frontiers in Research Metrics and Analytics

EMPIRICAL STUDY
published: 20 February 2018
doi: 10.3389/frma.2018.00006

# Tale of Three Databases: The Implication of Coverage Demonstrated for a Sample Query

*Judit Bar-Ilan**

*Department of Information Science, Bar-Ilan University, Ramat Gan, Israel*

Coverage is an important criterion when evaluating information systems. This exploratory study investigates this issue by submitting the same query to different databases relevant to the query topic. Data were retrieved from three databases: ACM Digital Library, Web of Science (with the Proceedings Citation Index) and Scopus. The search phrase was "information retrieval," publication years were between 2013 and 2016. Altogether 8,699 items were retrieved, out of which 5,306 (61%) items were retrieved by a single database only, and only 977 (11%) items were located in all three databases. These 977 items were further analyzed: citation counts retrieved from the three databases were compared. Citations were also compared to altmetric data of these publications, collected from Mendeley.

**Keywords: coverage, bibliometrics, citation databases, comparison, correlation**

OPEN ACCESS

***Edited by:***
*Neil R. Smalheiser,*
*University of Illinois at Chicago,*
*United States*

***Reviewed by:***
*Shenmeng Xu,*
*University of North Carolina*
*at Chapel Hill, United States*
*Philippe Mongeon,*
*Université de Montréal, Canada*

****Correspondence:***
*Judit Bar-Ilan*
*judit.bar-ilan@biu.ac.il*

***Received:*** *03 December 2017*
***Accepted:*** *01 February 2018*
***Published:*** *20 February 2018*

***Citation:***
*Bar-Ilan J (2018) Tale of Three Databases: The Implication of Coverage Demonstrated for a Sample Query.*
*Front. Res. Metr. Anal. 3:6.*
*doi: 10.3389/frma.2018.00006*

## INTRODUCTION

Cleverdon (1968) while working on the Cranfield project, stated that users judge information retrieval systems by six criteria:

(1) coverage
(2) recall
(3) precision
(4) response time
(5) presentation
(6) effort

Cleverdon defined *coverage* as "the proportion of the useful literature which is input to a system." Baeza-Yates and Ribeiro-Neto (1999) define *coverage* as a measure in a single database for a given query as the proportion of relevant documents known to the user which were retrieved for the query and the relevant docs known to the user.

In the evaluation of IR systems, the main measures are precision and recall and measures derived from them (e.g., Baeza-Yates and Ribeiro-Neto 1999; Manning et al., 2009) and coverage is often overlooked, even though it is of importance, especially if we are interested in a comprehensive view of the topic. In this paper, we concentrate on coverage by comparing three large databases on a test query both in terms of the number of publications indexed for a given query and also in terms of the citations these publications receive in each of these databases.

*Coverage* for us indicates how much of the existing relevant items (can be limited to articles, books, webpages, blog posts, news items, etc.) that exist in the universe is indexed in the specific database. Since we have no knowledge of the absolute number of relevant items in the universe, different databases indexing the search topic are compared, and the pooled results from the databases considered as an approximation of the universe.

# BACKGROUND

## Database Comparisons

Perry and Salisbury (1995) compared the coverage of dissertations in two sources: Dissertation Abstracts and WorldCat. Even though twice as many dissertations were located in WorldCat, dissertations from major research institutions that were not members of OCLC (the producers of WorldCat) were missing, while they were found in the Dissertation Abstracts. In another early paper (Ramos-Remus et al., 1994) Medline, BIOSIS and Embase were compared on a specific topic. They found that Medline had more records, unique items were found both on BIOSIS and Embase They suggested to search at least in two biomedical databases if comprehensive coverage is important. Examples of other early studies appear in Jacsó (1997) review article on content evaluation.

From 1963 until 2004, there was a single comprehensive citation database initiated by Eugene Garfield, first under the name of the ISI Citation Databases, and from 1996 onward known as the Web of Science (WoS) (Clarivate, 2018; Wikipedia, 2018a,b). It is currently owned by Clarivate. In 2004, two new competitors appeared, Elsevier's Scopus (Elsevier, 2004) and Google Scholar (Wikipedia, 2018c). Since 2004 a large number of coverage comparison studies appeared between the three databases. For a review of coverage studies of Google Scholar versus the other two databases, consult Halevi et al. (2017). Since Google Scholar is not source for this study, we concentrate on presenting a few of the numerous coverage studies of WoS and Scopus.

In 2005, Jacsó searched for items authored by Eugene Garfield, and found 1,522 items indexed by WoS versus 90 items on Scopus. In 2018, there are still more items indexed by WoS (1543) than Scopus (254), mainly because only WoS indexes Current Contents articles authored by Garfield (1063) and Scopus does not. In a later article, Jacsó (2009) showed that articles from the journal *Online Information Review* between 1977 and 2009 were covered much better in WoS than in Scopus. Today the numbers reported for the same period are almost equal.

Bar-Ilan (2008) searched data on a set of highly cited researchers in WoS, Scopus, and Google Scholar, and computed their h-indices (Hirsch, 2005). She found that the h-indices based on WoS data and Scopus data were quite similar, while the h-index computed from Google Scholar data was considerably higher, especially for computer scientists. The h-indices based on WoS and Scopus data were also computed for 30 researchers in Nursing, in all cases the h-index calculated from Scopus data was equal or higher than the WoS based h-index (De Groote and Raszewski, 2012).

Meho and Yang (2007) compared WoS, Scopus, and Google Scholar citations of 15 researchers in Library and Information Science. They found more citations on Scopus than on WoS, and if the unique citations found on Scopus were added to the WoS citations, there was an increase of more than 30% in the citation counts for all 15 researchers combined.

