Scientometrics (2015) 102:2059–2071
DOI 10.1007/s11192-014-1506-1

What is the best database for computer science journal articles?

Antonio Cavacini

Received: 10 September 2014 / Published online: 23 December 2014

Abstract We compared general and specialized databases, by searching bibliographic information regarding journal articles in the computer science field, and by evaluating their bibliographic coverage and the quality of the bibliographic records retrieved. We selected a sample of computer science articles from an Italian university repository (AIR) to carry out our comparison. The databases selected were INSPEC, Scopus, Web of Science (WoS), and DBLP. We found that DBLP and Scopus indexed the highest number of unique articles (4.14 and 4.05 % respectively), that each of the four databases indexed a set of unique articles, that 12.95 % of the articles sampled were not indexed in any of the databases selected, that Scopus was better than WoS for identifying computer science publications, and that DBLP had a greater number of unique articles indexed (19.03 %), when compared to INSPEC (11.28 %). We also measured the quality of a set of bibliographic records, by comparing five databases: Scopus, WoS, INSPEC, DBLP and Google Scholar (GS). We found that WoS, INSPEC and Scopus provided better quality indexing and better bibliographic records in terms of accuracy, control and granularity of information, when compared to GS and DBLP. WoS and Scopus also provided more sophisticated tools for measuring trends of scholarly publications.

Keywords Web of Science · Scopus · DBLP · INSPEC · Google Scholar

Introduction

Several bibliographic databases index computer science papers, allowing users to find information on publications. Bibliographic databases structure information in ways that allow users to retrieve it, by means of keywords, identifiers [for example, the Digital Object Identifier (DOI)], authors' affiliations and links to other records (like references and

A. Cavacini (✉)
Computer Science Department, University of Milan, Via Comelico 39/41, 20135 Milan, Italy
e-mail: antonio.cavacini@unimi.it

citation tracking). In this study we investigated the opportunity of using different databases for finding bibliographic information regarding journal articles in the computer science field. We compared two general and two specialized databases, by evaluating their bibliographic coverage. Our study is based on a manual search in each database of sampled articles related to computer science, followed by an analysis and comparison of the results. We aimed to determine which the most useful bibliographic database in the field of computer science was, and what the advantages of searching one, two or a combination of different databases for finding bibliographic information were. We selected two specialized databases: DBLP and INSPEC. We also selected two general databases: Scopus and Web of Science (WoS) Core Collection. Three out of four of the databases selected were subscribed by our Institution at the time of the searches. The remaining database, DBLP, was freely available. The research question was: is there a need of using multiple databases for searching computer science articles?

Literature review

Several studies compared two or more of the databases that we selected in our study, by analysing their coverage, subjects, nationality of the authors, nature of the publications indexed, or by using other parameters. Many studies focused on one or more of the three most important multidisciplinary databases: WoS, Scopus and GS. Some of these studies assessed the coverage of these and other databases by country. Vieira and Gomes (2009) sampled a set of scholarly articles consisting of the scientific production of 16 Portugese universities between 2000 and 2007, and used it to count the documents referenced in Scopus and WoS. Bartol et al. (2014) analyzed research documents published in Slovenia between 1996 and 2011, finding that Scopus lead over WoS in indexing documents and citations in all research fields. De Sutter and Van Den Oord (2012) correlated WoS, Scopus, ACM, GS and CiteSeer by analysing the publications of three Belgian computer scientists, found that the coverage of Scopus was comparable to that of WoS, and that GS could be useful for identifying and correcting under-citation problems in WoS and Scopus. Adriaanse and Rensleigh (2013) compared the coverage of South African environmental scholarly publication with WoS, Scopus and GS, finding that WoS performed better with coverage, unique items and number of citations, followed in citation counts by GS and then by Scopus; they also found that GS had the most inconsistencies and content quality problems. Bar-Ilan et al. (2007) compared the ranking of publications of highly cited Israeli researchers, measured by citation counts in GS, Scopus and WoS. They introduced a set of measures for comparing the three above mentioned databases, based on the number of bibliographic records and the citations received by a set of publications searched in each database. They found that GS retrieves more items and citations in computer science than in other fields, and that WoS and Scopus rankings were quite similar. Franceschet (2010) compared WoS to GS based on the citation analysis of a sample of publications of Italian science scholars, finding that GS identified five times the publications and eight times the citations of WoS, but provided less accurate data than WoS.

