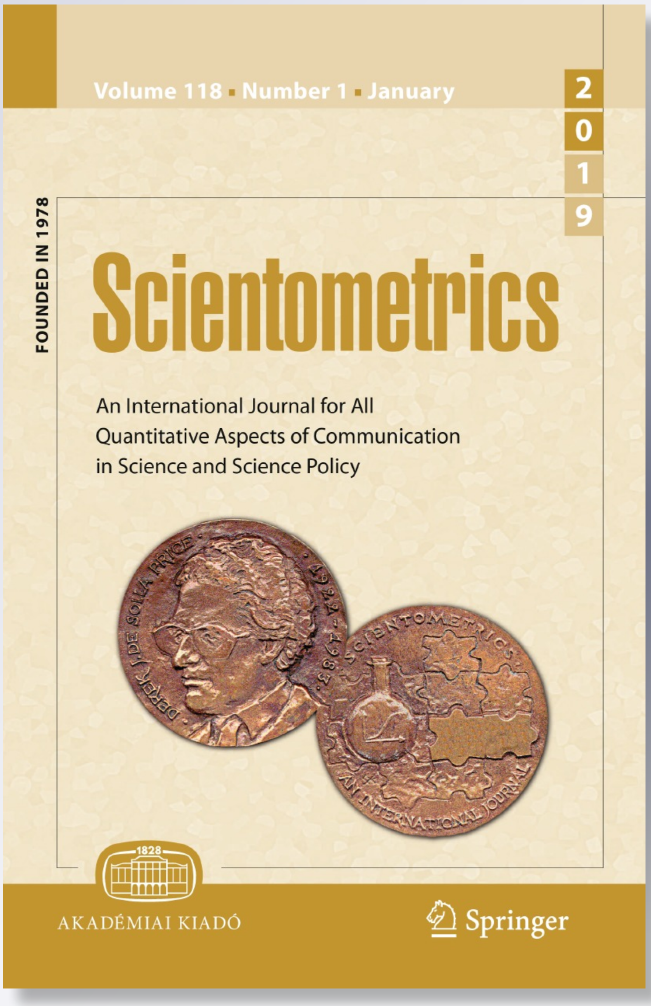
Google Scholar to overshadow them all? Comparing the sizes of 12 academic search engines and bibliographic databases

Michael Gusenbauer

Scientometrics
An International Journal for all Quantitative Aspects of the Science of Science, Communication in Science and Science Policy

ISSN 0138-9130
Volume 118
Number 1

Scientometrics (2019) 118:177-214
DOI 10.1007/s11192-018-2958-5


Volume 118 · Number 1 · January
2019
FOUNDED IN 1978
Scientometrics
An International Journal for All Quantitative Aspects of Communication in Science and Science Policy
1828
AKADÉMIAI KIADÓ
Springer

Scientometrics (2019) 118:177–214
https://doi.org/10.1007/s11192-018-2958-5

Google Scholar to overshadow them all? Comparing the sizes of 12 academic search engines and bibliographic databases

Michael Gusenbauer[1]

Received: 14 April 2018 / Published online: 10 November 2018

Abstract

Information on the size of academic search engines and bibliographic databases (ASEBDs) is often outdated or entirely unavailable. Hence, it is difficult to assess the scope of specific databases, such as Google Scholar. While scientometric studies have estimated ASEBD sizes before, the methods employed were able to compare only a few databases. Consequently, there is no up-to-date comparative information on the sizes of popular ASEBDs. This study aims to fill this blind spot by providing a comparative picture of 12 of the most commonly used ASEBDs. In doing so, we build on and refine previous scientometric research by counting query hit data as an indicator of the number of accessible records. Iterative query optimization makes it possible to identify a maximum number of hits for most ASEBDs. The results were validated in terms of their capacity to assess database size by comparing them with official information on database sizes or previous scientometric studies. The queries used here are replicable, so size information can be updated quickly. The findings provide first-time size estimates of ProQuest and EbscoHost and indicate that 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.

Keywords Academic search engine · Academic bibliographic database · Query hit count · Size · Iterative analysis · Metrics · Google Scholar

Introduction

Academic search engines and bibliographic databases (ASEBDs) are now the standard place from which to access up-to-date scientific publications. These services make an ever-increasing stock of scientific knowledge accessible for scientists by filtering the most relevant information. Students and scholars start their web searches with ASEBDs providing the lens through which they view science and conduct investigations (Haines et al. 2010).

✉ Michael Gusenbauer
michael.gusenbauer@jku.at

[1] Institute of Innovation Management, Johannes Kepler University Linz, Altenberger Straße 69, 4040 Linz, Austria

In the late 1990s, the rise of the internet saw ASEBDs become relevant and increasingly replace traditional offline systems of information retrieval (for an overview see Table 1). Existing data providers and publishers such as ProQuest, Ebsco, Thomson Reuters, and Elsevier entered the online realm to offer their information services. Nevertheless, only in the early 2000s did innovations in data access transform access to scientific information. Large crawler-based search engines such as Google Scholar, Microsoft Academic, and Scirus started to make huge volumes of scholarly data readily accessible to anyone at no cost (Ortega 2014). Google Scholar became the number one go-to information source in academia (van Noorden 2014) and is often used due to its convenience and users' familiarity with the search system (Georgas 2014; Jamali and Asadi 2010; Duke and Asher 2012). While not all documents were available in full-text form, Google Scholar could build up a significant resource of publicly available documents covering a large array of disciplines and languages. Google Scholar seems unrivalled in the efficient and effective provision of scholarly documents online. Yet, Microsoft Academic, after discontinuing its service, relaunched its academic search machine in 2017 to compete with Google Scholar once again (Harzing and Alakangas 2017). Beside Google Scholar and Microsoft Academic, there are however many other larger multidisciplinary search engines, bibliographic databases, and other information services that try to convince academic users of the validity of their unique information offering.

Search system scope

While academic users have a choice of which service to use, it is often unclear which search system serves them best. There are multiple criteria for evaluating the quality of search systems, such as relevance, objectivity, or accuracy (Jansen and Spink 2003; Brophy and Bawden 2005; Eastman and Jansen 2003). In this study we concentrate on one criterion, the *scope* of a search system in terms of its size, reflecting the number of accessible resources for a specific user (Lawrence and Giles 1999; Grigas et al. 2016; Hawking et al. 2001). The results an academic user obtains with a query are influenced, among other quality criteria, by the limits of the data available on the specific search engine or bibliographic database. When information overload is accounted for with relevance, a larger scope brings better search results than a smaller scope.

In addition to academic users, other groups interested in knowing the sizes of academic search systems include: information specialists at research institutions or libraries interested in knowing the sizes of search systems at a *particular* point of time to allow comparison, and in knowing the size of single search systems at *multiple* points of time to allow longitudinal assessment of performance and stability. Therefore, knowing the scope of a given search system is not only worthwhile for academic users, but also for information specialists.

Nevertheless, the growth in the ASEBD offering not only improved the way scholars accessed information, but also created drawbacks in transparency of scope (Halevi et al. 2017; Shariff et al. 2013; Aguillo 2012). Particularly Google Scholar's scope remains a mystery and a source of speculation, especially because Google Scholar's aim is to index the entire universe of scholarly information, estimating its size has attracted numerous academic works. Knowing Google Scholar's size and growth might be indicative of the size and growth of scholarly data as a whole (Orduña-Malea et al. 2015; Halevi et al. 2017): "[p]erhaps even Google Scholar does not know this "number"… a number that approximately represents the online scientific heritage circulating at present" (Orduña-Malea,

Table 1 Overview of the characteristics of 12 large multidisciplinary ASEBDs

Name of database	Type	Owner	Year of online launch	Types of documents covered	Subject coverage	Coverage dates	Search fields available	Content
AMiner	Search engine	Tsinghua University	2006	Journals publications, Conference proceedings	Multidisciplinary with a focus on Computer Science	Unknown to present	Keywords, Name, Organization	Upon request
Bielefeld Academic Search Engine (BASE)	Search engine	University Library Bielefeld	2004	Journals publications, Conference proceedings, Books, Theses, News articles, Patents, Working papers	Multidisciplinary	Unknown to present	Entire document, Title, Author, Subject, DOI, URL, Geography, Year, Access, Licence, Language	Open
CiteSeerX	Search engine	Pennsylvania State University	1997 and again in 2007	Journals publications, Conference proceedings, Books, Technical reports	Multidisciplinary with focus a on Computer and Information Science	Unknown to present	Text, Title, Author, Affiliation, Venue, Keywords, Abstract, Publication year, Min. citations	Open
EbscoHost	Aggregator (paywall)	EBSCO Publishing, Inc.	1996	Journals publications, Conference proceedings, Books, Theses, Magazines, News articles, Working papers	Multidisciplinary (depends on selected databases)	Depends on selected databases	All text, Author, Title, Subject terms, Source, Abstract, ISSN, ISBN, Language, Published date, Scholarly journal	Proprietary

Table 1 (continued)

Name of database	Type	Owner	Year of online launch	Types of documents covered	Subject coverage	Coverage dates	Search fields available	Content
Google Scholar	Search engine	Google LLC	2004	Journals publications, Conference proceedings, Books, Theses, Patents, Working papers, Case laws	Multidisciplinary	1700 (some books earlier) to current	Article, Title, Author, Publication, Date	Proprietary
Microsoft Academic	Search engine	Microsoft	2009 and again in 2017	Journals publications, Conference proceedings, Books, Theses, Patents, Working papers	Multidisciplinary	1800 to present	One for all (semantic search instead of keyword search)	Proprietary
ProQuest	Aggregator (paywall)	ProQuest LLC	1996	Journals publications, Conference proceedings, Books, Theses, Technical reports, Magazines, News articles, Working papers	Multidisciplinary (depends on selected databases)	Depends on selected databases (as early as 750 to current)	No full text, Abstract, Subjects, Organization, Person, Author, Text, Title, Publication, Accession number, ISSN, Peer reviewed, Dates, Language, Source type, Document type	Proprietary

Table 1 (continued)

Name of database	Type	Owner	Year of online launch	Types of documents covered	Subject coverage	Coverage dates	Search fields available	Content
Q-Sensei Scholar	Search engine	Q-Sensei Inc.	2008	Journals publications, Books	Multidisciplinary	Unknown to present	Title, Author, Publication date, Venue, Abstract, Keywords, Document type	Proprietary
Scopus	Bibliographic database (paywall)	Elsevier B.V.	2004	Journals publications, Conference proceedings, Books, Magazines	Multidisciplinary	1788 to present	Authors, 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	Proprietary
Semantic Scholar	Search engine	Allen Institute for Artificial Intelligence	2015	Journals publications, Conference proceedings	Multidisciplinary with a focus on Computer Science and Medicine	1936 to present	Publication type, Year, Author, Journals, and conferences	Proprietary

Table 1 (continued)

Name of database	Type	Owner	Year of online launch	Types of documents covered	Subject coverage	Coverage dates	Search fields available	Content
Web of Science	Bibliographic database (paywall)	Clarivate Analytics	1997	Journals publications, Conference proceedings, Books, Theses, Technical reports, News articles, Patents	Multidisciplinary (depends on selected databases)	1900 to present	Topic, Title, Author, Author identifiers, Group author, Editor, Publication name, DOI, Year published, Address, Open access, Highly cited, Research domain, Language, Research area, Funding agency, Group/Corporate authors, Document types	Proprietary
WorldWideScience	Search engine	U.S. Department of Energy	2007	Journals publications, Books, Theses, Patents	Multidisciplinary	Depends on federated databases	Full Record, Title, Author, Date, Language	Unclear

Ortega (2014), Pederson (2001) and Adamick and Reznik-Zellen (2010)

Ayllón, et al. 2014, p. 29). Researchers remain frustrated over Google Scholar's secrecy: "its secretiveness about every aspect of Google Scholar is on par with that of the North Korean government. The database is getting bigger and bigger but in the wrong way, through hoarding giga collections of irrelevant and/or non-scholarly content" (Jacsó 2012, p. 466). Google Scholar encourages scholarly research on its coverage to address such criticism, as shown on its FAQ pages: "all such questions [on search coverage] are best answered by searching for a statistical sample of papers that has the property of interest—journal, author, protein, etc. Many coverage comparisons are available if you search for [allintitle:"google scholar"], but some of them are more statistically valid than others". The suggestion illustrates that Google Scholar acknowledges the validity of some of the scientometric methods it is examined by.