Ball and Tunger (2006) studied citation distributions of WoS and Scopus in several subject areas and found WoS to be slightly better. Three years later Archambault et al., 2009 carried out a large study of citation and publication counts by countries and by subjects. In all cases the correlations between the two databases were extremely high, around 0.99.

Gavel and Iselid (2008) studied the overlap between several databases including WoS and Scopus based on the journal lists provided by each database. They found that the two databases together covered 15,157 journals out of which 49% of the titles were covered by both databases, with 10% of titles unique in WoS, and 41% unique in Scopus. This study was conducted 10 years ago, since then both databases expanded their coverage.

In 2014, WoS covered 13,605 journals and Scopus 20,346 journals (Mongeon and Paul-Hus, 2016). They reported overlap in the journal titles for four major fields Natural Sciences and Engineering (about 45% of the total number of titles covered by the two databases), Biomedicine (about 40%), Social Sciences (about 30%) and Arts and Humanities (about 45%). As of the end of 2017, Scopus covers active journal 23,507 titles. WoS covers 13,809 active journal titles in the three basic citation indexes plus another 7,171 journals in its newly introduced Emerging Sources Citation Index.

A number of studies compared coverage for specific subjects. Earth and Atmospheric Sciences were studied by Barnett and Lascar (2012). They found hundreds of unique titles in Scopus, and only a few in WoS, however most of the unique titles had low Impact Factors or SJR (http://www.scimagojr.com/journalrank.php). A similar conclusion was reached when comparing the WoS and Scopus journal titles in oncology (López-Illescas et al., 2008). In the field of business administration, Clermont and Dyckhoff (2012) compared four databases and found that the subject specific database, EconBiz, had better coverage than WoS and Scopus.

## Mendeley Reader Counts

In the previous section, we described studies that compared WoS and Scopus based journal lists or on publication and citation counts. Instead of citations one can use additional measures like usage data (if available), or the number of users of a reference manager who saved a particular document in their libraries. Mendeley, a free online reference manager, reports for each document in its database the number of users, called "readers" that downloaded the document. There are two major advantages in supplementing conventional indicators with Mendeley readership counts: (1) Readership counts accumulate much faster than citations, and can be early signals of future citations, (2) Not all readers are citers, many of the Mendeley members are students who may or may not publish journal articles. There are shortcomings of Mendeley reader counts and other altmetric indicators as well. The major concerns are that these indicators can be quite easily manipulated, and are not transparent (see, for example, Wilsdon et al., 2017). Studies show a correlation of about 0.5 between citation and readership counts in several disciplines (e.g., Haustein et al., 2014; Zahedi et al., 2014; Thelwall and Sud, 2016).

## Computer Science and Information Retrieval

We were not able to locate specific studies on database coverage in Computer Science. Hull et al. (2008) describe the general characteristics of to the ACM Digital Library. Hennessey (2012) reviews the features of the new interface and the enhanced integration of the ACM Guide to Computing with the Digital Library. Bar-Ilan (2010) studied the influence of the Proceedings Citation Index on publications and citations of highly cited researchers in computer science.

The most relevant article related to the topic of this article, was published by De Sutter and Van Den Oord (2012), where 17 computer scientists' publication and citations counts were retrieved from WoS, Scopus, the ACM Digital Library, and two other databases. Their goal was to study "undercitation." They introduced a measure called relative relevant undercitation, defined at the "fraction of all (cited, citing) paper pairs for which both cited and citing articles are indexed in the database but for which the database has no record of the citing paper in the cited-by list of the cited paper" (p. 71), and found undercitation in all databases.

## Study Goals

As can be seen from the above literature review there are differences between the coverage of databases. The aim of this study is to emphasize the influence of the varied coverage of the databases on various measures, like publication and citation counts, the h-index, most cited sources and most cited publications. Citation counts are compared with Mendeley readership counts for the subset of documents retrieved by the three studied databases.

In the following, we demonstrate the differences stemming from coverage for the term "information retrieval," by comparing three databases that provide citation counts, two of them comprehensive, WoS and Scopus, and one subject specific, the ACM Digital Library (ACM). Information retrieval is a topic relevant both for computer science and for information science. The query is not intended to cover "information retrieval" as a topic, it is only used to demonstrate the differences between the databases, and alerts users to search in multiple databases if there is need for comprehensive data.

*A priori* it was expected that the best coverage in terms of publication counts will be provided by the ACM Digital Library's Guide to Computing Literature, as it claims to be "the most comprehensive bibliographic database focused exclusively on the field of computing" (http://dl.acm.org/advsearch.cfm), and also because the coverage of papers appearing in proceedings is known to be spotty in Scopus and WoS (Bar-Ilan, 2010). The ACM Guide to Computer Literature also covers well the major information science sources related to information retrieval. In terms of citation counts there were no special expectations, because each database draws the citations only from the items covered by it, and it was not clear how much interest there is in information retrieval outside the field, which could only be captured by Scopus and WoS.

# METHODS

## Data Collection

For this study data were collected in May 2017, from three databases, ACM, Scopus and WoS. The search query was identical in all three cases: "information retrieval" as a phrase and so were the publication years, 2013–2016. Our aim was not to have a comprehensive view of the topic, but to have a fair comparison between the databases for a sample query. Fair means identical query, publication years and limitation where to search (e.g., title, abstract, keywords). However, because of the differences in the database search capabilities, there were slight differences in the search strategies as described below.

The ACM Digital Library allows to search in two sources: the ACM Full Text Collection and the more comprehensive (in terms of meta-data) ACM Guide to Computing Literature. The second option was chosen and we searched for term "information retrieval" in the abstract or in the title. After data cleansing (removal of duplicates, items with missing titles or authors), 3,937 items remained out of the initially retrieved 4,161 items. ACM Digital Library allows to download meta-data, but these do not include citation counts, which had to be added manually.

