patents, software, web sites, etc. Deciphering and making uniform their often cryptic citation content and formats is a Sisyphean task. The often careless attitude to references by us authors, as well as by editors comes back to haunt us when we miss a large number of citations which have wrong author name, journal title, and/or chronological-numerical designation. This must be calculated when interpreting the citedness score of papers as they may be widely scattered and hence overlooked.

Elsevier created *Scopus* by extracting records from its traditional indexing/abstracting databases, such as GEOBASE, BIOBASE, EMBASE, and enhanced them by cited references. This is a different approach from the one used for the citation index databases of ISI which were created from the grounds up with the cited references in the records (and in the focus of the whole project). Elsevier had to struggle with the same problems as ISI at an even larger scale given its wider source base of journals and conference proceedings with a wider variety of inconsistencies – although for a much shorter time span than in *WoS*.

G-S is a free service, and for many who consider it to be a gift for the world it may be anathema to say any but good words of it. It is also to be emphasized that it is a joint gift by some publishers and/or their digital facilitators (the content part), and *Google* (the software and the service operation part). If ISI or Elsevier could have received such unfettered access to the publishers' archives for harvesting their sites offering standard-compliant metadata, they could probably sell their services – if not for free – at a fraction of their current price. Building a multi-million record database incurs multi-million dollar investment just to subscribe to the journals, administer their processing, and record their standard bibliographic data, abstract, and descriptors, for about 1 million papers per year in the most recent period. Adding 20–22 million cited references per year increases those traditional costs enormously.

WoS and *Scopus* offer powerful features for browsing, searching, sorting and saving functions. *WoS* should increase the size of the sets which can be sorted, saved and exported. (*WoS* increased the maximum sortable set to 100,000 records as I went to press.) *Google* offers limited and sometimes dysfunctional search options for such well-structured data. The deficiencies of its search software for its database derived from the exceptionally metadata-rich archives of publishers prevent the searcher from performing efficiently even basic searches like finding articles published in *Current Science*. Bibliometric searches to explore the size, source base, breadth and composition of a database, or the literary genealogy of a specific subject are exceptionally well facilitated in *Scopus* and *WoS*, and are practically non-existent in *G-S*.

All the databases were updated and functionally enhanced when the test searches were run in April and May. The content of the databases are updated regularly in *WoS* and *Scopus*. *G-S* was updated in April after a 6-month dormant period. The numbers in the search examples and in the narrative part will change by the time of publication of this paper. Some of the shortcomings of software and content may also be corrected.

Database subject scope and composition

There are significant differences in the subject scope and composition of the three multidisciplinary citation databases. This may not be apparent just by looking at the publicity materials and journal base of a database. Understanding the different meanings of the time dimensions of coverage is also essential. In case of citation databases it is also essential to know from what year have been records enhanced by cited references. While *Scopus* makes it clear that this enhancement applies only to records for articles published in the past decade, this is not at all reflected by headlines like '*Scopus* has wider scope than *Science Citation Index*'[34]. Neither is it accurate to state[35] that '*Scopus* is all set to become the single largest scientometric database with more than 27 million abstracts and citations covering 14,500 journals from 4000 publishers and dating back to 1966'. Publicity materials and journals lists alone are not sufficient for dispensing such claims. For informed database selection decisions test searches must be done. According to my tests, about 67% of the 27.5 million records of *Scopus* have abstracts, which is a good ratio, and slightly better than the ratio in *WoS*, simply because Elsevier has been adding abstracts for articles in its role as an indexing/abstracting service provider, while ISI's clear policy is that it does not create abstract, but just uses it if available in the article itself. In addition, *WoS* makes it also clear that it started to add abstracts to *Science Citation Index* from 1971, and to *Social Sciences Citation Index* a year later.

As for citations, in my estimate there are citations for about one third of the *Scopus* records. Once again, it is not bad as many of the articles do not cite other works. What is important to know is that *Scopus* includes the citations for papers published in the past decade. Published corrections and explanations by informed searchers, like Arunachalam[36] can set the record straight in matters of database scope, size and composition, but this does not always happen. Professional searchers must do sample test searches and correctly interpret the results to corroborate claims and get factual information about databases.

WoS covers all disciplines one can think of or find in the curricula of universities in science, social sciences, arts and humanities. There are three major database components in *WoS* as shown in Figure 1. These can be searched in one fell swoop, and this is the default setting in *WoS*. There are also two chemical databases within *WoS* with less than 100,000 records: *Chemical Reactions*

Index, and *Index Chemicus* (not shown in the figure for their relatively small size). *WoS* comes the closest to be a genuinely pan-disciplinary database – at least for the past 30 years of the scholarly and professional literature. The Extended Science Citation Index is by far the largest component of *WoS*. There are several reasons for it. The most important is that *WoS* goes back the farthest (to 1945), while coverage in the Social Science Citation Index starts from 1956, and in the Arts & Humanities Citation Index from 1975.

The actual record distribution shown in Figure 1 is not identical to the journal distribution among the three major disciplinary domains. Another reason for the dominance of science literature is that science journals make up nearly 67% of the close to 9000 journals indexed from cover-to-cover, while social science journals and arts and humanities journals provide 20% and 13% of the fully covered journal base, respectively. For these latter two subsets thousands of additional journals are also scanned selectively.