Research on Google Scholar's size has a long tradition and is considered by some as the "golden fleece" (Orduña-Malea et al. 2015). Indeed, even just two years after Google Scholar's launch in late 2004, Mayr and Walter (2007) took up the challenge to be the first to assess its coverage. The study concluded that Google Scholar's coverage of Thomson Scientific Journal lists, Directory of Open Access Journals, and Journals from the SOLIS database was 78.5%. Later-on Aguillo (2012) found that Google Scholar might list a total of more than 86 million records. Two years later, Khabsa and Giles (2014) estimated that close to 100 million records were listed. Utilizing query hit count (QHC) methodology, Orduña-Malea et al. (2015) concluded that its size must extend beyond all previous estimates and concluded that Google Scholar is likely to contain 176 million documents, including articles, citations, and patents. Nevertheless, due to the opacity of Google Scholars' technical functionality "all methods [of assessing its coverage] show great inconsistencies, limitations and uncertainties" (Orduña-Malea et al. 2015, p. 947). In the face of these challenges, the question remains whether Google itself is only unwilling to report its size, or perhaps is in fact is incapable of doing so. This work intends to shed more light onto how large Google Scholar actually is and how it compares to other large multidisciplinary ASEBDs.

While Google Scholar is one of the most popular academic search engines, it is not the only one relevant for scientific enquiries (Orduña-Malea, Martín-Martín, et al. 2014). With an increasing number of search engines and bibliographic databases, so the competitive pressure increases to provide useful information. As the number of search systems increases, so the features and functionality offered in accessing search results diversifies. Hence, as ASEBDs became important gatekeepers of the provision of secondary information, and their role in science became increasingly relevant, research also became increasingly interested in investigating them. Since the millennium, research on the size of web search engines and other information search systems has featured in scientometric, informetric, bibliometrics, webometrics, and altmetrics journals (Orduña-Malea et al. 2015; Orduña-Malea and Delgado López-Cózar 2014; Hood and Wilson 2001; Thelwall 2008, 2009; Bar-Ilan 2008). Nevertheless, given the increase in ASEBDs, all differing in scope and functionality, research efforts have not caught up with their investigation. Currently there is no study to assess and compare major ASEBDs—a considerable gap in research this study aims to fill.

To monitor a larger set of ASEBDs requires a method capable of including all different ASEBDs. It is evident that all ASEBDs differ in qualities such as functionality, scope, data handling, and syntax. Previous studies assessed the size of ASEBDs with a variety of methods (Ortega 2014; Khan et al. 2017). These estimates of ASEBDs' sizes were predominantly performed for databases where this information was not officially reported. ASEBDs were assessed using queries against multiple journal lists (Mayr and Walter 2007), the overlap between ASEBDs (Khabsa and Giles 2014), the query of top-level domains (Aguillo 2012), and the use of blank or "absurd" queries to receive QHCs (Orduña-Malea et al. 2015; Orduña-Malea, Ayllón, et al. 2014). So far studies have examined ASEBDs individually (Aguillo 2012; Halevi et al. 2017; Mayr and Walter 2007; Orduña-Malea et al. 2015; Hug and Braendle 2017; Harzing 2014) or compared them in pairs or multiples (Meho and Yang 2007; Shultz 2007; Chadegani et al. 2013; Khabsa and Giles 2014; de Winter et al. 2014). Nevertheless, what has been missing so far is an up-to-date comparative overview of the sizes of the most popular ASEBDs. One reason for this shortcoming is the different estimating-methods employed that have made comparing the size of an ASEBD difficult.

Study objective

This study's aim is to estimate ASEBD sizes with a method that is applicable for most systems. We reasoned that all ASEBDs with a focus on the user would provide some form of query function. Hence, the goal of our analysis was to retrieve a maximum quantity of records of a given ASEBD with one single query. We investigated *scope* in terms of what information is *actually available* to the user, rather than the *theoretically indexed* knowledge. Even when databases might contain more articles in theory, the inaccessibility of these articles makes them irrelevant for the user. Hence the value of information systems in terms of scope lies in the knowledge stock it makes accessible through queries, not the stock it has theoretically stored or indexed on its servers but fails to list through query-based methods. To assess the quantity of knowledge actually accessible to users, we use the same tools available to the user. This means straightforward queries are assumed to retrieve the datasets that are effectively available to the searchers. While this query technique presents a *query bias*, as datasets that are not reached through regular query might be systematically disregarded, this limitation is the same as the user has. Hence, queries define the line between what data can and what data cannot be retrieved by the regular user (Bharat and Broder 1998). Nevertheless, it is noteworthy that *accessible* records do not mean *accessible unique* records. Indeed, search systems often include a significant portion of duplicates and indexing, or other cataloguing errors seemingly boost the total size of the system while not providing any new information to the user (Jacsó 2008; Valderrama-Zurián et al. 2015). Acknowledging the difficulty of assessing multiple multidisciplinary ASEBDs that vary in functionality this study tackles the need for up-to-date information on search system scope.

Method and data

Building on previous scientometric research, this study introduces an iterative method to compare the sizes of widely used multidisciplinary ASEBDs. These query-based size estimates are then assessed to discern their plausibility by comparing them to the official size information given by the database providers and the size information reported by other scientific studies.

Selection of search engines and bibliographic databases

We based our selection of academic search engines on the work of Ortega (2014) that presents a comprehensive guide to the landscape of academic search engines up until 2014. At that point the available search engines were: AMiner, Bielefeld Academic Search Engine

(BASE), CiteSeerX, Google Scholar, Microsoft Academic, Q-Sensei Scholar, Scirus and WorldWideScience. Of these eight search engines, Scirus could not be analysed as its services terminated in 2014. To this sample of seven we added a search engine that went online after Ortega's contribution (Semantic Scholar) as well as four large multidisciplinary bibliographic databases and aggregators (EbscoHost, ProQuest, Scopus, and Web of Science). Hence, this study analyses 12 ASEBDs. Their main characteristics such as "owner", "year of launch" and "coverage" are described in Table 1.

As ASEBDs are heterogenic in their functionality and data input formats, this study had to find a common method to access them. Previously researchers had been interested in the characteristics of single ASEBDs or a comparison of a few. Here a multitude of methods were applied including, webometric analysis (Aguillo 2012), capture/recapture methods (Khabsa and Giles 2014), citation analysis (Meho and Yang 2007; Hug and Braendle 2017), and search result comparison (Shultz 2007). However, as these methods are not practically applicable for *most* ASEBDs in a similar fashion, we introduced an iterative method to test the features of ASEBDs in our sample. This research builds on previous methodology developed and employed by Vaughan and Thelwall (2004) and Orduña-Malea et al. (2015) and advances their methods for finding ASEBD metrics. We implement an iterative element to identify the maximum QHC, meaning iterating towards a query that provides the maximum number of hits for a given search system.

Any given query of a search system is assumed to retrieve a set of records, and not retrieve others that lie outside of the query's scope. The sum of both retrieved and non-retrieved records amounts to the search system's coverage or its overall size. *Recall* denotes the search system's capability to retrieve all relevant records over a query (Croft et al. 2015). Our measure of QHCs denotes the number of retrieved records, while the total size of the database remains known only to the database provider. A given query retrieves either all records or a fraction of all records. QHCs therefore denote an estimation of a search system's minimally assumed size—the *least* number of records that it is expected to contain. This means an ASEBD *at least* covers this number of resources, and maybe more.

Accordingly, we included different resource formats and qualities, an approach similar to that of Orduña-Malea et al. (2015). Hence, QHCs reflect the scope of scholarly search engines and bibliographic databases as a determinant of their overall usefulness for scholarly work, while they do not state which database contains most of some particular academic resource type, such as peer reviewed articles. All ASEBDs analysed in this study were accessed between January 2018 and August 2018.

Search strategies and equations

Utilizing an iterative approach to find best estimates of the size of ASEBDs extends the methodologies used in scientometrics and information science. We followed the methodology employed recently by a number of studies in information metrics where ASEBD size is determined through queries with different search string designs (Halevi et al. 2017; Orduña-Malea et al. 2015; Orduña-Malea, Ayllón, et al. 2014). Earlier, this method was used to evaluate the scope of non-academic search engines (Vaughan and Thelwall 2004). In this study we build on these previous experiences and combine them with an iterative methodology that is, through variation of search strings, geared towards maximizing QHCs. In information metrics Orduña-Malea et al. (2015) already experimented with "direct queries" that searched with a specific filter and "absurd queries" that contained arbitrary characters. The logic of the "direct queries" was to utilize filter functions without including a search string, while the logic of “absurd queries” was to retrieve data with variations of a search string. With the latter, the idea was to select the most universal characters such as “1” or “a”, as almost any serious document would feature those characters at some point. Orduña-Malea, Ayllón, et al. (2014) note in relation to “absurd queries” that the method is “more accurate than it seems at first because the search engine is forced to check the entire database to answer the query, as the time responses are suggesting […] the final figures provided seem logical and coherent, and close to those achieved by other methods. […]” “Surprisingly, even though all methods seem invalid for various and diverse reasons, the external method and internal method based on absurd query (with all variants considered) return similar results despite being of a different nature, reinforcing the validity of the estimation performed” (Orduña-Malea et al. 2015, p. 947).