In Scopus, the searches were also in title and abstract, however in addition to limiting the publication years to 2013–2016, we had to limit the retrieved items to two subject areas computer science and social science (to include information science as well) to filter out noise. Out of the 5,635 items retrieved, 5,460 remained after data cleansing.

Web of Science does not allow to limit the search to abstract only, so we chose topic, which includes title, abstract and keywords. We had to exclude keywords from Scopus because inclusion of keywords added mainly noise (12,931 documents for a keyword search limited to publication years and subject area as above). An examination of a sample of the documents showed that the addition of keywords introduced a lot of noise, while in ACM the keyword search had a huge overlap with the title and abstract search. The search in WoS included the Science Citation Index, the Social Science Citation Index, the Arts and Humanities Citation Index, the Proceedings Citation Indexes, and the Emerging Journal Citation Index. The subject areas were limited to computer science and information science, 4,265 documents were retrieved. After retrieval were able to remove all items where the term "information retrieval" appeared in the keywords only 3,673 items retrieved from WoS remained in the dataset. Thus, we created three comparable datasets.

Next a list of unique documents was created from the items retrieved from the different data sources. This part was rather time consuming, because not all items had DOIs, and occasionally the DOIs were incorrect. Pairwise comparisons were conducted to discover overlap, and to collect the citation counts of the given item from the three databases. For items not matched by DOI, title and publication year were compared. These matches were manually checked, since in several cases items with identical titles and publication years were published in two different venues. It was impossible to automatically match items using the publication source as well, because there are no uniform naming conventions for proceeding titles, e.g., publication source for papers in SIGIR 2015, appear as:

- "Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval" in ACM.
- "SIGIR 2015—Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval" in Scopus and WoS.

and CIKM 2016 appears as

- "Proceedings of the 25th ACM International on Conference on Information and Knowledge Management" in ACM.
- "CIKM'16: PROCEEDINGS OF THE 2016 ACM CONFERENCE ON INFORMATION AND KNOWLEDGE MANAGEMENT" in WoS.

Web of Science retrieved items from the CIKM conference series only in 2016, while Scopus indexed only the 2014 proceedings, and ACM retrieved items from all 4 years covered in this study, however the source title for 2013 was slightly different, using "&" instead of "and."

Interestingly for conducting the manual check of items that were paired only by title and publication year the start and end page of the items were most useful. Altogether 8,699 unique items were identified.

It should be noted that it was not feasible to use Google Scholar or Microsoft Academic Search. In Google Scholar, one can search in the title, but not in the abstract, and appearance of the term "information retrieval" in the full text cannot serve as evidence that the paper is about information retrieval. In any case, even when conducting a title search Google Scholar reports as of May 2017, about 4,240 results published between 2013 and 2017, and for a general search about 45,400 results. Since Google Scholar does not allow to retrieve more than 1000 results, it was not feasible to include Google Scholar. Microsoft Academic Search reported more than 50,000 results for the time period, and 28,700 results for items published in 2013 alone.

The subset of documents appearing in all three databases, was further analyzed. Altmetric data were collected for this set to enhance the comparison between the databases. For the altmetric comparison, Mendeley was chosen and data were collected in September 2017.

Mendeley data were collected using Webometric Analyst, a free tool developed by Mike Thelwall (http://lexiurl.wlv.ac.uk/). The tool uses the Mendeley API and enables different types of input and retrieves mostly relevant records from Mendeley. We used two types of searches: article title as a query, and a doi query. The problems with Mendeley include multiple records for the same publication, missing or wrong dois (data are entered by users), problems with retrieving items with special characters. doi searches retrieve a single record for each query and not necessarily the record with the highest number of readers. The two types of searches together produce reasonable results, after the results are cleansed carefully. Manual searches were conducted for special cases (e.g., when the title contains special characters and the article has no doi).

## Data Analysis

Longitudinal publication trends for the whole set of publications and also for the individual databases were charted both in terms of number of publications and in terms of number of citations. The h-index of the topic in each database was computed. Most cited publications were identified.

The subset of items covered by all three databases underwent additional analysis of the citation patterns. The citation counts were compared with Mendeley reader counts and Spearman correlations were computed.

# RESULTS

## Longitudinal Trends

**Figure 1** shows the longitudinal trends in terms of the number of publications. Interesting to note that while the number of unique publications per year is nearly constant, the numbers are decreasing with slight fluctuations for Scopus and ACM, while mostly increasing for WoS. A possible explanation for the tendency of increase, could be the addition of the Emerging Sources Citation index, which indexes sources from 2015 only, it retrieved 96 items in 2015 and 117 items in 2016 from sources that were not covered before, which might explain, at least to some extent the differences. The total number of unique publications was 8,699, where 3,937 items were retrieved from ACM, 5,460 from Scopus and 3,673 from WoS. In order to test whether all items published in 2016 were indexed by the databases by May 2017, the query was rerun in all three databases, and the number of results reported was essentially the same.

**Table 1** shows the number of citations publications received from the time of publication until May 2017 per database. Scopus is highest for all years, ACM is second for 2013 and 2014 and third for 2015 and 2016. Citations to these items are collected from outside the subject area both for WoS and Scopus. Scopus is known to have better coverage than WoS (we saw this also in the number of items retrieved). The ACM Digital Library is subject specific and has wider coverage in the specific subject area than the other two databases, especially by covering a larger number of proceedings papers. Citations accumulate over the years; thus, it is not surprising that both total citation and citations per paper decrease as the time between date of publication and date of retrieval of citations counts decreases.

