# Citation Indexing and Indexes†

Paula Carina de Araújo*, Renata Cristina Gutierres Castanha** and Birger Hjørland***

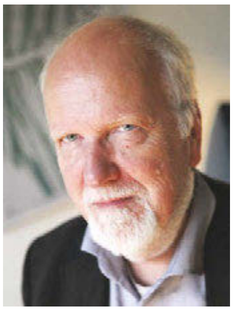
*Universidade Federal do Paraná, 632, Prefeito Lothário Meissner Avenue, Jardim Botânico – Curitiba,
Paraná, Brazil, 80210-070, <paulacarina@ufpr.br>

**Ambev, 1891, Antartica Ave, 13820-000, Jaguariúna, São Paulo, Brazil, <regutierres@gmail.com>

***University of Copenhagen, Department of Communication, 8 Karen Blixens Plads, DK-2300 Copenhagen S,
Denmark, <birger.hjorland@hum.ku.dk>

Paula Carina de Araújo is a professor at the Department of Information Science and Management, Universidade Federal do Paraná and at the Information Management Graduate Program at the Universidade do Estado de Santa Catarina since 2019 in Brazil. She holds a bachelor's degree in library science from the Universidade do Estado de Santa Catarina, a master's degree in information science, technology and management from the Universidade Federal do Paraná and a PhD in information science from the Universidade Estadual Paulista Júlio de Mesquita Filho (UNESP). She was a visiting scholar on a Fulbright Scholarship at the University of Washington researching epistemology of knowledge organization (2016-2017).

Renata Cristina Gutierres Castanha works as Data Product Manager for Ambev (Brazil, 2018-). She holds a bachelor's degree in mathematics from São Paulo State University (Unesp), a master's degree and PhD in information science from São Paulo State University (Unesp). She served as a visiting scholar at the School of Information Studies at the University of Wisconsin-Milwaukee (UWM), developing research on knowledge organization and multivariate statistical methods. Her research interests are scientometrics, domain analysis, academic genealogy, data science fundamentals and machine learning.



Birger Hjørland holds an MA in psychology and PhD in library and information science. He is professor emeritus in knowledge organization at the Department of Communication, University of Copenhagen (formerly Department of Information Studies and Royal School of Library and Information Science) since 2001 and at the University College in Borås 2000-2001. He was research librarian at the Royal Library in Copenhagen 1978-1990, and taught information science at the Department of Mathematical and Applied Linguistics at the University of Copenhagen 1983-1986. He is a member of ISKO Scientific Advisory Council, the editor-in-chief of the ISKO Encyclopedia of Knowledge Organization and a member of the editorial boards of Knowledge Organization, Journal of the Association for Information Science and Technology and Journal of Documentation. His h-index on 2020-05-30 is 49 in Google Scholar and 29 in Web of Science.

Araújo, Paula Carina de, Renata Cristina Gutierres Castanha and Birger Hjørland. 2021. "Citation Indexing and Indexes." *Knowledge Organization* 48(1): 72-101. 123 references. DOI:10.5771/0943-7444-2021-1-72.

**Abstract:** A citation index is a bibliographic database that provides citation links between documents. The first modern citation index was suggested by the researcher Eugene Garfield in 1955 and created by him in 1964, and it represents an important innovation to knowledge organization and information retrieval. This article describes citation indexes in general, considering the modern citation indexes, including Web of Science, Scopus, Google Scholar, Microsoft Academic, Crossref, Dimensions and some special citation indexes and predecessors to the modern citation index like Shepard's Citations. We present comparative studies of the major ones and survey theoretical problems related to the role of citation indexes as subject access points (SAP), recognizing the implications to knowledge organization and information retrieval. Finally, studies on citation behavior are presented and the influence of citation indexes on knowledge organization, information retrieval and the scientific information ecosystem is recognized.

Received: 5 August 2019; Revised: 1 September 2019; Accepted: 20 September 2019

Keywords: citation index, citations, databases, data, science, references, journals

† Derived from the article of similar title in the ISKO Encyclopedia of Knowledge Organization, version 1.0 published 2019-08-05. The authors wish to thank three anonymous reviewers for a detailed and important feedback on a former version of the article. Ziyoung Park provided up to date information on KCI. Initially Birger Hjørland was encyclopedia editor for this article, however he joined as co-author in the process and thereafter Claudio Gnoli served as encyclopedia editor of this article. The online version has been updated in relation to the printed version.

## 1.0 The idea of a citation database

Scientific and scholarly authors normally cite other publications. They do so by providing bibliographical references to other documents in the text and elaborating them in a special "list of references" (as in this encyclopedia article) or in footnotes. (Such references are also often in the bibliometric literature termed "cited references"). When reference is made to another document, that document receives a citation. As expressed by Narin (1976, 334, 337): "a citation is the acknowledgement one bibliographic unit receives from another whereas a reference is the acknowledgement one unit gives to another."[1] While references are made within documents, citations are received by other documents.[2] References contain a set of standardized information about the cited document, which allows its identification (as, for example, the references in the present article).[3]

A citation index is a paper-based or electronic database that provides citation links between documents. It may also be termed a reference index, but this term is seldom used,[4] and in the following we use the established term: citation index.[5]

It has always been possible to trace the references a given document makes to earlier documents (so-called backward searching). A citation index, however, makes it possible to trace the citations (if any) that a given document receives from later documents (so-called forward searching),[6] dependent of which documents has been indexed. Examples:

1. Nyborg (2005) is an article about sex differences in general intelligence (g), that concluded: "Proper methodology identifies a male advantage in g that increases exponentially at higher levels, relates to brain size, and explains, at least in part, the universal male dominance in society."
   If you would like to see whether this conclusion has been challenged or rejected by other researchers, you need only look up Nyborg's paper up in a citation index.[7]
2. If you would like to see whether somebody has used your published ideas in their research, you may look yourself up in a citation index.
3. If you would like to see whether a certain person or work is cited within a given field, you may look that person up and limit your search to that field. This way you may, for example, see which papers in library and information science have cited any work by Michel Foucault or the specific reference Garfield (1980).

No doubt, citation indexes are very important tools that have revolutionized the way we can search for information. This article focuses on the function of citation indexes to assist researchers identify useful and relevant research. Citation indexes are, however, increasingly used to evaluate research and researchers, and this function may influence how they are developing and thus also their functionality for document searching.

## 2.0 The principles and design of citation indexes

In the words of Weinstock (1971, 16): "a citation index is a structured list of all the citations in a given collection of documents. Such lists are usually arranged so that the cited document is followed by the citing documents." It is the scientist (or scholar) who creates the citations not the citation indexes as it has been claimed,[8] and the role of citation indexes is to make the citations findable.

McVeigh (2017, 941) explains that "a true citation index has two aspects [or parts]—a defined source index and a standardized/unified cited reference index." In Figure 1, on the left, two articles are shown. These articles are represented in the part of the citation index called the "source index." For each article, a long range of metadata is provided, including author names, title of article, title of journal, and the list of bibliographical references contained in the article. The source index is, therefore, a comprehensively described set of the indexed materials from which cited references will be compiled.

On the right in Figure 1, a list of references A-J, derived from the source index, are shown. These references represent the "cited reference index," where each reference points back to the article in which it occurs (and back to the source index). A citation index is thus—in the words of McVeigh (2017, 941) "derived from a two-part indexing of source material. Bibliographic entries are created for each source item; cited references are captured into a separate index, where identical references are unified. The resulting two-part structure is the basic architecture of a citation index."

Figure 2 shows an example: among Anders Ørom's many publications, two have been indexed by Web of Science, an article in *Knowledge Organization*, and another in *Journal of Documentation*. The figure also shows how many times each article has been cited (on July 8, 2019); the citing articles will be displayed by clicking on the number. However, as shown below, not all citing articles have been captured (in

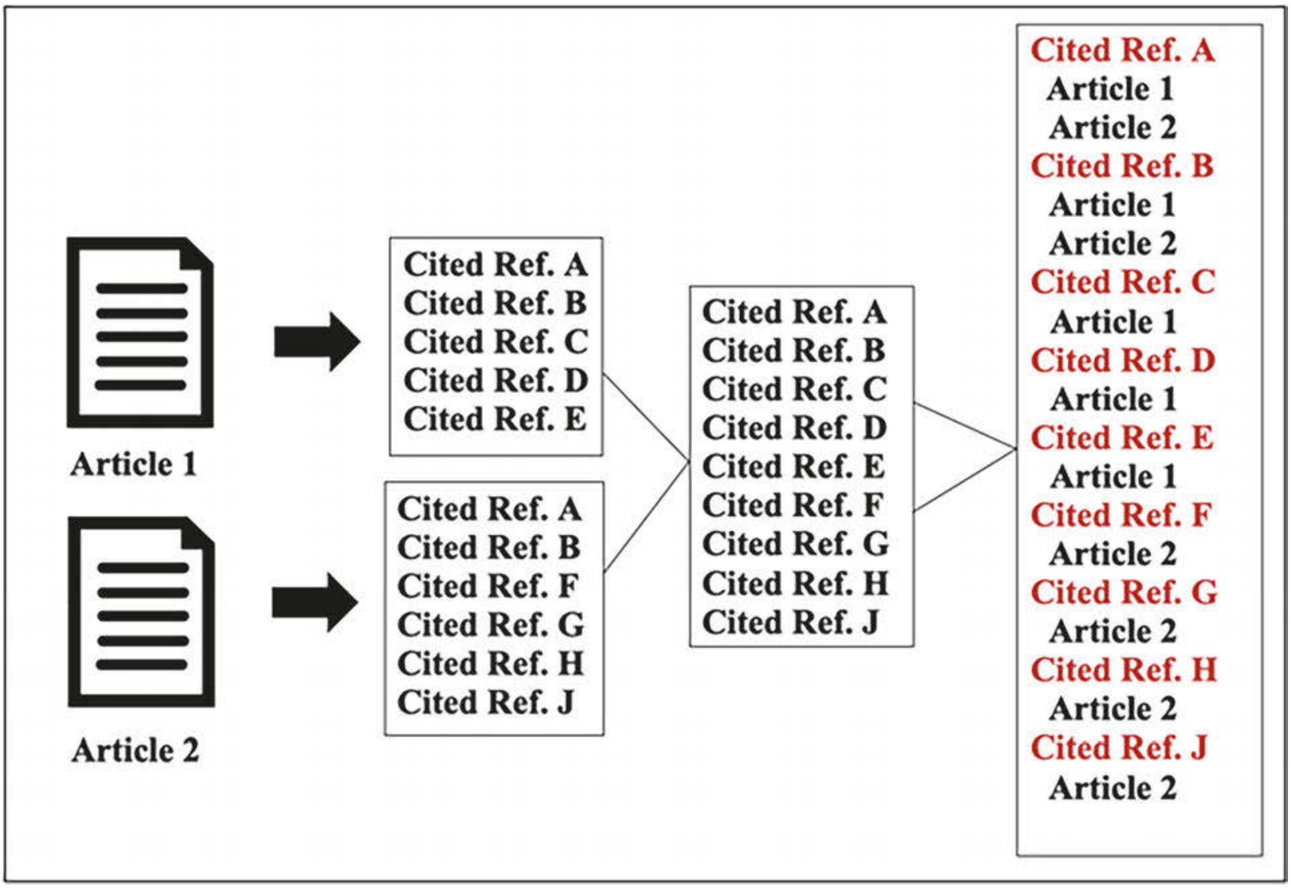
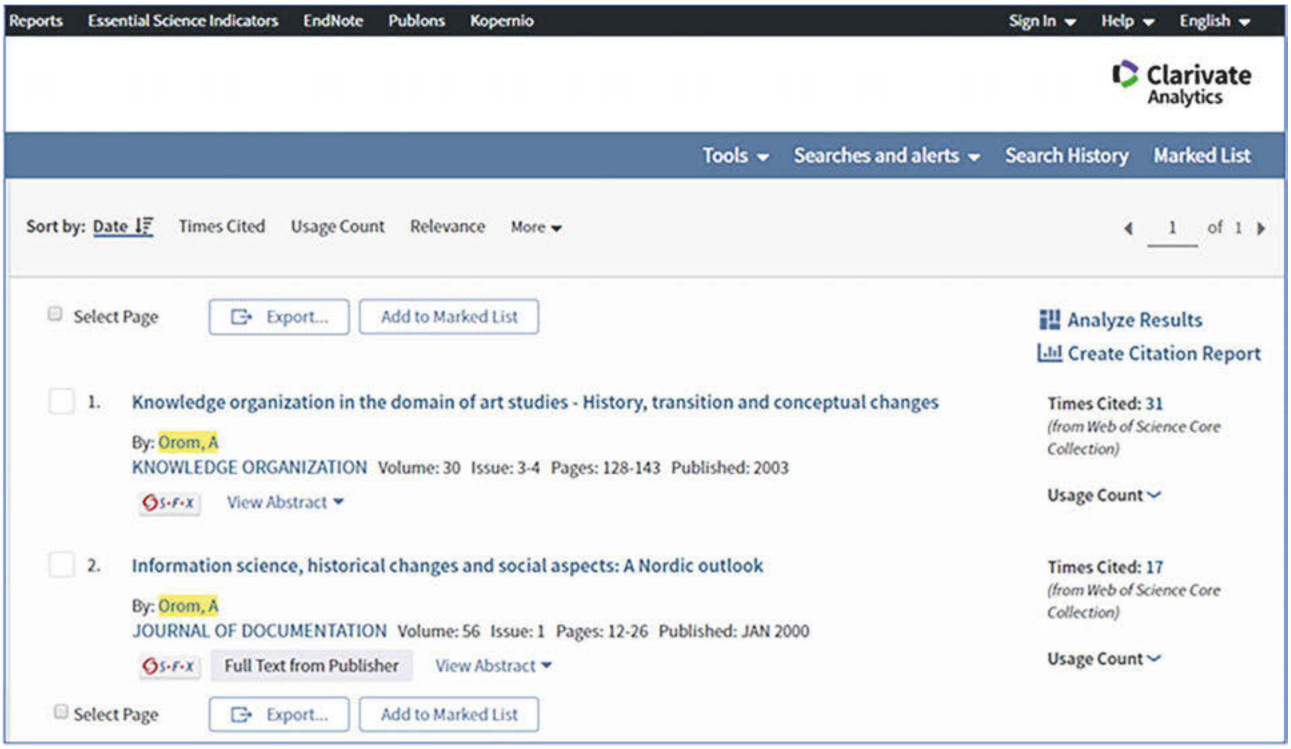


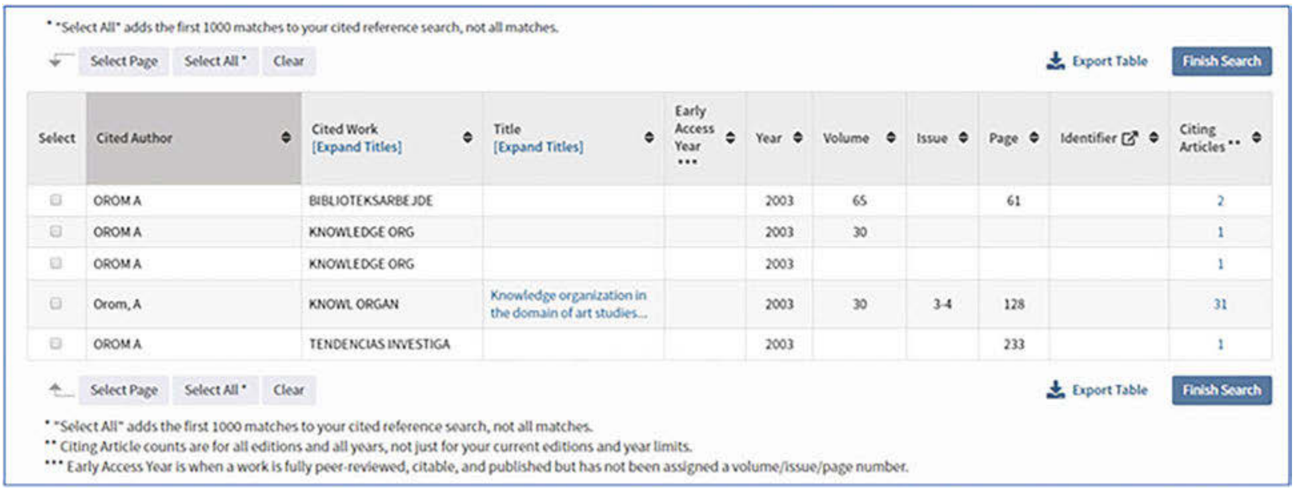
*Figure 1*. Citation index design based on McVeigh (2017, 941).