Following the motto *anything might work*, we iteratively tested five different categories of variations of search strings to formulate “direct queries” and “absurd queries” for each database: single characters, digits, terms, ANSI symbols, and also their cross-combinations and queries with wide data ranges. The query variations we utilized are outlined in Table 2. The reasoning was that almost all listed publications would contain at least one of these variations and therefore would be identified through these query-based methods. In particular we expected that most records would be written in English (Khabsa and Giles 2014) and that all of these would at least contain one of the most frequently used English words in its text. While this provides a language bias, it is not uncommon to focus on English articles as the largest ASEBDs seem to do so (Orduña-Malea et al. 2015). Accordingly, we consulted the 2008 Oxford Word List (Oxford University Press 2008) and interlinked sets of the top 100, top 50, top 25 or fewer most utilized English words with Boolean operators. To mitigate this language bias we also tested whether non-English-based variations, such as, digits, year ranges, and ANSI symbols, were capable of retrieving the maximum QHC. Whenever more than one character, digit, symbol, or term was used as input, the query was separated with Boolean “OR” operators. Furthermore, we performed queries by selecting exhaustive time spans in the expectation of covering the entire data set underlying the ASEBD. When all methods failed to produce a plausible QHC (as in the case of Q-Sensei Scholar) we tried queries with facets provided by the database. All queries were tested with and without using quotation marks. Queries were performed with Google Chrome in incognito mode and were tested under different paywall restrictions (i.e., university subscriptions) and locations (IP addresses). The exact composition of queries and the utilized preferences for each of the ASEBDs are illustrated in detail in “Appendix 2”.

Google Scholar

Google Scholar presents a special case among ASEBDs in that it is both one of the most frequently used, yet also one of the least understood and validated. This is why we dedicated particular effort to iterating some valid, stable method to obtain a good estimate of Google Scholar’s size. We started from the methodology of Orduña-Malea et al. (2015, p. 937) who collected hit count data through absurd queries: “[…]we ran test queries using the following syntax: <common_term -site:non-existent_site> The idea behind this is to query the number of occurrences of a very common term (likely to appear in almost all written records), and to filter out its appearances in a non-existent web site, which means that we are implicitly selecting every existing site. For example: <a -site:ssstfsffsdfasdfsf.com>, or <1 -site:ssstfsffsdfasdfsf.com>. The reason for including a term before the “-site” command is that this command does not work on its own”. In this study we tested

Table 2 Query methods

No.	Query method	Description
1	“Absurd queries” as used by Orduña-Malea et al. (2015)	a 1 the (most popular English word)
2	ANSI-coded symbols	:; ., # ! “ % & = ? + − * _
3	English alphabet	a b c d e f g h i j k l m n o p q r s t u v w
	Digits from zero to nine	0 1 2 3 4 5 6 7 8 9
4	The hundred most popular English words according to Oxford University Press (2008)	The and i to a was it my went we on he in they then of said had so she there is got you when one but that were with for day at her me mum up home go his all out saw because house time have dad came played are like after called going get back some him as school once two be did not what them very do fun next can play into big will am this an from weekend their people little ran could our friends about down didn’t put lived name night off bed see upon
5	“Direct queries” as used by Orduña-Malea et al. (2015, p. 936) in searching “the custom date range for the complete period of time”	A query with a wide date range was iteratively narrowed by 50-year increments until the amount of retrieved records started to decrease

the same proposed query, “1 -site:ssstfsffsdfasdfsf.com” and altered the search string (different “common_term” and different “non-existent_sites”), and the time frame (different time spans) according to our defined test search strings (see “Appendix 2”).

Plausibility assessment

Through varying queries iteratively, we received different QHC size estimations for every query. We took the maximum QHC value as the best estimate of the total number of a database’s records. In order to validate the QHCs obtained we performed two plausibility checks with our data. First, we collected official size statements provided by the ASEBD operators themselves. Second, other research studies might have previously examined ASEBD sizes using similar or different methods to ours. We compared the maximum QHCs with the size information of the ASEBDs themselves or of research conducted on the size of the ASEBD. The plausibility check was straightforward for most ASEBDs. When our maximum QHC was within plausible range of these comparative numbers we considered the QHC plausible. Plausible range was primarily determined by taking into account the time difference in size information. For subscription-based ASEBDs that provide access to *multiple* distinct bibliographic databases (i.e., EbscoHost, ProQuest, Web of Science) we also retrieved QHC data for *a specific* database where comparative size information was available from official sources. This way we could assess if and to what degree QHC data matched the official size statement of the provider. If the QHC was plausible for a single database, we reasoned QHCs would also be similarly plausible for *multiple* databases.

Results

Our analysis reveals the query hit counts of ASEBDs. We found that QHC sizes varied significantly from the smallest (CiteSeerX) containing 8,401,126 hits to the largest (Google Scholar) containing 389,000,000 hits. The results show that based on QHC, Google Scholar, WorldWideScience, and ProQuest (selection of 19 databases, see “Appendix 1”) are by far the largest systems providing scholarly information, with each containing about 300 million records. This leading group is followed by BASE, Web of Science (selection of ten databases, see “Appendix 1”), and EbscoHost (selection of 25 databases, see “Appendix 1”) each containing more than 100 million records; somewhat smaller ASEBDs are Scopus, Web of Science (Core Collection), and Q-Sensei Scholar each containing around 60 million records. In the case of those providers linking multiple databases—EbscoHost, ProQuest, and Web of Science—it is important to consider that the QHC reflects *a selection* of databases, and therefore, their QHCs are likely to be higher when *all* available databases of a provider are selected at once.

Two of the 12 ASEBDs—AMiner and Microsoft Academic—did not report numbers suitable for query-based size estimation. AMiner only reports QHCs of up to 1000 hits making it impossible to retrieve actual QHC data. Similarly, Microsoft Academic does not report data sets exceeding 5000 records since its relaunch in 2017. Earlier studies on Microsoft Academic were still able to report size numbers via simple queries (e.g., Orduña-Malea, Ayllón, et al. 2014; Orduña-Malea, Martín-Martín, et al. 2014).

We found that most of the query variations we employed proved successful in retrieving maximum QHCs for some ASEBDs. No query method returned the highest QHC of all databases. Most ASEBDs returned the highest QHCs via “direct queries” of wide time spans (5 times) or with symbol-queries (5 times). The asterisk (*) was the most successful symbol in retrieving maximum QHCs. In three cases—Google Scholar, ProQuest, and Web of Science—two methods simultaneously produced the same maximum QHC. Neither the single terms "a" and "the" nor character and number combinations provided a maximum QHC in our analysis. Only for one database (Scopus) did a *combination* of words prove successful in retrieving a maximum QHC, signifying that for this database alone a longer search string actually meant more retrieved records. In this case we therefore iteratively expanded the search string to see if the maximum QHC could be further increased. Indeed, a combination of the top 100 terms, all digits, and the English alphabet increased the QHC by almost 2% to 72 million records. To exclude a potential language bias, we additionally expanded the query with Russian and Chinese letters, but could not find any difference in maximum QHC. Table 3 presents the detailed outcomes.

Results of plausibility assessment

While maximum QHCs did in some cases diverge considerably from comparative measures, they were not necessarily implausible. In the case of CiteSeerX for example, official numbers were outdated and hence reported 17% fewer records than the QHC predicted. We hence assumed that the QHC probably reflected the search engine's size at that time. We also found that official size statements were frequently outdated or entirely unavailable for other databases. The plausibility assessment for all ASEBDs in our sample can be found in Table 4.

Overall, when comparison was possible, we found that QHCs were a plausible and therefore valid instrument for assessing the sizes of ASEBDs. Plausibility checks allowed us to conclude that QHC data was plausible for seven out of ten ASEBDs: Bielefeld Academic Search Engine (BASE), CiteSeerX, EbscoHost, Q-Sensei Scholar, ProQuest, Scopus, and Web of Science. In the case of BASE, the QHC exactly matched the official size information. Q-Sensei Scholar provided an exception as the maximum QHC was not identified through query but through selection of multiple facets. For this database we identified the maximum QHC by selecting all "year" or "type" facets. The resulting QHC only fell short by less than 1% compared to the updated official size information.

For EbscoHost, ProQuest, and Web of Science—which all adopt a subscription model—we found that the QHC depended significantly on which databases were searched. We found that QHCs for single databases were perfectly plausible (EbscoHost's ERIC, ProQuest Dissertations and Theses Global, and Web of Science's Core Collection). Hence, we reasoned that the QHCs were also plausible for multiple databases. Nevertheless, the QHCs for a joint search of all available scholarly databases fell short of official size numbers. This discrepancy can be explained by the limitation of the databases accessed because firstly, not all databases from these information services provide scientific content and some were thus excluded from our search; and secondly, we could not access all available databases ourselves because we lacked the necessary subscriptions. Therefore, the resulting QHCs reflect the volume of records available according to the *unique scope determined by the searcher*. Hence, for EbscoHost, ProQuest, and Web of Science maximum QHCs do not reflect the *total, objective size* of the service, but the aggregated size of the selected databases of that provider. The databases we selected are listed in "Appendix 1".

Only two QHCs were implausible: Semantic Scholar and WorldWideScience. We found that these two ASEBDs also provided inconsistent QHCs during the data retrieval process. Their QHCs were both significantly different from official size information and varied considerably when queries were repeated. Having presented the results of the QHC plausibility

Table 3 QHCs of search engines and bibliographic databases

Name of database	Single terms			Single symbols	Character combination	Number combination	Word combination			Time span/other
	a	1	the	ANSI	a–z	0–9	Top100	Top50	Top25	
AMiner	No report of QHCs exceeding 1000 hits.									
Bielefeld Academic Search Engine (BASE)	62,939,918	47,132,682	65,409,081	**117,762,309 (*)**	104,146,514	82,855,823	31,551,663	56,820,714	68,557,113	103,016,285 (0000–2018)
CiteSeerX	8,178,032	7,933,667	5,687,272	**8,401,126 (*)**	8,247,528	8,060,521	6,775,009	6,617,395	6,424,390	N/A (shows same result for varying time spans)
EbscoHost (selection of 25 databases—see "Appendix 1")	20,056,140	33,851,675	823	380,143 (")	41,092,238	37,390,677	56,597,398	48,888,668	35,458,081	**102,908,910 (1550–2018)**
Google Scholar	81,500,000	247,000,000	49,300,000	**389,000,000 (*)**	192,000,000					**389,000,000 (1700–2099)** (blank "common_term")
Microsoft Academic	No report of QHCs exceeding 5000 hits									
ProQuest (selection of 19 databases—see "Appendix 1")	199,100,118	150,864,723	189,797,217	**279,987,039 (;)**	250,422,353	247,542,112		270,846,888	270,178,534	**279,987,039 (750–2030)**

Table 3 (continued)

