UNIVERSITY OF SOUTHAMPTON

Faculty of Engineering and Physical Science
School of Electronics and Computer Science

Measuring the completeness of scholarly communications databases

by

Bartosz Paszcza
MSc
ORCiD: 0000-0001-6394-3573

A thesis for the degree of
Masters of Philosophy

July 2021

University of Southampton

Abstract

Faculty of Engineering and Physical Science
School of Electronics and Computer Science

Masters of Philosophy

Measuring the completeness of scholarly communications databases

by Bartosz Paszcza

As scholarly communication has been digitised and moved online, large streams of data are being generated by the millions of publications, citations, or viewership statistics. This data, gathered by a few specialised services, serves an important role in helping individual researchers conduct literature review, science policymakers to analyse the impact of research, and science as a whole to progress effectively.

This research is aiming to summarise the requirements regarding scope, quality, transparency and accessibility of scholarly communication databases, create a uniform methodology of analysis of these datasets based on these requirements. The methodology is then used to analyse Google Scholar, Microsoft Academic and Scopus and the results are compared to other studies of these datasets. High similarity of the results obtained using designed methodology to established publications show that the methodology may be a promising method of partially automated, cross-disciplinary analysis of scholarly databases. Finally, a method of conducting an automated overlap analysis of datasets is presented as a methodological contribution, alongside relevant statistics of precision and recall.

Contents

List of Figures

List of Tables

Declaration of Authorship

I declare that this thesis and the work presented in it is my own and has been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. None of this work has been published before submission

Signed:.. Date:..................

Acknowledgements

I would like to thank my supervisors: Leslie Carr, Stevan Harnad, and Jeremy Frey, and the whole Web Science DTC community, who helped me along the way. I would also like to thank my family and friends, who have relentlessly supported me regardless of whether they had any interest in the topic whatsoever.

I would also like to thank the EPSRC for funding this research.

Chapter 1

Introduction and motivation for the study

Data about the scientific publications and their citations have been rigorously collected in large volumes only since the second half of the 20th Century. However, the few decades have seen a rapid development in the methods of collection, the volume of information gathered, and the way the data is used. The emergence of the World Wide Web can be pointed as a key turning point. Firstly, the move of scholarly communication to the Web enabled timely indexing of every publication published, alongside with information regarding its use by researchers and the society in terms of citations, downloads or mentions in media. Secondly, online availability of the datasets via user-friendly search engines or Application Programming Interfaces has inspired creation of innovative services helping researchers, policymakers and companies. Scholarly communications data has become a high-value asset with a broad number of applications.

The aims of the project are to present the requirements for scholarly communications databases, create a methodology enabling in-depth study of the databases based on aforementioned requirements and use it to study Google Scholar, Microsoft Academic and Scopus, thus obtaining results enabling comparison with other studies of these datasets. The requirements discussed take into account not only the size and interdisciplinary scope of the datasets, but also openness, accessibility, transparency of data gathering, and quality of data. As a final methodological contribution of the study, a method for conducting database overlap analysis is presented with statistics on its precision and recall.

This study focuses on a critical analysis of available methods for analysis of the scholarly datasets. By analysing the deficiencies of currently used procedures, a novel method is created to enable a cross-disciplinary analysis of the quality, scope, methods of operation and data access of Microsoft Academic. The process is designed to enable large scale, semi-automated comparison of bibliographic databases, based on sampling of the datasets using disciplinary key-phrases sourced from systematic reviews. This methodology is designed to capture the variety of research publications types represented in databases, including not only journal publications, but also preprints, and grey literature such as policy papers. Hence, the newly created method enables a deeper insight into the decision-making of databases creators, reliability of the web crawlers indexing and the quality of resultant datasets.

Critical evaluation of the databases underlying online search engines (Google Scholar, Microsoft Academic), traditional databases collecting information about publications and citations (Scopus, Web of Science) and other datasets (PubMed, CrossRef, Dimesnions) becomes crucial as scholarly communication makes further progress in the digital age. It is of importance to individual researchers, who conduct literature reviews using those tools, science policymakers, who often evaluate scientists based on the data, and science as a whole, as being it may prevent from building research activities on unrepresentative data. The variety of online publications databases and fast-paced evolution of data-gathering platforms raises concerns regarding their representativeness with regard to certain sets of criteria, such as coverage of all scholarly disciplines, the definition of academic document indexed by the database, quality of metadata and references.

In answering specific research questions, users need to consider whether the data source used provides a balanced, comprehensive reflection of the discipline, publications, authors, or institutions considered. For example, it may be important to make sure non-English publications (Parekh-Bhurke et al., 2011) are properly taken into account by the chosen data source or that the search engine indeed indexes publications from the specific discipline in sufficiently representative manner (Moed and Visser, 2008). The number and selection of criteria varies depending on the aims of the search activity, but in general, it needs to be verified whether the proposed data sources provide knowledge necessary to address the research question.

At the same time, the changing nature of scholarly communications – with the growing popularity of preprints, not only journal publications - puts new challenges in front of evaluation of data sources usability. It means that the study should be conducted on a level of individual research outputs (articles, preprints, books) rather than traditional aggregators, such as journals, which no longer capture a complete picture of scholarly communication (Moed, 2005). Academic search engines, such as Google Scholar, Scopus or Web of Science, have become a commonly used tool in conducting literature review, even though their coverage and definitions of what constitutes a scholarly publication worthy of indexing differ to a large extent (Ortega, 2014).

In the process of adaptation to the digital age, scholarly community needs an efficient infrastructure for knowledge circulation (Borgman, 2010) with an understanding of the strength and weaknesses of online knowledge sources. The representativeness of the database has to be scrutinised with respect to a set of criteria. Because of the varying interfaces of the search engines, differing types of data, multiple methodologies are used to study them, making it difficult to compare more cases. Those analyses, however, have already pointed to some disadvantages of commonly used systems. For example, a recent study by van Eck and Waltmann (2019) has identified significant citation data errors in Web of Science (WoS) and Scopus. In the case of WoS the problem is created often by missing or incorrectly parsed references. In Scopus some inconsistencies are being created by duplicate publications. Study of publications in the field of international migrations conducted by Hassan, Visvizi and Waheed (2019) has come to a conclusion that a large number of publications is missing from Scopus due to lack of interoperability of the databases.

The research community needs a thorough understanding of the origin of the widely used datasets, data processing methods, limitations of use, as well as the scope and quality of the resultant datasets (Wilsdon, 2015). However, the observable lack of common rules of scholarly data source analysis indicates that the methodology of conducting studies of online search engines and databases has not kept pace with their rapid development. For example, despite the calls to present basic statistics regarding the effectiveness of cross-database overlap analysis methods, many studies in the area of scientometrics do not contain such information (Hug et al., 2017). To verify the representativeness of the source, we ought to look e.g. at the volume, differences in disciplinary coverage, timeliness of the records, and quality of the metadata. Unfortunately, the few studies focusing on that subject have not yet resulted in a coherent, widely acclaimed methodology to study online scholarly databases.

Analysis of the online publication and citation databases is crucial also because they are more and more commonly used to evaluate scientists, based on metrics such as the h-index. Calls for ’responsible metrics’, mentioned e.g. by Wilsdon (2015), the San Francisco Declaration on Research Assessment (ASCB, 2012), and the Leiden Manifesto (Hicks and Wouters, 2015), attempt to set necessary conditions for data sources used in metric-based evaluation. Wilsdon calls for an open and interoperable data infrastructure, enabling cross-database verification and supplementation of data; openly presented information on the methods of data collection and processing; usage of global identifiers (such as DOI or ORCiD); and common semantics. It has also been highlighted that metrics should be based on the best possible data in terms of accuracy and scope, showing the need for periodic evaluation of rapidly developing portals. Hence, reviews of data source needs to also analyse the openness, transparency, and interoperability of a given platform.

Online search for literature has become a common practice, enabling researchers to find more publications, sooner after they are published, and hence speeding up the research process. However, early analyses of scholarly communication databases show some of their potential flaws and limitations which we describe in Section 3.2. Designing a methodology for reviewing online data sources on scholarly communication is a crucial step in further development of scholarly communication for individual researchers, science policymakers and science as a whole. Such a methodology should enable to verify cross-disciplinary representativeness, capture of all types of research outputs, fair representation of non-English publications, verification of quality of data.

This thesis describes the current landscape of scholarly communication databases and services. It examines methods currently used in bibliometric studies to evaluate data sources and building on this knowledge, aims to propose and verify a generalised methodology.

The methodology needs to respect a few key properties, which would enable its usage to study various databases. We propose three key criteria for the methodology, as it should be:

- significantly automated, lowering the time and effort required to verify a data source,
- become discipline-agnostic;
- be based on a diverse and large sample of publications.

Hence the proposed method is based on sampling using discipline-specific keywords from all major fields of research. The data collection is conducted via the available Application Programming Interfaces (APIs). Data analysis is conducted using a standardised and reproducible overlap analysis methodology.

The proposed methodology aims to be adjustable to the specific needs of individual research questions, or even be used to study other kinds of databases, e.g. of media article aggregators. In this work, it is used to evaluate a new scholarly communication database, the Microsoft Academic.

Chapter 2

Literature review

2.1 The history of data on scholarly communications

In the second edition of his influential "The Structure of Scientific Revolutions", Kuhn (1970) stated that "the key to understanding the structure of research communities was to draw on the recent work in the sociology of science". The phrase "recent work" refers to breakthrough research conducted by Eugene Garfield and Derek De Solla Price, utilising data on scholarly publications and citations between them to study the evolution of science, as was mentioned by Leydesdorff et al. (2014). They have not been the first to quantitatively study relations in the scholarly community. Already in 1926, Alfred Lotka published an article, concluding that productivity of researchers (measured by the number of publication) follows an inverse-square law, hence meaning that there is a narrow group of scientists writing significant number of papers, whilst a much larger group in the "long tail" contributes by a much smaller number of research outputs (Lotka, 1926).

In the 1960s the collection of data regarding scholarly discourse was institutionalised and hence quantitative studies of scholarly discourse have begun. The process of collection of such data was initiated by Eugene Garfield in what he called Science Citation Index (Garfield, 1955), a service that has since evolved into the Web of Science (WoS). The intellectual-property and science division of Thomson Reuters, which included the service, has been acquired in 2016 by a pair of private equity funds for a reported sum of US$ 3.55 billion. The service now belongs to a company named Clarivate Analytics[1].