We observe even higher differences when considering the h-index of the publication set retrieved by each database. Although Hirsch (2005) defined the h-index for individuals, it can be easily extended to any data set, where a data set has h-index *h*, if there are *h* publications that received at least *h* citations each, and *h* is maximal. The h-index of the retrieved publications was 21 for WoS, 24 for ACM and 35 for Scopus—showing considerable differences between the databases. The h-index of the dataset retrieved by Scopus was 66% higher than the h-index of the data set retrieved from WoS.

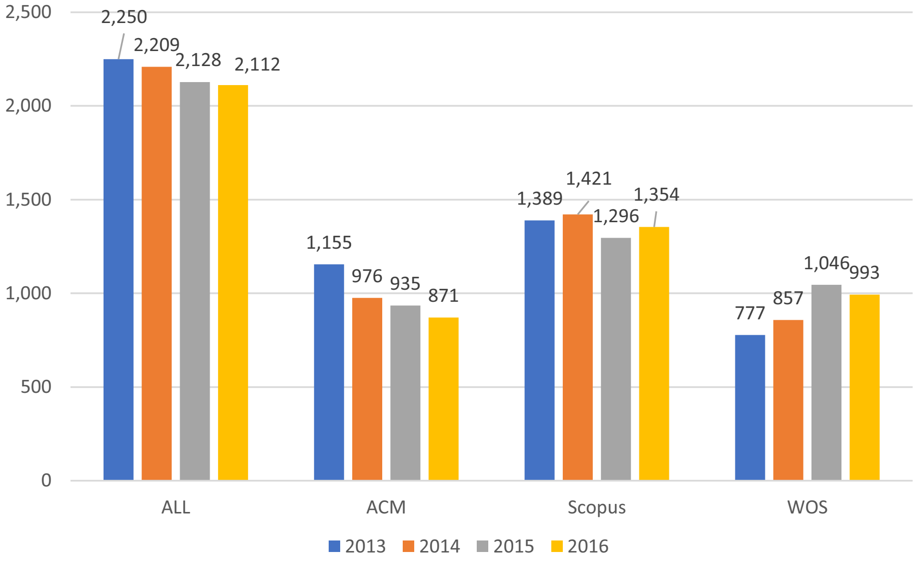

**FIGURE 1** | Number of publication per year and per database.

**TABLE 1** | Citations publications received from the time of publication until May 2017 per database, total number of citations and average number of citations.

| **Citations** | **ACM** | | **Scopus** | | **WoS** | |
|---|---|---|---|---|---|---|
| **Year** | **Total** | **Average per paper** | **Total** | **Average per paper** | **Total** | **Average per paper** |
| 2013 | 3,560 | 3.07 | 5,574 | 4.01 | 2,079 | 2.67 |
| 2014 | 2,168 | 2.24 | 3,746 | 2.64 | 1,412 | 1.66 |
| 2015 | 1,056 | 1.09 | 2,144 | 1.65 | 940 | 0.87 |
| 2016 | 319 | 0.38 | 623 | 0.46 | 236 | 0.25 |
| Total | 7,103 | 1.80 | 12,087 | 2.21 | 4,667 | 1.27 |

**TABLE 2** | Document types per databases, in absolute numbers and percentages.

| **Document types** | **ACM** | **%** | **Scopus** | **%** | **WoS** | **%** |
|---|---|---|---|---|---|---|
| Proceedings papers | 2,727 | 69.3 | 3,383 | 62.0 | 2,228 | 60.7 |
| Journal articles and reviews | 1,017 | 25.8 | 1,914 | 35.1 | 1,408 | 38.3 |
| Books and proceedings | 168 | 4.3 | 26 | 0.5 | 0 | 0.0 |
| Other | 25 | 0.6 | 137 | 2.5 | 37 | 1.0 |

## Document Types

In all three databases more than 60% of the retrieved items were proceedings papers, as can be seen in **Table 2**, which could be expected, since this is a computer science topic, where a large percentage of the publications are in proceedings. In WoS, this was mainly due to the inclusion of the Proceedings Citation Indexes, had we searched in the SCI and SSCI Citation indexes only we would have retrieved only 1,078 items.

Next, we examined which document types are cited more in each database and over time. The results are displayed in **Table 3**. The citations received by the categories books and proceedings and other are negligible and are not displayed. **Table 4** further emphasizes the differences between the three databases: In ACM the average number of citations received by proceedings papers and journal articles is nearly identical, while in WoS, journal articles receive 5.5 times more citations than proceedings papers on average. The average number of citations received was highest on ACM for proceedings papers, and on Scopus for journal articles.

## Overlap

The most interesting finding of this explorative study is the small overlap between the results retrieved by the databases as can be seen in **Figure 2**. We found only 977 documents (11% out of the total number of retrieved documents—8,699) that were retrieved by all three databases. On the other hand, 5,306 documents (61%) were found in a single database only. The largest overlap was between Scopus and WoS, 63% of the documents found by WoS were retrieved also by Scopus, and the smallest overlap was found between WoS and ACM, only 30% of the publication in WoS were found also by ACM. This finding shows that if one wants to retrieve comprehensive results on a topic, in most cases a single database is not sufficient, and more relevant or even partly relevant databases should be consulted.