1. Knowledge organization in the domain of art studies - History, transition and conceptual changes
By: Orom, A
KNOWLEDGE ORGANIZATION Volume: 30 Issue: 3-4 Pages: 128-143 Published: 2003
Times Cited: 31
(from Web of Science Core Collection)

2. Information science, historical changes and social aspects: A Nordic outlook
By: Orom, A
JOURNAL OF DOCUMENTATION Volume: 56 Issue: 1 Pages: 12-26 Published: JAN 2000
Times Cited: 17
(from Web of Science Core Collection)

*Figure 2*. Look-up Ørom A*[9] as author in WoS, source index.



* "Select All" adds the first 1000 matches to your cited reference search, not all matches.

| Select | Cited Author | Cited Work [Expand Titles] | Title [Expand Titles] | Early Access Year *** | Year | Volume | Issue | Page | Identifier | Citing Articles ** |
|---|---|---|---|---|---|---|---|---|---|---|
| □ | OROM A | BIBLIOTEKSARBE JDE | | | 2003 | 65 | | 61 | | 2 |
| □ | OROM A | KNOWLEDGE ORG | | | 2003 | 30 | | | | 1 |
| □ | OROM A | KNOWLEDGE ORG | | | 2003 | | | | | 1 |
| □ | Orom, A | KNOWL ORGAN | Knowledge organization in the domain of art studies... | | 2003 | 30 | 3-4 | 128 | | 31 |
| □ | OROM A | TENDENCIAS INVESTIGA | | | 2003 | | | 233 | | 1 |

* "Select All" adds the first 1000 matches to your cited reference search, not all matches.
** Citing Article counts are for all editions and all years, not just for your current editions and year limits.
*** Early Access Year is when a work is fully peer-reviewed, citable, and published but has not been assigned a volume/issue/page number.

*Figure 3*. A. Ørom as cited author and printing year = 2003 in WoS, cited reference index (forward searching).

the article in *Knowledge Organization*, only the thirty-one citations that have been "unified" are included).

If you would like to see which references are cited by Ørom (2003) (i.e., perform a backward search), you can get a copy of the article itself and see its list of references; there is also the possibility to select the full record in the source index (not shown) and here click at "cited references."[10]

Figure 3 shows a corresponding example of forward searching from WoS: five references that matched the query "references citing Ørom (2003)."

However, three of the five references in Figure 3 are to the same article in *Knowledge Organization*, in other words, there are three cited reference variants. This specific article has been cited 1+1+31 = 33 times (on July 8, 2019). One of the three reference variants provide the title of the paper, the issue and the starting page; also, in this reference the information that the article has thirty-one citing articles is given (the same thirty-one as in Figure 2). This means, that of the thirty-three citing references, the thirty-one have been "unified," but the unification algorithm has not been able to unify two of the variants. By selecting all three cited reference variants to the article in *Knowledge Organization*, a list of all thirty-three citing references can be displayed (this example is very simple with few cited reference variants, but often searches provide very many variants). We see that the unification makes the use of the citation index simpler.[11] McVeigh (2017, 941) writes that the data collected from those references are standardized to allow like citations to be collected or unified (McVeigh 2017, 941). We saw in Figure 3 what is meant by unification of citations (and that some citations were not unified).

It is important to understand how citations are not being unified—and that the concepts unification and control in bibliographic databases are relative concepts. Contrary to typical library catalogs, for example, citation indexes do not provide standardized author names.[12] The *Science Citation Index*, for example, is based on a rather mechanical indexing of both metadata in the source index and references in the cited reference index based on the data as given by the source documents themselves. This means that if an author sometimes uses two initials and sometimes only one initial, his writings are not unified (and both author searching and cited reference searching may be difficult). This is especially a problem when authors have common names like A. Smith or when names are spelled in many ways in source documents such as, for example, the names of many Russian authors like Lev Vygotsky (compare the concept "stray citations" in Section 6.3).

McVeigh (2017, 943) emphasizes that a citation index is more than just a bibliographic resource with linked cited references. It is the structured, standardized data in the cited reference index, independently of the source index, that for her defines a citation index.

## 3.0 Classifications of existing citation databases

In this article citation indexes are presented in the following order:

- Section 4 presents the *Science Citation Index* and later *ISI/Clarivate Analytics* citation indexes in chronological order;
- Section 5 presents citation databases from other database producers (except regional databases) in chronological order;
- Regional citation databases (and databases in other languages than English) are presented in Appendix 1 in chronological order;
- Section 6 presents comparative studies of six major competing citation indexes: Crossref, Dimensions, Google

Scholar (GS), Microsoft Academic (MA), Scopus and Web of Science (WoS);

- Predecessors to the *Science Citation Index* are presented in Section 7 (placed here because the interest in historical studies of former citation indexes was a response to the great interest that arose in the ground swell of Garfield's citation indexes.)

Citation indexes may also be classified as follows (see the endnotes for the specific titles in each category):

1) By subject coverage: a) universal citation databases;[13] b) databases covering science;[14] c) databases covering social sciences;[15] d) citation databases covering the humanities;[16] and, e) other subject specific citation databases.[17]
2) Citation databases covering specific document types: a) books;[18] b) conference proceedings;[19] and, c) data sets.[20]

The most important databases are placed in separately numbered sections. For each database, some standardized information is given (such as date of launching) together with presentation of relevant literature about that database. At the end of each description, there are links to the homepage of the database and list of journals or other sources covered (as bulled lists).

### 4.0 The *Science Citation Index* and other ISI/Clarivate Analytics citation indexes

The American government stimulated the development of scientific research soon after World War II. Considering the fast-growing volume of scientific literature and their concern regarded to the systems for information exchange capacity among scientists, the government sponsored many projects related to the improvement of methods for distributing and managing scientific information. Eugene Garfield was a member of the study team at Johns Hopkins Welch Medical Library sponsored by the Armed Forces Medical Library. Because of that experience "I [Garfield] became interested in whether and how machines could be used to generate indexing terms that effectively described the contents of a document, without the need for the intellectual judgments of human indexers"[21] (Garfield 1979, 6).

Garfield's experience working in that project, his experience doing voluntary abstracting work for *Chemical Abstracts* and the fact he learned that there was an index to the case literature of the law that used citations (Shepard's Citations see Section 7.1), led him to create the first modern citation index. He presented his idea of the citation index in Garfield (1955). Garfield's company, Institute for Scientific Information (ISI) was founded in 1960 in Philadelphia, and in 1964, Garfield published the first *Science Citation Index* (SCI; see further on SCI in Section 4.1). ISI has shifted ownership and name many times and is today known as Clarivate Analytics.[22] This company has over the years created a suite of citation indexes to be presented below. Because of the many name shifts, it may be referred to by different names, such as ISI's, Thomson Scientific's or Clarivate Analytic's citation indexes.

Web of Science (WoS)[23] is a platform created in 1997 consisting of databases designed to support scientific and scholarly research. It contains several databases, which can be searched together (but not all of them are citation indexes). They can be grouped as follows (see endnotes for lists of all databases in each group):

a) WoS Core Collection (e.g., SCI).[24]
b) Databases produced by ISI/Clarivate Analytic but not included in the Core Collection (e.g., Data Citation Index).[25]
c) Other databases hosted by WoS and produced by other data providers (e.g., Russian Science Citation Index).[26]

The WoS platform can be considered a modernized version of the SCI. Its citation databases are further described below in the order of their launching.

### 4.1 Science Citation Index (SCI) / Science Citation Index Expanded (SCIE)

SCI was officially launched in 1964. SCIE, is—as the name indicates—a larger version of the SCI.[27] After its launching, the SCI, and most other citation indexes, have expanded retroactively, so that year of launching does not tell what years are searchable (SCI contains in the time of writing references back to 1898). The SCI was founded on some ideas and practical considerations that contrasted it with the major subject bibliographies at the time:

1) It was covering all scientific disciplines, not just one (or a set of related fields), and it, therefore, had to be more selective in its selection of journals (and other documents) to be covered. This was done by applying "Bradford's law of scattering," which Garfield (1971) modified to "Garfield's law of concentration." Sugimoto and Larivière (2018, 25-26) emphasize that

> Garfield never sought for the SCI to be an exhaustive database; rather, he aimed to curate a selective list of journals that were the most important for their respective disciplines. He used Bradford's law of scattering as the core principle. This law dictates that the majority of cited literature in a domain is concentrated in a few journals and, conversely, that the majority of scholarly journals contain a minority of cited documents.

By implication, a citation index needs only to index relatively few journals in order to display a majority of the references. Initially, SCI covered 613 journals; today it indexes 9,046 journals showing data from 1898 to present with complete cited references (cf., Web of Science 2018). The journals are mainly selected by use of the "journal impact factor" (JIF), which is a bibliometric index that reflects the yearly average number of citations to recent articles published in a given journal. It is frequently used as a proxy for the relative importance of a journal within its field. It was devised by Eugene Garfield and is produced yearly by ISI/Clarivate Analytics; Scopus has developed an alternative called Scimago Journal Rank (SJR).

2) In contrast to the ordinary bibliographical databases, SCI did not use human indexers for individual articles (just for an overall classification of journals). It was based on information contained in the articles themselves (derived indexing), without any human assigned information, and it relied to a high degree of mechanical rather than intellectual indexing. It was assumed that searching the titles (translated to English) and the references/citation would be sufficient to make it an important tool (also, in contrast to the dominant documentation databases, SCI did not bring abstracts of the indexed articles at its start; however, abstracts taken from the articles were added from 1971, and thereby the difference between "abstract journals" and citation indexes was reduced).

3) Journals indexed in SCI were indexed "cover-to-covers" not selectively. This removed another source of subjective choice in the indexing process, as users could rely on journals being fully indexed.

Braun, Glänzel and Schubert (2000) examined how different disciplines, countries and publishers are represented in the SCI by comparing the subset of journals indexed here with the number of journals covered by Ulrich's International Periodicals Directory from 1998. They found that (254), in average, SCI covered 9.83% of the journals in Ulrich's, but some fields were overrepresented (physics 27.4%, chemistry 26.3%, mathematics 25.0%, biology 23.9%, pharmacy and pharmacology 15.0%, medical sciences 14.8%, engineering 14.5% and earth sciences 12.8%). 17 other fields were found to be underrepresented compared to the 9.83% (in decreasing order): environmental studies 9.4%, computers 8.9%, metallurgy 8.6%, energy 6.4%, public health and safety 5.9%, sciences: comprehensive works 5.0%, petroleum and gas 5.0%, agriculture 5.0%, food and food industries 4.7%, forests and forestry 4.4%, psychology 4.0%, aeronautics and space flight 3.6%, technology: comprehensive work 3.1%, geography 2.4%, gardening and horticulture 1.4%, transportation 0.6% and finally building and construction 0.4%. See also:

- *Web of Science Fact Book* (Clarivate Analytics, undated).
- Clarivate Analytics [2019]. Science Citation Index Expanded: Journal List.: http://mjl.clarivate.com/cgi-bin/jrnlst/jlresults.cgi?PC=D

### 4.2 Social Sciences Citation Index (SSCI)

SSCI was established in 1973 and indexes, according to Web of Science (2018), 3,300 journals showing data from 1900 to present with complete cited references. It is possible to search in an entire century of information in one place, across fifty-five disciplines of the social sciences.

While most physical science research papers are universal in their interest and published in international journals in English, much research from the social sciences tends to be of primary interest to readers from the authors' country, and often it is published in a national language and in journals not processed for the SSCI (cf., Lewison and Roe 2013); these authors examined SSCI's coverage of journals from different countries. They concluded that their results can only be regarded as rather approximate but that it is apparent that the shortfall in coverage is real and quite large and biggest for Russia, Poland and Japan, somewhat smaller for Italy, Spain and Belgium, less again for the Scandinavian countries and least for the Anglophone countries (Australia, Canada, the UK), as would be expected. Klein and Chiang (2004) found that there is evidence of bias of an ideological nature in SSCI coverage of journals. See also:

- SSCI website: http://mjl.clarivate.com/scope/scope_ssci/
- SSCI journal list: http://mjl.clarivate.com/publist_ssci.pdf

### 4.3 Arts & Humanities Citation Index (A&HCI)

The A&HCI was established in 1978. It demonstrated some ways in which the humanities differ from science and social sciences, for example in the use of many implicit citations which needs to be formalized by the indexing staff (Garfield 1980). It indexes today 1,815 journals showing data from 1975 to present with full cited references including implicit citations (citations to works found in the body text of articles and not included in the bibliography, e.g., works of art). Garfield (1977b) suggested—just before A&HCI was started—how this index might benefit the humanities.

It has long been felt that adequate coverage is more problematic in A&HCI compared to SCI and SSCI, and this was a major reason for the European Science Foundation to initiate the development of the European Reference Index for the Humanities (ERIH).[28] Sivertsen and Larsen (2012) considered the lower degree of concentration in the literature of the social sciences and humanities (SSH) and concluded that the concentration is strong enough to make citation indexes feasible in these fields. See also:

– Wikipedia. "A&HCI" at: https://en.wikipedia.org/wiki/Arts_and_Humanities_Citation_Index
– A&HCI journal list: http://mjl.clarivate.com/cgi-bin/jrnlst/jlresults.cgi?PC=H

### 4.4 Conference Proceedings Citation Index (CPCI)

The CPCI was established in 2008. It was preceded by some conference proceeding indexes from ISI, which were not citation indexes (Garfield 1970, 1977a, 1978 and 1981). CPCI indexes now 197,792 proceedings within two main sub-indexes: Conference Proceedings Citation Index: Science, (CPCI-S) and Conference Proceedings Citation Index: Social Science & Humanities (CPCI-SSH) (Web of Science 2018). The proceedings selection process is described by Testa (2012) (there is at present no specific homepage for this database at Clarivate Analytics).

### 4.5 Book Citation Index (BKCI)

BKCI was established 2011. According to Web of Science (2018), it currently indexes 94,066 books from 2005 to present. Clarivate Analytics (2018)[29] wrote:

> For the coverage of *Book Citation Index*, each book is evaluated on a case by case basis. The focus is on scholarly, research-oriented books for product. Once a book is selected, both the chapters and the book itself will be indexed. The index page is the guide for the book, so if available, the contents of the index page and all the references will be included. If the book is selected, the full book is indexed. There is no selective coverage. This means that there is no selective indexing of only a few chapters of a selected book into *Book Citation Index* - without indexing the whole book in BKCI.
> **Coverage of revised edition of a book in** *Book Citation Index*
> Revised edition is picked up only if there is new material to present. The revised version will be selected only if it presents new content. If the 2nd edition does not present 50% or more new content, it will not be selected.
> The focus is for new content, never published in the products. In fact, the revised version may be less than 1 percent of coverage.

Leydesdorff and Felt (2012a; 2012b) are two versions of the same study of BKCI. It found that books contain many citing references but are relatively less cited, which may find its origin in the slower circulation of books than of journal articles and that the reading of books is time consuming. The introduction of BKCI "has provided a seamless interface to WoS."

Torres-Salinas, Robinson-Garcia and López-Cózar (2012) analyzed different impact indicators referred to the scientific publishers included in the *Book Citation Index* for the social sciences and humanities fields during 2006-2011. They constructed "Book Publishers Citation Reports" and presented a total of nineteen rankings according to the different disciplines in humanities, arts, social sciences and law with six indicators for scientific publishers.

Gorraiz, Purnell and Glänzel (2013) wrote that BKCI was launched primarily to assist researchers identify relevant research that was previously invisible to them, because of the lack of significant book content in the WoS. The authors found that BKCI is a first step towards creating a reliable citation index for monographs, but that it is a very challenging issue because of the special requirements of this document type. Among the problems mentioned is, that books, in contrast to journal articles, seldom provide address information on authors. Therefore, in its current version (at the time of writing their article), the authors found that BKCI should not be used for bibliometric or evaluative purposes.

Torres-Salinas et al. (2013) used the BKCI to conduct analyses that could not have been done without this new index. The authors constructed "heliocentric clockwise maps" for four areas (disciplines): arts and humanities, science, social sciences and engineering and technology. For each area, citation average values for the dominant publishers are calculated and displayed. It was found, for example, that the area of engineering and technology is greatly unbalanced, because one publisher, Springer, dominates the area accumulating approximately 62% of the total share, that is, 28,000 book chapters of the total of 40,000 belong to this publisher. Other fields may also be unbalanced but not to such extent. Torres-Salinas, Robinson-Garcia, Campanario and López-Cózar (2014) provided descriptive information about BKCI and found:

> Humanities and social sciences comprise 30 per cent of the total share of this database. Most of the disciplines are covered by very few publishers mainly from the UK and USA (75.05 per cent of the books), in fact 33 publishers hold 90 per cent of the whole share. Regarding publisher impact, 80.5 per cent of the books and chapters remained uncited. Two serious errors were found in this database: the Book Citation Index does not retrieve all citations for books and chapters; and book citations do not include citations to their chapters.