The original aim of Garfield was to improve researchers' ability to review literature, with citations being a way to identify criticism or endorsement of published articles. It has since inspired creation of metrics to measure the development of science, a now

[1] https://www.nature.com/news/web-of-science-to-be-sold-to-private-equity-firms-1.20255

established field of research named bibliometrics, and a developing variety of commercial services.

The availability of the data has encouraged detailed studies. Garfield himself has begun a variety of quantitative analyses. The data has become a tool for creation of new measures. Garfield begun by analysing quantitatively the frequency of self-citations, number of citations of one journal by another or number of references per source paper (Garfield and Sher, 1963; Garfield, 1970). De Solla Price in his book ”Little Science, Big Science”, uses scientific methodology to analyse science itself. By utilising data on publications, he shows that science is developing logistically, rapidly developing until it will reach a certain balanced stadium. Looking at citation data, he hypothesises that the total scholarly literature existing at a given point in history (de Solla Price, 1963). Eugene Garfield, Derek De Solla Price and others have created tools to collect data on scholarly discourse and in turn created a novel, quantitative method of analysing the progress of science.

The interest in bibliographic databases, collecting data on scholarly publications and citations between them, has been steadily growing since the work of Garfield (Garfield, 1970, 1955) and their usage in analysis of evolution of science by De Solla Price (de Solla Price, 1983). It was, however, the invention of the World Wide Web, a system designed to improve the circulation of scholarly data and documents (Berners-Lee, 1989), that has enabled large-scale data gathering, accompanied by easy access by institutions and individual researchers.

Web of Science (formerly Science Citation Index) and its later born competitors - Scopus, Google Scholar, Microsoft Academic - are the tools collecting scholarly communication data. As large part of scholarly communication has transferred from paper to a digitised format, the variety and volume of data has also created alternative (to citation-based) metrics called ”altmetrics” (Priem et al., 2010). From early datasets of citations to online databases showing number of mentions of published article in social media, scholarly communication databases have been intensely developed - and have grown in importance - over the last decades. What started as a research curiosity about quantitative observations of the development of science has since developed into a growing field of study and valuable commercial properties and services, as is indicated for example by the price-tag of the recent buyout of WoS.

2.1.1 Bibliometric databases: Web of Science and Scopus

Scholarly communication has been defined by the American Association of College and Research Libraries as ”the system through which research and other scholarly writings are created, evaluated for quality, disseminated to the scholarly community, and preserved for future use. The system includes both formal means of communication, such as publication in peer-reviewed journals, and informal channels, such as electronic listservs" (ACRL Scholarly Communications Committee, 2003).

Scholarly communication data in this project is defined as publications (including non-traditional research outputs such as conference proceedings, blogs, social media posts), links between publications (citations, URL links), and other associated metadata (e.g. the number of downloads, number of shares in media or social media). Furthermore, authors, affiliations, venues of publication (e.g. journal) and links between them (e.g. co-authorship relationships) are also part of the scholarly communication data.

Garfield's Science Citation Index has been renamed Web of Science (WoS) and rapidly expanded in volume in the 1990s and 2000s, as the Web became the medium of scholarly communication. The database had been distributed to subscribers on CD-ROMs during this period. In 2002, an online portal to access the data was created. Shortly thereafter, the growing interest in scholarly databases has led to the creation of two competitors. The year 2004 has been a breakthrough point in the discipline. Then, the publishing giant Elsevier has created its own database and publication search engine, Scopus (Burnham, 2006).

Both Web of Science and Scopus rely on a selection of journals which meet their internal criteria. The data is then provided by the publishers, with rules of inclusion determined by owners of the datasets.

2.1.2 Web-based search engines: Google Scholar and Microsoft Academic

The advent of the Web has not only increased the number of publications and volume of data, but also has enabled new ways of indexing the scholarly communication. In 2004, the same year Scopus has been launched, the Web search engine leader Google has created Google Scholar (Harzing and van der Wal, 2008). Google Scholar has from the start primarily relied on Web crawlers and algorithmic identification of scholarly publications (Falagas et al., 2008). Crawlers, widely used for Web search, have hence found their way into the academic component of the WWW. Instead of relying on data provided by publishers, crawlers automatically traverse the web and scholarly documents, identifying their location (URL) and which other documents they are referencing.

Another software giant, Microsoft, has been experimenting with the creation of a citation search engine since 1996, with a system called Windows Live Academic. This, as well as a subsequent attempts such as Libra and Microsoft Academic Search, were repeatedly criticised because of limited coverage, lagging behind WoS and Scopus (Jacsó, 2011; Orduna-Malea et al., 2015). Microsoft Academic Search ceased to be updated around 2013. Microsoft Academic (MA), a new (although confusingly named) project launched in 2015, has been built from scratch by Microsoft's engineers (Sinha et al., 2015).

Microsoft Academic attempts to follow Google Scholar's strategy for approaching the problem of indexing academic articles: it is using Microsoft Bing crawlers to search for content, but is also using publication metadata obtained from the main bibliographic databases, repositories, and publishers to improve quality and broaden the database scope (Sinha et al., 2015). Those results have been initially confirmed by a variety of researchers, who also exposed some caveats of the service (Harzing and Alakangas, 2016b; Orduna-Malea et al., 2015; Hug et al., 2017; Herrmannova and Knoth, 2016). Unfortunately, in May 2021 Microsoft announced it will be retiring the Microsoft Academic website at the end of the calendar year[2].

The traditional bibliographic databases and the Web-based datasets differ in the way data is gathered, although hybrid models exist. Traditionally, what constitutes scholarly publication is defined by whether the publication venue is on an indexed list. Automatic crawlers, however, traverse the Web in search of publications and decide what is worthy of indexing on their own. The difference has profound implications on the number of research outputs indexed, citation figures, and quality of data, as we will present in 5.

2.1.3 Beyond the metadata: Open Access repositories

Regardless of the way the information is gathered, the above-mentioned services focus on metadata, information about the research publications, not their content itself. However, the World Wide Web has also created conditions for a new type of an open self-publication venue, the Open Access repository.

The spread of the World Wide Web at the end of the 20th Century meant that the cost of copying and distribution of publications has effectively diminished, as the majority of scholarly communication has been moved to the Internet. The initial attempts to create the so-called open-access journals (available without a necessary subscription, currently referred to as 'Gold Open Access' model), began in the late 1980s and early 1990s, using e-mail groups and volunteers (Jacobs, 2006).

In the meantime, however, a new way of storing and spreading the scholarly output was created: an open-access repository. Building on a tradition of self-archiving already spread in physics, arXiv.org was launched in 1991 by Ginsparg to enable storage of publications on a freely available website. In 1994, the Subversive Proposal (Harnad, 1995) aimed to extend the practice of self-archiving (recently often referred to as 'Green

[2] https://www.microsoft.com/en-us/research/project/academic/articles/microsoft-academic-to-expand-horizons-with-community-driven-approach/

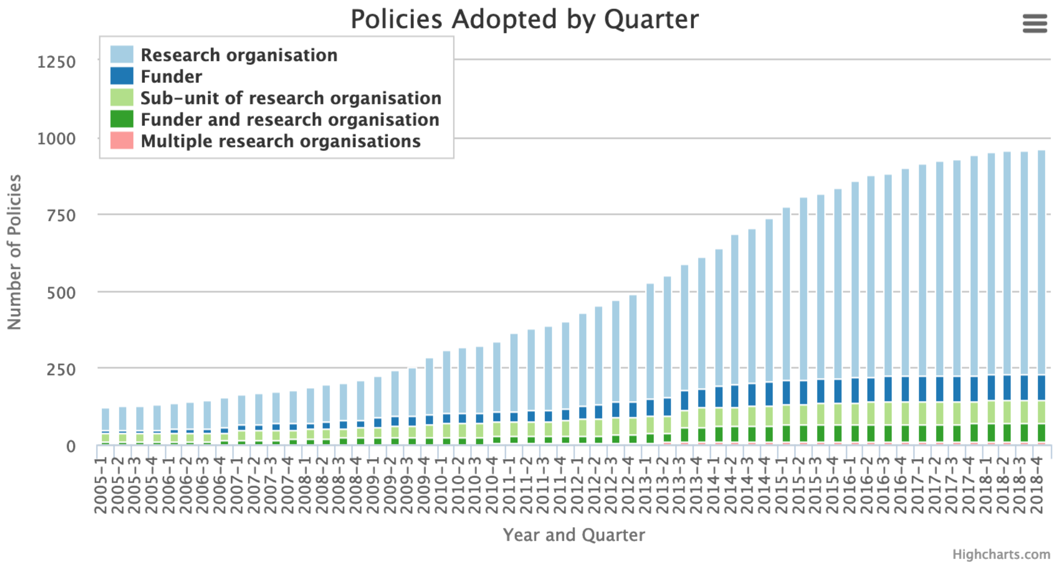


FIGURE 2.1: Number of OA mandates by type of organisation, 2005-2018, source: roarmap.eprints.org

Open Access' model) to other disciplines. It resulted in the launch of other repositories, such as CogPrints, and fundamentally - repository software Eprints, which enabled other institutions or individuals to create one's own repository in a simple way (Sponsler and Van de Velde, 2001).

At the verge of the 21st Century, the term "Open Access" (OA) was coined by the Budapest Open Access Initiative. Crucial in making the Open Access model widely adopted was the introduction of mandates requiring employees or grant recipients to deposit their work in an Open Access repository or publish in an OA journal. One of the key events was the outcome of the UK Parliamentary Committee in 2004[3]. The outcome recommended establishment of repositories by all UK higher education institutions, whilst research funding agencies were to demand deposing the outcomes of inquiries funded by them. Since then, the uptake of number of mandates can be observed on the ROARMAP[4]:

Although the direction towards free availability of outcomes of research, including publications, data, code, and other (by)products has been firmly established, a debate is continued regarding the way the openness should be achieved. Fundamentally, the two competing models are Green (self-archiving in repositories) and Gold (OA journals, with the Article Processing Charge paid by authors) (Mabe et al., 2012). Piwowar et al. (2018) estimated that at the year of publication of their article at least 28% of scholarly literature (19 million in total) was available in one of the OA models, but for articles published in 2015 (latest year analysed) the number rose to 45%.

[3] https://publications.parliament.uk/pa/cm200304/cmselect/cmsctech/399/39903.htm
[4] www.roarmap.eprints.org

The Open Access is not yet a universally adopted approach to publishing of research outputs, but it forms a vast pool of data for bibliometric analysis and research evaluation (Brody, 2006). For example, due to availability of the whole body of an article, categorisation and search practices could be greatly improved due to usage of advanced Natural Language Processing methods (Whalen et al., 2015).