Journals and patents have been the only resource types for many years, but the source base of *WoS* was extended by including some conference proceedings and monographic series with analytic records for conference papers and chapters. These represent a relatively small component in *WoS*. (The recently launched ISI Proceedings database has 3.8 million records about conference papers (covering the period of 1990–2005), but it is not (yet) part of *WoS*, and does not offer cited reference searching.)

Scopus does not cover sources in arts and humanities, and has modest coverage of the social sciences. *Scopus* covers 14,000 journals. The breadth of coverage varies widely. The distribution of the journals by major subject areas (as assigned by Elsevier's specialists) is reported on the FAQ page of *Scopus* referred to earlier (26) (Figure 2). It clearly shows that the focus of *Scopus* is life and health sciences (about 38%). Chemistry, physics, engineering and mathematics represent about 29% of the journals, while periodical sources in the social sciences make up about 17% of the journals. Journals classified under agriculture, biology, and earth and environmental sciences make up the rest (16%).

The distribution of the number of records (Figure 3) shows a much stronger concentration on health and life sciences (60% together), and far smaller presence of social science papers – merely 2.21%. It is to be noted that both the journals and the article records may be assigned to two or more subject areas in *Scopus*.

Beyond journal sources *Scopus* also covers books and conference proceedings. The coverage of about 20,000 books is limited to engineering and earth and environmental sciences, with a few books on chemistry and physics. Book records show a strange pattern with nearly 3500 books from 1985, but only 201 from 1989. Information about more than 1500 books is included from 1993 but only 71 from 2001, which is the most recent year for book records. The book records seem to have been removed as I went to press.

Conference materials are referred to as conference reviews, although they are not reviews, but either bibliographic records about conference proceedings or about conference papers. They are mostly from the fields of engineering, physics and chemistry. They show the same odd pattern as books. There are 103 records for items published in 1999, and close to 1400 records for conference materials published in 1982, 1987, 2003, 2004 each.

While the coverage of *Scopus* is claimed to go back forty years to 1966, it is an understatement. There is con-

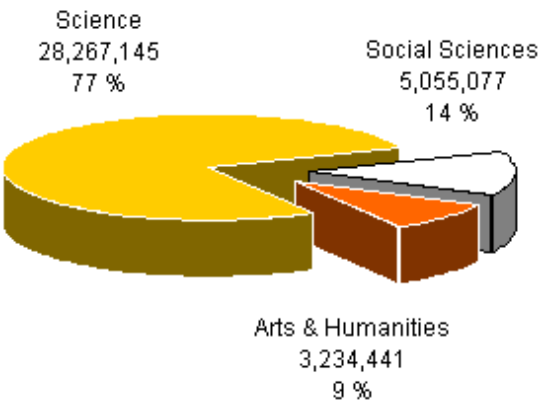

Figure 1. Distribution of records among the major components of *WoS*.

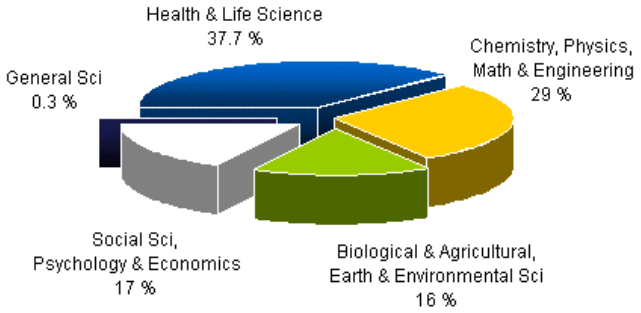

Figure 2. Journal distribution by subject areas in *Scopus*.

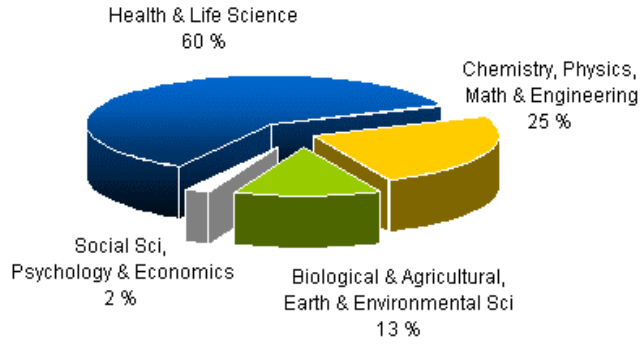

Figure 3. Distribution of records by subject areas in *Scopus*.

siderable coverage of the 1965 literature and even earlier materials to the tune of close to 90,000 records. Cited references are only included for articles published in the last decade according to *Scopus*. I found that there are also more than 7000 pre-1996 records with references.

Google does not offer publisher list, journal list, neither any clue about the time-span or the disciplinary distribution of records in *G-S*. Its search software prevents the users from finding any reliable data about these traits of the database. Compare this to the similarly free, professional site of HighWire Press which hosts the full text of nearly 2.5 million scholarly articles from high impact journals, including 900,000 open access articles. It provides not only well-structured and informative data about its publisher partners, journals, and archive composition, but also offers a sophisticated, still easy to use search and navigation software tailor-made for the exquisite archive.

Database size and dimensions

There are close to 35 million records in the edition of *WoS* which was used for this review. It covers the 1945–2005 period. There are other editions for different time periods, such as the largest one with coverage dating back to 1900 (also known as the ISI Century Database, with an additional set of 850,000 records), and a popular, less expensive edition with a time span from 1980 to 2005 with 25.5 million source items.