Zuccala, Breum, Bruun and Wunsch (2018) studied the metadata assigned to monographs in BKCI and found that many ISBNs are missing for editions of the same work, in particular "emblematic" (original/first) editions. The authors wrote:

> The purpose of including all ISBNs is to ensure that every physical manifestation of a monograph is recognized (e.g., print, paperback, hardcopy, e-print) and that each ISBN is indexed as part of the correct edition or expression. This, in turn, ensures that all monograph editions can clearly be identified as being part of the same intellectual contribution, or work. Thus, publication counts, and citation counts would be more accurate in the BKCI, and new metric indicators could be calculated more effectively.

See also:

- BKCI website: http://wokinfo.com/products_tools/multidisciplinary/bookcitationindex/
- BKCI master book list: http://wokinfo.com/mbl/

### 4.6 Data Citation Index

The Data Citation Index was established in 2012 by Thomson Reuters as a point of access to quality research data from repositories across disciplines. As data citation practices increase over the years, the new citation index based on research data is available through the WoS from Clarivate Analytics. The Data Citation Index is a tool designed to be a source of data discovery for sciences, social sciences and arts and humanities. Data Citation Index evaluates and selects repositories considering the content, persistence, stability and searchability. Then, data is organized into three document types: repository, data study and data set. In this index, descriptive records are created for data objects and linked to literature articles in the Web of Science.

The Data Citation Index emerges at a time in which data sharing is becoming a hot issue. Many researchers find, however, that data sharing is time consuming and too little acknowledged by colleagues and funding bodies. They are not sure whether the practice of sharing data is worth it as they are time-consuming and are not acknowledged by colleagues and funding bodies. Therefore, Force and Robinson (2014) explain that Data Citation Index aims to solve four key researcher problems: 1) data access and discovery; 2) data citation; 3) lack of willingness to deposit and cite data; and, () lack of recognition and credit.

Torres-Salinas, Martín-Martín and Fuentes-Gutiérrez (2014, 6) analyzed the coverage of the Data Citation Index considering disciplines, document types and repositories. Their study acknowledges that the Data Citation Index is heavily oriented towards the hard sciences and it uses three documents types: data set, data study and repository. Furthermore, four repositories represent 75% of the database, even though there are a total of 29 repositories that contain at least 4,000 records. We believe that the bias to hard sciences and the concentration on a few repositories is related to the data sharing practices that is relatively common in medicine, genetics biochemistry and molecular biology, for example.

A study presented in 2014 demonstrated that data citation practices are uncommon within the scientific community, since 88% of the data analyzed had received no citation. We can state that "data sharing practices are not common to all areas of scientific knowledge and only certain fields have developed an infrastructure that allows to use and share data" (Torres-Salinas, Martín-Martín and Fuentes-Gutiérrez 2014, 6). The pattern of citation also changes from one domain to another. "While in Science and Engineering & Technology citations are concentrated among datasets, in the Social Sciences and Arts & Humanities, citations are normally referred to data studies" (Torres-Salinas, Martín-Martín and Fuentes-Gutiérrez 2014, 6).

Data Citation Index is an initiative that "continues to build content and develop infrastructure in the interest of improving attribution for non-traditional research output and enabling data discoverability and access" (Force and Robinson 2014, 1048). It is a new tool that can help to argue with researchers about the importance of sharing their data in order to be cited. Furthermore, "encouraging data citation and facilitating connections between datasets and published literature, the resource elevates datasets to the status of citable and standardized research objects" (1048). See also:

- Data Citation Index website: https://clarivate.com/products/web-of-science/web-science-form/data-cita tion-index/
- Data Citation Index master data repository list: https://clarivate.com/master-data-repository-list

### 4.7 Emerging Source Citation Index (ESCI)

ESCI was established in 2015. It is a database that, according to Clarivate Analytics (2017), indexes 7,280 emerging journals (journals that are not yet considered to fulfill the requirements of SCI, SSCI & AHCI) from 2005 to present with complete cited references. "Journals in ESCI have passed an initial editorial evaluation and can continue to be considered for inclusion in products such as SCIE, SSCI, and AHCI, which have rigorous evaluation processes and selection criteria." ESCI is also (in 2019) described as covering "new areas of research in evolving disciplines, as well as relevant interdisciplinary scholarly content across rapidly changing research fields."

ESCI journals do not receive an impact factor but are evaluated regularly and those qualified will be transferred to the WoS and hence, will receive an impact factor.

Testa (2009) wrote:

> As the global distribution of *Web of Science* expands into virtually every region on earth, the importance of regional scholarship to our emerging regional user community also grows. Our approach to regional scholarship effectively extends the scope of the Thomson Reuters Journal Selection Process beyond the collection of the great international journal literature: it now moves into the realm of the regional journal literature. Its renewed purpose is to identify, evaluate, and select those scholarly journals that target a regional rather than an international audience. Bringing the best of these regional titles into the *Web of Science* will illuminate regional studies that would otherwise not have been visible to the broader international community of researchers.

ESCI thus seems to break with the original idea of SCI to include journals based on their impact factors. Perhaps it can be understood as a response for broader coverage in relation to research evaluation—as well as to the increasing competition from other producers of citation indexes?

See also:

- ESCI website: http://info.clarivate.com/ESCI?
- ESCI journal list: http://mjl.clarivate.com/cgi-bin/jrnlst/jlresults.cgi?PC=EX

## 5.0 Citation databases from other database producers

### 5.1 CiteSeerX

CiteSeerX (until 2006 called Cite Seer) is an autonomous and automatic citation indexing system introduced in 1997. It focusses primarily on the literature in computer and information science. CiteSeerX was an innovation from previous citation indexing systems, because the indexing process is completely automatic. The citation index autonomously locates, parses and indexes articles found on the world wide web. CiteSeerX was based on these features: actively acquiring new documents, automatic citation indexing and automatic linking of citations and documents. CiteSeerX is hosted by Pennsylvania State University since 2006 when it changed to its present name. See also:

- CiteSeerX website: https://citeseerx.ist.psu.edu/index

### 5.2 Crossref

Crossref (or CrossRef) was launched in early 2000 by the Publishers International Linking Association Inc. (PILA) as a cooperative effort among publishers to enable persistent cross-publisher citation linking in online academic journals by using the digital object identifier (DOI).[30] PILA is a non-profit organization that provides citation links for both open access journals and subscription journals for online publications by the contributing publishers (but access to the subscription journals via citation links depends, of course, on the user's or his library's subscription to the journals to which the links go).

Crossref citation data is made available on behalf of the Initiative for Open Citations (I4OC),[31] a project launched in 2017 to promote the unrestricted availability of scholarly citation data (however, Elsevier, a major academic publisher, did not join the initiative). Harzing (2019, 342) wrote: "the addition of open citation data in April 2017, making it possible to use Crossref for citation analysis through an API [application programming interface]. Since November 2017, *Publish or Perish* (Harzing 2007) has provided the option of searching for authors, journals and key words in Crossref." See also:

- Crossref home page: https://www.crossref.org/
- Initiative for Open Citations (I4OC), https://i4oc.org/

### 5.3 Scopus Citation Index (Scopus)

Elsevier released Scopus in 2004. Jacso (2005, 1539) wrote about it:

> 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).

Sugimoto and Larivière (2018, 30) wrote about this: "The establishment of Scopus is often heralded as a successful case of vertical integration, in which the firm that creates the citation index also owns the material it is indexing." The database presents itself as "the largest abstract and citation database of peer-reviewed literature: scientific journals, books and conference proceedings."[32] One important aspect described by Sugimoto and Larivière (2018, 31-32) is that Scopus is much younger than WoS in terms of coverage. Although it contains records going back to 1823, it is only consistent in indexing from 1996 onwards. Therefore, even if it is considered a high-quality source for contemporary analyses, it is of inferior quality to WoS for historical analyses.

Comparing Scopus versus WoS, Moed (2017, 200) reported a study from Leiden University's Centre for Science and Technology Studies (2007): "Scopus is a genuine alternative to WoS." This claim was supported by the following figures:

- Scopus tends to include all science journals covered by the WoS (since 1996)
- And Scopus contains some 40% more papers
- Scopus is larger and broader in terms of subject and geographical coverage
- Web of Science is more selective in terms of citation impact

In 2015, Scopus also added "The Scopus Article Metrics module," which added altmetrics data on the usage, captures, mentions and citations of each document indexed in the database. Altmetrics are scholarly communication indicators based on the social web;[33] a set of diverse metrics—for example, how many times a paper was shared in a social network like Twitter or, how many times it was saved in a reference manager like Mendeley (Souza 2015, 58). Those databases are not citation indexes but are using citation data.

The Scopus Article Metrics module includes new metrics based on four alternative metrics categories:

- Scholarly Activity: Downloads and posts in common research tools such as Mendeley and CiteULike
- Social Activity: Mentions characterized by rapid, brief engagement on platforms used by the public, such as Twitter, Facebook and Google
- Scholarly Commentary: Reviews, articles and blogs by experts and scholars, such as F1000 Prime, research blogs and Wikipedia
- Mass Media: Coverage of research output in the mass media (e.g., coverage in top tier media)

Other databases are also using altmetrics, including Dimensions (Section 5.6)—and more will probably do so in the future.

- Website: https://www.scopus.com
- Content coverage description: https://www.elsevier.com/__data/assets/pdf_file/0007/69451/0597-Scopus-Content-Coverage-Guide-US-LETTER-v4-HI-singles-no-ticks.pdf
- Journal list: https://www.scopus.com/sources.uri?zone=TopNavBar&origin=searchbasic

### 5.4 Google Scholar (GS)

GS was launched in November 2004 and was originally intended as a tool for researchers to find and retrieve the full text of documents; however, Sugimoto and Larivière (2018, 32) wrote that "few years after the introduction of the tool, bibliometric indicators were added to the online platform at the individual and journal level under the rubric Google Scholar Citation."[33] As documented in Section 6 below, GS has a much broader coverage than other citation databases, a fact also emphasized by Moed (2017, 115):

> Google Scholar does cover a large number of sources (journals, books, conference proceedings, disciplinary preprint archives or institutional repositories) that are not indexed in WoS or Scopus, and thus has a much wider coverage (e.g., Moed, Bar-Ilan & Halevi, 2016); its surplus is especially relevant for young researchers.

In addition to online journals (both acquired by agreement with publishers[35] and by crawling webpages), GS also use Google's own product Google Books to acquire citation information from books. Moed (2017, 207) summarized the attributes of GS in this way:

- Google Scholar is a powerful tool to search relevant literature
- It is also a fantastic tool to track one's own citation impact
- It is up-to-date, and has a broad coverage
- But there are many data quality issues
- Its online metrics features are poor
- Use in research evaluation requires data verification by assessed researchers themselves

GS itself has poor search facilities, but Anne-Will Harzing has developed the free software Publish or Perish (Harzing 2010; 2011) to gather data from Google Scholar and other citation databases.

Moed (2017), Sugimoto and Larivière (2018) and other researchers consider that here are data quality issues regarding GS, because its mechanical way of covering sources is available on the Internet if its algorithm identifies them as scholarly, which is based on certain characteristics, such as a reference list. By contrast, WoS and Scopus have an active content advisory board responsible for quality control. Moed (2017) also states that Google Scholar online metrics features are poor. The bibliometric aspect of the platform is limited to individual and journal-level metrics, as these are the only indicators it aggregates. Also, very little information about its coverage and citations count has been provided by the producer ( Kousha and Thelwall 2007; Gray et al. 2012; Sugimoto and Larivière 2018). These problems are also further addressed in Section 6.

– Website: https://scholar.google.com

### 5.5 Microsoft Academic (MA)

MA is a free public web search engine for academic publications re-launched by Microsoft Research in 2016. It replaces the earlier Microsoft Academic Search (MAS), which ended development in 2012.

In her study on the coverage of MA, Harzing (2016, 1646) concludes that "only Google Scholar outperforms Microsoft Academic in terms of both publications and citations." Hug, Ochsner and Brändle (2017) concluded that MA outperforms GS in terms of functionality, structure and richness of data as well as with regard to data retrieval and handling, but had reservations and pointed out that further studies are needed to assess the suitability of MA as a bibliometric tool. Further studies of MA include Harzing and Alakangas (2017a and 2017b), Hug and Brändle (2017) and Thelwall (2017; 2018a).

– Website: https://academic.microsoft.com/home

### 5.6 Dimensions

Dimensions was launched by Digital Science in January 2018. The database is offered in three different versions, a free version (Dimensions) and two paid versions (Dimensions plus and Dimensions analytics).[36]

Orduña-Malea and Delao-López-Cózar (2018) described its history and functionality. The most significant attribute is perhaps that it indexes individual articles (based on, e.g., Australian and New Zeeland Standard Research Classification). Harzing (2019) examined how Dimensions covered her own scholarly output and found that eighty-three of eighty-four journal articles were included (only surpassed by Google Scholar and Microsoft Academic), one out of four books and one out of twenty-five book chapters. Her overall conclusion—based on the small sample—was that Dimensions is better than Scopus and the Web of Science but is beaten by Google Scholar and Microsoft Academic. Thelwall (2018b) examined its coverage of journal articles (but not for other document types) and suggested on that basis that Dimensions is a competitor to the Web of Science and Scopus for non-evaluative citation analyses and for supporting some types of formal research evaluations. However, because it is currently indirectly spammable through preprint servers (e.g., by uploading batches of low-quality content), in its current form it should not be used for bibliometrics-driven research evaluations.

### 5.7 Other citation databases

A growing number of standard bibliographic database providers are now integrating citation data into their databases. For example, beginning in 2001, the PsycINFO database published by the American Psychological Association started to include references appearing in journal articles, books, and book chapters, and has since then also begun to include references appearing in some records for earlier years.

– Homepage: https://www.apa.org/pubs/databases/psycinfo/cited-references
– Journal Coverage List: https://www.apa.org/pubs/databases/psycinfo/coverage

SciFinder is a product from Chemical Abstracts Service (CAS) that was launched as a client-based chemistry database in late 1997; its citation analysis features were added in 2004, and in 2008, a web version was released. See further in Li et al. (2010).

– Homepage: https://www.cas.org/products/scifinder-n

BIOSIS Citation Index was created in 2010 and released on the Web of Knowledge. It combines the indexed life science coverage found in BIOSIS Previews (Biological Abstracts, Reports, Reviews, and Meetings) with the power of cited reference searching and indexes data from 1926 to the present.

– Homepage: https://support.clarivate.com/ScientificandAcademicResearch/s/article/BIOSIS-Citation-Index-Source-materials?language=en_US

### 6.0 Comparative studies of six major citation indexes

ISI and its successor Thomson ISI seem to have had a *de facto* monopoly on citation indexes from the early 1960s until about 2000 (disregarding Shepard's Citations, cf., Section 7.1). Thereafter, other citation indexes began to appear. As already stated, PsycINFO began adding citation information to its database in 2001. At that time, it was rather unthinkable that major competitors to the ISI databases should be developed because of the huge costs of establishing and managing databases with such a broad coverage. However, as we saw in Section 5, many surprising developments have taken place since then, and today we have a range of competing citation indexes. This section provides information of the relative strengths and weaknesses of six of the most important ones: Crossref, Dimensions, Google Scholar (GS), Microsoft Academic (MA), Scopus and WoS.

The first eye catching difference of these databases is that WoS and Scopus are proprietary databases (with paywalls) while Crossref, GS, MA and Dimensions all are free search services without paywalls (open access). However, this article will examine which databases perform best disregarding the costs associated with their use. As expressed by Jacso (2005, 1538) "open access should not provide excuse for ill-conceived and poorly implemented search options, and for convoluted, and potentially misleading presentation of information" (if such should be the case). It should also be said that this section focuses on the six databases' ability to track citations, whereas other issues, such as providing h-indices are not considered.

Another characteristic often used is to classify databases in "controlled" versus "non-controlled." WoS and Scopus, for example, have been characterized as controlled (or even "highly controlled") databases whereas GS is an uncontrolled (or low-controlled) database (e.g., by López-Cózar, Orduna-Malea and Martín-Martín 2019 and Halevi 2017). However, as we saw in Section 2, the concept "control" (as that of "unification") are relative concepts, and it should, therefore, always be specified in what way databases are controlled, as all databases in different ways make use of algorithms and mechanical procedures. It is probably better to characterize Crossref, GS, MA and Dimensions as crawl-based databases. This distinction seems to be important for their functionalities. Öchsner (2013, 31-46), for example, found that Web of Science and Scopus offer quite similar functionalities and coverage and maintain their real own databases. WoS has always made their databases from physical or electronic access to published journals. Scopus could do similar as described in Section 5.3. The other four (and free) citation search services depend on information from publishers' homepages and from other parts of the Web.