## Most Cited Publications Per Database

The set of the three most cited documents retrieved by each of the databases is displayed in order to highlight the differences in terms of citation between them. The top three documents ranked by citation counts are displayed in **Table 4** for ACM, Scopus, and WoS, respectively [five documents are listed for WoS, because there were three publications with exactly the same number of citations (34)]. **Table 4** shows, that although some of the articles were indexed by all three databases, there was a single overlap between WoS and Scopus—the second

**TABLE 3** | Citations per document type, database and publication year.

| Doc. types | ACM | | | Scopus | | | WoS | | |
|---|---|---|---|---|---|---|---|---|---|
| | # pubs | Total cits | Ave. cits/paper | # pubs | Total cits | Ave. cits/paper | # pubs | Total cits | Ave. cits/paper |
| **Proceedings papers** | | | | | | | | | |
| 2013 | 843 | 2,572 | 3.05 | 893 | 2,674 | 2.99 | 452 | 435 | 0.96 |
| 2014 | 674 | 1,542 | 2.29 | 880 | 1,819 | 2.07 | 545 | 320 | 0.59 |
| 2015 | 675 | 792 | 1.17 | 792 | 1,046 | 1.32 | 701 | 247 | 0.35 |
| 2016 | 535 | 190 | 0.36 | 818 | 283 | 0.35 | 530 | 27 | 0.05 |
| All years | 2,727 | 5,096 | 1.87 | 3,383 | 5,822 | 1.72 | 2,228 | 1,029 | 0.46 |
| **Journal articles and reviews** | | | | | | | | | |
| 2013 | 261 | 973 | 3.73 | 443 | 2,852 | 6.44 | 316 | 1,640 | 5.19 |
| 2014 | 229 | 575 | 2.51 | 509 | 1,892 | 3.72 | 298 | 1,083 | 3.63 |
| 2015 | 248 | 238 | 0.96 | 468 | 1,042 | 2.23 | 376 | 681 | 1.81 |
| 2016 | 279 | 126 | 0.45 | 494 | 328 | 0.66 | 418 | 205 | 0.49 |
| All years | 1,017 | 1,912 | 1.88 | 1,914 | 6,114 | 3.19 | 1,408 | 3,609 | 2.56 |

**TABLE 4** | Top-cited documents by database.

| Rank | Author | Title | Source | Year | Cits_ acm | Cits_sc | Cits_ wos |
|---|---|---|---|---|---|---|---|
| | | **Most cited ACM** | | | | | |
| 1 | Yuan et al. | Time-aware point-of-interest recommendation | SIGIR | 2013 | 68 | | |
| 2 | Xiao et al. | Expanding the input expressivity of Smartwatches … | SIGCHI | 2014 | 52 | 55 | |
| 3 | Panichella et al. | How to effectively use topic models for software engineering tasks? | ICSE | 2013 | 36 | 73 | 42 |
| | | **Most cited Scopus** | | | | | |
| 1 | Deng and Yu | Deep learning: Methods and applications | Found and Trends in Signal Proc. | 2013 | 22 | 145 | |
| 2 | Hussein et al. | Human action recognition using a temporal hierarchy… | IJCAI | 2013 | 26 | 89 | |
| 3 | Brehmer and Munzner | A multi-level typology of abstract visualization tasks | IEEE Tr. Visualization | 2013 | 29 | 77 | 53 |
| | | **Most cited WoS** | | | | | |
| 1 | Leaman et al. | DNorm: disease name normalization | Bioinformatics | 2013 | | 60 | 56 |
| 2 | Brehmer and Munzner | A multi-level typology of abstract visualization tasks | IEEE Tr. Visualization | 2013 | 29 | 77 | 53 |
| 3a | Benetos et al. | Automatic music transcription … | IEEE TR. FUZZY SYSTEMS | 2013 | | | 34 |
| 3b | Saha et al. | Improving bug localization … | ASE 2013 | 2013 | | | 34 |
| 3c | Srivastava and Salakhutdinov | Multimodal learning with … | J. Machine Learning Res. | 2014 | 19 | 59 | 34 |

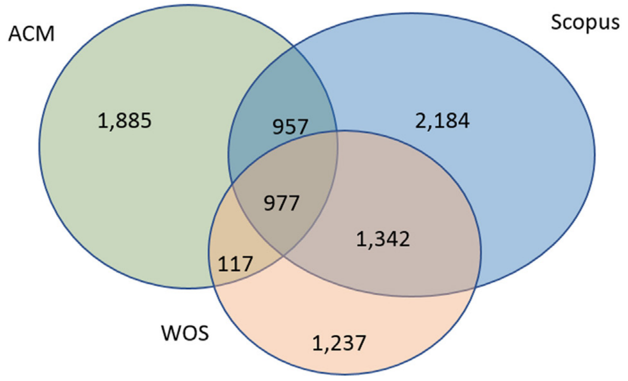

**FIGURE 2** | Overlap between the databases.

most cited publication in WoS was identical to the third most cited publication in Scopus. All three top items are proceedings papers in ACM (although it indexes journal articles as well), in Scopus two and in WoS four out of the five top cited items were journal publications.

To further our understanding of the differences between the databases, we took a closer look at the 977 publications that appeared in all three databases. In this subset, we can compare the citations received from each of the databases for each item.

## Documents Found in All Three Databases

**Table 5** displays the union of the top 10 articles cited by Scopus, ACM and WoS for the 977 documents appearing in all three databases. This table shows extreme differences between the number of citations of the top cited publications in the database. There is some overlap—instead of 30 rows in **Table 5** there are only 19, but only three are among the top-ten in all three databases. Perhaps the most striking example is the Paper by Maleszka et al. that received 29 and 26 citations from Scopus and WoS, respectively, but only a single citation from ACM. The huge differences seen in the rankings seen in **Table 5** are also a result of the differences in the coverage. Databases can only