Notably, two major services have been created to foster creation of a Web of linked research data and outputs. Crossref is an official Digital Object Identifier (DOI) registration agency run by Publishers International Linkining Association (PILA)[5]. As is observed in this study, the DOI, a unique identifier of a research output, is a crucial method of finding the same publication across databases or simply in the Web.

Dimensions.ai[6] is a novel linked data service and search engine run by Digital Science[7]. Despite being a recently created service, it indexes 120 million publications, alongside grants, datasets, or policy documents. Initial studies of the service indeed describe it as a service comparable in scope and quality to the likes of Web of Science or Scopus (Visser et al., 2021).

2.2 Metrics: using data to observe science, make decisions and measure impact

2.2.1 Bibliometrics: investigation of research

Interest in bibliometrics, a research field investigating academic publications and citations, has been steadily growing. From its origin as a tool for historical investigations of the development of science, created by Garfield and De Solla Price (Garfield, 1955; de Solla Price, 1963), it has evolved into a quantitative method of impact evaluation for individual researchers, scholarly institutions, and science policy (Wilsdon, 2015). Along with the growing importance of bibliometrics, the need for responsible use of them has also been highlighted. Criticism of the metrics-based evaluation is often directed at the scope and quality of data sources, especially with respect to non-English publications, documents other than journal articles, and varying coverage of individual disciplines (Moed and Visser, 2008).

As was mentioned in Chapter 1, in recent years we have observed growth in the number of portals collecting distributing scholarly communications data, novel methods of data collection, and in the variety of the types of data gathered. We have observed the introduction of new types of sources on academic communication, such as Open Access repositories and altmetric portals (Bornmann, 2015a; Amaro et al., 2015). Apart from

[5] www.crossref.org/02publishers/59pub_rules.html
[6] https://www.dimensions.ai/
[7] https://www.digital-science.com/

the traditional databases using publisher-provided information, we have observed the usage of Web-specific methods, such as crawlers, to find and identify academic documents online (Fagan, 2017).

Scholarly communications databases are fruitful sources of information for the scholarly community, policymakers, and scholarly communications innovators. Scientometricians and bibliometricians use citation data to gain quantitative insight into the process of science and communication of scientific results. Compared to the historical examples mentioned in Section 2.1, novel sources of data enable deeper and more fruitful studies. For example, such data enables to visualise and map scholarly networks, using co-citation, the co-occurrence of words, and co-authorship analysis (Mingers and Leydesdorff, 2015; Leydesdorff and Rafols, 2009).

2.2.2 Metrics and policymaking: the problem of evaluation of impact

Availability of scholarly communication data enables the creation of metrics, hence providing quantitative indicators of impact on the level of an individual article, journal, institution or researcher (Waltman, 2016a). Well-known examples of citation metrics include the Journal Impact Factor (Garfield, 1970) and the h-index (Hirsch, 2005).

The usage of Journal Impact Factor (JIF) has drawn significant criticism. It is a measure of an average number of citations of a given journal, originally intended as a guide for librarians regarding the choice of subscriptions (Garfield, 1999). However, the resultant number has been often used as a tool to evaluate researchers, an aim it was never designed to perform, which has drawn significant criticism (Seglen, 1997).

Usage of citation metrics in policy making and funding decisions by individuals, institutions, and the government has also drawn increasing attention (Bornmann, 2015b; Wilsdon, 2015). Scientometrics, apart from attempting to study the scientific procedure, has become a tool for shaping the future of science via usage of metrics in the decision-making process. Science policy has become dependent on the type, volume, and quality of data available (Waltman, 2016b), highlighting the need for reliable data sources as foundations of a fair science policy.

In his report regarding metrics usage in the British scientific evaluation system, “The Metrics Tide” (Wilsdon, 2015) has highlighted the need for “open and interoperable data infrastructure” underpinning indicators. Understanding the way data is collected and processed is “crucial” for the creation of transparent and fair metrics. Therefore, it is a principal aim of this research proposal to compare Microsoft Academic, the new source of data, to other existing databases regarding its’ transparency, scope, and data quality.

Publication and citation databases could also be a more fruitful source of input data for new services for scientists, but database owners' policies have hampered that opportunity. Subscription fees, restrictions on the volume of retrieved database records, and lack of ways to computationally access the data on large scale have been the limitations to the access to the citation data for research purposes and the creation of innovative applications (Waltman, 2016a). The new service studied in this project, Microsoft Academic, distinguishes itself because of less restrictive licensing and a promising, multifunctional Application Programming Interface enabling easier machine-readability (Hug et al., 2017). Therefore, the project is to evaluate whether the new source of data creates opportunities for improving scholarly communication studies.

2.2.3 Beyond citations: Altmetrics

Finally, the digitisation of scholarly discourse enables collection of new types of data and creation of new metrics. Altmetrics is a term describing proposed social web metrics, coined by Jason Priem and others (2010). Such metrics may help to deal with two of the upcoming issues of modern academia: the rise of the number of papers being published within a topic of inquiry to a level incomprehensible by individuals (a selection problem) and the lack of timeliness of traditional citation scores. The delay between a paper being published and it acquiring citations, often spanning a few years, prohibits the use of citation score to inform other researchers of the impact of the article directly after publication. The social web metrics - such as the number of views in bibliographic managers such as Mendeley or Zotero, number of tweets mentioning the paper, or mentions in media - provides information on the spread of the paper significantly faster.

Multiple studies investigating altmetrics have underlined their potential for broadening the scope of information about a paper. At the same time, usage of the altmetric scores raises significant questions. Firstly, usage of such information for evaluation or promotion of one's work raises questions about the feedback loop feeding underlying motivations of researchers, as number of downloads or retweets is a metric easily gamed (Gumpenberger et al., 2016). Even more than in the case of citation-based metrics, it should be asked whether it should be one of the researcher's aims to maximise the number of traditional and social media mentions? At a first glance, the question may seem rhetorical, as spreading knowledge is necessary as we conduct science. However, it is argued that often it is the process turning individual experiment result into a scientifically-proven established knowledge by the media that helps the rise of anti-scientific trends, such as the so-called anti-vaxxers community (Mnookin, 2012). Therefore, rewarding scientists for making their research newsworthy could create a scientific "click-bait" culture which is not desired.

Further questions arise around the the process of selection of metrics in this group. Researchers need to consider the quality and completeness of data, including proposed disciplinary normalisation (Sud and Thelwall, 2014). However, on a higher level, one also needs to scrutinise the inter-dependence of the metrics (Bornmann, 2014).

Altmetrics are seen as a future source of valuable information regarding individual publications, which remains to be further studied to uncover its' potential applications. The questions and dilemmas raised are inevitable at a relatively early stage of development of the novel type of measuring impact of publications.

2.3 Data-driven software innovation: helping researchers with their work

2.3.1 Turning information into knowledge

World Wide Web, a system built by Tim Berners-Lee to help scientific knowledge management and distribution in the international atomic physics laboratory CERN (Berners-Lee, 1989), has profoundly transformed scholarly communication. It has also enabled creation a number of novel tools helping researchers to search, access and distribute research outputs.

The scale of change induced by the Web can be compared only to the creation of scientific journals in 17th Century (Jinha, 2010), as it enabled faster and more convenient circulation of publications and online discussions. A study, attempting to measure the size of the Google Scholar dataset, returned an estimate of 160-165 million publications (Orduna-Malea et al., 2015), of which large part has been published after the dawn of the Web. Furthermore, the annual number of publications is constantly increasing, as is shown in 2.2, at least partially due to easier article sharing and management due to communication via the Internet. With around 7,000,000 articles being published annually around the World, the Web has become an indispensable and convenient medium for scientists.

On the other hand, scholarly communication has not taken full advantage of the opportunities provided by the digitisation of documents and the Web. Research outputs are still mainly published in a format designed for printing (PDF). Some limitations originating in the print era, such as the word limit for the article, remain (Bartling and Friesike, 2014). It can be argued that the situation can be substantially improved by taking advantage of the online nature of communication – including usage of data for innovative applications.

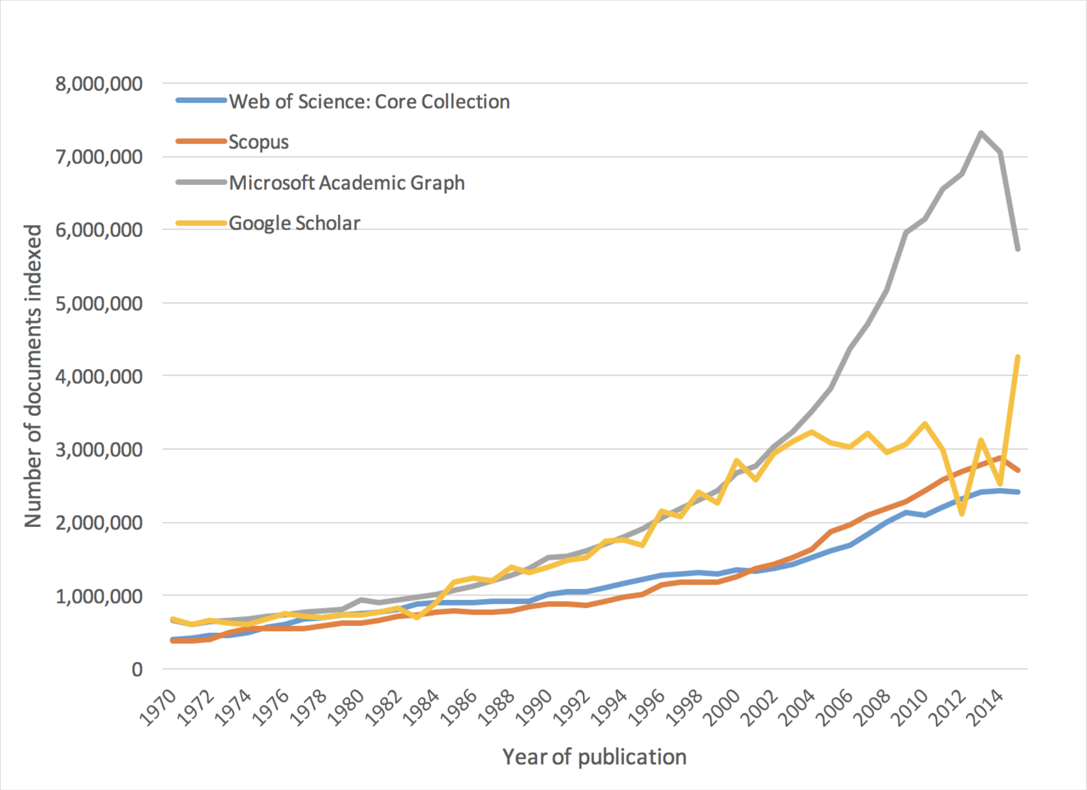


FIGURE 2.2: Comparison of number of papers indexed by each database per year of publication; numbers for Google Scholar are crude estimations only, especially after 2004, conducted during MSc dissertation research (Paszcza, 2016)

2.3.2 Overview of academic search services innovation

It comes as no surprise that new services attempting to improve the situation are being created. Bosman and Kramer are trying to crowdsource a list of innovative tools for researchers, named ”400+ tools and innovations in scholarly communications” [8]. In the category ”search (literature/data/patents/code)” 103 tools are listed, with eighty of these tools created since 2006. 2.3 presents the number of new tools for scholarly outputs search according to the year of their creation, indicating a growing trend for innovation in the area since the introduction of World Wide Web.