## 6.1 Coverage

Coverage may be described by the database itself or by bibliometric studies. Li et al. (2010) found that both Web of Science and Scopus missed some references from publications that they cover, while Google Scholar did not disclose which publications were indexed and the results are not downloadable, so it is difficult to determine which citations are missing. Also, Google Scholar included citations from websites and, therefore, one of its problem is the duplication of citing references.

Since the launching of Scopus and GS in 2004 there have been many studies comparing these databases with WoS, each other and other databases such as MEDLINE. As described by Martín-Martín et. al. (2018, 1161), these studies do not provide a clear view of the respective strengths of these databases:

> A key issue is the ability of GS, WoS, and Scopus to find citations to documents, and the extent to which they index citations that the others cannot find. The results of prior studies are confusing, however, because they have examined different small (with one exception) sets of articles ... For example, the number of citations that are unique to GS varies between 13% and 67%, with the differences probably being due to the study year or the document types or disciplines covered. The only multidisciplinary study (Moed, [Bar-Ilan and Halevi] 2016) checked articles in 12 journals from 6 subject areas, which is still a limited set.

The study by Martín-Martín et. al. (2018) is important for at least four reasons:

1) It is rather new, which is important because the databases have added many references since their start; therefore, many former studies may provide obsolete results;
2) It is the first large scale investigation;
3) This study is the broadest with respect to subject fields examined (252 subject categories);
4) This study reviews former studies and relate its own findings to the former ones.

The overall conclusion was that in all areas GS citation data is essentially a superset of WoS's and

Scopus's with substantial extra coverage. Martín-Martín et. al. (2018, 1175) wrote:

> This study provides evidence that GS finds significantly more citations than the WoS Core Collection and Scopus across all subject areas. Nearly all citations found by WoS (95%) and Scopus (92%) were also found by GS, which found a substantial amount of unique citations that were not found by the other databases. In the *Humanities, Literature & Arts*, *Social Sciences*, and *Business, Economics & Management*, unique GS citations surpass 50% of all citations in the area.

and:

> In conclusion, this study gives the first systematic evidence to confirm prior speculation (Harzing, 2013; Martín-Martín et al., 2018; Mingers & Lipitakis, 2010; Prins et al., 2016) that citation data in GS has reached a high level of comprehensiveness, because the gaps of coverage in GS found by the earliest studies that analysed GS data have now been filled.

This conclusion is reinforced by Gusenbauer (2019), who found: "Google Scholar's size might have been underestimated so far by more than 50%. By our estimation Google

Scholar, with 389 million records, is currently the most comprehensive academic search engine."

Concerning coverage, given the limitations of any study, GS was clearly shown to be superior compared with WoS and Scopus. Before we consider other aspects than coverage, let us consider how other databases compare. At the time of writing, GS and MA are considered the two largest citation databases. Hug and Brändle (2017, 1569) concluded:

> Our findings suggest that, with the exceptions discussed above, MA performs similarly to *Scopus* in terms of coverage and citations ... With its rapid and ongoing development, MA is on the verge of becoming a bibliometric superpower ... The present study and the studies of Harzing and Alakangas (2017a, b) provide initial evidence for the excellent performance of MA in terms of coverage and citations.

Hug and Brändle (2017) also provided the following figures: "MA is also progressing quickly in terms of coverage. According to the development team of MA, the database expanded from 83 million records in 2015 (Sinha et al. 2015) to 140 million in 2016 (Wade et al. 2016) and 168 million in early 2017 (A. Chen, personal communication, March 31, 2017). It is currently growing by 1.3 million records per month (Microsoft Academic 2017)." Again, these figures must be considered with reservations and should be followed by independent studies. However, the current evidence suggests that MA has almost (but not fully) the same coverage as GS, and because both databases develops quickly, it is impossible to predict what will be the case just in the short-term future.

These are relative figures, but how much of the scholarly documents available on the web do they cover? Our knowledge of this is limited and insecure. Moed (2017, 194) mentioned the following figures of the number of scientific/scholarly journals, and wrote "as Garfield has pointed out, perhaps the most critical issue is how one defines the concept of journal."

| | | |
|---|---|---|
| Derek de Solla Price | (1980) | 40,000 |
| Garfield | (1979) | 10,000 [37] |
| *Scopus*: covers in | 2014 about | 19,000 |
| *Web of Science*: covers in | 2014 | 11,000 |

*Ulrichsweb* (6 Febr 2015):

| | |
|---|---|
| Active, academic/scholarly journals | 111,770 (100%) |
| Peer-reviewed | 66,734 (60%) |
| Available online: | 47,826 (43%) |
| Open access | 15,025 (13%) |
| Included in Thomson-Reuters *Journal Citation Reports* (JCR) | 10,916 (10%) |

Another estimate is reported by Khabsa and Giles (2014) indicating that at least 114 million English-language scholarly documents are accessible on the web, of which Google Scholar at that time had nearly 100 million.

Harzing (2019) compared how Crossref, Dimensions, Google Scholar, Microsoft Academic, Scopus and the Web of Science covered her own complete bibliography (eighty-four journal articles, four books, twenty-five book chapters, 100+ conference papers, 200+ other publications, two software programs, in total 400+ papers). Of the journal articles, Scopus found seventy-nine and WoS sixty-one; all the other databases found all eighty-four or eighty-three of the journal articles. Other document types displayed greater variation, but the main tendency is clear; Scopus and WoS retrieve fewer references than the rest, and for all document types, none of the databases performed better than GS. If Scopus and WoS still have a role to play, it must be by other qualities than their coverage. We will, therefore, investigate such issues in the next subsections. Harzing's conclusion (2019, 341) was:

> Overall, this first small-scale study suggests that, when compared to Scopus and the Web of Science, Crossref and Dimensions have a similar or better coverage for both publications and citations, but a substantively lower coverage than Google Scholar and Microsoft Academic. If our findings can be confirmed by larger-scale studies, Crossref and Dimensions might serve as good alternatives to Scopus and the Web of Science for both literature reviews and citation analysis. However, Google Scholar and Microsoft Academic maintain their position as the most comprehensive free[38] sources for publication and citation data.

Concerning coverage, the speed of indexing is also a relevant factor (the coverage of the newest documents); how long does it take from publishing of a document until it has been indexed by a citation database? Clearly, GS and crawl-based databases are faster compared to WoS (which again is faster than databases depending on human indexing, such as PsycINFO).

### 6.2 Quality of indexed documents

We have seen that SCI/WoS based their selection mainly on journal impact factors (JIF) (and Garfield's law of concentration), presuming that journals with a high JIF are the most important. In addition, it describes its own selection in this way:[39]

> Since Clarivate Analytics is not a publisher, we are able to serve as an objective data provider. Web of Science Core Collection includes a carefully curated col-

> lection of the world's most influential journals across all disciplines. And, because quality and quantity aren't mutually exclusive, Clarivate Analytics has a dedicated team of experts who evaluate all publications using our rigorous selection process. The journal selection process is publisher neutral and applied consistently to all journals from our 3,300 publishing partners. With consistent and detailed selection criteria covering both quantitative and qualitative assessment, we select only the most relevant research from commercial, society, and open access publishers. Existing titles are constantly under review to ensure they maintain initial quality levels.

This quote contains a hidden criticism of Scopus (because Elsevier is both publisher of Scopus and of very many academic journals) and, in particular, of GS that is bypassing all the peer-reviewed journals' quality filters. The question of quality selection is perhaps just a special case of the question whether bibliographical databases (or libraries, encyclopedias etc.) need quality criteria and a selection process to meet such criteria, as, for example, also claimed by MEDLINE.[40] Such ideas are challenged today, and not only by GS.[41] So, a question is whether, for example, "Garfield's law of concentration" and other kinds of quality selection today is bypassed by attempts to increase coverage in citation indexes? WoS, of course, uses selection as an argument that users should prefer this database, but such claims need to be based in research. Probably we need both databases with quality control and databases with increased coverage. A study by Acharya et al. (2014) found, however, that important work is more and more published in non-elite journals. A selective citation index processing only journals with a higher citation impact could, therefore, miss important work. But such a claim contradicts dominant practices, e.g., to evaluate researchers by considering the quality level of the journals in which they publish.

Harzing and Alakangas (2016, 802) wrote: "Unlike the Web of Science and Scopus, Google Scholar doesn't have a strong quality control process and simply crawls any information that is available on academic related websites. Although most of Google Scholar's results come from publisher websites, its coverage does include low quality "publications" such as blogs or magazine articles." Sugimoto and Larivière (2018, 32) shared this view: "Google Scholar often aggregates nonjournal articles—from PowerPoint presentations to book chapters, basically anything that bears the trappings of an academic document." However, an anonymous reviewer wrote in this connection:

> Cassidy [Sugimoto] and Larivière (2018) [claim ...] that GS inflates citation counts with many worthless, non-scholarly citations. This is by now something of a myth. The faulty citations in GS tend to come from algorithmic misattributions and can be manually corrected. Both Cassidy [Sugimoto] and Larivière maintain GS pages, and I'd challenge them to point out citations they have received that are academically worthless, because the overwhelming majority of the citations Google lists come from perfectly respectable sources. (The same holds for any other researcher with a Google Scholar page.)

Hjørland (2019, Appendix 8) provides an example, that can illustrate the nature of citations provided by GS in addition to those provided by WoS. For the article Hjørland (1998), excluding self-citations, WoS found twenty-one citations and GS additional thirty-five citations. The additional citations were mostly university theses, conference proceedings and journals not covered by WoS. It confirms the anonymous reviewer's claim that nothing that can formally be considered academically worthless was added by GS. It should be added, however, as found by Martín-Martín et. al. (2018, 1175) that the extra citations found by GS tend to be much less cited compared to citing sources that are also in WoS or Scopus.

A study by Gehanno, Rollin and Darmoni (2013) found that of 738 original studies included in their gold standard database all were found in Google Scholar. They wrote (1):

> The coverage of GS for the studies included in the systematic reviews is 100%. If the authors of the 29 systematic reviews had used only GS, no reference would have been missed. With some improvement in the research options, to increase its precision, GS could become the leading bibliographic database in medicine and could be used alone for systematic reviews.

The article further stated (4):

> This 100% coverage of GS can be seen as amazing, since no single database is supposed to be exhaustive, even for good quality studies. For example, the recall ratios of Medline for randomized control trials (RCTs) only stand between 35% and 56% [Türp, Schulte and Antes 2002, Hopewell et al. 2002]. Since GS accesses only 1 million of the some 15 million records at PubMed, how can our results be explained? In fact, through agreements with publishers, GS accesses the "invisible" or "deep" Web, that is, commercial Web sites the automated "spiders" used by search engines such as Google cannot access.

Findings such as these suggests that the databases using automatic acquisition methods based on agreement with the publisher's challenge both citation databases like WoS and

Scopus and traditional documentation databases like MEDLINE, which so far have been considered the gold standard for serious document retrieval. It should be said, however, that there are still open questions and that also Gehanno, Rollin and Darmoni (2013) had reservations about their own study and found that GS requires improvement in the advanced search features to improve its precision. [42]

### 6.3 Control over and reliability of the search

A third issue concerns the users' control over the search process (related to the transparency of the databases) and the reliability of search. Hjørland (2015) argued for the importance of human decision-making during searches implying that information retrieval should be considered a learning process in which the genres and terminology of the domain is learned as well as the qualities of relevant databases. Many researchers consider, for example, lists of indexed journals and related information important for users. Such lists are produced by "classical" databases and by WoS and Scopus, but not for GS, for example. Gusenbauer (2019, 199) wrote in this connection:

> it remains unclear why Google Scholar does not report its size. Given the unstable nature of Google Scholar's QHC [query hit count] it might be possible that Google itself either has difficulties accurately assessing its size or does not want to acknowledge that its size fluctuates significantly. Perhaps it is important to Google to convey to those searching for information that it offers a structured, reliable, and stable source of knowledge. If Google maintains its policy of offering no information, scientometric estimation will have to remain the sole source of information on its size.

Gusenbauer (2019, 197) also wrote:

> The exact workings of Google Scholar's database remain a mystery. While our results remained stable during the examination period, we verified the results a few months later and found considerable differences. Our findings of Google Scholar's lack of stability and reliability of its reported QHC [query hit count] are in line with earlier research

The article further stated (197; italics in original):

> While some variation in QHCs seem to be commonplace among popular search engines, such as Bing or Google (Wilkinson and Thelwall 2013), it should not happen in the scientific context where study outcomes depend on the resources available in databases. Whenever QHC variations occur, the question remains whether they stem from *actual* variations in available records or mere *counting errors* by the search system. The former would be particularly problematic in the academic context where accuracy and replicability are important criteria. These problems seem to be shared only by search engines. We found that all of the bibliographic databases and aggregators we examined—EbscoHost, ProQuest, Scopus, and Web of Science—provide plausible QHC results. This is not surprising given these services access a stable and curated database over which they have extensive control.

Another drawback of using Google Scholar is the large number of duplicate papers retrieved by the citation index. These have been termed "stray citations,"[43] where minor variations in referencing lead to duplicate records for the same paper (compare the presentation of unification in Section 2).

### 6.4 Search options and metadata

The six databases discussed in this section have very different search options (and for some it is necessary to use software such as Publish or Perish (Harzing 2007 and later updates), which also have to be evaluated. Unfortunately, it has become difficult to find exact descriptions of search options. We are missing "blue sheets" as originally developed by the database host Dialog[44] (see Table 1).

WoS and Scopus offer at present the most advanced search possibilities and probably should be preferred for advanced searching in many situations. The most important alternative functions seem to be: 1) Google Scholar's possibility for full-text searching; and, 2) Dimensions' subject classification of single documents (as opposed to journal level classification only). However, there are many unique features, strengths as well as serious weaknesses in the different databases and qualified search decisions must be based on professional knowledge of the characteristics of all the citation indexes.

### 6.5 Conclusion of section

The six databases differ much, not just in their total number of documents indexed and citations provided, but also which fields, document types and time spans are covered, and all of this is changing over time. It has up to now been the consensus that the databases supplement each other. As Rousseau, Egghe and Guns (2018, 152) wrote:

> Many colleagues performed investigations comparing GS, Scopus and the WoS [among others]. Because of unique features related to each of these databases a general consensus, see e.g. (Meho and Yang 2006 [i.e.

| Database | Searchable fields |
|---|---|
| Crossref | "Since November 2017, Publish or Perish (Harzing 2007) has provided the option of searching for authors, journals and key words in Crossref" (Harzing 2019, 342). |
| Dimensions | Keyword searching in: Full data, title and abstract or DOI (+<br>Abstract searching: copy in an abstract).<br>Provides article level subject classification!<br>Filters: Publication year, researcher, fields of research, publication type, source title, journal list and open access. |
| Google Scholar | GS uses search options known from Google.[45] You may search by author name, by article title and/or by keywords, scholarly documents that cite a particular article, and look at the citation context of an article.[46]<br>Martín-Martín et al. (2018, 1175) found there is no reliable and scalable method to extract data from GS, and the metadata offered by the platform are still very limited (again, Publish or Perish and its user guide (Harzing 2007) may be helpful). |
| Microsoft Academic | MA apply "Semantic search", which, however, does not in itself indicates what search options are available.[47]<br>Harzing's Publish or Perish User's Manual, https://harzing.com/resources/publish-or-perish/manual provides the following information: Authors, years, affiliations, study field, full journal title, full article title, all of the words/any of the words (title), does not have an exact phrase search, cannot search for keywords in abstract or full-text of articles.<br>Search filters: Content types (no document type classification), people, research areas, published date (last week, month or year). |
| Scopus | Author, first author, source title, article title, abstract, keywords, affiliation (name, city, country), funding (sponsor, acronym, number), language, ISSN, CODEN, DOI, references, conference, chemical name, ORCID, CAS, year, document type, access type.<br>Scopus has "author identifier" to disambiguate author names.<br>Since 2015 Scopus has the Scopus Metrics Module (see Section 5.3) |
| Web of Science | Topic [i.e., title + abstract + author keywords + keywords plus], title, author, author identifiers, group author, editor, publication name, DOI, year published, address, open access, highly cited, research domain (only journal level subject classification), language, research area (only journal level subject classification), funding agency, group/Corporate authors, document types. (WoS can search exact phrases[48] and has other advanced search futures). WoS has a search option "author search" to help disambiguate author names.<br>The WoS platform does not allow combinations of advanced searches and cited reference searches.[49] |

*Table 1*. Searchable fields in the six databases (based on Gusenbauer 2019, 179-182, the databases themselves and the cited sources; all are citation databases and, therefore, allows searching citing references).