**TABLE 5** | The top-ten most cited documents in each of the databases.

| First author | Abbrev. Title | Source | Year | Scopus rank | ACM rank | WoS rank | Scopus cits | ACM cits | WoS cits |
|---|---|---|---|---|---|---|---|---|---|
| Brehme | A multi-level typology of abstract visualization tasks | IEEE Tr. Vis. and Comp. Graphics | 2013 | 1 | 3 | 1 | 77 | 29 | 53 |
| Carreno | Analysis of user comments | ICSE | 2013 | 2 | 12 | 3 | 71 | 16 | 32 |
| Srivastava | Multimodal learning with Deep Boltzmann Machines | J. Machine Learning Res. | 2014 | 3 | 10 | 2 | 59 | 19 | 34 |
| Bordes | A semantic matching energy function | Machine Learning | 2014 | 4 | 1 | 30–32 | 56 | 34 | 14 |
| Severyn | Learning to rank short text pairs | SIGIR 2015 | 2015 | 5 | 7 | 61–68 | 46 | 22 | 9 |
| Suominen | Overview of the ShARe/CLEF eHealth evaluation lab 2013 | LNCS | 2013 | 6 | 128–162 | 17–18 | 46 | 3 | 18 |
| Faro | The exact online string matching problem | ACM Comp. Surv. | 2013 | 7 | 16–18 | 15 | 44 | 14 | 20 |
| Suarez-Tangil | Dendroid: A text mining approach | Expert Sys w. Apps | 2014 | 8 | 13–15 | 5 | 43 | 15 | 30 |
| Ding | Collective matrix factorization | IEEE TVCG | 2014 | 9 | 2 | 10–11 | 42 | 30 | 23 |
| Amadeo | Enhancing content-centric networking | Comp. Networks | 2013 | 10 | 21–26 | 4 | 42 | 12 | 31 |
| Dit | Integrating inf. retrieval, execution and link analysis algorithms | Emp Soft. Eng. | 2013 | 11 | 6 | 12–13 | 41 | 23 | 22 |
| Sleiman | A survey on region extractors from web documents | IEEE T. Knowledge and Data Eng. | 2013 | 16 | 8 | 19–23 | 31 | 20 | 17 |
| Jones | Content-based retrieval of human actions | Inf. Sci. | 2013 | 22 | 16–18 | 6–9 | 28 | 14 | 26 |
| Deng | A study of supervised term weighting scheme | Expert Sys w. Apps | 2014 | 32 | 19–20 | 10–11 | 35 | 13 | 23 |
| Eickhoff | Increasing cheat robustness | Inf. Retr. | 2013 | 12–13 | 4 | 6–9 | 36 | 24 | 26 |
| Campos | Survey of temporal information retrieval | ACM Comp. Surv. | 2014 | 12–13 | 9 | 149–183 | 36 | 20 | 4 |
| Maleszka | A method for collaborative recommendation | Knowledge-Based Systems | 2013 | 19–21 | 277–432 | 6–9 | 29 | 1 | 26 |
| Hofmann | Balancing exploration and exploitation | Information Retrieval | 2013 | 23–25 | 5 | 46–60 | 27 | 23 | 10 |
| Li | A method for topological entity matching | Integrated Comp.-Aided Engineering | 2013 | 26–28 | 53–72 | 6–9 | 26 | 6 | 26 |

collect citations from documents indexed by them. The findings based on **Table 5** warrant further examination.

**TABLE 6** | General characteristics of the documents retrieved by all three databases.

| *N* = 977 | Scopus | WoS | ACM | Mendeley |
|---|---|---|---|---|
| Sum of citations/reads | 3,951 | 2,254 | 1,558 | 15,838 |
| No. cited/read docs | 644 | 507 | 434 | 910 |
| % cited/read | 66 | 52 | 44 | 93 |
| Average citations/reads[a] | 6.13 | 4.45 | 3.59 | 17.40 |
| Std citations/reads[a] | 8.74 | 5.58 | 4.42 | 22.67 |
| Median no. citations/reads[a] | 3 | 2 | 2 | 10 |
| Maximum no. citations/reads | 77 | 53 | 34 | 216 |

[a] *The values were computed only for the subset having at least one citation/read.*

**TABLE 7** | General characteristics of the documents published in 2013 and retrieved by all three databases with Mendeley reader counts added.

| *N* = 242 | Scopus | WoS | ACM | Mendeley |
|---|---|---|---|---|
| Sum of citations/reads | 1,795 | 1,051 | 701 | 4,831 |
| No. cited/read docs | 195 | 170 | 150 | 230 |
| % cited/read | 81 | 70 | 62 | 95 |
| Average citations/reads[a] | 9.21 | 6.18 | 4.67 | 21.00 |
| Std citations reads[a] | 11.20 | 7.19 | 5.14 | 24.87 |
| Median no. citations/reads[a] | 5 | 4 | 3 | 13 |
| Maximum no. citations/reads | 77 | 53 | 20 | 200 |

[a] *The values were computed only for the subset having at least one citation/read.*

**Table 6** displays the general characteristics of the set of documents covered by the three databases. We also explored the readership counts of the 977 documents in the online reference manager, Mendeley.

The distributions are heavily skewed, as can be seen from the huge SDs. Note, that if an item is indexed on Mendeley it has at least one reader, and quite amazingly, 93% of the documents in the dataset were saved to at least one Mendeley library. It is well-known that altmetric signals are earlier than actual citations, but even if we limit the dataset to publications in 2013 and consider the number of citations after nearly four years (should be sufficient to gather at least one citation), we see that Mendeley counts are higher (see **Table 7**).

The coverage of older documents is better in all four data sources, but Mendeley still has considerably higher counts than the other three databases. We see that the gap closes because older articles have more time to accrue citations.

Finally, we computed the Spearman correlations between pairs of data sources both for the whole period (2013–2015, 977 docs) and for 2013 only (242 docs) (see **Table 8**). The data sources were the three databases for citation counts and Mendeley for reader counts. For each pair of data sources correlations were computed only for the subset of documents cited/read by both data source. Correlations were computed only for items appearing in the pair of databases with citations/reader counts.