The novelty of the applications in this area manifests itself in various areas. Many of the tools are web-search interfaces for new or existing publication databases (e.g. OSF preprint search - enabling retrieval of records from “-xiv” disciplinary preprint repositories). Some of the novel services focus on new ways of visualisation of search results, often using topic modelling and clustering (Open Knowledge Maps, shown in 2.4) or citation network mapping (SciCurve, 2.5). Finally, some services focus on extraction of entities from the text itself, using Natural Language Processing and Semantic Web

[8] https://docs.google.com/spreadsheets/d/1KUMSeq_Pzp4KveZ7pb5rddcssk1XBTiLHniD0d3nDqo/

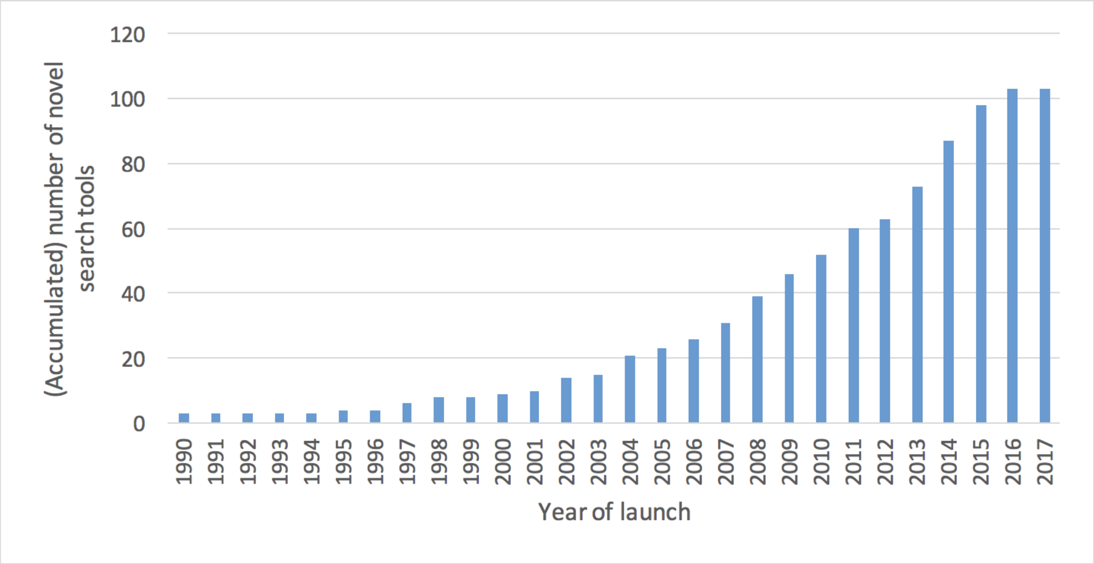


FIGURE 2.3: Accumulated number of new research outputs search tools for researchers by date of creation (data source: ”400+ tools and innovations in scholarly communications”survey)

technologies. Examples of such services include AMiner (Tang, 2016; Tang et al., 2007) or CiteSeerX (Caragea et al., 2014; Kodakateri et al., 2015).

2.3.3 The novel search engines

As shown in 2.6, researchers use the search services based on citation databases to find literature and filter results. In fact, the web portals of WoS, Scopus and Google Scholar have become the primary publication search tools for most of the researchers themselves.

JISC’s ”Researchers of Tomorrow” report (JISC, 2012) found out that 30% of PhD students born between 1984 and 1994 relied on Google and Google Scholar as primary sources for the literature search. The ”101 innovations in scholarly communications” survey found out that among 20,417 researchers, students, industry professionals and librarians who took the questionnaire, 89.1% respondents used Google Scholar, 41.5% Web of Science and 26.5% Scopus to search for publications (Bosman and Kramer, 2016). The question arises, whether these interfaces are adequate. Traditional databases, WoS and Scopus, do not index a large body of non-traditional research outputs published outside of selected peer-reviewed journals, for example in the Open Access repositories. On the other hand, Google Scholar has earned criticism for its’ lack of transparency and quality control, limited metadata, duplicate publications indexing (Jacsó, 2010; Prins et al., 2016; Harzing, 2016).

On the other hand, the availability of large science communication datasets enables the creation of innovative services for researchers and policymakers. Currently, the

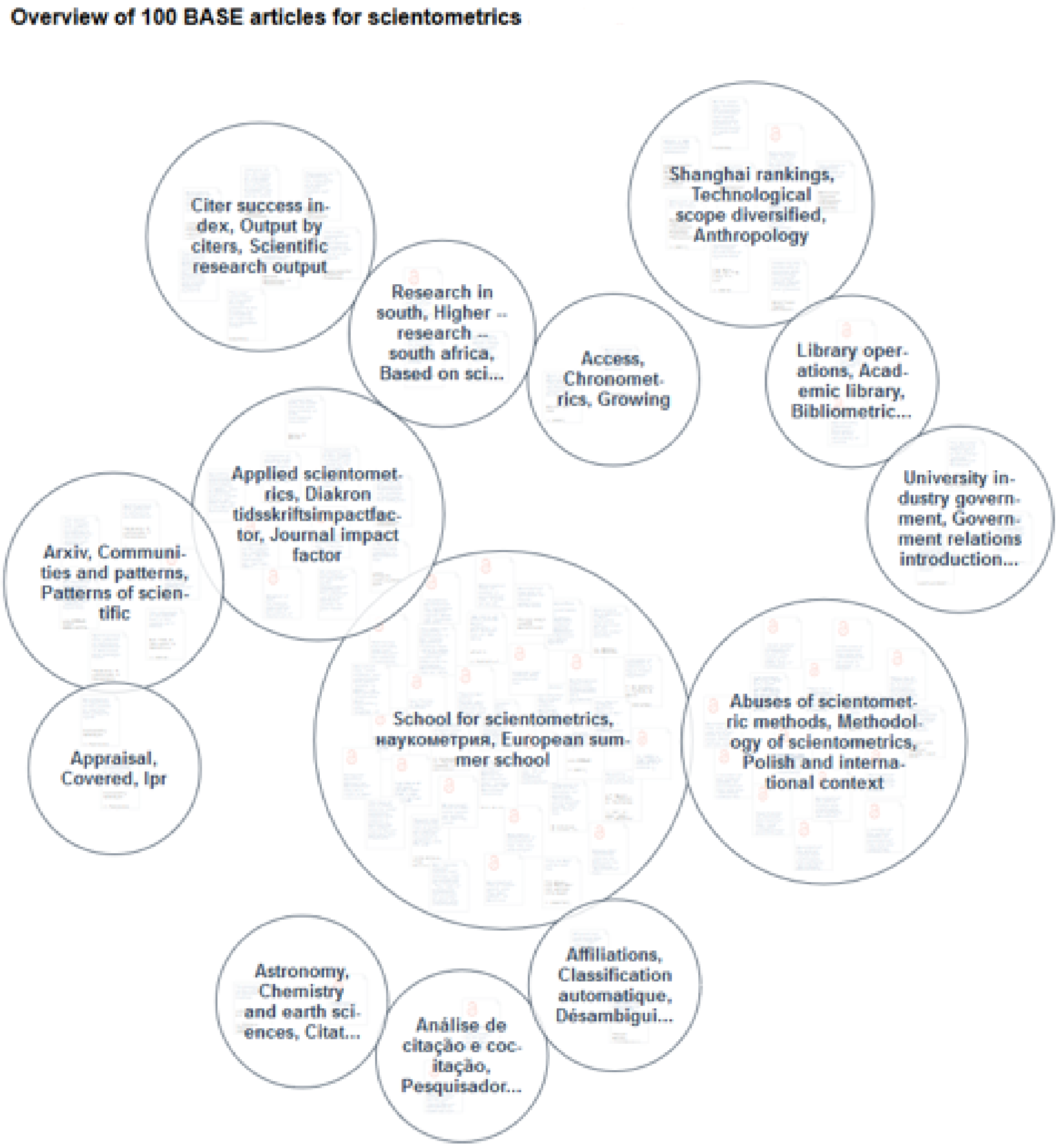

Overview of 100 BASE articles for scientometrics
Citer success index, Output by citers, Scientific research output
Shanghai rankings, Technological scope diversified, Anthropology
Research in south, Higher -- research -- south africa, Based on sci...
Access, Chronometrics, Growing
Library operations, Academic library, Bibliometric...
Applied scientometrics, Diakron tidsskriftsimpactfactor, Journal impact factor
University industry government, Government relations introduction...
Arxiv, Communities and patterns, Patterns of scientific
School for scientometrics, наукометрия, European summer school
Abuses of scientometric methods, Methodology of scientometrics, Polish and international context
Appraisal, Covered, Ipr
Affiliations, Classification automatique, Désambigui...
Astronomy, Chemistry and earth sciences, Citat...
Análise de citação e cocitação, Pesquisador...

FIGURE 2.4: Open Knowledge Maps - an example of search result of the phrase 'scientometrics'

researcher can specify a query – including an option to specify values for metadata entities, such as range of publication years – and obtain a list of documents with their total citation score. Navigating among the large body of results to identify relevant articles takes much effort, as we illustrate in Section 3.3, dedicated to literature search in systematic reviews. The rapid growth of the volume of publications, estimated to reach 8-9% annually for the last few decades, results in a need for a more efficient search and filtering tools and practices (Landhuis, 2016).