> 2007]) for one of the earliest studies, is that the three databases complement each other.

Is this consensus still valid? Rousseau, Egghe and Guns (2018) do not answer this question explicitly but seem to agree. The research reviewed in Section 6 indicates the same, although, as stated, the market is extremely dynamic and, therefore, needs to be monitored. GS almost always provides a larger total number of citations but have poor search possibilities and other problems, such as the lack of stability, e.g., that citations, once given, may later disappear. Such should not be the case, because citations are provided by publications, and a document, once published cannot cease to exist (either it never existed and, thereby, represents a bibliographic ghost, or its representation in the database has disappeared).

By now, the development in the market for citation indexes is extremely dynamic and, therefore, many conclusions drawn just a few years ago may be outdated. In addition, the relevant attributes are very complex, so that one should not just be careful to evaluate databases on single factors, such as the total number of references or citations they provide, but also consider the attributes of databases in relation to the specific task and domain for which it should be used. Section 6 provided some general information considered useful as general background knowledge, but it cannot provide detailed information about all relevant properties of all databases.

## 7.0 Predecessors to the *Science Citation Index*

It was the SCI that made citation indexes central to information science; only retrospectively it has been considered important to look back and recognize the earlier citation indexes, and, therefore, this section is placed after the others.

Garfield (1979, 7) acknowledges that "Shepard's Citations is the oldest major citation index in existence; it was started in 1873" (see further below, Section 7.1). However, it has later been found that the history of citation indexes is even older. Shapiro (1992, abstract) wrote:

> Historians of bibliometrics have neglected legal bibliometrics almost completely. Yet bibliometrics, citation indexing, and citation analysis all appear to have been practiced in the legal field long before they were introduced into scientific literature. Publication counts are found in legal writings as early as 1817. Citation indexing originated with "tables of cases cited," which date at least as far back as 1743. A full-fledged citation index book was published in 1860. Two ambitious citation analyses of court decisions appeared in 1894 and 1895.

Weinberg (1997; 2004) traced the citation index further back in time; nearly two centuries before Shepard's Citations, a legal citation index was embedded in the Talmud. Besides that, the citation index book *Mafteah ha-Zohar* was published in 1566, a Biblical citation index was printed in the prior decade and a Hebrew citation index to a single book is dated 1511. All of them three centuries before the full-fledged citation index book indicated by Shapiro (1992) as published in 1860. By describing the *Hebrew Citation Index*, Weinberg (1997, 318) states that the citations discussed in her paper "are not to the works of individuals, but to anonymous classics, mainly the Bible and Talmud." She affirms that prestige and promotion are not relevant factors, as they are in modern citation analysis and wrote (126): "many scientific indexing structures thought to have originated in the computer era were invented as much as a millennium earlier, in the domain of religion."

### 7.1 Shepard's Citations

This citation index started as lists in a series of books indexed to different jurisdictions in 1873 by Frank Shepard (1848–1902) (1951). The name *Shepard's Citations, Inc.* was adopted in 1951. In 1999, LexisNexis released an online version of the Shepard's Citation Service and the use of the print version is declining.

The verb "shepardizing" refers to the process of consulting *Shepard's* to see if a case has cited by later cases and, for example, been overturned, reaffirmed or questioned. It is also used informally by legal professionals to describe citations in general. (See also the description on Wikipedia, retrieved 2019). See further:

- Shepard's Citations on LexisNexis: https://www.lexisnexis.com/en-us/products/lexis-advance/shepards.page

## 8.0 Citations as subject access points (SAP)

There is a strong connection between the information retrieval domain and subject access points (SAP).[50] Two kinds of SAP should be distinguished for this discussion of citation indexes:

1) Terms, either assigned to documents by indexers (e.g., descriptors from thesauri) or derived from the document (e.g., words from titles, e.g., in so-called KWIC indexes or from the full text) (this is so-called "traditional indexing" or "conventional indexing") and
2) the use of bibliographical references as SAP (as used in citation databases).

Both terms and references may function as concept symbols as pointed out by Small (1978).

One purpose of research in this area is to shed light on the relative advantages and disadvantages of these two kinds of indexing and the SAP that they provide. About traditional indexing Garfield (1964, 144) stated that "the ideas expressed in a particular source article are reflected in the index headings used by some conventional indexing system. In that case, a display of the descriptors or subject headings assigned to that paper by the indexer constitutes a restatement of the subject matter of that paper in the indexer's terminology." [51] Garfield (1979, 2-3) further wrote:

> There also is a qualitative side to search effectiveness that revolves around how precisely and comprehensively an individual indexing statement describes the pertinent literature. The precision of the description is a matter of semantics, which poses a series of problems in a subject index. The basic problem is that word usages varies from person to person. It is patently impossible for an indexer, no matter how competent, to reconcile these personal differences well enough to choose a series of subject terms that will unfailingly communicate the complicated concepts in a scientific document to anyone who is searching for it.

Garfield here is writing about manual indexing, but in principle, also algorithmic indexing is a subjective process (cf., Hjørland 2011), therefore, his statement can be generalized to cover both intellectual and automatic forms of indexing. Garfield (1979, 3) wrote about bibliographical references as SAP in contrast to terms:

> Citations, used as indexing statements, provide these lost measures of search simplicity, productivity, and efficiency by avoiding the semantic problems ...

> In other words, the citation is a precise, unambiguous representation of a subject that requires no interpretation and is immune to changes in terminology.

We shall see below that research has painted a more complicated and nuanced picture about the relative strengths and weaknesses than those here quoted from Garfield. First, it should be recognized, that the problem of information retrieval is deeper than semantics. As Swanson concluded (1986, 114): "Any search function is necessarily no more than a conjecture and must remain so forever." It is not just about the meaning of terms assigned to documents, but at the deepest level it is about which documents are relevant for a specific query. By implication, both term indexing and citation indexing are no more than conjectures about what best facilitates retrieval.

Harter, Nisonger and Weng (1993) examined the "semantic relationships" between cited and citing articles by examining whether citing and cited documents were indexed similarly according to: a) *Library of Congress Classification* numbers; and, b) the descriptors assigned to them by three databases: ERIC, LISA and Library Literature. They found that the semantic relations between cited and citing papers are very low. However, it should be emphasized that inter-indexer consistency in general is low (see, for example, Soler Monreal and Gil-Leiva 2011 and Hjørland 2018, 614-5, Section 3.2), thereby, questioning whether the low degree of semantic relations is due to the difference between term indexing and citation indexing or to more general factors in indexing. Also, the author's use of the terms "semantic relations" and "subject relations" is unclear. Implicitly, Harter, Nisonger and Weng (1993) seem to assume a document-oriented view of the subject of documents rather than a request- or policy-oriented view (cf., Hjørland 2017, Section 3.1, 60-61).

Garfield's claim that "the citation is a precise, unambiguous representation of a subject" is problematic, because the subjectivity of determining meaning in subject terms is replaced with a new kind of subjectivity: the choice of documents to be cited (cf. Section 9 below: Studies of citation behavior). No doubt, however, citations are important; as Martyn (1965, 189) wrote: "We know that following up references cited in relevant papers is the scientist's most favored method of obtaining information; it is not therefore very surprising that a device [i.e., the citation index] which allows them to move forwards in time as well as backwards is greeted with enthusiasm."

McCain (1989, 113) found that citation-based retrieval was not equally successful for all topics, and there is a large variation in performance within the nine topics examined comparing to descriptor retrieval and "both approaches to retrieval are capable of identifying substantial numbers of relevant documents with which the user was not previously familiar." Research in information science tends to consider in the words of McCain (1989, 110) that "cited references and thesaurus terms are generally considered to be complementary approaches to the identification and retrieval of relevant documents in a bibliographic search." Although this statement provides valuable knowledge, it provides little information that can guide searchers in the choice of SAP and search strategy in specific cases. We have to provide more specific knowledge about the relative strengths and weaknesses of references as subject access points.

McCain (1989, 110) explains that there are a few published studies specifically comparing the results of descriptor retrieval and citation retrieval for the same set of queries. But "in the case of new interdisciplinary areas, citation retrieval may be much more effective than thesaurus terms" (seemingly suggesting that it may be otherwise in more established areas). This comes closer to the kind of information demanded above, but it has to be documented more precisely when one kind of SAP is more fruitful than another and in what way they supplement each other.

Pao and Worthen (1989, 231) conducted a pilot study on the relative effectiveness of "semantic relevance" (by terms) and "pragmatic relevance" (by citations). [52] They found that citation retrieval can identify relevant papers otherwise missed by semantic retrieval. And (234): "In most cases, semantic representations retrieved the major portion of relevant papers. Although citation representation retrieved far fewer relevant items, it was able to retrieve some relevant items in those instances when term retrieval failed." Furthermore, using both search methods to create a union set is likely to increase recall. And, considering the experiment, "for searches with few retrievals, the addition of citation searching tends to produce more relevant papers with reasonably good precision" (Pao and Worthen 1989, 226, 232).

Hjørland (2013) acknowledges that it is important to understand bibliometric techniques as a family of approaches to knowledge organization (KO) in order to illustrate their relative strengths and weaknesses. He conducted a comparative theoretical analysis of the assumptions in citation analysis compared with those in traditional forms of KO. By doing so, he gets to a main conclusion concerning citation analysis as an approach to KO, which presents the advantages and disadvantages of bibliographic references and citations as subject access points as it is presented in the Table 2.

Hjørland and Nielsen (2001, 276) concluded that "A given subject access point (e.g., descriptors, references) cannot be expected to have a fixed information value regardless of conventions in the knowledge domain and the writing culture." The relative value of references as SAP, therefore, must be studied in relation to terminological problems and citation behaviors in different domains.

| ADVANTAGES<br>of bibliographic references as subject access points (SAP) | DISADVANTAGES<br>of bibliographic references as subject access points (SAP) |
|---|---|
| References represent a form of "literary warrant" and are thus empirically based in the scholarly literature. | The relation between citations and subject relatedness is indirect and somewhat unclear (related to the difference between the social and the intellectual organization of knowledge). |
| Citations are provided by researchers (highly qualified subject specialists). | Bibliometric maps do not provide a clear logical structure with mutually exclusive and collectively exhaustive classes. |
| The number of references reflects the indexing depth and specificity (the average of scientific papers is about ten references per article). | Explicit semantic relations are not provided (e.g., genus—species relations and part—whole relations) (but future systems may distinguish between different kinds of citation links/motivations). |
| Citation indexing is a highly dynamic form of subject representation (each new document published and indexed updates the pattern). | Only derived indexing is provided: concepts not represented in the literary sample are not assigned. |
| References are distributed through papers, allowing the utilization of the paper structure in the contextual interpretation of citations. | There is a tendency to mix different theoretical structures due to the merging of literatures in the samples (rather than providing a system based on a pure theoretical basis). |
| Scientific papers form a kind of self-organization system. | Namedropping and other forms of imprecise citation may cause noise. |
| Citation based maps identify groups of researchers working in the same specialties. | |

*Table 2.* Bibliographic references as subject access points (SAP). Source: Hjørland (2013, 1324).

Hjørland and Nielsen (2001) also found (258) that "ordinary retrieval algorithms and citation practices seem simply to reflect different theories about subject relatedness" (i.e., ordinary retrieval algorithms tend to consider documents subject-related if they "are alike" by containing the same words or concepts, whereas citation practices tend to consider documents related if they are linked by citations). Hjørland (2013, 1321) took this difference a step further by suggesting: "The relations between papers in a certain tradition are used as criteria of subject relatedness rather than just classifying documents on the basis of shared properties." These different views correspond to different views in biological classification, in which the cladistic, genealogical approach is confronted by the numerical taxonomic approach. These approaches are again linked to different epistemologies: historicism versus empiricism. Thus Hjørland and Nielsen (2001, 258) concluded:

> our insight from citation indexes has profoundly changed not only the methods of IR but also the concept of subject relatedness itself and the basic aim of retrieving information.

## 9.0 Studies of citation behavior (citer motivations)

In scholarly communication, referencing previous works is an indispensable part of a document that reports research. As Ziman (1968, 58) wrote: "a scientific paper does not stand alone; it is embedded in the 'literature' of the subject." From the point of view of using the citations for information retrieval, it is important to consider the citation behavior or the citation culture of citing authors. Garfield (1965, 85) listed fifteen reasons (citation motives) to cite other documents:

1. Paying homage to pioneers
2. Giving credit for related work (homage to peers)
3. Identifying methodology, equipment and so on
4. Providing background reading
5. Correcting one's own work
6. Correcting the work of others
7. Criticizing previous work
8. Substantiating claims
9. Alerting to forthcoming work
10. Providing leads to poorly disseminated, poorly indexed or uncited work
11. Authenticating data and classes of facts, physical constants and so on
12. Identifying original publications in which an idea or concept was discussed
13. Identifying original publications or other work describing an eponymic concept or term
14. Disclaiming work or ideas of others (negative claims)
15. Disputing priority claims of others (negative homage)

Garfield (1979, 244-6) further challenges three criticisms of citation analysis: negative citations, self-citations and citations to methodology papers. For him, negative citations are as important as positive citations, because they are part of the process of scientific communication. If a work is so criticized as to be highly cited, it is a work that has some ideas that deserve the attention of other researchers, otherwise, it would be ignored by the scientific community. Many theories that are in force today were criticized initially, and, from these criticisms, were improved and became recognized. As for self-citation, Garfield uses a compelling argument: a researcher who aims to increase the number of citations he needs to publish to make his name appear. However, to generate a large number of publications, it is assumed that the researcher has much to say, otherwise the quality of the works will be smaller, and the author will only be able to publish works in peripheral journals, which are not indexed in the citation indexes. For this reason, Garfield believes that this is one of the criticisms that appear more in theory than in reality. The third point, about the high citation counts of some methodological papers is more difficult, but Garfield (1979, 245) says that such "a conclusion overlooks several important points. The most obvious one is the questionable validity of the judgment that methods are inherently less important than theories."

Issues such as negative citations, self-citations and methodological citations are often discussed in relation to the use of citations in research evaluation. However, here our question is about bibliographical references as SAP. For example, to the degree that it can be documented that scientists use to quote papers for methodological reasons rather than for theoretical reasons, such knowledge is directly useful; citation indexes may be better for identifying methodological papers than for identifying theoretical papers and other systems must be made for retrieving the last kind.

## 9.1 Citation theories

Bornmann and Daniel (2008, 48) wrote: "Two competing theories of citing behavior have been developed in past decades, both of them situated within broader social theories of science. One is often denoted as the normative theory of citing behavior and the other as the social constructivist view of citing behavior." The normative theory, following Robert K. Merton (e.g., Merton 1988) basically states that scientists give credit to colleagues whose work they use by citing that work. From this point of view, Small (2004, 71) termed "citations as the symbolic payment of intellectual debt" and furthermore wrote that citations represent "vehicles of peer recognition and constructed symbols for specific original achievements in science."

The social constructivist view was described this way (Bornmann and Daniel 2008, 49):[53]

> The social constructivist view on citing behavior is grounded in the constructivist sociology of science (see, e.g. Collins, 2004; Knorr-Cetina, 1981; Latour and Woolgar, 1979). This view casts doubt on the assumptions of normative theory and questions the validity of evaluative citation analysis. Constructivists argue that the cognitive content of articles has little influence on how they are received. Scientific knowledge is socially constructed through the manipulation of political and financial resources and the use of rhetorical devices (Knorr-Cetina, 1991). For this reason, citations cannot be satisfactorily described unidimensionally through the intellectual content of the article itself. Scientists have complex citing motives that, depending on the intellectual and practical environment, are variously socially constructed (e.g. to defend their claims against attack, advance their interests, convince others, and gain a dominant position in their scientific community).

Nicolaisen (2007) presented what a comprehensive discussion of theories of citation and included empirical tests of the social constructivist theory. His main conclusion (633) was:

> This chapter has sought to make clear that, in order to explain such behavior, we must cease taking the individual's knowledge structures as our starting point. Rather, we should focus our attention on knowledge domains, disciplines, or other collective knowledge structures. Attempts to explain citation behavior should thus refrain from psychologizing the act of citing and instead recognize it as embedded within the sociocultural conventions of collectives.