Name of database	Single terms			Single symbols	Character combination	Number combination	Word combination			Time span/other
	a	1	the	ANSI	a–z	0–9	Top100	Top50	Top25	
Q-Sensei Scholar	31,697,482	6,243,396	36,370,496	8,168,464 (&)	228 (a-y)	50,751			2,137 (Top8)	**47,195,502 (1969–2017;** all available years selected) **55,162,101** (all "type-facets" selected)
Scopus	60,433,436	40,156,373	63,411,230		70,995,663	66,812,009	71,091,362	70,890,200	70,683,693	**72,212,354** (Top100 OR 0–9 OR a–z)
Semantic Scholar	5,350,000	**5,410,000**	5,320,000		2,110,000	3,190,000	204,000	370,000	1,100,000	
Web of Science (Core Collection—see "Appendix 1")	33,832,256	10,204,360	41,791,314	**67,713,141 (":")**	39,567,990	19,942,632	63,718,875	63,206,556	62,080,068	**67,713,141 (1850–2018)**
Web of Science (selection of 10 databases incl. Core Collection—see "Appendix 1")	65,500,925	22,971,331	78,714,615	**105,519,854 (":")**	75,709,357	42,252,988	101,067,361	100,510,004	99,554,354	**105,519,854 (1800–2018)**
WorldWideScience	306,335,673	**323,202,597**	234,191,957	272,392,231 (:)	72,180,354	306,097,593	29,138,581	41,847,969	41,392,251	

Bold numbers indicate maximum QHCs

Table 4 Plausibility assessment of QHCs

Name of database	Max. QHC	Official size (website/report)	Estimated size (other academic research)	Plausibility assessment: whether max. QHC reflects the size of selected databases
AMiner		232,488,386 (01.2018, AMiner website)		
Bielefeld Academic Search Engine (BASE)	117,762,309 (01.2018)	117,762,309 (+0.00%) (01.2018, BASE website)		**Plausible**
CiteSeerX	8,401,126 (01.2018)	7,000,000 (–16.68%) (beginning of 2016, CiteSeerX report; Wu et al. 2016)	Close to 8,000,000 (–4.77%) (05.2013, Caragea et al. 2014)	**Plausible**
EbscoHost (selection of 25 databases—see "Appendix 1")	102,908,910 (08.2018)	132,000,000+ on 375+ databases (+28.27%) (03.2015, EbscoHost report)		**Plausible** (not all databases accessed in query)
Google Scholar	389,000,000 (01.2018)		330,804,940 (–14.96%) (03.2017, Delgado López-Cózar et al. 2018 with QHC methodology) 176,000,000 (–54.76%) (05.2014, Orduña-Malea et al. 2015 with QHC methodology)	**Questionable** (partially unstable hit count data)
Microsoft Academic		171,039,432 (01.2018, Microsoft Academic website)		
ProQuest (selection of 19 databases—see "Appendix 1")	279,987,039 (08.2018)			**Plausible** (not all databases accessed in query)
Q-Sensei Scholar	55,162,101 (01.2018)	55,573,606 (+0.75%) (01.2018, Q-Sensei website)		**Plausible**
Scopus	72,212,354 (08.2018)	69,000,000 (–4.45%) (2017, Elsevier website)	43,186,550 (–40.20%) (2017, Halevi et al. 2017 with QHC methodology)	**Plausible**
Semantic Scholar	5,410,000 (01.2018)	40,000,000+ (+639.37%) (2017, Semantic Scholar website)		**Implausible**

Table 4 (continued)

Name of database	Max. QHC	Official size (website/report)	Estimated size (other academic research)	Plausibility assessment: whether max. QHC reflects the size of selected databases
Web of Science (Core Collection—see "Appendix 1")	67,713,141 (01.2018)	68,000,000+ (+0.42%) (2017, Clarivate website)	65,171,195 (−3.75%) (2017, Halevi et al. 2017 with QHC methodology)	**Plausible**
Web of Science (selection of 10 databases incl. Core Collection—see "Appendix 1")	105,519,854 (01.2018)	145,000,000 (+37.41%) (2017, Clarivate website)		**Plausible** (not all databases accessed in query)
WorldWideScience	323,202,597 (01.2018)	500,000,000 (+54.70%) (2014, Deepwebtech report)		**Implausible**

The bold signifies the result of the assessment

assessment of nine ASEBDs, the remaining search engine Google Scholar seems to produce questionable QHCs owing to its lack of stability over query variations. Our QHC indicates that Google Scholar incorporated 389 million records in January 2018.

Discussion

This study has built on and extended previous scientometric research inquiring into the sizes of ASEBDs. It is novel in so far as it improves query-based methods for assessing ASEBDs and establishes those methods as adequate, fast predictors of the sizes of most ASEBDs. The methods used made it possible to assess a multitude of different ASEBDs and compare their sizes. The process not only delivered size information but also some insights into the diverse query functionalities of ASEBDs that prove to be the basis of the daily scientific enquiries of many researchers.

Size

We have obtained a QHC from ten of the 12 ASEBDs examined. Based on this QHC data we can assume that Google Scholar, with 389 million records, provides by far the greatest volume of scholarly information. Our maximum QHC in this regard seems plausible when compared to similar multidisciplinary search engines like Microsoft Academic that as of January 2018 covers more than 170 million records and is considered, with a ratio of 1:2.17, considerably smaller than Google Scholar (Orduña-Malea et al. 2015). If we apply the same ratio between the two search engines in January 2018, Google Scholar would amount to roughly 372 million records, a number close to our QHC of 389 million. Nevertheless, it is important to bear in mind that this size comparison might be flawed as Microsoft Academic has relaunched since the Orduña-Malea et al. (2015) research was conducted. This relaunch could have significantly impacted the structure and size of Microsoft Academic (Hug and Braendle 2017) and its performance in retrieving search results with high precision and recall (Thelwall 2018).

Comparing previous research findings with our QHC results we found that with 389 million records Google Scholar's maximum QHC in January 2018 amounts to an increase of 121% compared to the previously estimated size of 176 million by Orduña-Malea et al. (2015) in May 2014. The QHC estimation of both our study and that of Orduña-Malea et al. (2015) include articles, citations, and patents indexed on Google Scholar and thus can reasonably be compared. This size difference most likely stems from two factors: time difference and method difference. With regard to time difference, if we exactly replicate the query that resulted in the 176 million hits obtained by Orduña-Malea et al. (2015) (<1-site:ssstfsffsdfasdfsf.com> and wide year range), we arrive at a QHC of 247 million records. This indicates that in 44 months Google Scholar increased its size by 40% or an average growth rate of 1.6 million records per month. This monthly growth rate would only exceed Microsoft Academic's current monthly growth rate of 1.3 million records by a reasonable margin (Hug and Braendle 2017). Given this plausible increase in records over 44 months, it seems logical to assume that the same QHC method in May 2014 produced comparable results in January 2018 too.

With regard to method difference, as with all databases we iteratively tried other queries to identify a maximum QHC for Google Scholar. Indeed, two queries (asterisk and time span) resulted in significantly higher QHCs. This indicates that as of January 2018 Google Scholar's size was 389 million records. Accordingly, we believe that the second factor accounting for QHC differences between May 2014 (176 million) and January 2018 (389 million) is attributable to a difference in query method. We reason that it is plausible to assume that Google Scholar's QHC at 389 million is considerably higher than previously estimated. The question is whether Orduña-Malea et al. would also have obtained a higher maximum QHC had they used these same query methods in 2014.

Further, the most recent comparative data on Google Scholar's size stems from the work of Delgado López-Cózar et al. (2018), which estimates its size at 331 million records in March 2017. In comparison this would mean that our QHCs 10 months later indicate an increase of Google Scholar's total size (including articles, patents, and citations) of 18%. As Delgado López-Cózar et al. (2018) use a different estimation method that involved adding Google Scholar's yearly QHC to an overall total sum, we cannot compare our results directly, as we know from previous research (Orduña-Malea et al. 2015) that year-by-year queries might lead to slightly lower total QHCs than using wide year ranges. If one assumes the same percentage difference for 331 million records obtained by year-by-year estimation, one could calculate a hypothetical 343 million for wide year range estimation in March 2017. Then, the remaining difference of 46 million records ought to stem in part from an expansion of Google Scholar's database within these 10 months and in part from method difference. Using the previously calculated monthly growth rate of 1.6 million records, would leave 30 million records attributable to method difference, indicating that we found a specific absurd query variation that leads to a higher QHC. Hence, these findings suggest it is worthwhile employing iterative methodology to estimate Google Scholar's maximum QHC.

WorldWideScience seems to have the second largest QHC with 323 million records. However, its QHCs have to be considered highly unstable as identical queries result in entirely different QHCs if performed only seconds apart. QHCs are also comparatively significantly lower than the official size data. We therefore assume QHCs inadequately reflect WorldWideScience's total size. Further, according to Ortega (2014) WorldWideScience offers "more quantity than quality" as the system is assumed to produce "a large amount of duplicated results and is very time consuming". These downsides make it significantly less user-friendly compared to Google Scholar for example. Nevertheless, one significant advantage of WorldWideScience is its capacity to access data from the *deep web*, which cannot be harvested by search engines such as Google Scholar (Ortega 2014).

Our analysis of 19 databases provided by ProQuest revealed that its 280 million records place it among the most comprehensive ASEBDs. The scope of ProQuest, similar to EbscoHost and Web of Science, is probably even higher if *all available scientific databases* could be accessed. Hence, for these providers our QHCs ought to be seen as indicative of their *minimum* total size, assuming that unrestricted access results in even higher QHCs. Nevertheless, our QHCs are indicative of their dimensions relative to other providers. For example, ProQuest is one of the largest ASEBDs and EbscoHost and Web of Science, both with more than 100 million records, have similar sizes to BASE, yet are considerably larger than CiteSeerX, Q-Sensei Scholar, Scopus, and Semantic Scholar. In the end the *total size* of these providers is *theoretical*; users can only access a portion of the total volume due to subjective resource restrictions, compared to search engines such as Google Scholar that provide access to all indexed resources. In this regard this study is to our knowledge the first to offer a size measure to EbscoHost and ProQuest. The scope of Web of Science was estimated before, predominately for its popular product, the Core Collection (Orduña-Malea et al. 2015; Martín-Martín et al. 2015; Orduña-Malea, Ayllón, et al. 2014).

A size of 118 million records and the greatest portion of its content being open access (Ortega 2014) makes BASE a search engine especially valuable for users without access to paywalled content. Conversely, the focus on open access content means that large portions of the academic web are not represented. Nevertheless, if the user is aware of this shortcoming BASE is one of the most valuable multidisciplinary academic databases, especially when given its responsiveness and filtering options. The remaining ASEBDs, Q-Sensei Scholar (plausible QHC, 55 million records), Semantic Scholar (official data of 40 million records), and CiteSeerX (plausible QHC, 8 million records) are smaller than their counterparts, yet all of them draw legitimacy from having a distinct vision of how an academic search engine should function (Ortega 2014).

The ASEBDs in our sample without QHC data (AMiner and Microsoft Academic) provide updated information on their sizes themselves. While these sources provide large sets of resources—232 million in the case of AMiner and 171 million in that of Microsoft Academic—these systems are extremely difficult (and sometimes impossible) to access through a systematic query-based data retrieval, a criterion necessary for systematic literature reviews for example.