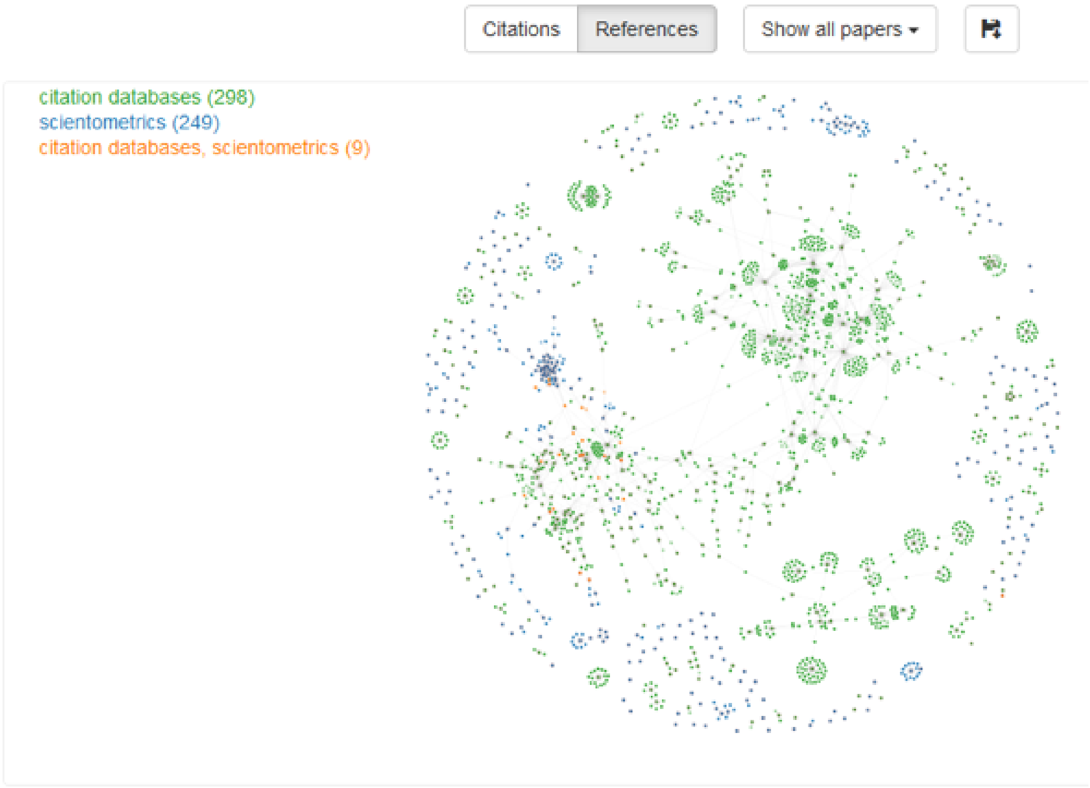


FIGURE 2.5: SciCurve: example of citation network graph for keywords 'scientometrics' and 'citation databases'

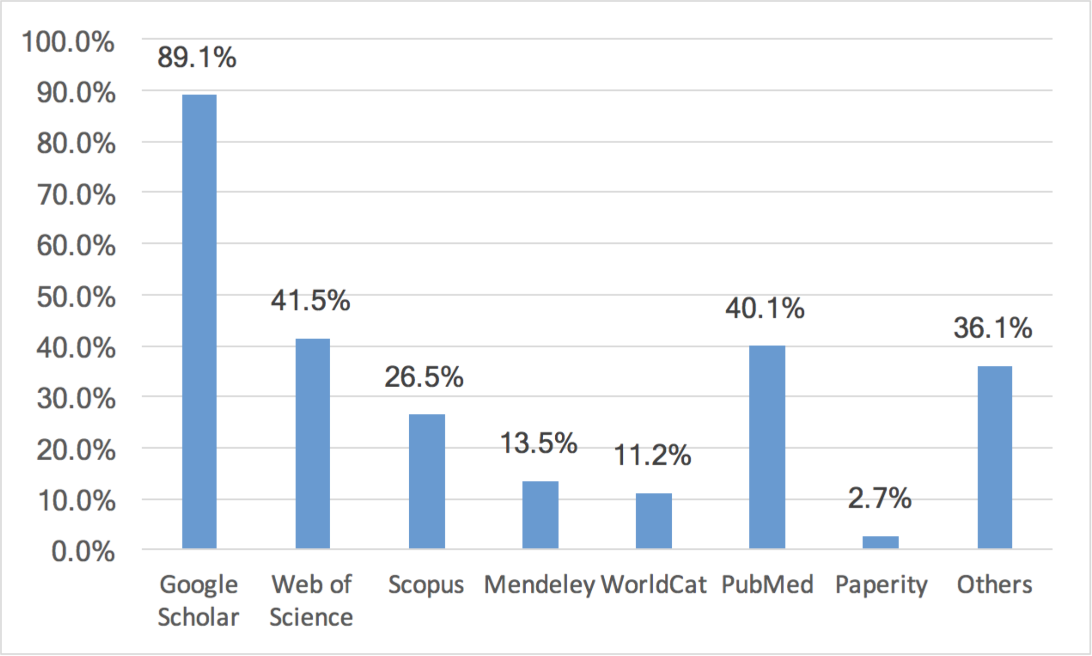


FIGURE 2.6: Tools for literature search - usage indicated by respondents to the "101 innovations in scholarly communication" survey (N = 20,417)

2.3.4 The deficiencies of emerging services

The drawbacks of online search services regard coverage of the individual portals, which may be leading to incomplete results, and narrow scope of innovation, without using many of the developments of other services.

Firstly, the data sources used are in many cases either limited in coverage to a single discipline or with their scope unverified. For example, SciCurve service uses PubMed, biomedical citation database as the only data source; Open Knowledge Maps service uses either PubMed or Bielefeld Academic Search Engine BASE. BASE is a multidisciplinary database claiming to index 100 million publications (Pieper and Summann, 2006), which only recently has been compared in size to other datasets, but without a deeper scrutiny of coverage of individual disciplines (Gusenbauer, 2019). Therefore, the reliability of the services created on top of these databases is effectively unknown and researchers have to judge whether their searches are complete by themselves.

Secondly, most of the new tools focus on one of the means of improving the service only, neglecting the advances done by others in other areas. The novel services are focused on improving the scholarly search in various ways: by using new data sources, improving analysis, filtering or visualisation, usually focusing on one of the aspects only. As noted by Reyhani Hamedani et al. (2016) text-based and link-based similarity measures individually present only partial information about the relationship between academic papers. A hybrid method, one taking into account both the document's content and its citations, should offer a thorough solution. Similarly, services advanced text and entity mining methods, such as AMiner or CiteSeerX, mainly display the information found in the form of a traditional list, with only a few (if any!) additional visualisation options.

It has to be mentioned that different use cases of the databases result in different requirements (Waltman and Larivière, 2020). In case of science policymaking, the scope and quality of the databases is of paramount importance. For example, precise reading e.g. of author's affiliations is needed to be able to enquire about a research output of an institution. On the other hand, for researchers conducting literature review it is important for a search engine to cover all related material, but also to provide links (citations) to other research outputs. Furthermore, search for keywords not only in titles and abstracts, but also full body of an article and including similar terms may be beneficial. Hence, this research is aiming to combine a diverse list of requirements for scholarly communications databases coming from these two main use cases to create a methodology enabling an in-depth assessment of usability of the datasets for researchers or policymakers.

The novel tools' innovativeness is hampered by the limitations in the coverage of the available datasets, but furthermore - by the lack of in-depth understanding of the nature of deficiencies of underlying databases. The creators and users of the services should be able to analyse whether the service is based on sources of data diverse enough to be able to answer their queries in a representative manner. Hence, this research is aiming to create a semi-automatic, cross-disciplinary methodology to conduct the review of reliability of the underlying datasets.

Chapter 3

Requirements of scholarly data sources

3.1 Aims of the study

The aim of this study is to compile a list of requirements posed by researchers, bibliometricians and scholarly policymakers regarding scope, quality and accessibility of scholarly communication databases, which is presented in Section 2.2. The study aims to propose a methodology based on those requirements, enabling to study databases and provide in-depth information regarding their disciplinary coverage, indexing of individual research outputs (documents) and their citations, overview of the methods of data gathering, management and accessibility. Then, a comparison of three databases - Microsoft Academic, Scopus, and Google Scholar - is to be conducted using the proposed methodology, enabling the results of the study to be compared to other studies of these datasets.

3.2 Currently used methodologies and their limitations

The developing area of bibliometric data sources investigation uses varying methodologies to study the scope and quality of databases and portals. The crucial problem of these studies can be defined as finding a ”golden standard set” - an independently verified set of records that could serve as a control group in the study.

Multiple approaches to create a control group are considered. Studies investigate publications from a chosen journal, e.g. (Thelwall, 2017a), from a single institution (Hug et al., 2017; Vieira and Gomes, 2009), or from a set of authors from a single university or discipline (Harzing and Alakangas, 2017; Meho and Yang, 2007), enabling to study the data source on a broader basis. Some studies combine the above-mentioned methods, but focus on reviewing coverage of a single discipline - e.g. Business and Economics in the case of Levine-Clark et. al. (2009).

Another issue is the proliferation of non-traditional ways of publications, including preprints, which some of the above-mentioned studies omit. Hence, to verify that a given data source can truthfully represent the scope of publications in a chosen discipline, the method of sampling should to incorporate journal publications from various sources (as some publishers may be indexed better than others), but also non-journal publications, such as books, conference proceedings, preprints, and other research outputs, such as data (Kousha and Thelwall, 2008). Limiting oneself to the output of traditional journals does not allow to study the representativeness of a data source regarding the whole of scholarly communication.

Furthermore, relying on a research output of a single entity, whether it is a researcher, an institution, or a set of authors, also has an impact on the kind of knowledge gained. If one aim is to study database coverage of scholarly communication in general, the method should allow for inclusion of some under-represented types of research outputs, such as non-English publications (Thelwall, 2018). Relying on a single entity, commonly an author or institution from one of the leading scholarly institutions, may provide limited knowledge, for example, regarding the coverage of the developing countries' contribution.

We need to note that studying a single discipline, although helpful in discipline-related studies and as an information tool for interested researchers and librarians, does not provide enough information about disciplinary biases among the databases, preventing conclusions from being drawn regarding general coverage of the scholarly communication as a whole.

Although a significant number of studies of scholarly datasets have been conducted, no generalised, commonly accepted methodologies have been created. Furthermore, many of the studies focus on a narrow sample based on an individual researcher, institution, discipline or a set of journals, which has a potential to skew the comparison result.

3.2.1 Database overlap analysis

There is a lack of a 'golden standard' control group ideally representing the scholarly communication landscape, due to each of the databases used in comparison being imperfect. Unfortunately, obtaining a reference set is a necessity to compare multiple data sources (Moed, 2005). Hence, in this study we have chosen to create the set by analysing the overlap of the three studied databases: Scopus, Google Scholar, and Microsoft Academic.