Other studies compared citations from multidisciplinary and specialized databases in different fields of the exact sciences. For example, Whitley (2002) compared the uniqueness and duplication of citation records in WoS and Chemical Abstracts, for a sample from 30 researchers in chemistry. Bakkalbasi et al. (2006) used a set of articles from sample journals from two disciplines (oncology and physics) and 2 years (1993 and 2003) to carry out searches and analyse the overlap of citation counts between GS, WoS

and Scopus. They concluded that each database contained some unique material, which varied depending on the subject of the publication. Abrizah et al. (2013) compared the coverage of Library and Information Science journals in WoS and Scopus, finding that 45 journals (27.8 %) were overlapping, 72 were unique to Scopus, and 23 were unique to WoS. Kousha and Thelwall (2007) found strong positive correlations in several disciplines between citation counts extracted from GS and WoS, including the computing field. Gorraiz and Schloegel (2008) compared the Scopus and WoS databases for the coverage of pharmacology and pharmacy.

Meho and Yang (2007) compared WoS, GS and Scopus based on the number of citations obtained by a set of sampled documents, and concluded that all three databases should be used to get a comprehensive picture of scholarly literature within different fields. Bosman et al. (2006) also compared WoS, GS and Scopus by testing the coverage of specific journals. They also compared WoS and Scopus in specific fields of research, finding that in the computer science field Scopus had 10 % more coverage when compared to WoS. Bar-Ilan (2008) compared WoS, Scopus and GS by sampling authors based on their nationality and on their h-index scores. Bar-Ilan (2010) analysed in detail the citations to a scholarly work in WoS, GS and Scopus, investigating the causes of the unique citations found in WoS (for example: journals uniquely indexed, or mistakes in indexing) when compared to Scopus (which provided citation data only from 1996 onwards, which indexed uniquely some journal titles, and which had indexing mistakes), and concluding that no single citation database can replace all the others. This study also found that GS was better than WoS and Scopus in finding unique citations, though it highlights that only 1/4 of those GS citations came from journals, and only a fraction of the citing journals were peer reviewed (18 out of 28), while many other citations (43 out of 108) came from sources that one might want to dismiss, such as reports, manuscripts, newsletters, encyclopaedia entries and patents. The study also found that most of the unique citations by WoS and Scopus were also indexed by GS, but either GS did not extract the references (51.1 % of the cases), or GS did not access the full text of the items (32.2 % of the cases). Zhang (2014) found that Scopus identified more journal articles than WoS in the computer science field. Wainer et al. (2011) sampled computer science publications from the personal websites of 50 US computer scientists (and 150 more scholars from different areas), checked how many were not indexed by WoS or Scopus (calling them the "Invisible Work"), and found that Scopus was better than WoS in representing computer science researchers. Bosman et al. (2006) also found that Scopus had a better coverage than WoS in the field of computer science. When comparing WoS and GS based on citation counts, de Winter et al. (2014) found that GS had a higher retroactive growth than WoS, and that GS had also limitations compared to WoS (like duplications, fakes, non-peer reviewed documents, and false positives). Other studies also found that GS required extra analyses of the retrieved citing sources, to single out the irrelevant and non-scholarly materials (Gasparyan et al. 2013). Ze and Bo (2012) suggested that GS provides many more citations that WoS and Scopus in the computer science field, because GS indexes extensively conference papers which reference other scholarly articles.

There are many specialized databases which cover the field of computer science. Several studies reviewed and compared one or more of them. For example, Petricek et al. (2005) compared two online computer science databases, DBLP and CiteSeer [which was replaced by CiteSeerX (Giles et al. 1998)], they estimated that DBLP database covered approximately 24 % of the entire literature of computer science, and that the number of papers present in CiteSeer decreased over the years, when compared to the papers indexed in DBLP. Fiala (2012) found that the information indexed by CiteSeerX is often

incomplete and lacks precision (about 10 % of the information stored in the database might have been erroneous). This was linked to the fact that CiteSeerX crawled the web and extracted bibliographic information from documents freely accessible on the Web.

Databases used

We used the database AIR to extract our sample of articles. According to the AIR Website, "AIR (Archivio Istituzionale della ricerca) is the institutional repository of the University of Milan and has been gathering the whole scientific production of the institution since 2004: journal articles, monographs, book chapters, proceedings, PhD Theses and patents. Upload of metadata is mandatory since 2009, and full-text is recommended where it is possible" (AIR 2014). The University of Milan has chosen to devote AIR to the function of storing data on the research conducted by the Institution itself; therefore we found that it contained more bibliographic information than full-text papers. In fact, the AIR instructional repository preserved the full text of the scholarly literature only when the articles were open access and the full-text could be added to the bibliographic records (Cassella and Morando 2012).

We used Scopus, WoS Core Collection, DBLP and INSPEC to do our bibliographic coverage searches. All these databases are used for basic bibliographic searches, but they include extra features, for example citation counts, lists of documents citing the documents searched, bibliometric analysis tools, and access to the full text or links to publishers' full-text.