Queries

The results show that ASEBDs are diverse in their functionality and features, so their analysis requires an overarching comparative methodology. Most of the different query variations employed successfully retrieved a maximum QHC in at least one case. This first shows that academic services function differently and second underlines the validity of our broad iterative approach of testing a multitude of query variations. We found that employing "absurd" or "direct" queries (Orduña-Malea et al. 2015) is not absurd after all, as we could produce plausible QHCs for seven ASEBDs: BASE, CiteSeerX, EbscoHost, ProQuest, Q-Sensei Scholar, Scopus, and Web of Science. Specifically, the results show that for most ASEBDs, queries with varying symbols were most effective in terms of retrieving a maximum QHC.

The only ASEBD in our sample where QHCs exactly matched official size information was BASE. In some cases, the resulting QHC was higher than the number provided by the ASEBD operator, illustrating the problem that size statements are frequently outdated. In two cases (Q-Sensei Scholar and Web of Science Core Collection) official numbers were only slightly higher than maximum QHCs, indicating that not all of the providers' database's records can effectively be accessed via query or at least not via the queries that were tested in this study.

Despite the QHC proving a relevant tool to assess the sizes of most ASEBDs, it was not suitable in *all* cases. In fact, for four search engines in our sample (AMiner, Microsoft Academic, Semantic Scholar, and WorldWideScience), the QHC proved to be inadequate to a greater or lesser degree. AMiner and Microsoft Academic did not report their QHC while providing up-to-date size information on their websites. Queries on Semantic Scholar and WorldWideScience returned variable results and could not be verified. It remains uncertain whether the outdated official size information for these two search engines correctly indicates the volume of records *actually accessible* to the user.

We found that Google Scholar's QHC for identical queries seemed reliable and precise at some points of time and unreliable and imprecise at other times. This issue was identified by Jacsó as early as 2008 and again by Orduña-Malea et al. (2015). To examine Google Scholar's query results, we made an effort to discern patterns of reliability and precision. The current analysis benefits from that of Orduña-Malea et al. (2015), which found that the introduction of a limiter "non-existent_site" produced more plausible and stable results. We confirmed their findings in so far as Google Scholar produces significantly fewer results with straightforward queries not using any other limiters. Following our iterative approach, we did not however just replicate the queries of Orduña-Malea et al. (2015)

but also tested different search strings to verify if the QHC was indeed the maximum value. We found that "non-existent_site" produced the same results, while changes to the "common_term" altered the QHC significantly. Keeping the "non-existent_site" the same, we identified differences in the QHCs as we changed the terms from "1" to "a" or "the" or to other symbols. Queries with more than 30 s loading time resulted in a time out notification. To reduce the server load, we limited the length of queries. The process of iteration revealed a set of characters that produced the maximum QHCs (see Table 3). It also made it possible to record a maximum QHC of 389 million for the time span of 1700–2099. The fact that we received this maximum QHC with two methods (asterisk or time span only) could indicate that the QHC results are valid. Without the operator "non-existent_site", the same query however produced a QHC of only 710,000.

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 are in line with earlier research (Martín-Martín et al. 2017; Mingers and Meyer 2017; Aguillo 2012; Orduña-Malea and Delgado López-Cózar 2014; Jacsó 2005, 2008, 2012; Orduna-Malea et al. 2017). Despite these irregularities, employing the identical method as 4 years earlier (Orduña-Malea et al. 2015), we could replicate a reasonable QHC of Google Scholar. This could indicate that "absurd queries" can be a valid instrument to assess and replicate the QHC of Google Scholar over long periods of time. The current difficulties in replicating QHC results notwithstanding, our findings indicate that QHC methods can be reliable estimators of Google Scholar's size. Compared to other major databases, Google Scholar seems to provide a multidisciplinary database outperforming the coverage of competitors such as Web of Science and Scopus (Martín-Martín, Orduna-Malea, and Delgado López-Cózar 2018; Martín-Martín, Orduna-Malea, Thelwall, et al. 2018).

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.

Further, this study highlights another important issue in academic document searching. While EbscoHost, ProQuest, and Web of Science seem to provide plausible QHC results, the scope of these services is often not clear for the user, as the volume of retrieved information depends on the specific settings of the user accessing it. In these three cases, academic institutions subscribe to different databases that are hosted by these providers. Therefore, what a user captures varies according to the subscriptions held. Users' search scope might be suboptimal owing to limited institutional access, but those users might also not be aware of this limitation. Inexperienced users might think that these bibliographic databases and aggregators in fact only consist of a single, unitary database. The significant volume of academic research that mentions ProQuest or EbscoHost as its search frame, without stating the specific databases accessed, is indicative of this issue. In such cases the exact scope of the search remains unclear to readers and reviewers, which is especially worrying when research-synthesis studies are concerned. For reasons of scientific rigour, we suggest researchers should be educated on the issues around accurately reporting search scope.

Limitations and future research

This research found that the QHC measure a consistent methodology and seems a valid predictor of the sizes of most ASEBDs in our sample. Nevertheless, we will point out four limitations that at the same time provide avenues for future research.

First, following earlier research (Orduña-Malea et al. 2015; Khabsa and Giles 2014) some queries employed in this study focus on records that at least in some part use the English alphabet or English terms, that is, in using “the” and word combinations of the Oxford word list. While this procedure seemingly focuses on English documents only it is rarely the case that non-English documents use non-English letters or type only. The word “a” for example is used in multiple different languages; or, further a significant number of Chinese documents include some translation of the title or abstract or use single keywords or letters that makes these documents identifiable via English-based query methods. To provide an alternative to language-based queries, we employed queries that would work irrespective of language, such as digits and ANSI symbols. For many ASEBDs, these non-language-based queries proved successful in providing maximum QHCs. Building on these language-issues in queries, we suggest future research assesses search systems comparatively with regards to the scope of each language’s coverage. Longitudinal analyses might prove particularly productive in quantifying the development of English versus non-English scientific publication activity.

Second, the actual number of records a database contains might never be known with absolute certainty. While our method of using QHCs was compared against size numbers from official and research sources, the assessment of size is ultimately always based on some information on provision stemming from the ASEBD itself. While it is possible to expose irregularities in this information through plausibility-checking methods such as those employed in this study, validation of the numbers is another question. Validating the accuracy of this information with absolute certainty is most likely impossible without having access to the full dataset. While we know that Google Scholar’s metrics are problematic to some degree, we can never be sure if BASE’s, for example, are not also. The latter is a search engine that updates and publishes its information of its knowledge stock in real time, but being sure that information is accurate would involve downloading all records and counting them, which is not only impractical, but in most (if not all) cases impossible. For example, Google Scholar limits visible records to a maximum of 1000 and Web of Science sets a limit of 100,000. Such lack of transparency means researchers have to work with the information that is available, while constantly challenging the validity of the numbers concerned. This study has tried to minimize these limitations through triangulation of data through multiple query variations and comparative size numbers, and accordingly, we believe that the QHCs reported in this study are a good proxy of the actual database sizes available to users.

Third, QHCs reflect the number of *all* indexed records on a database, not the number of *unique* records indexed. This means duplicates, incorrect links, or incorrectly indexed records are all included in the size metrics provided by ASEBDs. Hence, the number of *unique* records contained by ASEBDs, especially by larger multidisciplinary search engines with automated curation processes, is likely to be systematically exaggerated by QHCs. It is estimated that Scopus for example contains 12.6% duplicates (Valderrama-Zurián et al. 2015) and Google Scholar is assumed to list up to 10% erroneous, undated records (Orduña-Malea et al. 2015). These estimates show that duplicates constitute a significant proportion of total records in both search engines and other database types. As the ratio of unique records to duplicates or other erroneous records differs among ASEBDs, this factor is likely to affect their comparative size if assessed in terms of *unique* records. Hence, this limitation shows it is important to consider the types of records available behind the size numbers.

Fourth, the *size* of a database is only one of multiple criteria that need to be assessed jointly to get an overall picture. For users a balance of these criteria, weighted for their unique preferences and requirements, influences the choice over which database best fits a given task. To assess databases further, especially concerning their suitability for academia, research would need to consider aggregate measures consisting of multiple variables such as relevance, objectivity, functional scope, or the user interface.

Conclusion

We conclude that the QHC measure is in most cases adequate to discern the sizes of ASEBDs. The iterative method used in this study has proven useful to receive plausible up-to-date information on the sizes of eight of the 12 ASEBDs examined: BASE, CiteSeerX, EbscoHost, Google Scholar, ProQuest, Q-Sensei Scholar, Scopus, and Web of Science. While BASE and Q-Sensei Scholar provide updated size information on their websites, the other six ASEBDs do not, making QHCs relevant and necessary size predictors. For ASEBDs, where comparative numbers were entirely missing, our study is the first to introduce these sizes numbers.

Specifically, we found that it is plausible to assume that Google Scholar's size has been underestimated by between 8% (compared to Delgado López-Cózar et al. 2018) and 55% (compared to Orduña-Malea et al. 2015) so far owing to method difference, not time difference. It is certainly the most comprehensive academia search engine, but nevertheless, it remains unclear why Google Scholar does not report its size. Given the unstable nature of Google Scholar's QHC 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.

For all ASEBDs for which QHCs have been shown to function plausibly, they provide a simple and quick insight into ASEBD scope, particularly compared with other scientometric methods that require more statistics and data manipulation. The method presented here is replicable and permits anyone to quickly obtain updated information that can be tailored to specific categories of content, provided a specific ASEBD supports such filters. For example, Web of Science can be searched via the ":" operator; the resulting records can then be refined according to document type, organization, content category, et cetera. Using that approach makes the volume of the available content easily divisible and researchable. For the exceptions where QHCs are not plausible, other scientometric methods might bring more satisfactory results.

Furthermore, our methodology of QHC-based size estimates will prove useful for longitudinal analysis of ASEBD growth and time series monitoring. The method also makes it possible to compute the time lags between date of publication and indexing of items on the respective ASEBD, which allows the enquirer to assess *freshness* (Croft et al. 2015) of the ASEBD's underlying data. The simplicity of the QHC method in requiring no statistical calculations reduces the workload tremendously, a quality that should prove critical for further application (Prins et al. 2016). Monitoring ASEBDs is even more necessary in times of exponential growth of information and scientific output (Bornmann and Mutz 2014). Ideally metrics ought not only to track from time to time but monitor continuously. Hence, the QHC method could bring easy replicability to receive updated size metrics, thus increasing data relevance. While we have focused only on large multidisciplinary ASEBDs, the QHC can also be relevant to check the sizes of other (and particularly smaller) information systems such as repositories, digital libraries, library catalogues, bibliographic databases, and journal platforms.

Acknowledgements Open access funding provided by Johannes Kepler University Linz. I would like to thank the two anonymous reviewers for their valuable comments.