Studies vary in the usage of these metadata entities to identify records referring to the same research document in another database (Prins et al., 2016), a practice referred to as analysis of datasets' overlap. The methodology of matching records referring to the same documents across multiple databases is often vaguely described in bibliometric studies. Most of the bibliographic databases offer a variety of metadata attributes allowing for cross-database record matching: DOIs, URLs (address of the full-text or abstract of the indexed article), title, authors list, date of publication, venue of publication (including its title, volume, issue, page), and other identifiers (such as PubMed ID).

The two metadata attributes often used in this process are the article's title and authors list (Zuccala and Cornacchia, 2016). The documents' titles are often cleansed before comparison – deleting blanks, punctuation, and special characters (Mikki, 2010; Walters, 2007) to improve accuracy. Year of publication is sometimes used as a third criterion for document matching (Cavacini, 2015). However, this metadata entity has been found to be less effective due to a large proportion of parsing errors (Bar-Ilan, 2008) and the growing availability of online preprints, which can be recorded with an earlier date than the journal publication. Finally, the name of venue (journal) of publication is sometimes used in overlap analysis (Bornmann et al., 2009; Cavacini, 2015; de Winter et al., 2014).

Unfortunately, describing the methods used to find overlap between databases in detail and presenting some basic precision and recall statistics regarding the accuracy of the process is still not a common practice in the field, which was highlighted by Hug et. al. (2017). This, in turn, hinders the ability to estimate confidence in the results obtained, especially in cases where automatically-generated by crawlers and hence more error-prone databases are being investigated. Recently, some major studies have begun to present data, methodology and outcome datasets regarding overlap analysis, further signalling a change in transparency of scientometric studies Martín-Martín et al. (2021); Visser et al. (2021).

This study aims to first analyse the accuracy of overlap identification methods to then describe an automatic or semi-automatic means of overlap identification which would enable faster and more knowledgeable comparison of bibliographic databases.

3.2.2 Microsoft Academic

The newly designed methodology is to be put to a test by verifying the representativeness of a new data source, Microsoft Academic. It is a scholarly search engine and bibliographic database launched in 2015 (Wang et al., 2019). It is the latest in a series of Microsoft's attempts to create a scholarly communications search engine, being preceded by the discontinued Windows Live Academic and Microsoft Academic Search (Jacsó, 2011).

Early research on the service has concluded that it is a promising source of information because of its size reaching approximately 150 million records in 2017 (Hug and Brändle, 2017) and a relatively open API service and licensing enabling easier data access (Hug et al., 2017; Harzing and Alakangas, 2016b, 2017; Paszcza, 2016; Herrmannova and Knoth, 2016), leading to a conclusion in a review of academic web search engines by Fagan (2017) that Microsoft Academic ”shows real potential to compete with Google Scholar in coverage and utility”.

The current incarnation of the service is not the first attempt of Microsoft to construct a scholarly search engine and underlying database, provoking Harzing (Harzing, 2016) to call it 'a Phoenix rising from ashes'. The company has experimented with creation of a citation database since 1996, but even in 2011 coverage of the previous generation of such system - the Microsoft Academic Search - has been found lagging behind the three above-mentioned competitors (Jacsó, 2011). Furthermore, the service was discontinued and was found to have stopped indexing new articles around year 2011 (Orduña-Malea et al., 2014). Four years later, Microsoft launched an entirely new service - Microsoft Academic (MA). The new database and search engine is based on data gathered both by Microsoft's Bing web crawlers and documents from online libraries, repositories, scientific publishers, and other scholarly databases (Sinha et al., 2015). Reports highlight its volume as larger than that of WoS or Scopus, and the depth and quality of its metadata as better than that of Google Scholar (Harzing and Alakangas, 2016b; Paszcza, 2016; Herrmannova and Knoth, 2016).

A recent study by Visser, van Eck, and Waltmann (2021) has not only conducted a statistical analysis of Microsoft Academic's coverage, but also an in-depth analysis of overlap of Scopus, CWTS database, Dimensions, Crossref and MA. It has concluded that MA offers ”by far the most comprehensive coverage” among studied datasets, with particularly good indexing of Humanities and Social Sciences. Another study, by Martin-Martin et. al. (2021) further confirms the latter finding with regard to Microsoft Academic, whilst also providing detailed statistics on the overlap analysis and methodology.

Unfortunately, in May 2021 Microsoft announced it will be retiring the Microsoft Academic website at the end of the calendar year[1]. Some software underlying Microsoft Academic, such as language and network similarity packages[2] or core search engine

[1] https://www.microsoft.com/en-us/research/project/academic/articles/microsoft-academic-to-expand-horizons-with-community-driven-approach/

[2] https://docs.microsoft.com/en-us/academic-services/graph/network-similarity

MAKES[3], have been open-sourced on GitHub. As a reason for such step, Microsoft cites community-developed databases and algorithms as fulfilling the role Microsoft Academic Search was aiming to perform.

3.2.2.1 Coverage

Studies attempting to measure the scope and quality of the Microsoft Academic database have found that it outperforms Scopus and Web of Science (Harzing, 2016; Harzing and Alakangas, 2016b; Paszcza, 2016) or at least performing on-par with them (Hug et al., 2017), hence being second in terms of scope only to coverage of Google Scholar.

Harzing and Alkangas (2016b) in their study based on research output of 146 academics from the University of Melbourne concluded that "only Google Scholar outperforms Microsoft Academic regarding both publications and citations". They established that MA provided either similar or higher citation scores than Scopus and Web of Science in each of the five analysed broad disciplinary areas.

Finally, in a largest study to date, Hug and Braendle (2017) analysed the coverage of articles from the institutional Zurich Open Archive and Repository (ZORA). Firstly, they found that 52% of documents stored in ZORA could be found in Microsoft Academic. Then, upon restricting the sample to traditional research outputs only and publications from years 2008-2015, they compared the performance of MA to Web of Science and Scopus. In this study, MA performed on par with the two other bibliographic data sources. Furthermore, the comparison probably lowered Microsoft's database performance by neglecting non-traditional research outputs (grey literature, such as dissertations, published research reports or newspaper articles), which are not indexed in Scopus or Web of Science, but potentially could be found in Microsoft Academic.

In a comparison of citation counts of MA with Scopus and Mendeley, Thelwall (2017a) found out that MA found 6% more citations than Scopus overall and 51% more for current-year citations, indicating faster pace of database updating. However, MA performance has recorded significantly less citations (59%) than Mendeley. In a further study (Thelwall, 2018) it was found that search in Microsoft Academic can be conducted with high precision and recall, in a variety of disciplines. Furthermore, the correlation of citations between Microsoft Academic and Scopus was found to be 'universally high'.

[3] https://docs.microsoft.com/en-us/academic-services/knowledge-exploration-service/?view=makes-3.0

3.2.2.2 Quality

Further studies have focused on the quality of data collected in Microsoft Academic. Herrmanova and Knoth (2016) analysed the downloadable version of the database, named Microsoft Academic Graph, providing exploratory statistics, e.g. of the total number of documents (120 million), establishing MAG as the largest publicly available papers and citations database. On the other hand, they found that a large number of indexed articles lacked some of the metadata, as only 30 million publications in the dataset had citation information.

However, because their study was based on the last version of the dataset published in the downloadable form of tab-separated files in February 2016, there is a high possibility of it being outdated. During the time of my MSc dissertation research (August 2016), we found that significantly better results could be obtained using the API service than using downloadable Microsoft Academic Graph dataset, indicating a continuous improvement of the database.

Haunshild et. al. (2017) analysed the number of linked references in Microsoft Academic in greater detail. They found that the number of linked references that can be obtained is limited by Microsoft to 50 per paper - a limit which does not influence how the citation counts are calculated. Compared to Web of Science, there were 11.1% of documents which had no reference in MA, but had more than zero linked references in WoS, indicating problems with quality of citation linkage between database records in MA.

Interestingly, the usage of web crawlers by Microsoft Academic does not give the service an advantage in finding early citation. Such a result was obtained in a comparison with Scopus conducted by Thelwall (2017b), based on a sample of nine journals, largely from the field of library studies, but also including Nature and Science.

3.2.2.3 Development of the database

To provide evidence for the improvement of performance of Microsoft Academic over the period of a year, we have compared the numbers of document indexed by their year of publication in the original, downloadable dataset (Microsoft Academic Graph, published 05/02/2016) and data we obtained via the API in April 2017. As shown in Figure 10 and Figure 11, during these 14 months the number of indexed documents published after 1970 has increased by 10.8% in total. Interestingly, there has been a decrease in the number of publications published after 2010, as shown in 3.2, which may be explained by removal of duplicate entries and reclassification of non-academic writings.

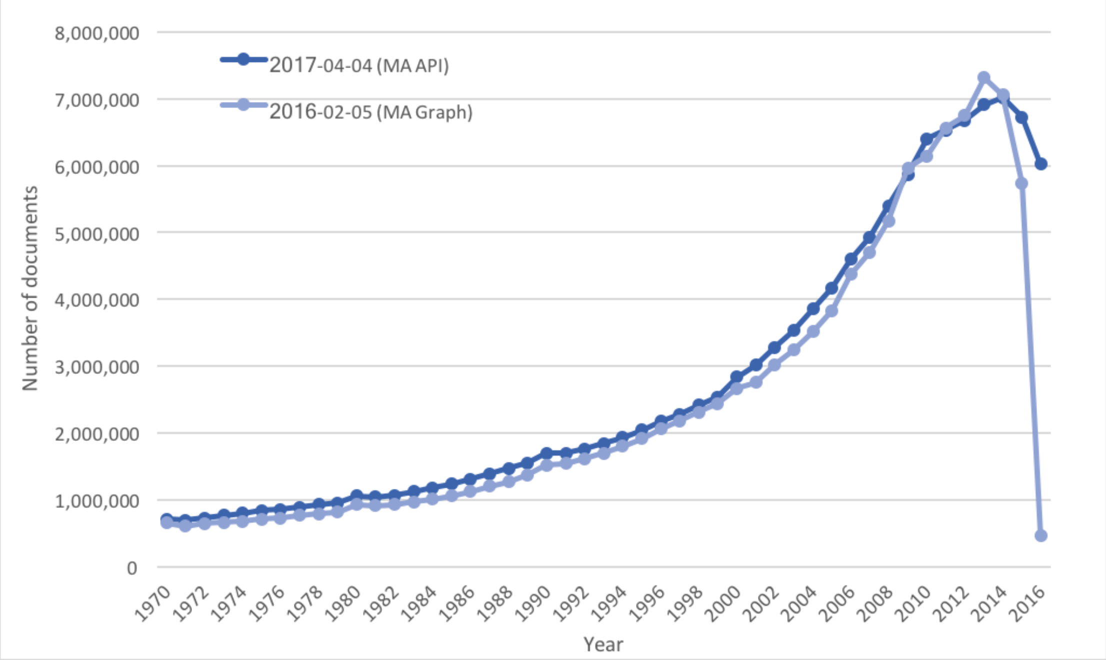


FIGURE 3.1: Development of Microsoft Academic: number of documents indexed in February 2016 vs April 2017