WoS was the only large scope database providing citation data, until 2004, when two new comprehensive databases were launched: Scopus and Google Scholar (GS; Bar-Ilan 2008). At present, WoS and Scopus are the most widely used bibliographic databases in academic institutions (Zhang 2014). GS, WoS and Scopus are multidisciplinary bibliographic databases, combining a range of common (like the citation data) and unique features (like metrics and tracking services). Scopus indexes more than 20.000 peer-reviewed journals (http://www.elsevier.com/online-tools/scopus/content-overview), and also a significant number of book series and conference proceedings. Scopus continually reviews source titles for inclusion. In WoS, journals are also frequently added or deleted from the databases. In fact "Reuters editorial staff reviews over 2,000 journal titles for inclusion in Web of Science. Around 10–12 % of the journals evaluated are accepted for coverage" (http://wokinfo.com/essays/journal-selection-process). The WoS Core Collection provides access to multidisciplinary information on more than 12,000 high impact scholarly journals in the world. Additional databases are available through subscription, including the INSPEC archive, which is produced by the Institution of Engineering and Technology and covers, among other subjects, computer science. Therefore through the WoS interface, one can search more than one database (in addition to the WoS Core Collection, like for example INSPEC) if the affiliated Institution subscribes to additional databases, and if the user selects to search across two or more of the Databases subscribed to by the Institution. For these reasons, we made separate searches in the WoS Core Collection database and in the INSPEC database. We also combined the results of the WoS and INSPEC databases, to compare the combined results with the Scopus database.

Scopus and WoS provide the user with advanced search features, which make it easier to search, refine and retrieve the bibliographic material. On the contrary, GS lacks sophisticated advanced search options, and allows the user to refine the results only by a few parameters (like the year). GS uses web crawlers to automatically extract bibliographic

information describing scholarly literature from the Internet, including journal websites, authors' personal websites, digital archives and institutional repositories. This might cause errors in metadata, and in the indexing of non-scientific work. GS also lacks clarity on how it collects and processes bibliographic data (de Winter et al. 2014). Over the past few years GS has significantly expanded its indexing of full texts of scholarly literature through agreements with publishers (like Elsevier), online libraries and repositories (Gasparyan et al. 2013). As a consequence, the bibliographic information provided by GS is less consistent and reliable than that provided by WoS and Scopus and requires additional manual control to make sure that bibliographic records match.

In May 2014 we sampled articles published in the years 2011–2013, when at least one of the subjects assigned to the articles was computer science (Subject: INF/01—Informatica), as indexed in the Institutional repository of the University of Milan (AIR: http://air.unimi.it/). The set was cleaned by correcting inaccurate or incomplete information. This left us with three sets of 77, 93 and 70 articles, for the years 2011–2013 respectively.

GS database indexed on average 96.96 % of the articles from our sample. This was a consequence of the fact that the bibliographic records which we sampled from the AIR database were harvested by Google Scholar. In fact the University of Milan (Università degli Studi di Milano) Institutional Archive of Research (AIR) is ranked no. 26 for GS visibility among the Institutional repositories accessible via Web (Top Institutionals 2014).

For these reasons, and to avoid bias or errors in selecting our sample, we did not to use GS to measure the coverage of the bibliographic records, though we compared it to the other databases to test the quality of the bibliographic records.

Other notable bibliographic databases for computer science are ACM Digital Library, IEEEXplore digital library, Microsoft Academic Search and CiteSeerX. ACM Digital Library provides bibliographic information, and access to the full text content (for members who subscribe) for scientific works (journals, conference proceedings, books and technical reports) published by the Association for Computing Machinery. The IEEEXplore digital library provides bibliographic information and access (for paid subscribers) to the full text of journals, conference proceedings, technical standards, eBooks and educational courses published by the IEEE (Institute of Electrical and Electronics Engineers) and its publishing partners, in the field of engineering and computer science. Because both of these databases contain only bibliographic information regarding documents published by the two organizations, we chose not to use them.

We sampled in October 2014 all the articles published in 2013 (84), when at least one of the subjects assigned to the articles was computer science (Subject: INF/01—Informatica), as indexed in the Institutional repository of the University of Milan (AIR: http://air.unimi.it/). We found that approximately 2.38 % were indexed in Microsoft Academic Search, 7.14 % were indexed in CiteSeerX, and 57.14 % were indexed in DBLP. These results seem to confirm previous findings regarding a decrease over the years of the number of papers present in CiteSeer, when compared to the papers indexed in DBLP (Petricek et al. 2005). In addition, recent studies (Orduña-Malea et al. 2014) reported that Microsoft Academic Search lacks updates since 2013, that its coverage plummeted starting from 2011, that it was virtually ignored by bibliometricians and users, and therefore even its disappearance (or lack of updates) has been ignored. For these reasons, we decided not to use MAS (Microsoft Academic search) nor CiteSeerX, and to use DBLP. The DBPL computer science bibliography is a free online bibliographic database for computer science publications, which currently indexes more than 2.6 million publications. The DBLP is supported mainly by the Schloss Dagstuhl, Germany, and the University of Trier, Germany (Ley 2009).