It seems obvious that theories of citing behavior are important in relation to considering references as SAP. Hjørland (2002) introduced a view that can be understood as representing social epistemology as a third position between Merton and social constructivism; on the one hand Merton is right that scientists cite what they consider the most valuable documents in relation to their argumentation. On the other hand, different scientific perspectives (traditions, perspectives or paradigms) may differ with respect to what are considered the most valuable documents. In the literature about schizophrenia, for example, psychoanalytic journals tend to cite other sources than neuroscientific journals (cf., Hjørland 2002, 266). By conclusion: scholars citation behavior may at the deepest level be explained by their theoretical and epistemological commitments and the study of citations should, therefore, consider traditions and paradigms as a high priority.

## 10.0 General conclusion

Orduña-Malea and Delao-López-Cózar (2018) expressed some views, which we find serve as a proper conclusion for this article. They mentioned the importance of tools such as telescopes and microscopes for the development of science and they related the importance of citation indexes for the understanding of the ecosystem of scientific information with such tools. Here we can specify that these tools function at two levels: 1) they are tools for the scholars seeking knowledge; and, 2) they are tools for the scholars studying science (including scientometricians and information scientists). The authors had the following additional views:

- The appearance of *Science Citation Index* (SCI) and *Social Science Citation Index* (SSCI) allowed for the first time the realization of studies that had been impossible and unthinkable before, contributing to acquaintaince with aspects that until then had remained invisible to the eyes of researchers.
- The appearance in 2004 of both Scopus and Google Scholar represents a turning point in the story. However, while Scopus brought forth evolution (extended coverage, new journal's topic classification, new journal indicators, innovative visualization techniques and lately an integration of altmetrics), Google Scholar implied a revolution by its automatic means to locate, discover and organize academic material with simplicity and speed, entering the bibliographic databases in the big data universe.
- The story presents a new milestone on January 15, 2018, the day on which the Dimensions platform was officially launched. At the time of its launch, Dimensions was made up of 128 million documents apart from information on funding (3.7 million of awarded grants) and approximately four billion connections between them. Additionally, these data are enriched with impact information, both in terms of citations received (connections among cited / citing documents, available for fifty million records) and altmetrics (available for nine million documents approximately), academic profiles (twenty million profiles), Global Research Identifier Database (GRID) geotagging, as well as a classification of subject areas based on machine learning techniques.

While we find it too early to grant Dimensions this position in the development of citation indexing, Orduña-Malea and Delao-López-Cózar (2018) nonetheless provided a fine description of the importance of citation indexing. To this description may be added that among the most important contributions of citation indexing is the development of search engines like Google; it is worth considering that the major factor behind the success of Google is its utilization of links between documents and the number of in-links to documents. In this way, principles of bibliometrics and citation studies takes a prominent role in front-end technologies today.

## Notes

1. Price (1970, 7) expressed this terminology in the following way: "if Paper R contains a bibliographic footnote using and describing Paper C, then R contains a reference to C, and C has a citation from R." This way "the number of references a paper has is measured by the number of items in its bibliography as endnotes and footnotes, etc., while the number of citations a paper has is found by looking it up in some sort of citation index and seeing how many other papers mention it" (Price 1970, 7).
   Sugimoto and Larivière (2018, 67) also explain that the distinction between citations and references is conceptual. "First, while (almost) all research documents contain references, not all documents are cited. The second difference relates to time: references are always made to past literature and are static; that is, the reference list will never grow or change over time. Citations, on the other hand, come from documents written in the future. Therefore, citations are dynamic."
2. Although this way of distinguishing "reference" and "citation" has got a certain impact in the field of bibliometrics, it is not common practice. For example, the influential *Chicago Manual of Style* (17th edition, 2017, 743) uses the term "source citation" for what in this article are called "references," and Nicolaisen (2007, 609) wrote: "Unless stated otherwise, the term citation is used synonymously with the term bibliographic reference."
3. References contain a set of standardized information. However, there are different standards, and authors of scholarly papers have to apply the standard used in a specific journal or by a specific publisher. For example, this encyclopedia (*IEKO*) uses *The Chicago Manual of Style,* whereas many other publications in information science use *Publication Manual of the American Psychological Association*. Although such standards differ, there are certain basic elements they all cover, for example, author names, printing years and name of journal in which an article is published.
4. *European Reference Index for the Humanities* (ERIH) was a project for developing a citation index initiated by European Science Foundation from 2008. Websites: http://archives.esf.org/hosting-experts/scientific-review-groups/humanities-hum/erih-european-reference-index-for-the-humanities.html and https://dbh.nsd.uib.no/publiseringskanaler/erihplus/. The term "reference in-

dex" seems logically the best term because what is being indexed are the references in the documents indexed by the database.

5. The term "citation index" is, unfortunately, also ambiguous. For example, MEDLINE database calls itself "the principal online bibliographic citation database of the NLM [National Library of Medicine]." https://www.nlm.nih.gov/lstrc/jsel.html. However, it is not a citation database as understood in this article (and in general). MEDLINE just provides references, not citations for the indexed documents. It is, by the way, strange, that MEDLINE has not added an index of cited references, as done, for example, by the PsycINFO database.
6. Backward searching is also termed "searching antecedents," while forward searching is "searching descendants" of a given paper.
7. A short look at the cited references seems not to identify any critical articles in this case. If you are not happy with what you find, you may question: 1) the coverage of the database (are certain kinds of research favored?); and, 2) the epistemology of the field of intelligence research (are the IQ test really testing biological issues as claimed or are they testing how certain people respond due to their status and to socio-cultural issues? Feminist epistemology and feminist philosophy of science—along with other epistemological positions—have made objections to this way of doing research; this can also be found in the citation indexes, although probably not in the set of references citing specific empirical investigations such as Nyborg (2005). Nyborg's research has been very controversial and heavily discussed in, among other places, the media (see, for example, https://en.wikipedia.org/wiki/Helmuth_Nyborg). Empirical evidence that contradicts Nyborg' claim about the superiority of male intelligence can be found, for example, in Flynn and Rossi-Case (2011). The reception of this reference may of course also be traced by citation indexes, e.g., in Google Scholar.
8. As already described, when an author cites a document, he or she also provides a reference to another work. Both citations and references are produced by the very same act, whether it is called "referencing" or "citing." By implication, we do not need two theories: one theory about references and another about citations, as claimed by Wouters (2016, 73-4):
   "Wouters (1999) concluded that a theory of referencing behavior should be seen as fundamentally distinct from a theory of evaluative bibliometrics. This was based on the statement that there is a fundamental distinction between reference and citation. By analyzing references and citations as different signs, they were essentially positioned as different objects. Their relation is one of descent: the citation emerges in an act of "semiosis" (the creation of a novel sign) from the reference. This has an important implication: it is no longer the scientist who creates the citation. Its source lies in the citation index and the producer of that index is the creator of the sign citation."
9. Remark that the Danish letter Ø and similar non-English letters are not used in Web of Science but is replaced by "O." Remark also that in the beginning this Citation Index did not provide first names, only initials; although today both initials and first names are used (in two different fields, initials in the AU field and first names in AF (author full name) it is nonetheless necessary to use the AU field to retrieve older records and thus all documents authored by a given author such as Anders Ørom. This makes it very difficult to disambiguate common names like "A Smith," although WoS has a specific "author search" facility that may help solving the problem.
10. In this case, the use of the source index to search back is not as good as using the article itself: in the original article, there were thirty-seven cited references and in WoS only thirty-one are listed, and their bibliographical information are not as full or accurate as that in Ørom (2003, 142-3).
11. Atkins (1999) wrote: "The standardized capture and further unification of references has three main benefits: 1) [for the producer of the database]: it allows for speed in data capture since the key needed for matching is quite short; 2) it enables the presentation of more consistent references to customers, regardless of the number of variants presented in the source journals; and 3) it enables internal and external links in ISI products."
12. In reference lists, the same reference may be referred to in many different ways and these are not standardized. The Russian psychologist Lev Semyonovich Vygotsky, for example, can be spelled in many ways, and it is a complicated task to make a search for documents citing him by collocating the different spellings. This is contrary to the library tradition in which "authority control" is applied indicating a standard name form and a unique name for each person (e.g., by adding birth date when needed). In the world of scholarly communication and citation indexes, this problem may in the future be solved in a similar way by ORCID (Open Researcher and Contributor ID) and other services, which provide a persistent digital identifier that distinguishes a researcher from every other researcher and support automated linkages between the researcher and the professional activities (see https://en.wikipedia.org/wiki/ORCID). There is also another standardization problem, however. When scholars cite a certain document (books in particular), they may cite different editions with different printing years and different translations. This of-

ten poses problems for citation searching and scholars should prefer to cite original editions or standard editions. But many do not, and, therefore, the problem persist.

13. Web of Science, Scopus Citation Index; Google Scholar; Cite Seer; Korean Journal Database; Scientific Electronic Library Online Citation Index (SciELO); Emerging Source Citation Index (ESCI); Crossref; Microsoft Academic; Dimensions.
14. *Science Citation Index* (*SCI*); BIOSIS Citation Index; Chinese Science Citation Database (CSCD); Russian Science Citation Index (RSCI).
15. Social Sciences Citation Index (SSCI); Chinese Social Sciences Citation Index (CSSCI).
16. Arts & Humanities Citation Index (A&HCI).
17. *Shepard's Citations* (law); PsycINFO (psychology); CiteSeerX (computer and information science); SciFinder (Chemestry).
18. Book Citation Index (BKCI).
19. Conference Proceedings Citation Index (CPCI).
20. Data Citation Index.
21. The statement "without the need for the intellectual judgments of human indexers" is important. However, as discussed in Section 8, citation indexes introduce a new kind of subjectivity: the choice of references made by the authors.
22. Moed (2005, 11) wrote: "*Eugene Garfield Associates* was founded in 1954 and launched numerous editions of *Current Contents* by 1960. In that year, the company name was changed to the *Institute for Scientific Information* (ISI). In 1964, ISI launched the Science Citation Index (SCI), as a quarterly multidisciplinary index." In 1992, ISI was bought by The Thomson Corporation and changed names to Thomson ISI. In 2006, it changed name to Thomson Scientific and in 2008 Thomson and Reuters merged under the name Thomson Reuters. From 2016, it was bought by Onex and Baring Asia under the name Clarivate Analytics, as which it is still known, although in May 13, 2019, Clarivate merged with Churchill Capital.
23. Web of Science (WoS) has been a part of Web of Knowledge (WoK) and WoS and WoK have also been used as synonyms. Today, the term WoK is seldom used and seems to be replaced by WoS.
24. WoS Core collection is opposed to the Specialist Collection (BIOSIS Citation Index, BIOSIS Previews, Biological Abstracts, Zoological Record, Medline, CAB Global Health, Inspec and FSTA) and the Regional Collection (see Appendix 1). The contents of the Core Collection (depends on subscription agreements) includes the following databases:
    - *Science Citation Index Expanded* (1900-present)
    - Social Sciences Citation Index (1956-present)
    - Arts & Humanities Citation Index (1975-present)
    - Conference Proceedings Citation Index-Science (1990-present)
    - Conference Proceedings Citation Index-Social Science & Humanities (1990-present)
    - Book Citation Index– Science (2005-present)
    - Book Citation Index– Social Sciences & Humanities (2005-present)
    - Emerging Sources Citation Index (2015-present)
    - Current Chemical Reactions (1986-present) (includes Institut National de la Propriété Industrielle structure data back to 1840)
    - Index Chemicus (1993-present)
25. P.t. the following databases:
    - Current Contents Connect (1998-present)
    - Data Citation Index (1900-present)
26. The WoS homepage in addition hosts the following databases produced by other companies:
    - Derwent Innovations Index (1969-present)
    - KCI-Korean Journal Database (1980-present)
    - MEDLINE® (1950-present)
    - Russian Science Citation Index (2005-present)
    - SciELO Citation Index (2002-present)
    - Zoological Record (1976-present)
27. The "expanded" version of the SCI (and formerly also of the SSCI) were created in order for subscribers to have a choice between a relatively cheap and an expensive version. The expanded versions are superset of the non-expanded versions. There are no other theoretical issues involved in this distinction.
28. ERIH, see: http://archives.esf.org/hosting-experts/scientific-review-groups/humanities-hum/erih-european-reference-index-for-the-humanities.html
29. Clarivate Analytics (2018). Book Citation Index: https://support.clarivate.com/ScientificandAcademicResearch/s/article/Web-of-Science-Core-Collection-Book-Citation-Index---Coverage-is-of-the-full-book-and-not-selective-chapters?
30. Digital Object Identifier (DOI) is a persistent identifier or handle used to identify objects uniquely, standardized by the International Organization for Standardization (ISO)
31. I4OC should not be confused with OpenCitations (http://opencitations.net/ ), another initiative, established in 2010.
32. About Scopus: https://www.elsevier.com/solutions/scopus
33. The concept of altmetrics was cited the first time in a tweet from Jasom Priem from University of North Carolina-Chapel Hill in 2010. A manifesto was presented in the same year (Priem, Taraborelli, Growth 2010). Altmetrics are fast, using public APIs to gather data in days or weeks. They're open—not just the data but the

scripts and algorithms that collect and interpret it. Altmetrics look beyond counting and emphasize semantic content like usernames, timestamps and tags. Altmetrics are not citations, nor are they webometrics; although these latter approaches are related to altmetrics, they are relatively slow, unstructured and closed. (Priem, Taraborelli and Growth 2010)

34. Journal level indicators on GS is, however, quite limited.
35. It is not published which publishers are indexed by GS; however, for now, Elsevier publications are not indexed by GS—probably because Elsevier wants users to go to their own portal (ScienceDirect) which is not free.
36. Dimensions plus is adding patents, clinical trials, grants and policy documents and their connections. Additionally, it allows the search of new entities (organizations and financing agents). It also includes advanced analysis tools, such as the comparison between organizations or financing agents, the generation of advanced reports as well as the possibility of integrating custom implementations.
37. Garfields law of concentration (Garfield, 1979, 160) stated: “The core literature for all scientific disciplines involves a group of no more than 1,000 journals, and may involve as few as 500.” Garfield added (160): “Though larger collections certainly can be justified in many cases, the single function of providing resonable cost-effective coverage of the literature most used by research scientists requires no more than 500 to 1,000 journals.”
38. The word “free” seems to be misplaced here, since, whether free or not, Google Scholar and Microsoft Academic maintain their position as the most comprehensive sources for publication and citation data.
39. Clarivate Analytics (undated): Web of Science Fact Book. https://clarivate.com/wp-content/uploads/2017/05/d6b7faae-3cc2-4186-8985-a6ecc8cce1ee_Crv_WoS_Upsell_Factbook_A4_FA_LR_edits.pdf and: “With the Web of Science platform, you can access an unrivalled breadth of world-class research literature linked to a rigorously selected core of journals and uniquely discover new information through meticulously captured metadata and citation connections ... Find out what makes Web of Science Core Collection the most accurate, objective, and complete resource available.” https://clarivate.com/products/web-of-science/.
40. MEDLINE’s journal selection criteria are discussed here: https://www.nlm.nih.gov/lstrc/jsel.html.
41. In libraries, the idea of curated and quality-controlled collections is challenged by the principle of “patron-driven acquisitions” and in Wikipedia and social media the concept “wisdom of the crowd” has challaged the idea of edited works.
42. A special question is whether prepublications (“in press” publications) are indexed and should be indexed. Some databases (e.g., WoS) do not index preprints but wait until a document is formally published before including it in the database, GS and Scopus, on the other hand, also index prepublications. In this way some statistical figures about GS may be inflated compared to WoS. Arguments can be given, however, in favor for the inclusion of preprints: it can be important to find documents fast, and, therefore, indexing of prepublications can be fruitful. A counter argument is that it is important to keep a well-defined notion of “a publication” and to avoid possible duplicate representations of the same document.
43. Harzing (2016): “Stray Citations” [blog post]: https://harzing.com/resources/publish-or-perish/tutorial/google-scholar/stray-citations
44. Archived link to Dialog’s “blue sheets” http://web.archive.org/web/20000816215552/http://library.dialog.com/bluesheets/html/bln.html.
45. Google and GS use the following search keys:
    Boolean operators (AND, OR, NOT)
    “+” Searches stop words
    “-” A minus before a word ignores documents in which that word appears.
    “intitle:” restrict the results to documents containing that word in the title other words in the query will return documents that mention the word anywhere in the document
    “allintitle:” restrict the results to those with all the query words in the title.
    “site:” searches for the word in the site/domain name. Limits searches to a special domain or site.
    “inurl:” searches for the word in the URL
    “allinurl:” searches for all the words in the URL
    “author:” searches for the word in the author’s name
    “filetype:” limits file type
    ““ searches the phrase
    “*” in phrase searching “*” replaced by any single word.
    “..” Number range
46. Noruzi (2005, 175) mentions that you may also search journal title. However, what can be done is to use journal title as a filter: “Return articles published in.” Among the drawbacks of Google Scholar is that a Google Scholar search only includes the first 1000 hits. GS does not provide metadata on the document type and the language of the document that it covers.
47. Microsoft wrote: “Microsoft Academic understands the meaning of words, it doesn’t just match keywords to content. For example, when you type “Microsoft,” it knows you mean the institution, and shows you papers authored by researchers affiliated with Microsoft. Similarly, Microsoft Academic knows journal titles, conference names, and many research topics.” https://academic.microsoft.com/home;

However, such functions can only be made by a built-in knowledge organization system (e.g., a thesaurus or an ontology) or another semantic technology. But many systems (also old systems like MEDLINE) have KOSs (or can be combined with KOSs) that allow for kinds of semantic searches, e.g., ending the EXPLODE operator (!) to a valid thesaurus term will result in retrieval of narrower terms. For example, the statement SELECT DEMENTIA! in MEDLINE® will retrieve narrower (i.e., more specific terms) such as ALZHEIMER DISEASE, CREUTZFELDT-JAKOB SYNDROME, etc.