Size of the database in itself is not the decisive factor. The composition and source base, and the time span mentioned above are as important. The question is in what sense is the database large. It is obvious this 1945–2005 edition of *WoS* is larger in its first 20 years simply because *Scopus* starts it coverage substantially in 1965. (For space reason the chart does not include the 721,500 records in *WoS* between 1945 and 1954.) Figure 4 illustrates at a glance that *WoS* remains larger then *Scopus* until 1996, when *Scopus* catches up with *WoS* in terms of the number of records added per year, and gets slightly ahead of *WoS* in 2000.

The significant differences in the 2005 intake at the end of the first quarter of the year when the snapshot was taken, suggest that *WoS* is more current, i.e. adds records sooner to the database than *Scopus*.

In *G-S* the year range searches bring up unlikely hit numbers, and including the false results would have been misleading. The total number of records reported by the software is only 1,360,000 even when using the largest feasible time range from 868 to 2005 (in case Google, Inc. got the right for the Google Print project to digitize the Buddhist Diamond Sutra, believed to be the first printed book). This is about the number of items added to *Scopus* and *WoS* each, for papers published in 2003 alone.

In *G-S* there is no way to search reliably by publication year range, let alone by a combination of publication year and author name even though the advanced search template offers such a feature. A series of searches for each year between 1955 and 2005 seemed to provide a reasonable count of the records in *G-S* (although many records may not have the publication year, and for many others the page numbers and other 4-digit numbers seem often interpreted by *G-S* as publication year), and with much reservations it was included in Figure 4. (The chart does not include the roughly 230,000 records *G-S* may have for the period before 1955.) Results of sample searches could provide some clues, but the hit numbers reported by *G-S* for most test searches were implausible. For example, the number of records reported in *G-S* is more for the past 10 years than for the past 20 years. The search was done repeatedly, and yielded the same numbers. Going further back by years and decades produced similarly implausible results. Typically one would not make such year range searches without other criteria, but exactly such plausibility test searches reveal underlying problems in handling precious metadata. It also reveals absurd citedness scores, and/or misleadingly presented citedness scores.

The casual user would learn sooner or later that the result lists show short entries, displaying the title of the article, the name of the author(s), and the citedness of their paper. The top line for each result is typically linked to the abstract or full text of the article. The bottom line of each entry shows the name of the journal, or the imprint in case of books, and the additional links to sites where the abstract or the full text of the paper is (or may be) also available. Entries which start with the label [Citation] indicate that the information about the paper was extracted from the cited reference list of other articles.

The entries are redundant and are not easy to decipher. Instead of an informative abstract they often repeat the same information already shown in the hot-linked title of the entry. In Figure 5, it is enigmatic why there are two links to a site called hooklee.com in the third record. Both links have the same URL of the site of a scholar who works for the Center of Chaos Control and Synchronization. Searchers could relate to the chaos part but perhaps not to the control and synchronization parts in figuring out why *G-S* chose to crawl it and grace us with two identical links when it did not crawl millions of pages of legitimate open access archives of its partner publishers (such as *Nature*), and of other huge repositories. It is a common practice of *G-S* to display identically named links with identical URLs which does not reduce the infoglut. Such redundancy adds to the searchers' confusion. Still, at least these entries are not misinforming the searchers. They just discombobulate them.

But there are also absurd entries in many of the result lists, although not in such a prominent position as in Figure 6. The first search was run to return articles published between 1995 and 2005, and the second one for the 1985–2005 time period. It is not immediately clear how could there be fewer records for the almost twice as long second

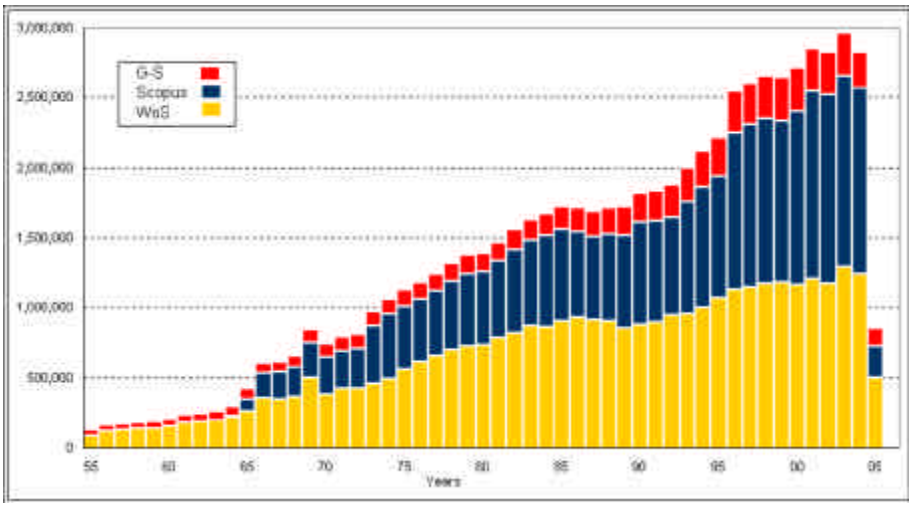

Figure 4. Distribution of records for source items in *Scopus*, *WoS* and *G-S*. (*G-S* data are rough estimates.)