Methods used

We selected all the journal articles published, when at least one of the subjects assigned to the article was computer science (Subject: INF/01—Informatica), as indexed in the Institutional repository of the University of Milan (AIR: http://air.unimi.it/). A total of 1,135 articles were found in October 2014. Of these, 1,075 (94.71 %) were written in English, 56 (4.93 %) were in Italian, 3 (0.26 %) were in Spanish, and 1 (0.09 %) was in French. English-language journals are placed at the top of the impact factor ranking in the field of computer science; therefore it is understandable that authors from non-English speaking countries want to publish their best works in these journals, rather than journals with lower visibility, and which are not indexed in the most used bibliographic databases (Ugoljni and Casilli 2003; González-Alcaide et al. 2012). In fact, studies showed that non-English journals receive fewer citations than pure English journals in selected fields like, for example, chemistry and physics (Liang et al. 2013).

The set of metadata related to each article was extracted from AIR, downloaded and stored in Excel files. The spreadsheets stored information on authors, titles, publication dates and journals. The set was cleaned by correcting inaccurate or incomplete information, and by removing duplicates. No duplicate was found. This left us with a set of 1,135 articles, published between 1979 and 2014. Our units of analysis were the articles. Each article title from the sample was searched individually in the following databases: Scopus, WoS Core Collection, DBLP and INSPEC. We searched in the WoS search interface, selecting the WoS Core Collection, in the basic search field, using the Title option. We used the WoS search interface, selecting the INSPEC database search option, in the basic search field, using the Title option. We selected the Document search field of the Scopus database, using the Article title option. We also used the Complete Search DBLP in the DBLP interface. These choices increase significantly the precision of our searches, when compared to the lower precision that we got from the default "Topic" searches.

We counted the instances in which, by searching a title in a database, we found a bibliographic record which matched the metadata of our sample (by matching authors, years of publication, and journal title). We recorded the instances in which the documents were indexed in each database. We used the count to calculate the percentage of articles indexed by each database, by two or more databases, and by none of the databases. We also used our results to conduct statistical analysis. We reported the number of articles indexed in each database. Table 1 shows the percentage of articles indexed by each database sampled in our research.

Elsevier's Scopus retrieved the highest percentage of articles (75.86 %; Table 1). Scopus was followed by WoS Core Collection (indexing 64.49 %). The larger coverage of Scopus, when compared to WoS, confirmed the findings from similar researches (for example, Vieira and Gomes 2009). Also, WoS combined with INSPEC (through the "all databases search" option) indexes a number of publications (74.80 %) comparable to those indexed by Scopus (75.86 %).

Table 1 Percentage of articles indexed by the selected databases

Database	Scopus	WoS Core Collection	DBLP	INSPEC	WoS + INSPEC
Total	75.86	64.49	61.15	53.39	74.80

Among the specialized databases, DBLP had the highest average percentage of articles indexed (61.15 %), followed by INSPEC (53.39 %). We used statistical analysis to find correlations among the results from each database, which suggested the most relevant comparisons to be carried out. Also, we dichotomized the set of databases in multidisciplinary (Scopus, and WoS) and specialized (DBLP and INSPEC) databases, in order to carry out additional comparisons. We entered our results as values 1 (indexed) and 0 (not indexed) in spreadsheets, and exported these values for analysis, using SPSS version 21. We tested Pearson bivariate correlations among the findings of our searches in the databases.

We found that Scopus strongly correlates with WoS (Core Collection) at the 0.01 level (2-tailed) over our 1,135 publication analysis (0.640). Previous studies (Bar-Ilan et al. 2007; Archambault et al. 2009) found remarkably strong correlations between Scopus and WoS. For these reasons, we compared WoS (Core Collection) and Scopus, and the combined search of Wos and INSPEC databases ("all databases" option) to Scopus. We also compared the two specialized databases: DBLP and INSPEC, which correlated positively at the 0.01 level (2-tailed) over our 1,135 publication analysis (0.389).

Overlapping and uniqueness of database coverage

In order to determine the amount of articles uniquely indexed by each database, the amount of overlapping coverage, and the amount of articles not indexed by any database, we recorded the instances, in which the articles from our samples were indexed in each of the four databases selected, in spreadsheets. We assigned the value 1 to the articles indexed, and 0 to the articles not indexed. We used the results to compare the coverage of any two, three or four databases w, x, y and z, through matching algorithms. We calculated the sum of the values when the conditions were satisfied, and the percentages. The results are displayed in Tables 2, 3, 4 and 5.