Appendix 1: Coverage of EbscoHost, ProQuest and Web of Science in this research

EbscoHost

1. AMED—The Allied and Complementary Medicine Database
2. Anthropology Plus
3. ATLA Religion Database with ATLASerials
4. Audiobook Collection (EBSCOhost)
5. British Education Index
6. Business Source Alumni Edition
7. Business Source Premier
8. Child Development and Adolescent Studies
9. CINAHL Plus
10. eBook Collection (EBSCOhost)
11. EconLit
12. Education Abstracts (H.W. Wilson)
13. Educational Administration Abstracts
14. Ergonomics Abstracts
15. ERIC
16. European Views of the Americas: 1493–1750
17. GreenFILE
18. Historical Abstracts
19. Humanities Abstracts (H.W. Wilson)
20. Library, Information Science and Technology Abstracts
21. MEDLINE
22. Regional Business News
23. RILM Abstracts of Music Literature (1967 to present only)
24. SPORTDiscus
25. Bibliography of Asian Studies

ProQuest

1. ABI/INFORM Global
2. ABI/INFORM Trade and Industry (1971–present)
3. British Periodicals (1681–1939, 1869–2005)

4. The Cecil Papers
5. Colonial State Papers (1574–1757)
6. Digital National Security Archive
7. Documents on British Policy Overseas (1898–1990)
8. GeoRef (1693–present)
9. Humanities Index (1962–present)
10. Index Islamicus (1906–present)
11. MLA International Bibliography (1926–present)
12. Nursing and Allied Health Database
13. Periodicals Archive Online
14. Periodicals Index Online
15. Physical Education Index (1970–present)
16. PILOTS: Published International Literature on Traumatic Stress (1871–present)
17. ProQuest Dissertations and Theses Global
18. SciTech Premium Collection (1946–present)
19. Social Science Premium Collection (1914–present)

Web of Science: Core collection

1. Science Citation Index Expanded (1900–present)
2. Social Sciences Citation Index (1900–present)
3. Arts and Humanities Citation Index (1975–present)
4. Conference Proceedings Citation Index—Science (1990–present)
5. Conference Proceedings Citation Index—Social Science and Humanities (1990–present)
6. Book Citation Index—Science (2010–present)
7. Book Citation Index—Social Sciences and Humanities (2010–present)
8. Emerging Sources Citation Index (2015–present)
9. Current Chemical Reactions (2010–present) (Includes Institut National de la Propriete Industrielle structure data back to 1840)
10. Index Chemicus (2010–present)

Web of Science: All databases

1. Web of Science Core Collection
2. BIOSIS Citation Index 2010–present)
3. BIOSIS Previews (1968–2008)
4. Data Citation Index (2010–present)
5. Derwent Innovations Index (2010–present)
6. KCI-Korean Journal Database (1980–present)
7. MEDLINE (1950–present)
8. Russian Science Citation Index (2005–present)
9. SciELO Citation Index (1997–present)
10. Zoological Record (2010–present)

Appendix 2: Search queries

Name of search engine	Query settings	Query							Time span	Other
		Single terms	Single symbols	Character combination	Number combination	Word combination				
			ANSI	a–z	0–9	Top100	Top50	Top25		
AMiner	No report of QHCs exceeding 1000 hits.									
Bielefeld Academic Search Engine (BASE)	Advanced search: exact search; entire document	a 1 the	: ; . , # ! “ % & = ? + - * _	a OR b OR c OR d OR e OR f OR g OR h OR i OR j OR k OR l OR m OR n OR o OR p OR q OR r OR s OR t OR u OR v OR w OR x OR y OR z	0 OR 1 OR 2 OR 3 OR 4 OR 5 OR 6 OR 7 OR 8 OR 9	the OR and OR i OR to OR a OR was OR it OR my OR went OR we OR on OR he OR in OR they OR then OR of OR said OR had OR so OR she OR there OR is OR got OR you OR when OR one OR but OR that OR were OR with OR for OR day OR at OR her OR me OR mum OR up OR home OR go OR his OR all OR out OR saw OR because OR house OR time OR have OR dad OR came OR played OR are OR like OR after OR called OR going OR get OR back OR some OR him OR as OR school OR once OR two OR be OR did OR not OR what OR them OR very OR do OR fun OR next OR can OR play OR into OR big OR will OR am OR this OR an OR from OR weekend OR their OR people OR little OR ran OR could OR our OR friends OR about OR down OR didn’t OR put OR lived OR name OR night OR off OR bed OR see OR upon	the OR and OR i OR to OR a OR was OR it OR my OR went OR we OR on OR he OR in OR they OR then OR of OR said OR had OR so OR she OR there OR is OR got OR you OR when OR one OR but OR that OR were OR with OR for OR day OR at OR her OR me OR mum OR up OR home OR go OR his OR all OR out OR saw OR because OR house OR time OR have OR dad OR came OR played	the OR and OR i OR to OR a OR was OR it OR my OR went OR we OR on OR he OR in OR they OR then OR of OR said OR had OR so OR she OR there OR is OR got OR you OR when	0000–2018	N/A

Name of search engine	Query settings	Query							Time span	Other
		Single terms	Single sym-bols	Character combination	Number combina-tion	Word combination				
			ANSI	a–z	0–9	Top100	Top50	Top25		
CiteSeerX	Advanced search: default	a 1 the	: ; . , # ! “ % & = ? + - * _	a OR b OR c OR d OR e OR f OR g OR h OR i OR j OR k OR l OR m OR n OR o OR p OR q OR r OR s OR t OR u OR v OR w OR x OR y OR z	0 OR 1 OR 2 OR 3 OR 4 OR 5 OR 6 OR 7 OR 8 OR 9	the OR and OR i OR to OR a OR was OR it OR my OR went OR we OR on OR he OR in OR they OR then OR of OR said OR had OR so OR she OR there OR is OR got OR you OR when OR one OR but OR that OR were OR with OR for OR day OR at OR her OR me OR mum OR up OR home OR go OR his OR all OR out OR saw OR because OR house OR time OR have OR dad OR came OR played OR are OR like OR after OR called OR going OR get OR back OR some OR him OR as OR school OR once OR two OR be OR did OR not OR what OR them OR very OR do OR fun OR next OR can OR play OR into OR big OR will OR am OR this OR an OR from OR weekend OR their OR people OR little OR ran OR could OR our OR friends OR about OR down OR didn’t OR put OR lived OR name OR night OR off OR bed OR see OR upon	the OR and OR i OR to OR a OR was OR it OR my OR went OR we OR on OR he OR in OR they OR then OR of OR said OR had OR so OR she OR there OR is OR got OR you OR when OR one OR but OR that OR were OR with OR for OR day OR at OR her OR me OR mum OR up OR home OR go OR his OR all OR out OR saw OR because OR house OR time OR have OR dad OR came OR played	the OR and OR i OR to OR a OR was OR it OR my OR went OR we OR on OR he OR in OR they OR then OR of OR said OR had OR so OR she OR there OR is OR got OR you OR when	1000–2018	N/A

Name of search engine	Query settings	Query							Time span	Other
		Single terms	Single symbols	Character combination	Number combination	Word combination				
			ANSI	a–z	0–9	Top100	Top50	Top25		
EbscoHost	Advanced search: all text (TX)	a 1 the	: ; . , # ! “ % & = ? + - * _	a OR b OR c OR d OR e OR f OR g OR h OR i OR j OR k OR l OR m OR n OR o OR p OR q OR r OR s OR t OR u OR v OR w OR x OR y OR z	0 OR 1 OR 2 OR 3 OR 4 OR 5 OR 6 OR 7 OR 8 OR 9	the OR and OR i OR to OR a OR was OR it OR my OR went OR we OR on OR he OR in OR they OR then OR of OR said OR had OR so OR she OR there OR is OR got OR you OR when OR one OR but OR that OR were OR with OR for OR day OR at OR her OR me OR mum OR up OR home OR go OR his OR all OR out OR saw OR because OR house OR time OR have OR dad OR came OR played OR are OR like OR after OR called OR going OR get OR back OR some OR him OR as OR school OR once OR two OR be OR did OR not OR what OR them OR very OR do OR fun OR next OR can OR play OR into OR big OR will OR am OR this OR an OR from OR weekend OR their OR people OR little OR ran OR could OR our OR friends OR about OR down OR didn’t OR put OR lived OR name OR night OR off OR bed OR see OR upon	the OR and OR i OR to OR a OR was OR it OR my OR went OR we OR on OR he OR in OR they OR then OR of OR said OR had OR so OR she OR there OR is OR got OR you OR when OR one OR but OR that OR were OR with OR for OR day OR at OR her OR me OR mum OR up OR home OR go OR his OR all OR out OR saw OR because OR house OR time OR have OR dad OR came OR played	the OR and OR i OR to OR a OR was OR it OR my OR went OR we OR on OR he OR in OR they OR then OR of OR said OR had OR so OR she OR there OR is OR got OR you OR when	1550–2018	N/A

Name of search engine	Query settings	Query							Time span	Other
		Single terms	Single symbols	Character combination	Number combination	Word combination				
			ANSI	a–z	0–9	Top100	Top50	Top25		
Google Scholar	Basic search settings: search pages written in any language Preferences: articles incl. patents and citations	a 1 the + <–site:ssstfsffsdfasdfsf.com> + Time span: 1700–2099	: ; . , # ! “ % & = ? + - * _	a OR b OR c OR d OR e OR f OR g OR h OR i OR j OR k OR l OR m OR n OR o OR p OR q OR r OR s OR t OR u OR v OR w OR x OR y OR z	0 OR 1 OR 2 OR 3 OR 4 OR 5 OR 6 OR 7 OR 8 OR 9	N/A	N/A	the OR and OR i OR to OR a OR was OR it OR my OR went OR we	<–site:ssstfsffsdfasdfsf.com> + Time span: 1700–2099	N/A
Microsoft Academic	No report of QHCs exceeding 5000 hits									

Name of search engine	Query settings	Query							Time span	Other
		Single terms	Single symbols	Character combination	Number combination	Word combination				
			ANSI	a–z	0–9	Top100	Top50	Top25		
ProQuest	Advanced search; anywhere	a 1 the	: ; . , # ! “ % & = ? + - * _	a OR b OR c OR d OR e OR f OR g OR h OR i OR j OR k OR l OR m OR n OR o OR p OR q OR r OR s OR t OR u OR v OR w OR x OR y OR z	0 OR 1 OR 2 OR 3 OR 4 OR 5 OR 6 OR 7 OR 8 OR 9	N/A	the OR and OR i OR to OR a OR was OR it OR my OR went OR we OR on OR he OR in OR they OR then OR of OR said OR had OR so OR she OR there OR is OR got OR you OR when OR one OR but OR that OR were OR with OR for OR day OR at OR her OR me OR mum OR up OR home OR go OR his OR all OR out OR saw OR because OR house OR time OR have OR dad OR came OR played	the OR and OR i OR to OR a OR was OR it OR my OR went OR we OR on OR he OR in OR they OR then OR of OR said OR had OR so OR she OR there OR is OR got OR you OR when	750–2030	N/A