48. Exact phrase searching is not possible according to Harzing's Publish or Perish User's Manual, https://harzing.com/resources/publish-or-perish/manual (retrieved 2019-07-24); this is a mistake, however. See https://images.webofknowledge.com/images/help/WOS/hs_search_rules.html
49. Combinations of advanced searches and cited reference searches were formerly possible for the ISI citation databases in the database host Dialog. It is a great disadvantage that they cannot be performed directly on WoS, but must be done in downloaded sets using other software tools.
50. For a broader introduction to SAP, see Hjørland and Nielsen (2001).
51. What Garfield called a restatement may also be called an interpretation based on both the indexer's subjectivity and the properties of the indexing language, see further in Hjørland (2018).
52. The terms "semantic relevance" (by terms) and "pragmatic relevance" (by citations) can be discussed, because a pragmatic view of semantics considers the meaning and application of terms determined by pragmatic principles.
53. See also the discussion by Zuckerman (1987).

## References

Acharya, Anurag, Alex Verstak, Helder Suzuki, Sean Henderson, Mikhail Iakhiaev, Cliff Chiung Yu Lin and Namit Shetty. 2014. "Rise of the Rest: The Growing Impact of Non-Elite Journals." https://arxiv.org/pdf/1410.2217.pdf

Atkins, Helen. 1999. "The ISI® Web of Science—Links and Electronic Journals." *D-Lib Magazine* 5, no. 9. doi:10.1045/september99-atkins

Bornmann, Lutz and Hans-Dieter Daniel. 2008. "What do Citation Counts Measure? A Review of Studies on Citing Behavior."*Journal of Documentation* 64: 45-80. doi:10.1108/00220410810844150

Clarivate Analytics 2019. *Web of Science Fact Book*. https://clarivate.com/wp-content/uploads/2017/05/d6b7faae-3cc2-4186-8985-a6ecc8cce1ee_Crv_WoS_Upsell_Factbook_A4_FA_LR_edits.pdf

Clarivate Analytics. 2015. *Russian Science Citation Index*. http://wokinfo.com/products_tools/multidisciplinary/rsci/?utm_source=false&utm_medium=false&utm_campaign=false

Clarivate Analytics. 2016. *Korean Journal Database*. http://wokinfo.com/products_tools/multidisciplinary/kci_kjd/?utm_source=false&utm_medium=false&utm_campaign=false

Clarivate Analytics. 2017. *Web of Science Core Collection - Emerging Sources Citation Index*. http://wokinfo.com/media/pdf/ESCI_Fact_Sheet.pdf?utm_source=false&utm_medium=false&utm_campaign=false

Clarivate Analytics. 2018a. *Chinese Science Citation Database*. http://wokinfo.com/products_tools/multidisciplinary/cscd/?utm_source=false&utm_medium=false&utm_campaign=false

Clarivate Analytics. 2018b. "Web of Science Core Collection: Book Citation Index - Coverage is of the Full Book and Not Selective Chapters." https://support.clarivate.com/ScientificandAcademicResearch/s/article/Web-of-Science-Core-Collection-Book-Citation-Index---Coverage-is-of-the-full-book-and-not-selective-chapters?

Collins, Harry. 2004. *Gravity's Shadow. The Search for Gravitational Waves*. Chicago: University of Chicago Press.

Flynn, James R. and Lilia Rossi-Case. 2011. "Modern Women Match Men on Raven's Progressive Matrices." *Personality and Individual Differences* 50: 799-803. doi:10.1016/j.paid.2010.12.035

Force, Megan M. and Nigel J. Robinson. 2014. "Encouraging Data Citation and Discovery with the Data Citation Index." *Journal of Computer-Aided Molecular Design* 28: 1043-48. doi:10.1007/s10822-014-9768-5

Garfield, Eugene. 1955. "Citation Indexes for Science: A New Dimension in Documentation through Association of Ideas." *Science* 122, no. 3159: 108-11. doi:10.1126/science.122.3159.108

Garfield, Eugene. 1964. "Science Citation Index: A New Dimension in Indexing." *Science* 144, no. 3619: 649-54. doi:10.1126/science.144.3619.649

Garfield, Eugene. 1965. "Can Citation Indexing be Automated?" In *Statistical Association Methods for Mechanized Documentation*, ed. Mary Elizabeth Stevens, Vincent E. Giuliano and Laurence B. Heilprin. Washington: National Bureau of Standards, 189-92.

Garfield, Eugene. 1970. "A New ISI Program for Dissemination and Retrieval of Conference Papers." *Current Contents/Life Sciences* 13, no. 10: 4-5.

Garfield, Eugene. 1971. "The Mystery of the Transposted Journal Lists. Wherein Bradford's Law of Scattering is Generalized According to Garfield's Law of Concentration." *Current Contents* no. 17: 222-3.

Garfield, Eugene. 1977a. "ISI's New Index to Scientific and Technical Proceedings lets You Know What Went on at a Conference Even if You Stayed at Home." *Current Contents* no. 40: 5-10.

Garfield, Eugene. 1977b. "Will ISI's Arts & Humanities Citation Index Revolutionize Scholarship?" *Current Contents* no. 2: 5-9.

Garfield, Eugene. 1978. "Introducing Index to Social Sciences and Humanities Proceedings: More Help in Locating and Acquiring Proceedings." *Current Contents* no. 33: 5-9.

Garfield, Eugene. 1979. *Citation Indexing: Its Theory and Application in Science, Technology and Humanities*. New York, NY: John Wiley & Sons.

Garfield, Eugene. 1980. "Is Information Retrieval in the Arts and Humanities Inherently Different from that in Science? The Effect That ISI®'s Citation Index for the Arts and Humanities Is Expected to Have on Future Scholarship." *Library Quarterly* 50: 40-57. doi:10.1086/629874

Garfield, Eugene. 1981. "Introducing ISI-ISTP and B (Index to Scientific and Technical Proceedings and Books): Online Access to the Conference Literature and Multi-Authored Books." *Current Contents* no. 34: 5-9.

Gehanno, Jean-François, Laetitia Rollin and Stefan Darmoni. 2013. Is the Coverage of Google Scholar Enough to be Used Alone for Systematic Reviews?" *BMC Medical Informatics and Decision Making* 13, no. 7. http://www.biomedcentral.com/1472-6947/13/7

Gorraiz, Juan, Philip J. Purnell and Wolfgang Glänzel. 2013. "Opportunities for and Limitations of the Book Citation Index." *Journal of the American Society for Information Science and Technology* 64: 1388-98. doi:10.1002/asi.22875

Gray, Jerry E., Michelle C. Hamilton, Alexandra Hause, Margaret M. Janz, Justin P. Peters and Fiona Taggart. 2012. "Scholarish: Google Scholar and Its Value to the Sciences." *Issues in Science and Technology Librarianship* 70. doi:10.5062/F4MK69T9

Gusenbauer, Michael. 2019. "Google Scholar to Overshadow Them All? Comparing the Sizes of 12 Academic Search Engines and Bibliographic Databases." *Scientometrics* 118: 177-214. doi:10.1007/s11192-018-2958-5

Halevi, Gali, Henk Moed and Judit Bar-Ilan. 2017. "Suitability of Google Scholar as a Source of Scientific Information and as a Source of Data for Scientific Evaluation -- Review of the Literature." *Journal of Informetrics* 11: 823-4. doi:10.1016/j.joi.2017.06.005

Harter, Stephen P., Thomas E. Nisonger and Aiwei Weng. 1993. "Semantic Relationships Between Cited and Citing Articles in Library and Information Science Journals." *Journal of the American Society for Information Science* 44: 543-52. doi:10.1002/(SICI)1097-4571(199310)44:9<543::AID-ASI4>3.0.CO;2-F

Harzing, Anne-Wil. 2007. "Publish or Perish." https://harzing.com/resources/publish-or-perish

Harzing, Anne-Wil. 2010. *The Publish or Perish Book: Your Guide to Effective and Responsible Citation Analysis*. Melbourne: Tarma Software Research.

Harzing, Anne-Wil. 2011. *The Publish or Perish Book, Part 2: Citation Analysis for Academics and Administrators*. Melbourne: Tarma Software Research.

Harzing, Anne-Wil. 2013. "A Longitudinal Study of Google Scholar Coverage Between 2012 and 2013." *Scientometrics* 98: 565-75. doi:10.1007/s11192-013-0975-y

Harzing, Anne-Wil. 2016. "Microsoft Academic (Search): A Phoenix Arisen from the Ashes?" *Scientometrics* 108: 1637-47. doi:10.1007/s11192-016-2026-y

Harzing, Anne-Wil. 2019. "Two New Kids on the Block: How do Crossref and Dimensions Compare with Google Scholar, Microsoft Academic, Scopus and the Web of Science?" *Scientometrics* 120: 341-9. doi:10.1007/s11192-019-03114-y

Harzing, Anne-Wil and Satu Alakangas. 2016. "Google Scholar, Scopus and the Web of Science: A Longitudinal and Cross-Disciplinary Comparison." *Scientometrics* 106: 787-804. doi:10.1007/s11192-015-1798-9

Harzing, Anne-Wil and Satu Alakangas. 2017a. "Microsoft Academic: Is the Phoenix Getting Wings?" *Scientometrics* 110: 371-83. doi:10.1007/s11192-016-2185-x

Harzing, Anne-Wil and Satu Alakangas. 2017b. "Microsoft Academic is One Year Old: The Phoenix is Ready to Leave the Nest." *Scientometrics* 112: 1887-94. doi:10.1007/s11192-017-2454-3

Hjørland, Birger. 1998. "The Classification of Psychology: A Case Study in the Classification of a Knowledge Field." *Knowledge Organization* 25: 162-201.

Hjørland, Birger. 2002. "Epistemology and the Socio-Cognitive Perspective in Information Science." *Journal of the American Society for Information Science and Technology* 53: 257-70. doi:10.1002/asi.10042

Hjørland, Birger. 2011. "The Importance of Theories of Knowledge: Indexing and Information Retrieval as an Example." *Journal of the American Society for Information Science and Technology* 62: 72-7. doi:10.1002/asi.21451

Hjørland, Birger. 2013. "Citation Analysis: A Social and Dynamic Approach to Knowledge Organization." *Information Processing & Management* 49: 1313-25. doi:10.1016/j.ipm.2013.07.001

Hjørland, Birger. 2015. "Classical Databases and Knowledge Organization: A Case for Boolean Retrieval and Human Decision-making During Searches." *Journal of the Association for Information Science and Technology* 66: 1559-75. doi:10.1002/asi.23250

Hjørland, Birger. 2017. "Subject (of Documents)." *Knowledge Organization* 44: 55-64.

Hjørland, Birger. 2018. "Indexing: Concepts and Theory." *Knowledge Organization* 45: 609-39.

Hjørland, Birger. 2019. "The Classification of Psychology: A Case Study in the Classification of a Knowledge Field." https://www.isko.org/cyclo/psychology#app8

Hjørland, Birger and Lykke Kyllesbech Nielsen. 2001. "Subject Access Points in Electronic Retrieval." *Annual Review of Information Science and Technology* 35: 249-98.

Hopewell, Sally, M. Clarke, A. Lusher, C. Lefebvre and M. Westby. 2002. "A Comparison of Hand Searching Versus MEDLINE Searching to Identify Reports of Randomized Controlled Trials." *Statistics in Medicine* 21: 1625-34.

Hua, Weina. 2001. "The Development of the Chinese Social Sciences Citation Index." *Indexer* 22, no. 3: 128-9.

Hug, Sven E., Michael Ochsner and Martin P. Brändle. 2017. "Citation Analysis with Microsoft Academic." *Scientometrics* 111: 371-8. doi:10.1007/s11192-017-2247-8

Hug, Sven E. and Martin P. Brändle. 2017. "The Coverage of Microsoft Academic: Analyzing the Publication Output of a University." *Scientometrics* 113: 1551-71. doi:10.1007/s11192-017-2535-3

Jacso, Peter. 2005. "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* 89: 1537-47. https://www.jstor.org/stable/24110924

Khabsa, Madian and C. Lee Giles. 2014. "The Number of Scholarly Documents on the Public Web." *PLoS ONE* 9, no. 5: e93949. doi:10.1371/journal.pone.0093949

Klein, Daniel B. and Eric Chiang. 2004. "The Social Science Citation Index: A Black Box - With an Ideological Bias?" *Econ Journal Watch* 1: 134-65. https://econjwatch.org/articles/the-social-science-citation-index-a-black-box-with-an-ideological-bias

Knorr-Cetina, Karin. 1981. *The Manufacture of Knowledge: An Essay on the Constructivist and Contextual Nature of Science*. Oxford: Pergamon Press.

Knorr-Cetina, Karin. 1991. "Merton Sociology of Science: The First and the Last Sociology of Science." *Contemporary Sociology: A Journal of Reviews* 20: 522-6.

Kousha, Kayvan and Mike Thelwall. 2007. "Google Scholar Citations and Google Web/URL Citations: A Multi-Discipline Exploratory Analysis." *Journal of the American Society for Information Science and Technology* 58: 1055-65. doi:10.1002/asi.20584

Latour, Bruno and Steve Woolgar. 1979. *Laboratory Life: The Social Construction of Scientific Facts*. London: Sage.

Lewison, Grant and Philip Roe. 2013. "The Shortfall in Coverage of Countries' Papers in the Social Sciences Citation Index Compared with the Science Citation Index. In *Proceedings of ISSI 2013 Vienna: 14th International Society of Scientometrics and Informetrics Conference; Vienna, Austria 15th to 20th July 2013,* ed. Juan Gorraiz, Edgar Schiebel, Christian Gumpenberger, Marianne Hörlesberger and Henk Moed. Vienna: AIT Austrian Institute of Technology, 2: 1601-12.

Leydesdorff, Loet and Ulrike Felt. 2012a. "'Books' and 'Book Chapters', in the Book Citation Index (BKCI) and Science Citation Index (SCI, SoSCI, A&HCI)." *Proceedings of the American Society for Information Science and Technology* 49: 1-7. doi:10.1002/meet.14504901027

Leydesdorff, Loet and Ulrike Felt. 2012b. "Edited Volumes, Monographs and Book Chapters in the Book Citation Index (BKCI) and Science Citation Index (SCI, SoSCI, A&HCI)." *Journal of Scientometric Research* 1: 28-34. doi:10.5530/jscires.2012.1.7

Li, Jie, Judy F. Burnham, Trey Lemley and Robert M. Britton. 2010. "Citation Analysis: Comparison of Web of Science®, Scopus™, Scifinder® and Google Scholar." *Journal of Electronic Resources in Medical Libraries* 7: 196-217.

López-Cózar, Emilio Delgado, Enrique Orduna-Malea and Alberto Martín-Martín. 2019. "Google Scholar as a Data Source for Research Assessment." In *Springer Handbook of Science and Technology Indicators*, ed. Wolfgang Glaenzel, Henk Moed, Ulrich Schmoch and Michael Thelwall. Springer Handbooks. Cham: Springer, 95-127. doi:10.1007/978-3-030-02511-3

Martín-Martín, Alberto, Enrique Orduna-Malea, Mike Thelwall and Emilio Delgado López-Cózar. 2018. "Google Scholar, Web of Science and Scopus: A Systematic Comparison of Citations in 252 Subject Categories." *Journal of Informetrics* 12: 1160-77.

Martyn, John. 1965. "An Examination of Citation Indexes." *Aslib Proceedings* 17, no. 6: 184-96. doi:10.1108/eb050021

McCain, Katherine W. 1989. "Descriptor and Citation Retrieval in the Medical Behavioral Sciences Literature: Retrieval Overlaps and Novelty Distribution." *Journal of the American Society for Information Science* 40: 110-14. doi:10.1002/(SICI)1097-4571(198903)40:2<110::AID-ASI5>3.0.CO;2-T.