We see that all the correlations are significant and medium high to high between Scopus, WoS, and ACM, and medium strength between Mendeley and the other databases. This finding is supported by previous studies (e.g., Haustein et al., 2014; Zahedi et al., 2014). All the correlations increased when the citation/read window is longer. For Mendeley a possible explanation is that readers come early and citations come later (see also Thelwall and

**TABLE 8** | Spearman correlations between the data sources for documents retrieved by all three citation databases and included in Mendeley—all years and in 2013 only.

| All | ACM | Scopus | WoS | All 2013 | ACM 2013 | Scopus 2013 | WoS 2013 |
|---|---|---|---|---|---|---|---|
| Mendeley | 0.493[a] | 0.532[a] | 0.505[a] | M2013 | 0.550[a] | 0.670[a] | 0.574[a] |
| *N* | 418 | 616 | 489 | *N* | 147 | 187 | 164 |
| ACM | | 0.735[a] | 0.581[a] | ACM2013 | | 0.776[a] | 0.652[a] |
| *N* | | 410 | 349 | *N* | | 145 | 135 |
| Scopus | | | 0.857[a] | Scopus2013 | | | 0.907[a] |
| *N* | | | 489 | *N* | | | 165 |

[a]*Correlation is significant at the 0.01 level (two-tailed).*

Sud, 2016) and for the other databases, the reason might be that it takes time for the databases to stabilize.

## DISCUSSION AND CONCLUSION

As stated in the introduction we expected the ACM Digital Library to have the best coverage, however this assumption was shown to be wrong, as Scopus had the highest number of publications, citations and average number of citations per paper. This is different from the finding for business administration, where the subject specific database had the best coverage (Clermont and Dyckhoff, 2012) based on journal titles.

The major goal of this paper was to highlight the importance of coverage for comprehensive data retrieval. Coverage is one of the parameters in information retrieval evaluation (Cleverdon, 1968), and it has major implications in research assessment as well. WoS and Scopus are selective databases and this is the reason for the varied coverage. However, the small overlap between the databases is worrying. When considering overlap based on journal titles (Gavel and Iselid, 2008; Mongeon and Paul-Hus, 2016), both papers report 45–50% overlap between WoS and Scopus, in this case study out of the 6,814 unique items retrieved by WoS or Scopus, only 2,319 appear in both databases (34%) (see **Figure 1**). Of course, it is not possible to generalize based on the results of a single query, but this issue should be further studied.

Coverage also has a direct impact on citations as well. The fairest comparison is the average number of citations per paper. Here the picture is less clear, when considering proceedings papers (a major document type in computer science), the highest average citations per proceedings paper is by ACM (1.87), closely followed by Scopus (1.72) and WoS lags far behind (0.46). On the other hand, for journal articles, Scopus is highest (3.19), followed by WoS (2.56) and ACM is third with 1.88 average citations per journal article (see **Table 3**).

We also studied Mendeley reader counts for the set of 977 items covered by all three citation databases. We see that the number of readers is considerably higher than the number of citations received, both for the whole dataset (3 times higher than the average number of citation by Scopus), and for papers published in 2013 (twice as high), allowing citations to catch up with reader counts. It should also be noted that even items not cited by ACM, Scopus and WoS have readers on Mendeley. When comparing citation counts from the three databases per paper, the highest correlation is between Scopus and WoS, around 0.9 both for all years and for 2013 only. The reader citation correlations are around 0.5, in line with previous studies (e.g., Haustein et al., 2014; Zahedi et al., 2014).

The results emphasize the need for searching in multiple databases in order to increase recall as recommended by previous studies (e.g., Ramos-Remus et al., 1994; Meho and Yang, 2007; De Groote et al., 2012).

The study is exploratory in its nature and has its limitations. It should be extended to try to understand the meaning of these differences, i.e., why does each database tell us a different story? A single query is not enough for far reaching conclusions, but enough to raise interest to further explore the issue. In addition, the relevance of the retrieved documents should be assessed. In the current study, we relied on the databases, and have not checked relevance manually. The query was not intended to cover IR, but serves as a demonstration of the differences between the databases and also shows that altmetrics (in this case Mendeley reader counts) provide additional insights, like what the users of Mendeley, who are not all citers, are interested in.

## AUTHOR CONTRIBUTIONS

This is a single authored paper. JI is responsible for all parts of the work.

## ACKNOWLEDGMENTS

Mendeley data were gathered with Webometric Analyst, provided by Mike Thelwall. This paper extends a preliminary version presented as a short paper at BIRNDL'17 (Bar-Ilan, 2017). The author holds the copyright of the conference paper.

## REFERENCES

Archambault, É, Campbell, D., Gingras, Y., and Larivière, V. (2009). Comparing bibliometric statistics obtained from the Web of science and Scopus. *J. Assoc. Inf. Sci. Technol* 60, 1320–1326. doi:10.1002/asi.21062

Baeza-Yates, R., and Ribeiro-Neto, B. (1999). *Modern information retrieval*, Vol. 463. New York: ACM Press.