[CITATION] Quality of abstracts
C Tenopir, **P Jacso** - Cited by 11 - Web Search
Online, 1993

[BOOK] Build your own database
P Jacso, FW Lancaster - Cited by 6 - Library Search - Web Search
Chicago: American Library Association, 1999

A deficiency in the algorithm for calculating the impact factor of scholarly journals: the journal ...
P Jacso - View as HTML - Cited by 5 - Web Search
Page 1. A DEFICIENCY IN THE ALGORITHM FOR CALCULATING THE IMPACT FACTOR OF SCHOLARLY JOURNALS: THE JOURNAL IMPACT FACTOR Péter Jacsó ...
Cortex, 2001 - masson.it - hooklee.com - hooklee.com - ncbi.nlm.nih.gov

Figure 5. Redundant entries in the result list.

time span than for the first one. Beyond that, the confusing and inconsistent layout of the result list, gives the impression that the article published in 2005 in *Infection and Immunity* is the one which already garnered 10,338 citations in its very short lifetime. Actually, it is the Towbin *et al*. article published 25 years ago, and known to be among the top cited papers from *PNAS*. That is where the link in the title, and the other links take the user to. It is an article cited by the paper in *Infection and Immunity*. It should not have been brought up for the first query which specified 1995–2005 for the data range. Such mix-up appears in many result lists of *G-S*.

Little pictures, big picture

Getting the big picture of the composition, size and dimension of databases is important for an overall sense of the databases. Getting little pictures for well-defined, unambiguous searches by subjects, authors, journals with or without additional limiters (such as document type, language) brings the issue to a human scale. Small result sets can be compared with relative ease. Knowing the subject field, its journals and/or the author's ouvre can help in finding the reasons for the differences. Together, the little pictures may corroborate the validity of the big picture – if the numbers are taken with a grain of salt. Sample searches by author, journal names, exact article titles, publication years are common in comparing traditional databases. Such searches can be used – as illustrated above – in the citation-based and citation-enhanced databases.

However, as their most special assets are the cited and citing references, these were the primary targets of most of the test searches for this article. Many of them landed on the cutting floor for reasons of the generous, but still limited space in this special issue, and the incomparability of the results. For example, there is not much to compare and analyse when searching for Eugene Garfield as an author. *WoS* brings up 1522 records, *Scopus* finds 90, and *G-S* reports 806 hits, settling to 781 as you proceed in the result list which include a variety of other authors with E as one of their middle initials, such as RE Garfield. At least these can be excluded with some efforts.

Searching by Garfield as cited author showed the prowess of the wider source base of *Scopus* combined with the inclusion of cited references from the past decade. It also showed the lasting impact of many of

Figure 6. Implausible hits and citedness score in the confusing results display.

Garfield's articles still intensely cited after decades. *WoS* found 1657 records, *Scopus* reported 1103. It was too large a set to analyse item by item, but it still was informative – especially when looking at the distribution of the set of citing journals, the range of citing years, and the diversity of subject fields. *G-S* offers only one search option on its template, not distinguishing between author and cited author.

Searching for documents citing a specific paper

Searching for papers citing the 1955 *Science* paper of Garfield (the test article) yielded insightful results. *WoS* had 215 hits, *Scopus* had 71, and *G-S* had 100. In *G-S* there were a few duplicates (such as same paper from different archives with slightly different titles of the citing article), and one undecipherable hit, which reduced the result to 95 (Figure 7). There were no efforts to try to herd all the citations with misspelled author name, volume, issue and page numbers.

The overall picture is somewhat familiar. *WoS* is the old reliable, and it is also coming in as the most comprehensive for the test article even after 1996 until 2000, then again in 2003. It is interesting to see that *G-S* got competitive results from 1996 (except for 1999), and then got ahead of the other two databases in 2001, 2002 and 2004. The chart indicates that in this case *Scopus* catches up with (and gets ahead of) *WoS* only in 2002 – still taking 2nd place after *G-S* two years in a row. An item-by-item analysis of the results provides more insights about the overlap, and the lack of it in some cases.

For the 1996–2005 period *WoS* found 83 records, *Scopus* found 76 and *G-S* found 82. This is quite close. However, when comparing the results item by item only 33 citing papers were common in the three result sets. There are several reasons for the many unique records (not detailed here for lack of space), but the morale of it is important: a single database cannot provide comprehensive citing coverage. It also shows clearly that Garfield's vision has been acknowledged more intensely in the past few years than ever before.

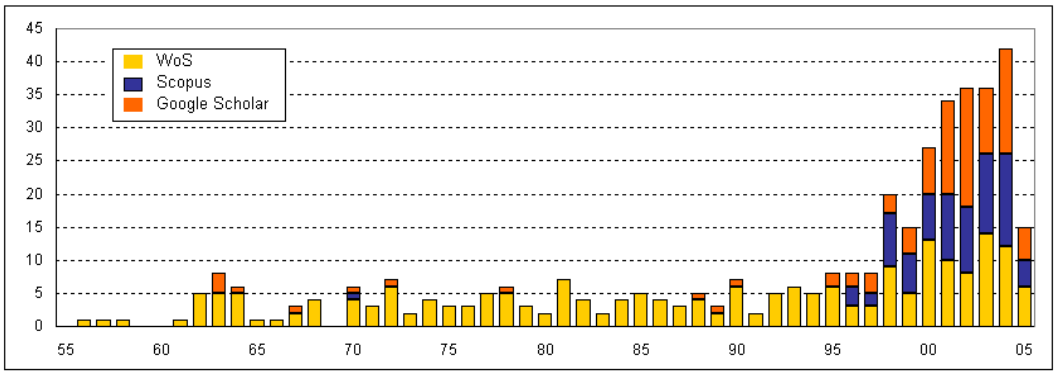

Figure 7. Number of articles in *WoS*, *Scopus* and *G-S* citing the Citation Indexes for Science' article.