Overall, DBLP and Scopus were the databases that indexed the highest number of unique articles (4.14 and 4.05 % respectively), when compared to the other databases (Table 2). Also, the highest average number of unique articles was indexed by Scopus (13.83 %), when compared to WoS (2.47 %; Table 3). Other studies found similar results (Zhang 2014; Wainer et al. 2011; Bosman et al. 2006). We also found that the average number of unique articles indexed by Scopus (8.11 %) was almost similar to the average number of the combined searches in WoS and INSPEC (7.05 %; Table 4). We also found that DBLP had a greater number of unique articles indexed (19.03 %), when compared to INSPEC (11.28 %; Table 5).

Scopus and WoS overlapped, on average, on 62.03 % of the articles indexed (Table 3). The overlap of Scopus, WoS and INSPEC combined, averaged 67.75 % of the articles indexed (Table 4). These results are similar to the findings of Vieira and Gomes (2009),

Table 2 Percentage of articles indexed in only one of the four databases selected and not in the others; in none of the databases, and in all of the four databases selected

Database	Only Scopus	Only WoS	Only DBLP	Only INSPEC	None of the databases	All databases
Total	4.05	1.06	4.14	1.32	12.95	34.45

Table 3 Comparison of WoS (Core Collection) and Scopus in %

Database	Not Scopus not WoS	Scopus and WoS	Scopus only	WoS only
Total	21.67	62.03	13.83	2.47

Table 4 Comparison of WoS combined with INSPEC (WoS + I) and Scopus in %

Database	Not Scopus not WoS + I	Both Scopus and WoS + I	Scopus only	WoS + I only
Total	17.09	67.75	8.11	7.05

Table 5 Comparison of DBLP and INSPEC in %

Database	Only DBLP	Only INSPEC	Not DBLP not INSPEC	Both DBLP and INSPEC
Total	19.03	11.28	27.58	42.11

who concluded that about 2/3 of their sample of documents could be found in both databases.

Overall, Scopus indexed the highest average number of articles in our sample, and the highest average set of unique articles, when compared to WoS. Our results also show that each of the four databases indexed a set of unique articles (from 4.05 % of Scopus to 1.06 % of WoS; Table 2), suggesting that the four databases complement one another, and that one should be cautious in restricting the searches to only one of them. Also, 12.95 % of the articles sampled were not indexed in any of the databases selected (Table 2), suggesting that neither of the selected databases provides a comprehensive gathering of the targeted literature.

The quality of the bibliographic records

In 391 cases (34.45 %) the articles of our sample were indexed in all four databases (Table 2). We used these articles to carry out an analysis of the quality of the bibliographic records. We searched each of the 391 overlapping articles in all four databases and also in Google Scholar, we inspected the bibliographic records retrieved from each database, and measured their quality, by identifying a set of 20 common structured elements. We selected the following 20 elements from the bibliographic records retrieved:

Table 6 Quality, precision, ranking position and citation counts of the five databases

	SCOPUS	WoS Core Collection	DBLP	INSPEC	GS
Quality	18.00	19.17	7.00	19.00	5.74
Cited by	18.75	11.03	0.00	10.54	35.01
Ranking position	14.16	9.42	1.10	32.70	1.01
Precision	0.04	0.07	0.79	0.02	0.000020

Titles, authors names, authors' information, abstract, keywords (controlled), keywords (uncontrolled), publication date, DOI, "cited by" feature, journal title, ISSN, volume (or issue) number, pages, references, publisher, classification, language (of the article), metrics, funding, full text links (options).

We checked the completeness, accuracy and consistency of the 20 selected elements for the records, by measuring each field, by assigning values of 0 (when we detected issues) and 1 (when no problems were found), and by summing the values. Similar methods can be found in Bellini and Nesi (2013) and Palavitsinis et al. (2014). We ranked the databases based on the scores of the elements measured.

WoS ranked at the top with a score of 19.17 (see Table 6), and it provided two extra elements, when compared to Scopus: Classification, and Funding information (Funding Agency and Grant Number, in 16.62 % of the articles).GS provided the full list of authors in 74.42 % of the articles (due to the layout of the records). No quality issues were found in the Scopus, DBLP and INSPEC records.

We counted the average precision of the databases, defined as the ratio of relevant records retrieved, to the total number retrieved. With higher precision, fewer non-relevant documents were retrieved (Wolfram 2003). We searched one article at a time, therefore we reported the ratio of 1 to the total number of articles retrieved, 1 out of 1 being the best possible score (see Table 6). When searching in Scopus the "title, abstract and keywords" field (which was the default one) we retrieved on average 23.85 articles, with a precision of 0.04. When searching in WoS the "Topic" field (which is the default one in the WoS interface) we retrieved on average 14.48 articles for the Core Collection, and on average 64.20 articles for the INSPEC, with a precision of 0.07 and of 0.02 respectively. When searching DBLP with the Complete Search interface, the precision was 0.79: the best score among all databases. The precision became 100 % when we searched the full title in the "Title" field of the WoS, Scopus and INSPEC interfaces. No option for a "Title" search was provided by the DBLP database.