Ball, R., and Tunger, D. (2006). Science indicators revisited – science citation index versus SCOPUS: a bibliometric comparison of both citation databases. *Inf. Serv. Use* 26, 293–301. doi:10.3233/ISU-2006-26404

Bar-Ilan, J. (2008). Which h-index? – A comparison of WoS, Scopus and Google Scholar. *Scientometrics* 74, 257–271. doi:10.1007/s11192-008-0216-y

Bar-Ilan, J. (2010). Web of science with the conference proceedings citation indexes: the case of computer science. *Scientometrics* 83, 809–824. doi:10.1007/s11192-009-0145-4

Bar-Ilan, J. (2017). *Bibliometrics of "Information Retrieval" – A Tale of Three Databases*. Available at: http://ceur-ws.org/Vol-1888/paper7.pdf

Barnett, P., Lascar, C. (2012). Comparing unique title coverage of Web of Science and Scopus in Earth and Atmospheric Sciences. *Issues in Science & Technology Librarianship*. Available at: http://www.istl.org/12-summer/refereed3.html

Clarivate. (2018). *History of Citation Indexing*. Available at: http://clarivate.com/essays/history-citation-indexing/

Clermont, M., and Dyckhoff, H. (2012). "Coverage of business administration literature in Google Scholar: analysis and comparison with EconBiz, Scopus and

Web of science," in *Bibliometrie – Praxis und Forschung*, Vol. 1, 5–1. Available at: http://www.bibliometrie-pf.de/

Cleverdon, C. W. (1968). *The Critical Appraisal of Information Retrieval Systems*. Available at: http://dspace.lib.cranfield.ac.uk/bitstream/1826/1366/1/1968c.pdf

De Groote, S. L., and Raszewski, R. (2012). Coverage of Google Scholar, Scopus, and Web of science: a case study of the h-index in nursing. *Nurs. Outlook* 60, 391–400. doi:10.1016/j.outlook.2012.04.007

De Sutter, B., and Van Den Oord, A. (2012). To be or not to be cited in computer science. *Commun. ACM* 55, 69–75. doi:10.1145/2240236.2240256

Elsevier. (2004). *Scopus Comes of Age*. Available at: http://www.elsevier.com/about/press-releases/science-and-technology/scopus-comes-of-age

Gavel, Y., and Iselid, L. (2008). Web of Science and Scopus: a journal title overlap study. *Online Inf. Rev.* 32, 8–21. doi:10.1108/14684520810865958

Haustein, S., Peters, I., Bar-Ilan, J., Priem, J., Shema, H., and Terliesner, J. (2014). Coverage and adoption of altmetrics sources in the bibliometric community. *Scientometrics* 101, 1145–1163. doi:10.1007/s11192-013-1221-3

Halevi, G., Moed, H., Bar-Ilan, J. (2017). Suitability of Google Scholar as a source of scientific information and as a source of data for scientific evaluation—Review of the Literature. *J. Informetrics* 11, 823–834. doi:10.1016/j.joi.2017.06.005

Hennessey, C. L. (2012). ACM digital library. *Charleston Advis.* 13, 34–38. doi:10.5260/chara.13.4.34

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proc. Natl. Acad. Sci. U.S.A.* 102, 16569–16572. doi:10.1073/pnas.0507655102

Hull, D., Pettifer, S. R., Kell, D. B. (2008). Defrosting the digital library: bibliographic tools for the next generation web. *PLoS Computational Biology* 4, e1000204. doi:10.1371/journal.pcbi.1000204

Jacsó, P. (1997). Content evaluation of databases. *Annu. Rev. Inf. Sci. Technol. (ARIST)* 32, 231–267.

Jacsó, P. (2009). Database source coverage: hypes, vital signs and reality checks. *Online Inf. Rev.* 33, 997–1007. doi:10.1108/14684520911001963

López-Illescas, C., de Moya-Anegón, F., and Moed, H. F. (2008). Coverage and citation impact of oncological journals in the Web of science and Scopus. *J. Informetrics* 2, 304–316. doi:10.1016/j.joi.2008.08.001

Manning, C., Raghavan, P., and Schütze, H. (2009). *Introduction to Information Retrieval*. Cambridge, MA: Cambridge University Press.

Meho, L. I., and Yang, K. (2007). Impact of data sources on citation counts and rankings of LIS faculty: web of science versus Scopus and Google Scholar. *J. Assoc. Inf. Sci. Technol.* 58, 2105–2125. doi:10.1002/asi.20677

Mongeon, P., and Paul-Hus, A. (2016). The journal coverage of Web of science and Scopus: a comparative analysis. *Scientometrics* 106, 213–228. doi:10.1007/s11192-015-1765-5

Perry, S., and Salisbury, L. (1995). Access to information in both CitaDel and first-search: a comparative study of dissertation coverage. *Inf. Technol. Libr.* 14, 17–29.

Ramos-Remus, C., Suarez-Almazor, M., Dorgan, M., Gomez-Vargas, A., and Russell, A. (1994). Performance of online biomedical databases in rheumatology. *J. Rheumatol.* 21, 1912–1921.

Thelwall, M., and Sud, P. (2016). Mendeley readership counts: an investigation of temporal and disciplinary differences. *J. Assoc. Inf. Sci. Technol.* 67, 3036–3050. doi:10.1002/asi.23559

Wikipedia. (2018a). *Science Citation Index*. Available at: http://en.wikipedia.org/w/index.php?title=Science_Citation_Indexandoldid=819147066

Wikipedia. (2018b). *Institute of Scientific Information*. Available at: http://en.wikipedia.org/w/index.php?title=Institute_for_Scientific_Informationandoldid=813831910

Wikipedia. (2018c). *Google Scholar*. Available at: https://en.wikipedia.org/w/index.php?title=Google_Scholar&oldid=823752212

Wilsdon, J. R., Bar-Ilan, J., Frodeman, R., Lex, E., Peters, I., and Wouters, P. (2017). *Next-Generation Metrics: Responsible Metrics and Evaluation for Open Science*. Available at: http://eprints.whiterose.ac.uk/113919/

Zahedi, Z., Costas, R., and Wouters, P. (2014). How well developed are altmetrics? A cross-disciplinary analysis of the presence of 'alternative metrics' in scientific publications. *Scientometrics* 101, 1491–1513. doi:10.1007/s11192-014-1264-0

**Conflict of Interest Statement:** The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.