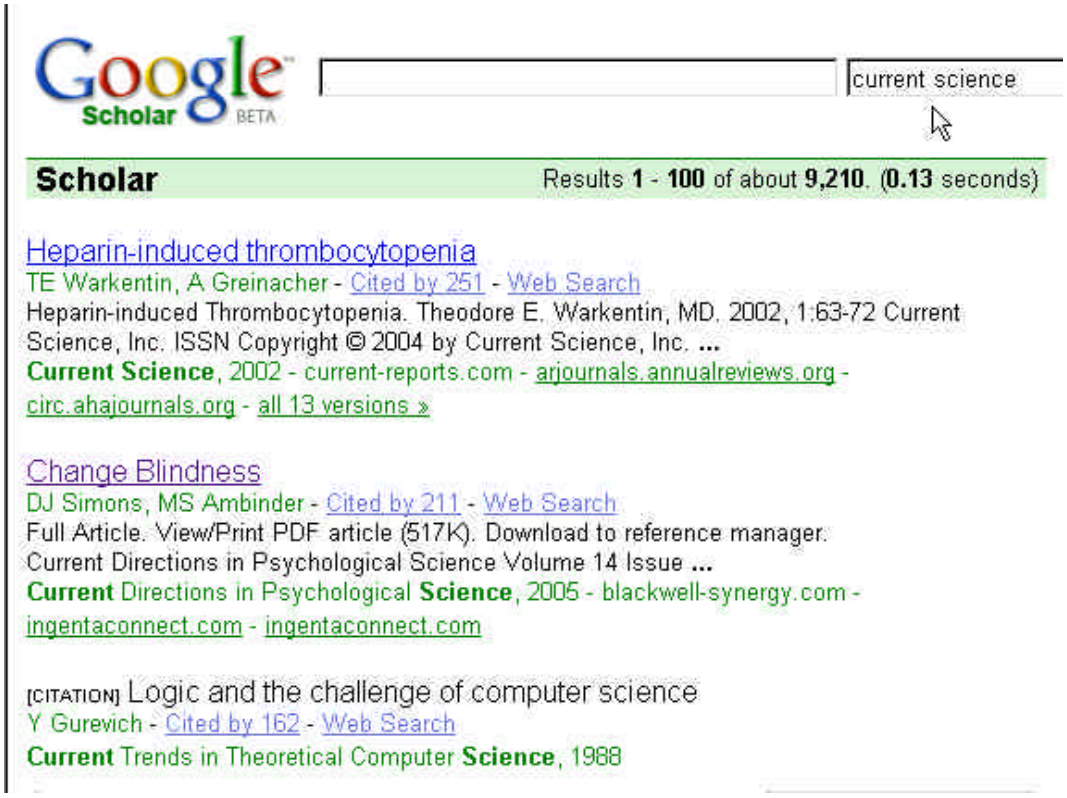

Figure 8. Thousands of false 'hits' for *Current Science*, the journal.

True, some of the hits from *G-S* were non-traditional and not particularly important literature sources, but the majority of the unique records in all databases were relevant and substantial. Importantly, *G-S* had links to the free full text versions for almost half of its find. Researchers at a well-endowed library with good link resolver software and blessed with competent and motivated systems librarians could achieve a better link-through rate to the primary documents from *WoS* and *Scopus*. Such an environment would also increase the rate of *G-S* which cannot deliver even the abstracts for non-subscribers from the archives of IEEE and ACM Digital Library.

Searching for a specific journal

Originally, I wanted to test the breadth of coverage of *Current Science*. However, the difference in the number of records was so great that it would have made it a futile exercise. Suffice it to say that *WoS* had 26,020 records, *Scopus* had 3657 records. They are comparable only for 2003 and 2004 when *WoS* had 771 and 709 records, and *Scopus* had 711 and 672 records, respectively. For 2005 articles *Scopus* had 36 records and *WoS* had 327. This reflects the overall currency of *WoS* at the time of the test. *Scopus* started the coverage of *Current Science* only in

Table 1.