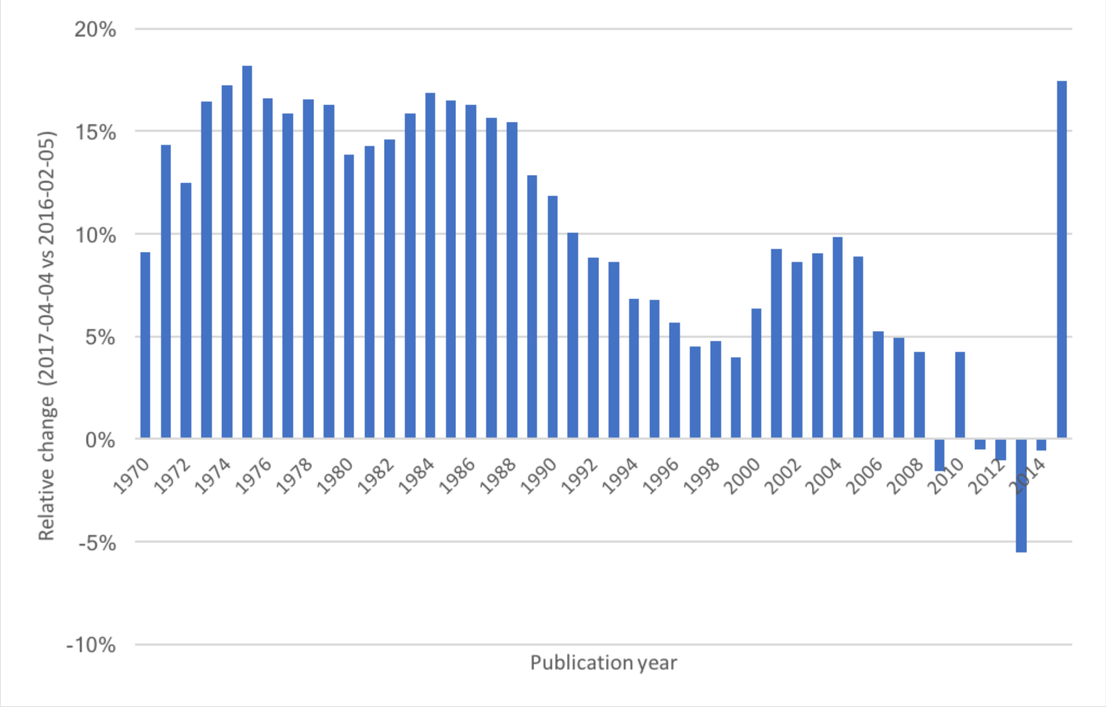


FIGURE 3.2: Development of Microsoft Academic: relative change in number of document indexed in February 2016 vs April 2017

Furthermore, Alkangas and Harzing (2017) findings confirm improvement of quality of the data. They found that between 2015 and 2017, some problems, such as title splits and incorrect year allocations, were resolved. Furthermore, MA does not suffer from common Google Scholar problems, e.g. author and journal truncation, thus providing a 'cleaner' data.

3.2.2.4 Access to data - the Application Programming Interface

In their previous publication, Hug, Ochsner and Braendle (2017) focused on analysing the MA API. They praised the service for structured and rich metadata, multiple querying functionalities, postulating that the service "the potential to be used for full-fledged bibliometric analyses (e.g. field-normalization, co-citation, bibliographic coupling, co-authorship relations, the co-occurrence of terms)". For example, the API allows creating services using both content-based similarity metrics and citation network graphs using Similarity (paper similarity calculation) and GraphSearch (retrieving a set of documents found using initial seed and rules of proceeding between entities) methods. The API also allows retrieval of bags-of-words from publications' abstracts and titles for further natural language processing analysis (Hug et al., 2017).

However, the three authors also draw attention to problems regarding the quality of the metadata: wrongly associated publication years and dynamic, automatically created fields of study classification make field-normalisation of metrics considerably harder than in Web of Science or Scopus. The system lacks an interface enabling records search using documents DOIs, severely limiting access to the database. DOI, as a unique identifier of a document, is a more precise way to find a given publication than search by a non-unique title, or keywords.

What distinguishes MA from its competitors is the relative open licensing of the API (Wang et al., 2019). The other three rival services have very limited or proprietary data usage capabilities. Google Scholar does not provide any API service and can be accessed only via the website or Publish or Perish software. Scopus and Web of Science API services are either limited in volume, destined for non-commercial purposes only, or available only to paying subscribers to the database. The Microsoft Academic Knowledge Exploration Service provides an accessible API. Although the current pricing seems not to have been yet made publicly available[4], in 2018 it used to be provided free-of-charge for the first 10,000 queries per month, then priced at $0.25 per 1,000 transactions[5]. Therefore Microsoft Academic potentially provides an opportunity for both large-scale, data-driven scientometric studies of scholarly communication and for building innovative services for researchers.

[4] https://docs.microsoft.com/en-us/academic-services/knowledge-exploration-service/resources-pricing

[5] https://stackoverflow.com/questions/54596766/how-to-subscribe-to-the-non-free-tier-version-of-academic-knowledge-api

The availability of robust metadata in Microsoft Academic enables relatively easy comparison with other data sources. Data obtained via the API includes article's DOI, (often multiple) links to full-texts, title, venue of publication and list of all authors. Such set allows to compare a set of papers to data obtained via Scopus API (using title, main author, publication venue) or from Google Scholar (title, authors, links to full-texts) (Sinha et al., 2015). The robustness of Microsoft Academic, in fact, helps overcome some limitation of the data sources. Whilst Google Scholar does not provide publication's DOI and Scopus limits its' metadata to just one author per paper, Microsoft Academic has both DOI, multiple full-text links and (supposedly) complete authors list.

3.3 Researchers' literature search practices as an example for database evaluation methodology

Any future search service needs to answer the needs of real-life search practices of the research community in order to be useful for the community. Unfortunately, only a very limited number of studies have been dedicated to the analysis of academic search practices using search engines usage data (Khabsa et al., 2016). Therefore, we will consider literature review practices recorded by researchers conducting systematic reviews as a 'golden standard' for the literature search.

Systematic reviews are secondary studies produced to provide an answer to a research question based on a consideration of all relevant primary research papers. They contain a well-documented literature search practices, including databases used, methods and individual queries (Higgins and Green, 2008; Moher et al., 2015; Liberati et al., 2009).

Reviews need to present a complete, to the extent possible, overview of a research area, which makes them similar in aims to literature reviews conducted in primary studies. Indeed, such similarity is visible in research guides for researchers produced by various university libraries, such as Leeds[6] or Birmingham[7], or guides for researchers[8].

Although originally used in medicine and life sciences, the practice of systematic reviews has spread to fields including economics, education, criminal justice and are often used for informed policy making (Tranfield et al., 2003). The spread of the practice highlights the usefulness of such secondary studies and its literature review practices to present a coherent, complete review of a research question. Therefore, search documentation in systematic reviews should represent the good practices of the scientific community - or at least reflect its idealised vision.

[6] https://library.leeds.ac.uk/researcher-literature-search-strategy

[7] http://www.birmingham.ac.uk/Documents/college-social-sciences/social-policy/hsmc-library/guide/Performing-a-literature-search.pdf

[8] http://www.howtodoaliteraturereview.com/describing-your-literature-search/

3.3.1 Search methods in systematic reviews

The data sources proposed for literature reviews include various types of the scholarly communication databases. For example, the widely used "Cochrane Handbook for Systematic Reviews of Interventions" lists bibliographic databases (e.g. MEDLINE, EMBASE, Cochrane Central Register), citation indexes (WoS, Scopus), online full-text journals sites (BioMed Central, PLoS), conference abstracts and proceedings datasets (ISI Proceedings), secondary sources such as reviews and reference lists, trial registers, and websites (among others) as proposed sources of information (Higgins and Green, 2008). Such a variety of data sources often demands laborious and time-consuming search by reviewers. Hence, in the future, unification of some of these sources in one service should be profitable for the research community. Since both Google Scholar and Microsoft Academic, apart from indexing journal publications, use web crawlers to identify non-traditional scientific output online, it should be verified whether MA API can help address the problem.

Practices for literature search in systematic reviews related to documents and citation databases can be divided into a few categories.

3.3.1.1 Search using keyword or keyphrase

Beginning with an initial seed of articles known or obtained by the reviewers, some of whom should be experts in the field, researchers define a set of discipline-specific keywords (or key phrases, e.g. "body mass index") that are common in the initial seed of articles and are related to question asked (Higgins and Green, 2008; Khan et al., 2003). Afterwards, an iterative process of queries construction and their refinement is performed to retrieve an optimal set of documents, as shown in Figure 3.4. The goal is to capture an adequate scope of relevant literature, but narrow enough to enable efficient manual review by researchers (Lichtenstein et al., 2008).

As mentioned before, multiple databases are used to perform such search (Liberati et al., 2009; Major et al., 2007). The keywords, often joined by Boolean operators to form key phrases, are used to perform "triangulation", enabling reviewers to identify precisely the literature referring to the specific research question among the body of research in related disciplines and areas – as shown in 3.5.