GS was the least precise database of all, with an average of 44,537.89 articles retrieved and precision of 0.00002 (see Table 6). However, Precision didn't seem an objective measure of the accuracy of the databases, especially for GS, as the articles searched were almost always on the top of the GS search results, although this search engine retrieved many thousands of articles on average. Also, GS used a feature called "Showing the best result for this search", which displays the best results, providing the option of searching additional results if the article searched for is not listed among those provided. For these reasons, we added the measure of Ranking Position (or Search result ranking position; Table 6), which counted the average position in which the searched for article appeared among the list the articles displayed in the result pages (1 being the top position). Using this measurement, GS ranked at the top of the score (1.01), closely followed by DBLP (1.10). WoS (9.42) was better than Scopus (14.16), while INSPEC was at the bottom of the list (32.70). These results show that although GS was less precise than Scopus, INSPEC and WoS, it was far better though in ranking position.

We also recorded the citation counts per article from each database (GS, Scopus, WoS and INSPEC), using our subset of 391 articles, and compared the average results. GS provided on average 35.01 citations per article, WoS Core Collection provided 11.03 citations per article, and Scopus provided 18.75 citations per article (see Table 6). These results are consistent with the findings by Bar-Ilan (2010).

Overall, DBLP provided records in a homogeneous and controlled way, but its records provided less features and information when compared to those of WoS, INSPEC and Scopus. The information provided by GS was sometimes partial, like in the case of authors

and abstracts. Also, GS provided the title of the article, the names of the authors (up to 3, but not of the remaining authors if the article had more than three authors), the publication date, and the journal source. GS relied on other databases or online information to find additional bibliographic information and any available full texts. In fact, GS listed web sources where to find more information about the article (on average it listed 8.33 sources, called "versions").

When taking into account the features of the three multidisciplinary searching tools, WoS and Scopus provided bibliographic records in a more standardized fashion, which was more usable and required less manual control, especially when compared to the lack of consistency and accuracy of the GS bibliographic data (Franceschet 2010). Other studies highlighted the errors found in GS, the noisiness of the database (retrieving informal material, draft papers, unpublished writings) and the lack of quality control; Aguillo (2012) suggested that GS seemed a cheap substitute for more reliable bibliographic databases like Scopus and WoS. Overall, GS had more citation counts, less precision and the quality of the data indexed was remarkably lower than that of bibliographic databases like WoS and Scopus, was of limited use and couldn't replace the other two databases due to quality issues (Meho and Yang 2007). However, GS had the best ranking positions, and it was the only free multidisciplinary database among those tested (while DBLP was the only free specialized database), and a powerful and outstanding search engine for searching scholarly articles in the computer science field.

Conclusions

This study investigated the difference in coverage of bibliographic records in the computer science field, sampled from an Italian university repository, and searched in four bibliographic databases: Scopus, WoS, INSPEC and DBLP.

When comparing the coverage of the four databases, DBLP and Scopus indexed the highest number of unique articles (4.14 and 4.05 % respectively), and each of the four databases indexed a set of unique articles (the lowest 1.06 % was of WoS), suggesting that the four databases complement one another, and that one should be cautious in restricting the searches to only one of them. Also, 12.95 % of the articles sampled were not indexed in any of the databases selected, suggesting that neither of the selected databases provides a comprehensive gathering of the targeted literature.

We found significant positive correlations at the 0.01 level (2-tailed) between WoS and Scopus. We found that when INSPEC was searched together with WoS, it had a coverage of computer science articles which was comparable to that of Scopus.

When comparing the coverage of WoS and Scopus, we found that there was a significant overlap between the two databases, that they both indexed a significant number of unique articles (Scopus 13.83 %; WoS 2.47 %), and that Scopus was better than WoS for identifying computer science publications. We found that the average number of unique articles indexed by Scopus (8.11 %) was close to the average number of the combined searches in WoS and INSPEC (7.05 %). We also found that Scopus and WoS overlapped, on average, 62.03 % of articles indexed, and that the overlap of Scopus, WoS and INSPEC combined, averaged 67.75 % of the articles indexed. We also found that DBLP had a greater number of unique articles indexed (19.03 %), when compared to INSPEC (11.28 %). These results suggested that the four databases complemented each other, in a way that neither one could replace the other.