McVeigh, Marie E. 2017. "Citation Indexes and the Web of Science." *Encyclopedia of Library and Information Sciences 4*. ed. John D. McDonald and Michael Levine-Clark. Boca Raton, FL: CRC Press, 2: 940-50.

Meho, Lokman I. and Kiduk Yang. 2007. "Impact of Data Sources on Citation Counts and Rankings of LIS Faculty: Web of Science Versus Scopus and Google Scholar." *Journal of the American Society for Information Science and Technology* 58: 2105-25. doi:10.1002/asi.20677

Merton, Robert K. 1988. "The Matthew Effect in Science, II: Cumulative Advantage and the Symbolism of Intellectual Property." *Isis* 79: 606-23. doi:10.1086/354848

Microsoft Academic. 2017. “Some Facts About the Current Size of our Data Stop & meet us at #kdd2017 @MLatMSFT.” *Twitter*, August 15. ·https://twitter.com/MSFTAcademic/status/897494672200921088

Mingers, John and Evangelia A. E. C. G. Lipitakis. 2010. “Counting the Citations: A Comparison of Web of Science and Google Scholar in the Field of Business and Management.” *Scientometrics* 85: 613-25. doi:10.1007/s11192-010-0270-0

Moed, Henk F. 2005. *Citation Analysis in Research Evaluation*. Information Science and Knowledge Management 9. Dordrecht: Springer.

Moed, Henk F. 2017. *Applied Evaluative Informetrics.* Cham: Springer. doi:10.1007/978-3-319-60522-7_14

Moed, Henk F., Judit Bar-Ilan and Gali Halevi. 2016. “A New Methodology for Comparing Google Scholar and Scopus.” *Journal of Informetrics* 10: 533-51. doi:10.1016/j.joi.2016.04.017

Moskaleva, Olga, Vladimir Pislyakov, Ivan Sterligov, Mark Akoev and Svetlana Shabanova. 2018. “Russian Index of Science Citation: Overview and Review.” *Scientometrics* 116: 449-62. doi:10.1007/s11192-018-2758-y

Narin, Francis. 1976. *Evaluative Bibliometrics: The Use of Publication and Citation Analysis in the Evaluation of Scientific Activity*. Cherry Hill, NJ: Computer Horizons.

Nicolaisen, Jeppe. 2007. “Citation Analysis.” *Annual Review of Information Science and Technology* 41: 609-41.

NIAS (Nordic Institute of Asian Studies). 2018. “Korean Citation Index.” *AsiaPortal: A Nordic Information Resource Portal for Asian Studies*. http://www.asiaportal.info/database/korean-citation-index/

Noruzi, Alireza. 2005. “Google Scholar: The New Generation of Citation Indexes.” *Libri* 55, no. 4: 170-80. doi:10.1515/LIBR.2005.170

Nyborg, Helmuth. 2005. “Sex-related Differences in General Intelligence *g*, Brain Size and Social Status.” *Personality and Individual Differences* 39: 497-509. doi:10.1016/j.paid.2004.12.011

Öchsner, Andreas. 2013. *Introduction to Scientific Publishing: Backgrounds, Concepts, Strategies*. Berlin: Springer.

Orduña-Malea, Enrique and Emilio Delao-López-Cózar 2018. “Dimensions: Re-discovering the Ecosystem of Scientific Information.” *El profesional de la información* 27: 420-31.

Ørom, Anders. 2003. “Knowledge Organization in the Domain of Art Studies: History, Transition and Conceptual Changes.” *Knowledge Organization* 30: 128-43.

Packer, Abel L., Nicholas Cop and Solange M. Santos. 2014. “A rede SciELO em perspectiva.” In *SciELO - 15 anos de acesso aberto: Um estudo analítico sobre acesso aberto e comunicação científica.* Paris: UNESCO, 41-66.

Pao, Miranda Lee and Dennis B. Worthen. 1989. “Retrieval Effectiveness by Semantic and Citation Searching.” *Journal of the American Society for Information Science* 40, no. 4: 226-35. doi:10.1002/(SICI)1097-4571(198907)40:4<226::AID-ASI2>3.0.CO;2-6

Price, Derek J. de Solla. 1970. “Citation Measures of Hard Science, Soft Science, Technology and Nonscience.” In *Communication among Scientists and Engineers*, ed. Derek J. de Solla Price, Carnot.E. Nelson and Donald K. Pollock. Lexington, MA: Heath, 3-22.

Priem, Jason, Dario Taraborelli, Paul Growth and Cameron Neylon. 2010. ”Altmetrics: A Manifesto” *Altmetrics* (blog), 26 October. http://altmetrics.org/manifesto

Prins, Ad A. M., Rodrigo Costas, Thed N. van Leeuwen and Paul F. Wouters. 2016. “Using Google Scholar in Research Evaluation of Humanities and Social Science Programs: A Comparison with Web of Science Data.” *Research Evaluation* 25: 264-70. doi:10.1093/reseval/rvv049

Rousseau, Ronald, Leo Egghe and Raf Guns. 2018. *Becoming Metric-Wise: A Bibliometric Guide for Researchers*. Oxford: Chandos.

Shapiro, Fred R. 1992. “Origins of Bibliometrics, Citation Indexing and Citation Analysis: The Neglected Legal Literature.” *Journal of the American Society for Information Science* 43: 337-39. doi:10.1002/(SICI)1097-4571(199206)43:5<337::AID-ASI2>3.0.CO;2-T

“Shepard's Citations.” 2020. Wikipedia. https://en.wikipedia.org/wiki/Shepard%27s_Citations

Sinha, Arnab, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-june (Paul) Hsu and Kuansan Wang. 2015. “An Overview of Microsoft Academic Service (MAS) and Applications.” In *Proceedings of the 24th International Conference on World Wide Web, ed.* Aldo Gangemi, Stefano Leonardi and Alessandro Panconesi. New York: ACM, 243-6. doi:10.1145/2740908.2742839

Sivertsen, Gunnar and Birger Larsen. 2012. “Comprehensive Bibliographic Coverage of the Social Sciences and Humanities in a Citation Index: An Empirical Analysis of the Potential.” *Scientometrics* 91: 567-75. doi:10.1007/s11192-011-0615-3

Small, Henry G. 1978. “Cited Documents as Concept Symbols.” *Social Studies of Science* 8: 327-40.

Small, Henry. 2004. “On the Shoulders of Robert Merton: Towards a Normative Theory of Citation.” *Scientometrics* 60: 71-9. doi:10.1023/B:SCIE.0000027310.68393.bc

Soler Monreal, M. Concha and Isidoro Gil-Leiva. 2011. “Evaluation of Controlled Vocabularies by Inter-Indexer Consistency.” *Information Research* 16, no. 4. http://informationr.net/ir/16-4/paper502.html

Souza, Iara Vidal Pereira de. 2015. “Altmetrics or Alternative Metrics: Concepts and Key Features.” *AtoZ*: *Novas práticas em informação e conhecimento* 4, no. 2: 58-60. doi:10.5380/atoz.v4i2.44554

Sugimoto, Cassidy R. and Vincent Larivière. 2018. *Measuring Research: What Everyone Needs to Know*. Oxford: Oxford University Press.

Swanson, Don R. 1986. "Undiscovered Public Knowledge." *Library Quarterly* 56: 103-18.

Testa, James. 2009. "Regional Content Expansion in Web of Science®: Opening Borders to Exploration." *GlobalHigherEd* (blog), January 15. https://globalhighered.wordpress.com/2009/01/15/regional-content-expansion-in-web-of-science/

Testa, James. 2012. "Clarivate Analytics Conference Proceedings Selection Process." http://wokinfo.com/products_tools/multidisciplinary/webofscience/cpci/cpciessay/

Thelwall, Mike. 2017. "Microsoft Academic: A Multidisciplinary Comparison of Citation Counts with Scopus and Mendeley for 29 Journals." *Journal of Informetrics* 11: 1201-12. doi:10.1016/j.joi.2017.10.006

Thelwall, Mike. 2018a. "Microsoft Academic Automatic Document Searches: Accuracy for Journal Articles and Suitability for Citation Analysis." *Journal of Informetrics* 12: 1-9. doi:10.1016/j.joi.2017.11.001

Thelwall, Mike. 2018b. "Dimensions: A Competitor to Scopus and the Web of Science?" *Journal of Informetrics* 12: 430-5. doi:10.1016/j.joi.2018.03.006

Torres-Salinas, Daniel, Evaristo Jiménez-Contreras, Nicolas Robinson-García. 2014. "How Many Citations Are There in the Data Citation Index?" Paper presented at Science and Technology Indicators (STI) Conference, Leiden, 3-5, September 2014. https://digibug.ugr.es/bitstream/handle/10481/32931/1409.0753.pdf?sequence=1&isAllowed=y

Torres-Salinas, Daniel, Alberto Martín-Martín and Enrique Fuente-Gutiérrez. 2014. "Analysis of the Coverage of the Data Citation Index - Thomson Reuters: Disciplines, Document Types and Repositories." *Revista española de documentación científica* 37: 1-6. doi:10.3989/redc.2014.1.1114

Torres-Salinas, Daniel, Nicolas Robinson-Garcia, Juan Miguel Campanario and Emilio Delgado López-Cózar. 2014. "Coverage, Field Specialization and the Impact of Scientific Publishers Indexed in the Book Citation Index." *Online Information Review* 38: 24-42.

Torres-Salinas, Daniel, Nicolás Robinson-García and Emilio Delgado López-Cózar. 2012. "Towards a 'Book Publishers Citation Reports'. First approach using the 'Book Citation Index'." *Revista española de documentación científica* 35: 615-620. http://redc.revistas.csic.es/index.php/redc/article/view/766/848

Torres-Salinas, Daniel, Rosa Rodríguez-Sánchez, Nicolás Robinson-García, J. Fdez-Valdivia and J. A. García. 2013. "Mapping Citation Patterns of Book Chapters in the Book Citation Index." *Journal of Informetrics* 7: 412-24. doi:10.1016/j.joi.2013.01.004

Türp, Jens C., Jutta-Maria Schulte and Gerd Antes. 2002. "Nearly Half of Dental Randomized Controlled Trials Published in German are not Included in Medline." *European Journal of Oral Sciences* 110: 405-11.

Wade, Alex D., Kuansan Wang, Yizhou Sun and Antonio Gulli. 2016. "WSDM Cup 2016: Entity Ranking Challenge." In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, ed. Paul N. Bennet, Vanja Josifovski, Jennifer Neville and Filip Radlinski. New York: ACM, 593-4.

Web of Science. 2018. "Web of Science Core Collection: Descriptive Document." https://clarivate.libguides.com/ld.php?content_id=45175981

Weinberg, Bella Hass. 1997. "The Earliest Hebrew Citation Indexes." *Journal of the American Society for Information Science* 48, no. 4: 318-30. https://doi.org/10.1002/(SICI)1097-4571(199704)48:4<318::AID-ASI5>3.0.CO;2-Z

Weinberg, Bella Hass. 2004. "Predecessors of Scientific Indexing Structures in the Domain of Religion." In *The History and Heritage of Scientific and Technological Information Systems: Proceedings of the 2002 Conference*, 65:126-34. Philadelphia, PA: American Society for Information Science & Technology.

Weinstock, Melvil. 1971. "Citation Indexes." *Encyclopedia of Library and Information Science*, ed. Allen Kent, Harold Lancour and Jay E. Daily. New York: Marcel Dekker, 5: 16-41.

Wilkinson, David and Mike Thelwall. 2013. "Search Markets and Search Results: The Case of Bing." *Library & Information Science Research* 35, no.4: 318-25. doi:10.1016/j.lisr.2013.04.006

Wouters, Paul F. 1999. "The Citation Culture." PhD diss., University of Amsterdam. https://dare.uva.nl/search?identifier=b101b769-100f-43e5-b8d2-cac6c11e5bbf

Wouters, Paul F. 2016. "Semiotics and Citations." In *Theories of Informetrics and Scholarly Communication: A Festschrift in Honor of Blaise Cronin*, ed. Cassidy R. Sugimoto. Berlin: Walter de Gruyter, 72-92.

Ziman, John M. 1968. *Public Knowledge: An Essay Concerning the Social Dimension of Science*. London: Cambridge University Press.

Zuccala, Alesia, Mads Breum, Kasper Bruun and Bernd T. Wunsch. 2018. "Metric Assessments of Books as Families of Works." *Journal of the Association for Information Science and Technology* 69, no. 1: 146-57. doi:10.1002/asi.23921

Zuckerman, Harriet. 1987. "Citation Analysis and the Complex Problem of Intellectual Influence." *Scientometrics* 12, no. 5-6: 329-38. doi:10.1007/BF02016675

**Appendix 1: Regional Citation Databases**
(listed according to data of establishment)

**App. 1.1:** *Chinese Science Citation Database* (CSCD)

CSCD was established in 2000 as a result of a partnership between Clarivate Analytics and Chinese Academy of Sciences. The database is hosted on the *Web of Science*. CSCD is fully integrated and searchable with WoS and covers approximately 1,200 top scholarly publications from China, with nearly 2 million records in total. (Clarivate Analytics, 2018a).

– Website: http://wokinfo.com/products_tools/multidisciplinary/cscd/
– Journal List: https://support.clarivate.com/ScientificandAcademicResearch/s/article/Chinese-Science-Citation-Database-Journal-List?language=en_US

**App. 1.2:** *Chinese Social Sciences Citation Index* (CSSCI)
CSSCI was established in 2000 as an interdisciplinary citation index, developed by Nanjing University. It covers about 500 Chinese academic journals of humanities and social sciences. CSSCI has bridged a gap in the field of Chinese social science research and has become a useful database for information retrieval and an important tool for evaluating research work and social science journals (Hua 2001).

– Website: http://cssci.nju.edu.cn/login_u.html

**App. 1.3:** *Korean Journal Database* (KCI)

The *Korean Journal Database*, KCI was established in 2010. It "provides a comprehensive snapshot of the most influential regional content from researchers in South Korea. Using citation connections from the Web of Science™, regional work is framed within the broader context of global research." It indexes about 5,600 journals (at present) from 1980 to present (the list can be downloaded in MS-Excel format from the website) and is a result from a collaboration with the National Research Foundation of Korea. Subject coverage includes: Arts & Humanities, Life Sciences & Biomedicine, Physical Sciences, Social Sciences, and Technology (Clarivate Analytics 2016). Citation information, statistical data and bibliographic information on domestic journals from Korea and indicators like the number of citations and h-index are available through KCI (NIAS 2018).

– Website: https://www.kci.go.kr/kciportal/main.kci?locale=en
– Journal List: https://www.kci.go.kr/kciportal/landing/index.kci

**App. 1.4:** *Scientific Electronic Library Online Citation Index* (SciELO)
SciELO was established in 2014. It is a multidisciplinary citation index from Latin America. *SciELO Citation Index* is part of the SciELO Network wich was created in 1998 and, its first collection was *SciELO* Brazil. We may indicate some of the roles of SciELO Collections: journals indexing based on specific criteria; identify statistics (access, downloads, citation); the publication of full text in open access and the guarantee of interoperability of collections and journals (Packer, Cop, and Santos 2014).

– Website: http://www.scielo.org/php/index.php
– Journal List: http://www.scielo.org/applications/scielo-org/php/secondLevel.php?xml=secondLevelForSubjectByLetter&xsl=secondLevelForSubjectByLetter

**App. 1.5:** *Russian Science Citation Index* (RSCI)
RSCI was established in 2016. This database includes papers from selected Russian journals and is based on the data from the national citation index *Russian Index of Science Citation* (RISC). RISC was launched in 2005 but it is scarcely known to the English-language audience. It is a government-funded project primarily aimed at creating a comprehensive bibliographic/citation database of Russian scholarly publishing for evaluation purposes based on Scientific Electronic Library (further eLibrary.ru) which started as a fulltext database of scholarly literature for grant holders of Russian Basic Research Foundation (Moskaleva et al. 2018). In collaboration with the Scientific Electronic Library (eLibrary.ru), the Russian Science Citation Index on Web of Science™ enables discovery of new insights from Russian publications. It covers fields such as engineering, material science, and ecology and indexes over 600 titles (Clarivate Analytics 2015).

– Website: http://wokinfo.com/products_tools/multidisciplinary/rsci/
– Journal List: https://clarivate.com/wp-content/uploads/2017/09/RSCI_Journal_List.pdf?utm_source=false&utm_medium=false&utm_campaign=false