Name of search engine	Query settings	Query							Time span	Other
		Single terms	Single symbols	Character combination	Number combination	Word combination				
			ANSI	a–z	0–9	Top100	Top50	Top25		
QSensei Scholar	Basic search: default	a 1 the	: ; . , # ! “ % & = ? + - * _	a OR b OR c OR d OR e OR f OR g OR h OR i OR j OR k OR l OR m OR n OR o OR p OR q OR r OR s OR t OR u OR v OR w OR x OR y (max. computable length)	0 OR 1 OR 2 OR 3 OR 4 OR 5 OR 6 OR 7 OR 8 OR 9	N/A	N/A	the OR and OR i OR to OR a OR was OR it OR my (max. computable length)	1969–2017 (all available years selected)	All “type-facets” selected

Name of search engine	Query settings	Query							Time span	Other
		Single terms	Single symbols	Character combination	Number combination	Word combination				
			ANSI	a–z	0–9	Top100	Top50	Top25		
Scopus	Document search: all fields	a 1 the	: ; . , # ! “ % & = ? + - * _	a OR b OR c OR d OR e OR f OR g OR h OR i OR j OR k OR l OR m OR n OR o OR p OR q OR r OR s OR t OR u OR v OR w OR x OR y OR z	0 OR 1 OR 2 OR 3 OR 4 OR 5 OR 6 OR 7 OR 8 OR 9	the OR and OR i OR to OR a OR was OR it OR my OR went OR we OR on OR he OR in OR they OR then OR of OR said OR had OR so OR she OR there OR is OR got OR you OR when OR one OR but OR that OR were OR with OR for OR day OR at OR her OR me OR mum OR up OR home OR go OR his OR all OR out OR saw OR because OR house OR time OR have OR dad OR came OR played OR are OR like OR after OR called OR going OR get OR back OR some OR him OR as OR school OR once OR two OR be OR did OR not OR what OR them OR very OR do OR fun OR next OR can OR play OR into OR big OR will OR am OR this OR an OR from OR weekend OR their OR people OR little OR ran OR could OR our OR friends OR about OR down OR didn’t OR put OR lived OR name OR night OR off OR bed OR see OR upon	the OR and OR i OR to OR a OR was OR it OR my OR went OR we OR on OR he OR in OR they OR then OR of OR said OR had OR so OR she OR there OR is OR got OR you OR when OR one OR but OR that OR were OR with OR for OR day OR at OR her OR me OR mum OR up OR home OR go OR his OR all OR out OR saw OR because OR house OR time OR have OR dad OR came OR played	the OR and OR i OR to OR a OR was OR it OR my OR went OR we OR on OR he OR in OR they OR then OR of OR said OR had OR so OR she OR there OR is OR got OR you OR when	N/A	a OR b OR c OR d OR e OR f OR g OR h OR i OR j OR k OR l OR m OR n OR o OR p OR q OR r OR s OR t OR u OR v OR w OR x OR y OR z OR 0 OR 1 OR 2 OR 3 OR 4 OR 5 OR 6 OR 7 OR 8 OR 9 OR the OR and OR i OR to OR a OR was OR it OR my OR went OR we OR on OR he OR in OR they OR then OR of OR said OR had OR so OR she OR there OR is OR got OR you OR when OR one OR but OR that OR were OR with OR for OR day OR at OR her OR me OR mum OR up OR home OR go OR his OR all OR out OR saw OR because OR house OR time OR have OR dad OR came OR played OR are OR like OR after OR called OR going OR get OR back OR some OR him OR as OR school OR once OR two OR be OR did OR not OR what OR them OR very OR do OR fun OR next OR can OR play OR into OR big OR will OR am OR this OR an OR from OR weekend OR their OR people OR little OR ran OR could OR our OR friends OR about OR down OR didn’t OR put OR lived OR name OR night OR off OR bed OR see OR upon + OR 的; +OR и

Name of search engine	Query settings	Query							Time span	Other
		Single terms	Single symbols	Character combination	Number combination	Word combination				
			ANSI	a–z	0–9	Top100	Top50	Top25		
Semantic Scholar	Default search: all fields	a 1 the	N/A	a OR b OR c OR d OR e OR f OR g OR h OR i OR j OR k OR l OR m OR n OR o OR p OR q OR r OR s OR t OR u OR v OR w OR x OR y OR z	0 OR 1 OR 2 OR 3 OR 4 OR 5 OR 6 OR 7 OR 8 OR 9	the OR and OR i OR to OR a OR was OR it OR my OR went OR we OR on OR he OR in OR they OR then OR of OR said OR had OR so OR she OR there OR is OR got OR you OR when OR one OR but OR that OR were OR with OR for OR day OR at OR her OR me OR mum OR up OR home OR go OR his OR all OR out OR saw OR because OR house OR time OR have OR dad OR came OR played OR are OR like OR after OR called OR going OR get OR back OR some OR him OR as OR school OR once OR two OR be OR did OR not OR what OR them OR very OR do OR fun OR next OR can OR play OR into OR big OR will OR am OR this OR an OR from OR weekend OR their OR people OR little OR ran OR could OR our OR friends OR about OR down OR didn't OR put OR lived OR name OR night OR off OR bed OR see OR upon	the OR and OR i OR to OR a OR was OR it OR my OR went OR we OR on OR he OR in OR they OR then OR of OR said OR had OR so OR she OR there OR is OR got OR you OR when OR one OR but OR that OR were OR with OR for OR day OR at OR her OR me OR mum OR up OR home OR go OR his OR all OR out OR saw OR because OR house OR time OR have OR dad OR came OR played	the OR and OR i OR to OR a OR was OR it OR my OR went OR we OR on OR he OR in OR they OR then OR of OR said OR had OR so OR she OR there OR is OR got OR you OR when	N/A	N/A

Name of search engine	Query settings	Query							Time span	Other
		Single terms	Single symbols	Character combination	Number combination	Word combination				
			ANSI	a–z	0–9	Top100	Top50	Top25		
Web of Science	Advanced search: TS field code→TS=(…)	a 1 the	: ; . , # ! “ % & = ? + - * _	a OR b OR c OR d OR e OR f OR g OR h OR i OR j OR k OR l OR m OR n OR o OR p OR q OR r OR s OR t OR u OR v OR w OR x OR y OR z	0 OR 1 OR 2 OR 3 OR 4 OR 5 OR 6 OR 7 OR 8 OR 9	the OR and OR i OR to OR a OR was OR it OR my OR went OR we OR on OR he OR in OR they OR then OR of OR said OR had OR so OR she OR there OR is OR got OR you OR when OR one OR but OR that OR were OR with OR for OR day OR at OR her OR me OR mum OR up OR home OR go OR his OR all OR out OR saw OR because OR house OR time OR have OR dad OR came OR played OR are OR like OR after OR called OR going OR get OR back OR some OR him OR as OR school OR once OR two OR be OR did OR not OR what OR them OR very OR do OR fun OR next OR can OR play OR into OR big OR will OR am OR this OR an OR from OR weekend OR their OR people OR little OR ran OR could OR our OR friends OR about OR down OR didn’t OR put OR lived OR name OR night OR off OR bed OR see OR upon	the OR and OR i OR to OR a OR was OR it OR my OR went OR we OR on OR he OR in OR they OR then OR of OR said OR had OR so OR she OR there OR is OR got OR you OR when OR one OR but OR that OR were OR with OR for OR day OR at OR her OR me OR mum OR up OR home OR go OR his OR all OR out OR saw OR because OR house OR time OR have OR dad OR came OR played	the OR and OR i OR to OR a OR was OR it OR my OR went OR we OR on OR he OR in OR they OR then OR of OR said OR had OR so OR she OR there OR is OR got OR you OR when	1850–2018	N/A

Name of search engine	Query settings	Query							Time span	Other
		Single terms	Single symbols	Character combination	Number combination	Word combination				
			ANSI	a–z	0–9	Top100	Top50	Top25		
WorldWideScience	Preferred language: English (one language needs to be selected)	a 1 the	: ; . , # ! “ % & = ? + - * _	a OR b OR c OR d OR e OR f OR g OR h OR i OR j OR k OR l OR m OR n OR o OR p OR q OR r OR s OR t OR u OR v OR w OR x OR y OR z	0 OR 1 OR 2 OR 3 OR 4 OR 5 OR 6 OR 7 OR 8 OR 9	the OR and OR i OR to OR a OR was OR it OR my OR went OR we OR on OR he OR in OR they OR then OR of OR said OR had OR so OR she OR there OR is OR got OR you OR when OR one OR but OR that OR were OR with OR for OR day OR at OR her OR me OR mum OR up OR home OR go OR his OR all OR out OR saw OR because OR house OR time OR have OR dad OR came OR played OR are OR like OR after OR called OR going OR get OR back OR some OR him OR as OR school OR once OR two OR be OR did OR not OR what OR them OR very OR do OR fun OR next OR can OR play OR into OR big OR will OR am OR this OR an OR from OR weekend OR their OR people OR little OR ran OR could OR our OR friends OR about OR down OR didn’t OR put OR lived OR name OR night OR off OR bed OR see OR upon	the OR and OR i OR to OR a OR was OR it OR my OR went OR we OR on OR he OR in OR they OR then OR of OR said OR had OR so OR she OR there OR is OR got OR you OR when OR one OR but OR that OR were OR with OR for OR day OR at OR her OR me OR mum OR up OR home OR go OR his OR all OR out OR saw OR because OR house OR time OR have OR dad OR came OR played	the OR and OR i OR to OR a OR was OR it OR my OR went OR we OR on OR he OR in OR they OR then OR of OR said OR had OR so OR she OR there OR is OR got OR you OR when	N/A	N/A

References

Adamick, J., & Reznik-Zellen, R. (2010). Trends in large-scale subject repositories. *D-Lib Magazine, 16*(11/12), 3.

Aguillo, I. F. (2012). Is Google Scholar useful for bibliometrics? A webometric analysis. *Scientometrics, 91,* 343–351. https://doi.org/10.1007/s11192-011-0582-8.

Bar-Ilan, J. (2008). Informetrics at the beginning of the 21st century—A review. *Journal of Informetrics, 2,* 1–52. https://doi.org/10.1016/j.joi.2007.11.001.

Bharat, K., & Broder, A. (1998). A technique for measuring the relative size and overlap of public Web search engines. *Computer Networks and ISDN Systems, 30,* 379–388. https://doi.org/10.1016/s0169-7552(98)00127-5.

Bornmann, L., & Mutz, R. (2014). Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology, 66,* 2215–2222. https://doi.org/10.1002/asi.23329.

Brophy, J., & Bawden, D. (2005). Is Google enough? Comparison of an internet search engine with academic library resources. *Aslib Proceedings, 57,* 498–512. https://doi.org/10.1108/00012530510634235.