Date	Document title	Author(s)	Cited by		
			WoS	Scopus	G-S
2001	The 2001 Kutch (Bhuj) earthquake:	Rajendran, K. ...	27	29	12
2000	First results from a new observational system	Premkumar, K. ...	17	17	
2000	Nanostructured thin films by self-assembly of	Sastry, M.	19	20	
2000	Small-angle neutron scattering diffractometer	Aswal, V.K. ...	30	30	
2000	Fundamental genomic unity of ethnic India is	Roychoudhury, S. ...	19	17	19
2000	Geodetic contributions to the study of	Bilham, R. ...	19	19	6
1999	Microbial lipases: Potential biocatalysts for	Saxena, R.K. ...	20	23	
1999	Solid state fermentation for the production of	Pandey, A.. ...	80	85	44
1999	Immobilized enzymes in bioprocess	D'Souza, S. F.	21	20	
1999	Persistence in nonequilibrium systems	Majumdar, S. N.	82	58	45
1999	Reactive oxygen species: Oxidative damage	Bandyopadhyay, U. ...	24	37	22
1999	Surface-enhanced Raman scattering: A new	Kneipp, K. ...	25	28	
1998	Detailed study report of Samta, one of the	Biswas, B.K.. ...	26	28	
1998	Calcutta's industrial pollution: Groundwater	Chakraborti, D. ...	14	17	
1998	Microwave radiation as a catalyst for	Sridar, V.	28	26	
1998	Geodetic constraints on the translation and	Bilham, R. ...	31	30	
1997	Groundwater arsenic calamity in Bangladesh	Dhar, R.Kr....	87	92	62
1997	Late Quaternary vegetational and climatic	Rajagopalan, G	25	26	7
1997	Applications of geographic information	Menon, S. ...	21	23	12
1997	Geomorphology and surficial geology of the	Rao, V.P. ...	17	20	
1996	Documenting diversity: An experiment	Gadgil, M.	20	24	13
1996	Microsatellites in plants: A new class of	Gupta, P. K. ...	48	51	55
1996	Arsenic in groundwater in seven districts of	Mandal, B. K. ...	99	107	58
1996	40Ar-39Ar ages of Anjar Traps, Western	Venkatesan, T. R. ...	25	22	1
1996	Recent advances in the chemistry of water-	Katti, K. V.	23	18	
1996	Graphical analysis of DNA sequence	Nandy, A.	19	19	
1996	Winter cooling in the northern Arabian Sea	Prasanna Kumar, S. ...	25	25	2
1996	Phytoplankton production and chlorophyll	Bhattathiri, P M.A.. ...	35	32	
1996	Bacterial abundance and production in the	Ramaiah, N. ...	23	20	
1996	Micelles: Self-organized surfactant	Moulik, S. P.	28	29	
Total source items found			**30**	**30**	**14**
Total citing references			**977**	**992**	**358**

the mid-1990s; before 1994 it had occasional records in some of the years. From 1996 to 2003 it had an intake of 40–60% of *WoS*. The test still served as a warning that the breadth of coverage of journals can be very different among databases in toto, and can widely fluctuate across the years.

G-S made it impossible to search for *Current Science* in an acceptable way, let alone browse the source journal index to pick variant names as it is possible in both *Scopus* and *WoS*. The latter also allows the browsing of the index of the cited journal names, including the abbreviated and variant spellings, such as the abbreviations (Curr Sci, Curr Science), and the place of publication qualifier (India and Bangalore) for *Current Science*.

The underlying problem is that the journal name index cannot be searched for an exact phrase. *G-S* removes the quotation marks around the search term when used in the 'published in' cell on the template. (This was fixed as I went to press.) This made it impossible to search for *Current Science* without retrieving a very large number of records from journals whose title includes the words Current and Science, such as *Current Directions in Psychological Science*, or *Current Trends in Theoretical Computer Science*. *G-S* delivers the coup de grace by mistaking the name of Current Science, Inc. for the journal name and floods the searchers with thousands of records for the various *Current Reports* series of Current Science, Inc. (Figure 8). This is another example for do-

ing a great disservice by ignoring the rich metadata, and essential search requirements. *G-S* does not allow the obvious solution of searching for the journal by using the site: ias.ac.in/currsci which option has become available recently in the regular *Google*.

Searching for the most cited articles from a journal

The top 30 most cited papers of *Current Science* were selected from *Scopus*, and known-item searches were made in the two other databases to see how the citedness of these articles compare. It favors *Scopus* if we assume that *WoS* may have articles from earlier years which give them a better chance to garner more citations. Sampling several years in *WoS* did not find higher cited articles from *Current Science*. Coverage of *Current Science* by *G-S* is abysmal. Had it been used for creating the most cited 30 articles would have brought in Harnad's excellent article from *Current Science* with 26 citations received (versus 10 and 8 in *WoS* and *Scopus*), but would have made no other remarkable difference except for lowering the threshold.

Table 1, which is sorted in reverse chronological order, speaks for itself. *G-S* has the source records for less than half of the articles. It is apparent that *WoS* and *Scopus* have almost identical citedness scores for many of the articles. It is also clear that *Scopus* has the edge for articles related to life sciences, while *WoS* leads for articles in chemistry and physics just as the pie chart about the record distribution in the entire database (Figure 1) suggested. The largest difference is in an economics paper. No wonder as *WoS* has much better coverage of the social sciences than *Scopus*.

The total number of citations received by these 30 articles illustrates even more pointedly the lack of predictability and breadth of *G-S* even for recent years. Except for the article about microsatellites in plants, where *G-S* has the most citations, it is always a distant third, and often an extremely distant third.

Conclusions

As Garfield[5] pointed out, traditional indexing has serious limits, and adding more indexers would not be a panacea. 'Were an army of indexers available, it is still doubtful that the proper subject indexing could be made.' He added that 'by using authors' references in compiling the citation index, we are in reality utilizing an army of indexers, for every time an author makes a reference, he is in effect indexing that work from his point of view'. Technology certainly improved the efficiency of some parts of human indexing, but the ever increasing indexing quotas of indexers, often makes the intellectual process look like an assembly line operation resulting in declining quality. Current developments validate his vision big time.