In measuring the quality of the bibliographic records, we compared five databases: Scopus, WoS, INSPEC, DBLP and GS. We found that WoS, INSPEC and Scopus provided better quality indexing and better bibliographic records in terms of accuracy, consistency, control, relevance and granularity of information, when compared to GS and DBLP. WoS and Scopus also provided more sophisticated tools for measuring trends of scholarly publications. GS couldn't be considered as a substitute for bibliographic databases such as WoS and Scopus, as it needed to improve the metadata, the search functions and the control of information (de Winter et al. 2014). However, our results also suggest that GS was a powerful search engine for searching scholarly articles in the computer science field.

The fact that we sampled journal articles indexed in an Institutional repository seems to limit the findings. In fact, we used a sample of articles which represented the scientific production of researchers who worked in an Italian university. The journal articles indexed in the years 1979–2014 were on average only 29.64 % (1,135) of the total number of documents indexed by the AIR repository (3,828 documents as of October 2014, when at least one of the subjects assigned to the articles was computer science). In fact, in the computer science field, proceedings are usually a prime avenue of publication, and selected conference proceedings are as prestigious as journal articles (Franceschet 2010), though on average proceedings papers receive less citations when compared to research articles (Zhang and Glänzel 2012). Therefore, the conference objects published as journal articles, and the proceeding papers included in monographic issues and indexed in AIR, require additional analysis, taking into account that some high quality conferences papers are as prestigious as journal articles in the field of computer science (Freyne et al. 2010; Sicilia et al. 2011). As a consequence, caution should be exercised in light of the fact that we sampled only journal titles, and that conference proceedings represent a large portion of the computer science literature (Zhang 2014). Also, WoS and Scopus constantly update their source list, therefore the number of scientific publications indexed keeps changing throughout the year. As a consequence, identical searches might give different results if repeated over time in WoS and Scopus, as journals and other sources are frequently added or deleted from the databases. Similar variations occur with the other databases that we used. For these reasons, further research is needed to confirm the above mentioned findings over a prolonged period of time, and with a more representative sample.

References

Abrizah, A., Zainab, A. N., Kiran, K., & Raj, R. G. (2013). LIS journals scientific impact and subject categorization: A comparison between Web of Science and Scopus. *Scientometrics, 94*(2), 721–740.

Adriaanse, L. S., & Rensleigh, C. (2013). Web of Science, Scopus and Google Scholar a content comprehensiveness comparison. *Electronic Library, 31*(6), 727–744.

Aguillo, I. F. (2012). Is Google Scholar useful for bibliometrics? A webometric analysis. *Scientometrics, 91*(2), 343–351.

Archambault, É., Campbell, D., Gingras, Y., & Larivière, V. (2009). Comparing bibliometric statistics obtained from the Web of Science and Scopus. *Journal of the American Society for Information Science and Technology, 60*(7), 1320–1326.

Archivio Istituzionale della Ricerca. (2014). Retrieved April 11, 2014 from http://air.unimi.it.articles/000553/article.pdf

Bakkalbasi, N., Bauer, K., Glover, J., & Wang, L. (2006). Three options for citation tracking: Google Scholar, Scopus and Web of Science. *Biomedical Digital Libraries, 3*(1), 7.

Bar-Ilan, J. (2008). Which h-Index? A comparison of WoS, Scopus and Google Scholar. *Scientometrics, 74*, 257–271.

Bar-Ilan, J. (2010). Citations to the "introduction to informetrics" indexed by WoS. Scopus and Google Scholar. *Scientometrics, 82*(3), 495–506.

Bar-Ilan, J., Levene, M., & Lin, A. (2007). Some measures for comparing citation databases. *Journal of Informetrics, 1*, 26–34.

Bartol, T., Budimir, G., Dekleva-Smrekar, D., Pusnik, M., & Juznic, P. (2014). Assessment of research fields in Scopus and Web of Science in the view of national research evaluation in Slovenia. *Scientometrics, 98*(2), 1491–1504.

Bellini, E., & Nesi, P. (2013). Metadata quality assessment tool for open access cultural heritage institutional repositories.In *Proceeding of the ECLAP 2013 Conference, 2nd International Conference on Information Technologies for performing arts, media access and entertainment, ECLAP 2013, Lecture Notes in Computer Science, LNCS 7990* (pp. 90–103) Springer.

Bosman, J., van Mourik, I., Rasch, M., Sieverts, E., & Verhoeff, H. (2006). Scopus reviewed and compared. The coverage and functionality of the citation database Scopus, including comparisons with Web of Science and Google Scholar, Utrecht: Utrecht University Library. Retrieved from http://dspace.library.uu.nl/handle/1874/18247

Cassella, M., & Morando, M. (2012). Fostering new roles for librarians: Skills sets for repository managers—Results of a survey in Italy. *LIBER Quarterly*, *21*, 407–428. Retrieved from http://liber.library.uu.nl/publish/

De Sutter, B., & Van Den Oord, A. (2012). To be or not to be cited in Computer Science. *Communications of the ACM, 55*(8), 69–75.

de Winter, J. C. F., Zadpoor, A. A., & Dodou, D. (2014). The expansion of Google Scholar versus Web of Science: A longitudinal study. *Scientometrics, 98*(2), 1547–1565.