Caragea, C., Wu, J., Ciobanu, A., Williams, K., Fernández-Ramírez, J., Chen, H.-H., et al. (2014). CiteSeer x: A Scholarly Big Dataset. In M. de Rijke, A. P. de Vries, C. Zhai, F. de Jong, K. Radinsky, & K. Hofmann (Eds.), *36th European conference on IR research, Amsterdam, The Netherlands, April 13–16 2014* (Vol. 8416, pp. 311–322), LNCS sublibrary. SL 3, Information systems and application, incl. Internet/Web and HCI). Cham: Springer. https://doi.org/10.1007/978-3-319-06028-6_26.

Chadegani, A. A., Salehi, H., Yunus, M., Farhadi, H., Fooladi, M., Farhadi, M., et al. (2013). A comparison between two main academic literature collections: Web of Science and Scopus databases. *Asian Social Science, 9*(5), 18–26.

Croft, W. B., Metzler, D., & Strohman, T. (2015). *Search engines: Information retrieval in practice*. Boston: Pearson.

de Winter, Joost C. F., Zadpoor, A. A., & Dodou, D. (2014). The expansion of Google Scholar versus Web of Science: A longitudinal study. *Scientometrics, 98,* 1547–1565. https://doi.org/10.1007/s11192-013-1089-2.

Delgado López-Cózar, E., Orduna-Malea, E., & Martín-Martín, A. (2018). Google Scholar as a data source for research assessment. In W. Glaenzel, H. Moed, & U. Schmoch (Eds.), *Springer handbook of science and technology indicators*. Berlin: Springer.

Duke, L. M., & Asher, A. D. (Eds.). (2012). *College libraries and student culture: What we now know*. Chicago: American Library Association.

Eastman, C. M., & Jansen, B. J. (2003). Coverage, relevance, and ranking. *ACM Transactions on Information Systems, 21,* 383–411. https://doi.org/10.1145/944012.944015.

Georgas, H. (2014). Google vs. the library (part II): Student search patterns and behaviors when using Google and a federated search tool. *Portal: Libraries and the Academy, 14*(4), 503–532.

Grigas, V., Juzėnienė, S., & Veličkaitė, J. (2016). 'Just Google it': The scope of freely available information sources for doctoral thesis writing. *Information Research: An International Electronic Journal, 22*(1), n1.

Haines, L. L., Light, J., O'Malley, D., & Delwiche, F. A. (2010). Information-seeking behavior of basic science researchers: Implications for library services. *Journal of the Medical Library Association: JMLA, 98,* 73–81. https://doi.org/10.3163/1536-5050.98.1.019.

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. *Journal of Informetrics, 11,* 823–834. https://doi.org/10.1016/j.joi.2017.06.005.

Harzing, A.-W. (2014). A longitudinal study of Google Scholar coverage between 2012 and 2013. *Scientometrics, 98,* 565–575. https://doi.org/10.1007/s11192-013-0975-y.

Harzing, A.-W., & Alakangas, S. (2017). Microsoft Academic is one year old: The Phoenix is ready to leave the nest. *Scientometrics, 112,* 1887–1894. https://doi.org/10.1007/s11192-017-2454-3.

Hawking, D., Craswell, N., Bailey, P., & Griffiths, K. (2001). Measuring search engine quality. *Information Retrieval, 4,* 33–59. https://doi.org/10.1023/a:1011468107287.

Hood, W. W., & Wilson, C. S. (2001). The literature of bibliometrics, scientometrics, and informetrics. *Scientometrics, 52,* 291–314. https://doi.org/10.1023/a:1017919924342.

Hug, S. E., & Braendle, M. P. (2017). The coverage of Microsoft academic: Analyzing the publication output of a university. *Scientometrics, 113,* 1551–1571. https://doi.org/10.1007/s11192-017-2535-3.

Jacsó, P. (2005). Google Scholar: the pros and the cons. *Online Information Review, 29,* 208–214. https://doi.org/10.1108/14684520510598066.

Jacsó, P. (2008). Google Scholar revisited. *Online Information Review, 32,* 102–114. https://doi.org/10.1108/14684520810866010.

Jacsó, P. (2012). Using Google Scholar for journal impact factors and the h-index in nationwide publishing assessments in academia—Siren songs and air-raid sirens. *Online Information Review, 36,* 462–478. https://doi.org/10.1108/14684521211241503.

Jamali, H. R., & Asadi, S. (2010). Google and the scholar: The role of Google in scientists' information-seeking behaviour. *Online Information Review, 34,* 282–294. https://doi.org/10.1108/14684521011036990.

Jansen, B. J., & Spink, A. (2003). An analysis of web documents retrieved and viewed. In P. Langendoerfer & O. Droegehorn (Eds.), *4th International conference on internet computing, Las Vegas, Nevada, 23–26 June* (pp. 65–69).

Khabsa, M., & Giles, C. L. (2014). The number of scholarly documents on the public web. *PLoS ONE, 9,* 1–6. https://doi.org/10.1371/journal.pone.0093949.

Khan, S., Liu, X., Shakil, K. A., & Alam, M. (2017). A survey on scholarly data: From big data perspective. *Information Processing and Management, 53,* 923–944. https://doi.org/10.1016/j.ipm.2017.03.006.

Lawrence, S., & Giles, C. L. (1999). Accessibility of information on the web. *Nature, 400,* 107–109. https://doi.org/10.1038/21987.

Martín-Martín, A., Orduña-Malea, E., Ayllón, J. M., & López-Cózar, E. D. (2015). Does Google Scholar contain all highly cited documents (1950–2013)? *Granada: EC3 Working Papers* (19).

Martín-Martín, A., Orduna-Malea, E., & Delgado López-Cózar, E. (2018a). Coverage of highly-cited documents in Google Scholar, Web of Science, and Scopus: A multidisciplinary comparison. *Scientometrics, 116,* 2175–2188. https://doi.org/10.1007/s11192-018-2820-9.

Martín-Martín, A., Orduna-Malea, E., Harzing, A.-W., & Delgado López-Cózar, E. (2017). Can we use Google Scholar to identify highly-cited documents? *Journal of Informetrics, 11,* 152–163. https://doi.org/10.1016/j.joi.2016.11.008.

Martín-Martín, A., Orduna-Malea, E., Thelwall, M., & López-Cózar, E. D. (2018b). Google Scholar, Web of Science, and Scopus: A systematic comparison of citations in 252 subject categories. *Journal of Informetrics, 12,* 1160–1177. https://doi.org/10.31235/osf.io/42nkm.

Mayr, P., & Walter, A.-K. (2007). An exploratory study of Google Scholar. *Online Information Review, 31,* 814–830. https://doi.org/10.1108/14684520710841784.

Meho, L. I., & 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,* 2105–2125. https://doi.org/10.1002/asi.20677.

Mingers, J., & Meyer, M. (2017). Normalizing Google Scholar data for use in research evaluation. *Scientometrics, 112,* 1111–1121. https://doi.org/10.1007/s11192-017-2415-x.

Orduña-Malea, E., Ayllón, J. M., Martín-Martín, A., & Delgado López-Cózar, E. (2014). About the size of Google Scholar: playing the numbers. *EC3 Working Papers, 18*(23).

Orduña-Malea, E., Ayllón, J. M., Martín-Martín, A., & Delgado López-Cózar, E. (2015). Methods for estimating the size of Google Scholar. *Scientometrics, 104,* 931–949. https://doi.org/10.1007/s11192-015-1614-6.

Orduña-Malea, E., & Delgado López-Cózar, E. (2014). Google Scholar Metrics evolution: An analysis according to languages. *Scientometrics, 98,* 2353–2367. https://doi.org/10.1007/s11192-013-1164-8.

Orduña-Malea, E., Martín-Martín, A., Ayllon, J. M., & Delgado López-Cózar, E. (2014b). The silent fading of an academic search engine: The case of Microsoft Academic Search. *Online Information Review, 38,* 936–953. https://doi.org/10.1108/oir-07-2014-0169.

Orduna-Malea, E., Martín-Martín, A., & López-Cózar, E. D. (2017). Google Scholar as a source for scholarly evaluation: A bibliographic review of database errors. *Revista española de Documentación Científica, 40*(4), 185.

Ortega, J. L. (2014). *Academic search engines: A quantitative outlook (Chandos information professional series)*. Oxford: Chandos Publishing.

Oxford University Press. (2008). *Oxford wordlist*. Oxford: Oxford University Press.

Pederson, J. P. (2001). *International directory of company histories (International Directory of Company Histories Ser)* (Vol. 40). Farmington Hills: Saint James Press. **(Imprint); Cengage Gale**.

Prins, A. A. M., Costas, R., van Leeuwen, T. N., & Wouters, P. F. (2016). Using Google Scholar in research evaluation of humanities and social science programs: A comparison with Web of Science data. *Research Evaluation, 25,* 264–270. https://doi.org/10.1093/reseval/rvv049.

Shariff, S. Z., Bejaimal, S. A., Sontrop, J. M., Iansavichus, A. V., Haynes, R. B., Weir, M. A., et al. (2013). Retrieving clinical evidence: A comparison of PubMed and Google Scholar for quick clinical searches. *Journal of Medical Internet Research*. https://doi.org/10.2196/jmir.2624.

Shultz, M. (2007). Comparing test searches in PubMed and Google Scholar. *Journal of the Medical Library Association: JMLA, 95,* 442–445. https://doi.org/10.3163/1536-5050.95.4.442.

Thelwall, M. (2008). Bibliometrics to webometrics. *Journal of Information Science, 34,* 605–621. https://doi.org/10.1177/0165551507087238.

Thelwall, M. (Ed.). (2009). *Introduction to webometrics: Quantitative web research for the social sciences (Synthesis lectures on information concepts, retrieval, and services)* (Vol. 4). San Francisco: Morgan & Claypool Publishers.

Thelwall, M. (2018). Microsoft Academic automatic document searches: Accuracy for journal articles and suitability for citation analysis. *Journal of Informetrics, 12,* 1–9. https://doi.org/10.1016/j.joi.2017.11.001.

Valderrama-Zurián, J.-C., Aguilar-Moya, R., Melero-Fuentes, D., & Aleixandre-Benavent, R. (2015). A systematic analysis of duplicate records in Scopus. *Journal of Informetrics, 9,* 570–576. https://doi.org/10.1016/j.joi.2015.05.002.

van Noorden, R. (2014). Online collaboration: Scientists and the social network. *Nature, 512,* 126–129. https://doi.org/10.1038/512126a.

Vaughan, L., & Thelwall, M. (2004). Search engine coverage bias: Evidence and possible causes. *Information Processing and Management, 40,* 693–707. https://doi.org/10.1016/s0306-4573(03)00063-3.

Wilkinson, D., & Thelwall, M. (2013). Search markets and search results: The case of Bing. *Library & Information Science Research, 35,* 318–325. https://doi.org/10.1016/j.lisr.2013.04.006.

Wu, J., Liang, C., Yang, H., & Giles, C. L. (2016). CiteseerX Data: Semanticizing scholarly papers. In F. Özcan & G. Koutrika (Eds.), *2016 ACM SIGMOD/PODS Conference, San Francisco, California, 26 June–01 July 2016. New York, New York, USA: ACM.* https://doi.org/10.1145/2928294.2928306.