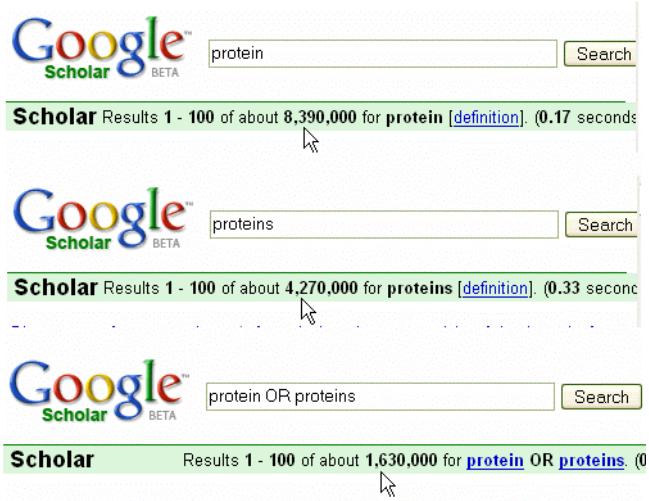

Figure 9. Fatal mistake in the simplest Boolean operation.

Gene Garfield stepped down as the president and chairman of ISI a decade ago, but he did not retire with his emeritus title. His latest project[37] with Russian scientists and programmers is to make the genesis of an idea and the invisible college of authors very visible by postprocessing results from *WoS* searches to show the direct and indirect intellectual relationships between authors based on the citations their papers gave and received.

I tested a beta version of the HistCite software[38] – among others – on the subject of citation indexes. It gave a remarkable view even in a flat matrix about the network of the 215 papers which cited the seminal article in *Science*. The graphic representation made the lians, vines and epiphytes of the citation forest highly visible.

This is not only the 50th anniversary of publishing his landmark *Science* article but also Garfield's 80th birthday. I am already looking forward to what shall I write about in 2015 which will be not only his 90th birthday but also the 50th anniversary of his conference paper in which he pondered about the feasibility of automating citation indexing[39]. This idea recently has become one of the hot issues in information science. Unfortunately, *G-S* gives a bad name to autonomous citation indexing. It shows lack of competence, and understanding of basic issues of citation indexing. *G-S* fails even in implementing the most basic Boolean OR operation correctly (Figure 9). Riding on the waves of the regular *Google* software which is great for processing the unstructured heap of billions of Web pages, *G-S* cannot handle even the meticulously tagged, metadata-enriched few million journal articles graciously offered to it by many publishers for free.

Some bright minds who designed and implemented autonomous citation indexes[40–42], and citation parsing tools[43] clearly proved that citation indexing can be automated successfully if one has some of the intellect, foresight, drive and stamina of Gene Garfield to whom I wish Happy 50th/80th.