Fiala, D. (2012). Bibliometric analysis of CiteSeer data for countries. *Information Processing and Management, 48*(2), 242–253.

Franceschet, M. (2010). A comparison of bibliometric indicators for computer science scholars and journals on Web of Science and Google Scholar. *Scientometrics, 83*(1), 243–258.

Freyne, J., Coyle, L., Smyth, B., & Cunningham, P. (2010). Relative status of journal and conference publications in Computer Science. *Communications of the ACM, 53*(11), 124–132.

Gasparyan, A. Y., Ayvazyan, L., & Kitas, G. D. (2013). Multidisciplinary bibliographic databases. *Journal of Korean Medical Science, 28*(9), 1270–1275.

Giles, C. L., Bollacker, K. D., & Lawrence, S. (1998). CiteSeer: an automatic citation indexing system. *Proceedings of the third ACM conference on Digital libraries*: 89–98. doi:10.1145/276675.276685. ISBN 0-89791-965-3. CiteSeerX:10.1.1.30.6847.

González-Alcaide, G., Valderrama-Zurián, J. C., & Aleixandre-Benavent, R. (2012). The impact factor in non-English-speaking countries. *Scientometrics, 92*(2), 297–311.

Gorraiz, J., & Schloegl, C. (2008). A bibliometric analysis of pharmacology and pharmacy journals: Scopus versus Web of Science. *Journal of Information Science, 34*(5), 715–725.

Kousha, K., & Thelwall, M. (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*(7), 1055–1065.

Ley, M. (2009). DBLP: Some lessons learned. *Proceedings of the VLDB Endowment, 2*(2), 1493–1500.

Liang, L., Rousseau, R., & Zhong, Z. (2013). Non-English journals and papers in physics and chemistry: Bias in citations? *Scientometrics, 95*(1), 333–350.

Meho, L., & Yang, K. (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*(13), 2105–2125.

Orduña-Malea, E., Ayllón, J. M., Martín-Martín, A., & López-Cózar, E. D. (2014). Empirical evidences in citation-based search engines: Is microsoft academic search dead? Granada: EC3 Reports, 16: May 21, 2014. arXiv:1404.7045 [cs.DL]. Retrieved from http://arxiv.org/abs/1404.7045

Palavitsinis, N., Manouselis, N., & Sanchez-Alonso, S. (2014). Metadata quality in digital repositories: Empirical results from the cross-domain transfer of a quality assurance process. *Journal of the Association for Information Science and Technology, 65*(6), 1202–1216.

Petricek, V., Cox, I. J., Han, H., Councill, I. G., & Giles, C. L. (2005). A comparison of on-line computer science citation databases. In A. Rauber, S. Christodoulakis, & A. M. Tjoa (Eds.) ECDL 2005. *LNCS* (vol. 3652, pp. 438–449). Heidelberg: Springer.

Sicilia, M., Sánchez-Alonso, S., & García-Barriocanal, E. (2011). Comparing impact factors from two different citation databases: The case of Computer Science. *Journal of Informetrics, 5*(4), 698–704.

Top Institutionals. (2014). Retrieved April 11, 2014 from http://repositories.webometrics.info/en/top_Inst?sort=asc&order=scholar

Ugoljni, D., & Casilli, C. (2003). The visibility of Italian journals. *Scientometrics, 56*(3), 345–355.

Vieira, E. S., & Gomes, J. A. N. F. (2009). A comparison of Scopus and Web of Science for a typical university. *Scientometrics, 81*(2), 587–600.

Wainer, J., Billa, C., & Goldenstein, S. (2011). Invisible work in standard bibliometric evaluation of Computer Science. *Communications of the ACM, 54*(5), 141–148.

Whitley, K. M. (2002). Analysis of SciFinder Scholar and Web of Science citation searches. *Journal of the American Society for Information Science and Technology, 53*(14), 1210–1215.

Wolfram, D. (2003). *Applied informetrics for information retrieval research*. Westport: Libraries Unlimited Inc.

Ze, H., & Bo, Y. (2012). Mining Google Scholar citations: An exploratory study. ICIC 2012, *LNCS* 7389, pp. 182–189.

Zhang, L. (2014). The impact of data source on the ranking of computer scientists based on citation indicators: A comparison of Web of Science and Scopus. *Issues in Science and Technology Librarianship*, 75. Retrieved from http://www.istl.org/14-winter/refereed2.html.

Zhang, L., & Glänzel, W. (2012). Proceeding papers in journals versus the "regular" journal publications. *Journal of Informetrics, 6*(1), 88–96.