1. Bush, V., As we may think. *Atlantic Monthly*, 1945, 101–108. http://sloan.stanford.edu/mousesite/Secondary/Bush.html
2. Thomson — Web of science databases. http://wos02.isiknowledge.com/help/h_database.htm
3. About SCOPUS. http://www.info.scopus.com/aboutscopus/contentcoverage/index.shtml
4. About Google Scholar. http://scholar.google.com/scholar/about.html
5. Garfield, E., Citation indexes for science. *Science*, 1955, **122**, 108–111. http://www.garfield.library.upenn.edu/essays/v6p468y1983.pdf
6. Garfield, E., Association-of-ideas techniques in documentation: Shepardizing the literature of science. Submitted as course work to: Research Information Center, National Bureau of Standards, 1954. http://www.garfield.library.upenn.edu/papers/assocofideasy1954.html
7. Weinberg, B. H., Who invented the index? — An agenda for research on information access features of Hebrew and Latin manuscripts. In Conference Proceedings of 66th IFLA Council and General, Jerusalem, Israel, 2000. http://www.ifla.org/IV/ifla66/papers/081-174e.htm
8. Garfield, E., 'Science Citation Index' – A new dimension in indexing. *Science*, 1964, **144**, 649–654. http://www.garfield.library.upenn.edu/essays/v7p525y1984.pdf
9. Cronin, B. and Barsky Atkins, H. (eds), *The Web of Knowledge: a Festschrift in Honor of Eugene Garfield.* ASIS Monograph Series. Medford, NJ, Information Today, 2000.
10. About Scirus. http://www.scirus.com/srsapp/aboutus/
11. Deis, L. and Goodman, D., Web of Science (2004 version) and Scopus. *Charleston Advisor* [online], 2005, **6**. http://www.charlestonco.com/comp.cfm?id=43
12. LaGuardia, C., E-views and reviews: Scopus vs. Web of Science. *Library J.* [online] (Jan 15, 2005). http://www.libraryjournal.com/index.asp?layout=articlePrint&articleID=CA491154
13. Myhill, M., Google Scholar review. *Charleston Advisor* [online], 2005, **6.** http://www.charlestonco.com/review.cfm?id=225
14. Jacsó, P., Web of science citation indexes. *Gale — Reference Reviews* [online] (Aug, 2004). http://www.gale.com/servlet/HTMLFileServlet?imprint=9999&region=7&fileName=/reference/archive/200408/webscience.html
15. Jacsó, P., Scopus. *Gale — Reference Reviews* [online] (Sep 2004). http://www.galegroup.com/servlet/HTMLFileServlet?imprint=9999&region=7&fileName=reference/archive/200409/scopus.html
16. Jacsó, P., Google Scholar. *Gale — Reference Reviews* [online] (Dec 2004). http://www.gale.com/servlet/HTMLFileServlet?imprint=9999&region=7&fileName=/reference/archive/200412/googlescholar.html
17. Jacsó, P., Google Scholar: the pros and the cons. *Online Inf. Rev.*, 2005, **29**, 208-214. DOI: 10.1108/14684520510598066 ABS
18. Jacsó, P., PsycINFO. *Gale — Reference Reviews* [online] (Jan 2004). http://www.gale.com/servlet/HTMLFileServlet?imprint=9999&region=7&fileName=reference/archive/200401/psycinfo.html
19. Jacsó, P., Citation enhanced indexing/Abstracting Databases. *Online Inf. Rev.*, 2004, **28**, 235 238. DOI**:** 10.1108/14684520410543689 ABS
20. Jacsó, P., E-psyche. *Gale – Reference Reviews* [online] (Jan 2004). http://www.gale.com/servlet/HTMLFileServlet?imprint=9999&region=7&fileName=reference/archive/200401/epsyche.html
21. Jacsó, P., Link-enabled cited references. *Online Inf. Rev.*, 2004, **28**, 306-311. DOI: 10.1108/14684520410553804 ABS
22. Jacsó, P., Citedness scores for filtering information and ranking search results. *Online Inf. Rev.*, 2004, **28**, 371–376. DOI:10.1108/14684520410564307 ABS
23. Jacsó, P., Citation Searching. *Online Inf. Rev. Rew.*, 2004, **28**, 454–460. DOI: 10.1108/14684520410570580 ABS
24. Jacsó, P., Browsing indexes of cited references. *Online Inf. Rev.*, 2005, **29,** 107–112. DOI**:** 10.1108/14684520510583972 ABS
25. Thomson — ISI citation products. http://www.isinet.com/cit/
26. SCOPUS FAQs http://www.info.scopus.com/aboutscopus/faqs/index.shtml
27. SCOPUS Info — List of journals. http://www.info.scopus.com/aboutscopus/documents/title_list.xls
28. Thomson — ISI journal list. http://www.isinet.com/journals/
29. SCOPUS — List of publishers. http://www.info.scopus.com/aboutscopus/documents/publisher_list.xls
30. SCOPUS — Open access journal list. http://www.info.scopus.com/aboutscopus/documents/oa_list.xls
31. Garfield, E., The concept of citation indexing: A unique and innovative tool for navigating the research literature. [online] http://scientific.thomson.com/knowtrend/essays/citationindexing/concept/
32. Garfield, E., History of citation indexing. [online] http://scientific.thomson.com/knowtrend/essays/citationindexing/history/
33. Google Scholar Help. http://scholar.google.com/scholar/help.html
34. Jain, N. C., *Scopus*™ has wider scope than *Science Citation Index*. *Current Science*, 2005, **88**, 331. http://www.ias.ac.in/currsci/feb102005/331.pdf
35. Prathap, G., Who's Afraid of Research Assessment? *Current Science*, 2005, **88**, 14-17. http://www.ias.ac.in/currsci/jan102005/14.pdf
36. Arunachalam, S., On publication indicators. *Current Science*, 2004, **86**, 629–632. http://www.ias.ac.in/currsci/mar102004/629.pdf
37. Garfield E., Pudovkin, A. I. and Istomin, V. S., Why Do We Need Algorithmic Historiography? *JASIST*, 2003, **54**, 400-412. http://www.garfield.library.upenn.edu/papers/jasist54(5)400y2003.pdf
38. Garfield E., Pudovkin, A. I. and Istomin, V. S., Algorithmic Citation-Linked Historiography — Mapping the Literature of Science. Presented the ASIS&T 2002: Information, Connections and Community. 65th Annual Meeting of ASIST in Philadelphia, PA. Nov. 18–21, 2002. http://www.garfield.library.upenn.edu/papers/asis2002/asis2002presentation.html
39. Stevens, M. E., Giuliano, V. E. and Heilprin, L. (eds), Can citation indexing be automated? In Symposium Proceedings, Statistical Association Methods for Mechanized Documentation, National Bureau of Standards Miscellaneous Publication, 1965, **269**, Washington, pp. 189–192. http://www.garfield.library.upenn.edu/essays/V1p084y1962-73.pdf
40. Bollacker, K. D., Lawrence, S. and Lee, C., CiteSeer: An autonomous web agent for automatic retrieval and identification of interesting publications. In Proceedings of 2nd International ACM Conference on Autonomous Agents, ACM Press, 1998, pp. 116–123. http://citeseer.csail.mit.edu/cache/papers/cs/209/http:zSzzSzwww.neci.nj.nec.comzSzhomepageszSzgileszSzpaperszSzACM98.Digital.Libraries.CiteSeer.pdf/giles98citeseer.pdf
41. Harnad, S. and Carr, L., Integrating, navigating, and analysing open eprint archives through open citation linking (the *OpCit* Project). *Curr. Sci.*, 2000, **79**, 629–638. http://www.ias.ac.in/currsci/sep102000/629.pdf
42. Hitchcock, S., Woukeu, A., Brody, T., Carr, L., Hall, W. and Harnad, S., Evaluating Citebase, an Open Access Web-based Citation-ranked Search and Impact Discovery Service. [online], 2003. http://opcit.eprints.org/evaluation/Citebase-evaluation/evaluation-report.html
43. Jewell, M., ParaTools Reference Parsing Toolkit - Version 1.0 Released. *D-Lib*, [online] 2003, **9**. http://www.dlib.org/dlib/february03/02inbrief.html#JEWELL