3.3.1.2 "Snowballing" - search for citing and referenced publications

Reviewers then consider documents citing or being referred to by the retrieved papers to identify more related literature (Greenhalgh et al., 2005; Higgins and Green, 2008). Since this type of search relies on citation indexing, it is often performed using citation

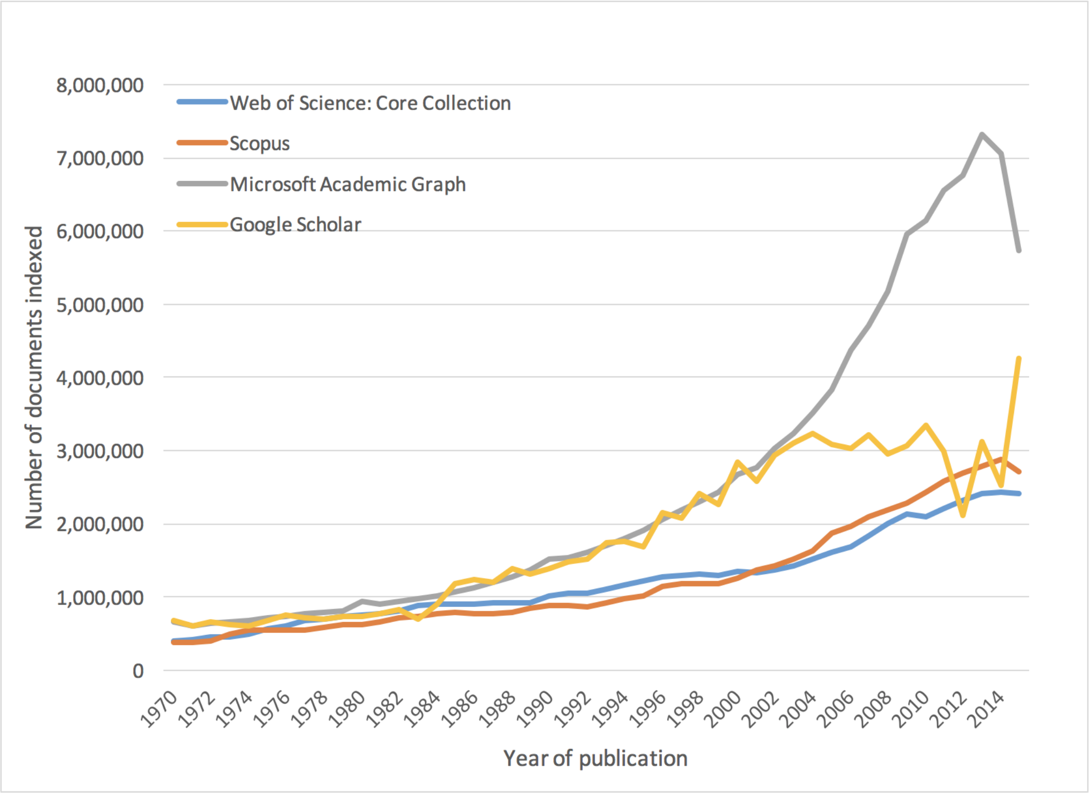
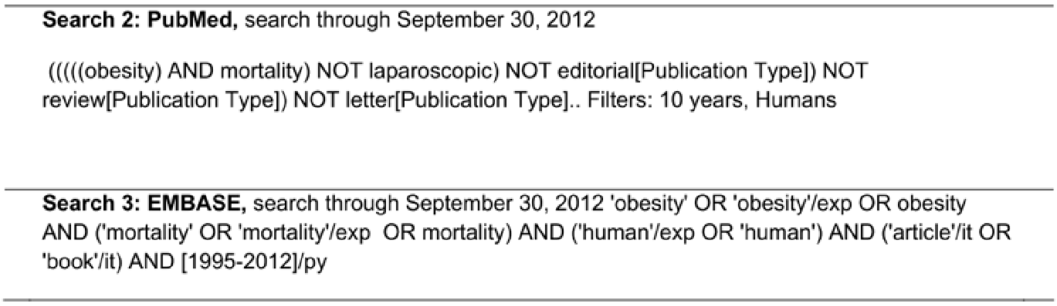


FIGURE 3.3: Comparison of number of papers indexed by each database per year of publication; numbers for Google Scholar are crude estimations only, especially after 2004 (conducted during MSc dissertation research, (Paszcza, 2016))



Search 2: PubMed, search through September 30, 2012

(((((obesity) AND mortality) NOT laparoscopic) NOT editorial[Publication Type]) NOT review[Publication Type]) NOT letter[Publication Type].. Filters: 10 years, Humans

Search 3: EMBASE, search through September 30, 2012 'obesity' OR 'obesity'/exp OR obesity AND ('mortality' OR 'mortality'/exp OR mortality) AND ('human'/exp OR 'human') AND ('article'/it OR 'book'/it) AND [1995-2012]/py

FIGURE 3.4: Example of search queries used for systematic reviews in PubMed and EMBASE (Flegal et al., 2013)

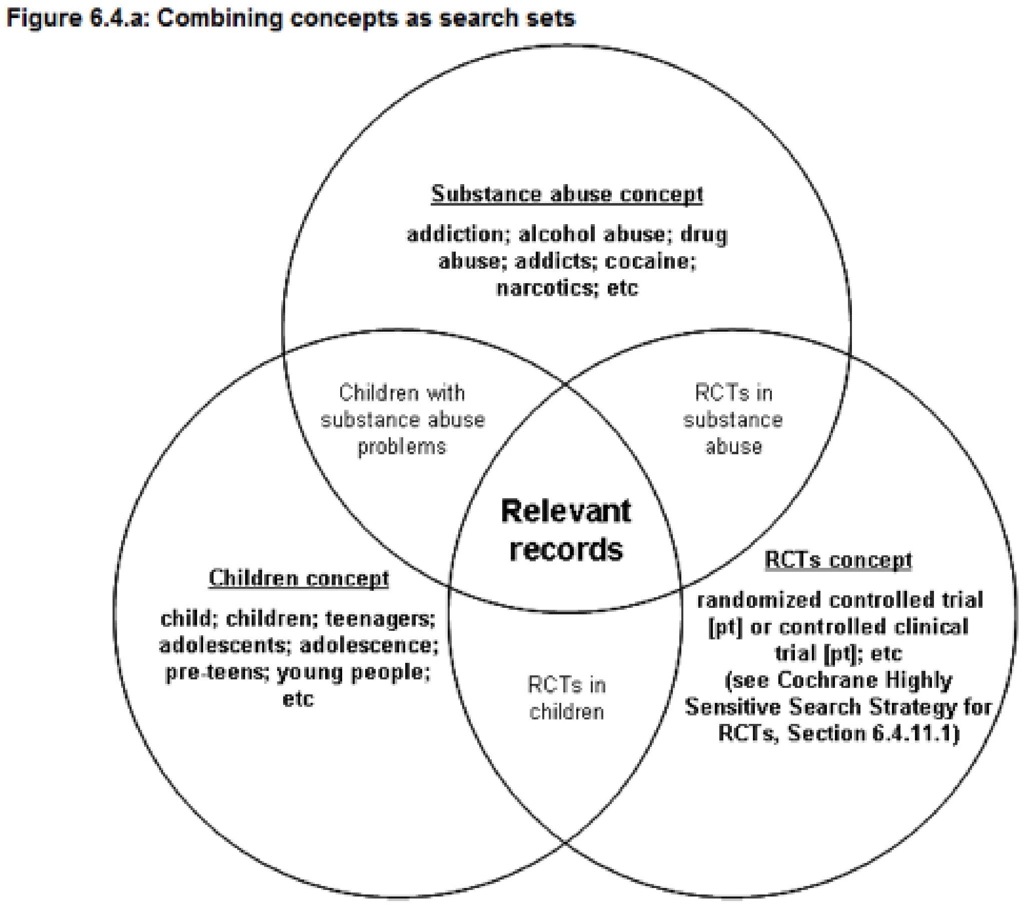


Figure 6.4.a: Combining concepts as search sets
Substance abuse concept
addiction; alcohol abuse; drug abuse; addicts; cocaine; narcotics; etc
Children with substance abuse problems
RCTs in substance abuse
Relevant records
Children concept
child; children; teenagers; adolescents; adolescence; pre-teens; young people; etc
RCTs in children
RCTs concept
randomized controlled trial [pt] or controlled clinical trial [pt]; etc
(see Cochrane Highly Sensitive Search Strategy for RCTs, Section 6.4.11.1)

FIGURE 3.5: Triangulation: identifying a set of papers relevant to the research question using keywords joined by Boolean operators (Higgins and Green, 2008)

databases. For example, in a study by (Niyibizi et al., 2017), Google Scholar and Web of Science was used to identify citing publications. Therefore, the quality of links between entities in MA has to be verified if it is to become a representative data source.

3.3.1.3 Search for non-traditional research output, such as conference proceedings and abstracts, books, grey literature

Such literature is not indexed in the core of traditional citation databases (Scopus, Web of Science) as they index information from a set of selected journals. Microsoft Academic and Google Scholar, which index publications using web-crawlers, should in principle help address that problem. However, many studies of Google Scholar have highlighted problems with classification of articles as scientific by automatic crawlers (Harzing and van der Wal, 2008; Prins et al., 2016). Whether Microsoft Academic is more reliable in this perspective remains to be verified.

The systematic literature search methods are a well-described source of information regarding the ways researchers are searching for knowledge. Such information serves as an inspiration for creation of a novel methodology to study scholarly data sources, as it should help us verify the databases with regards to the ways they are used by individual researchers.

This research is aiming to design a new methodology learning from the deficiencies of currently used methods. Inspired by the process of systematic reviews, the new methodology is to utilise keyword search as a way of sampling the space of scholarly publications. The problem of establishing a "golden reference set" also needs to be overcome to ensure a sensible overlap analysis.

Chapter 4

Proposed Completeness Methodology

4.1 Overview

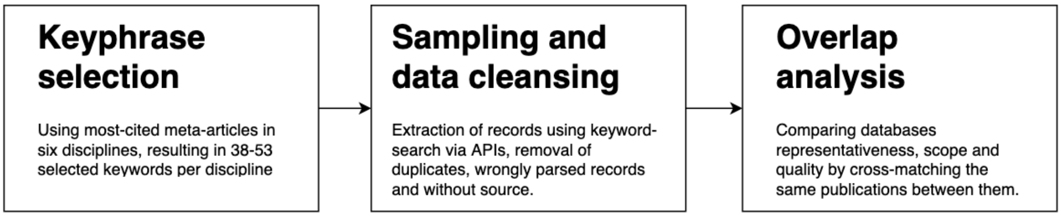

Keyphrase selection
Using most-cited meta-articles in six disciplines, resulting in 38-53 selected keywords per discipline
Sampling and data cleansing
Extraction of records using keyword-search via APIs, removal of duplicates, wrongly parsed records and without source.
Overlap analysis
Comparing databases representativeness, scope and quality by cross-matching the same publications between them.

FIGURE 4.1: Overview of the methodology

Figure 4.1 presents the overview of the proposed methodology for quantitative analysis of bibliographic databases. It begins with the creation of a ”seed” set of keywords, later used in the querying process in Google Scholar, Microsoft Academic, and Scopus, to find publications with these keywords in tiles. This process aims to create a reference set - sample of documents representative for broad fields of study used in comparison.

The cleansing process and methods of analysing the overlap between databases are also presented. Then, a comparison between the three databases was conducted, including an analysis of overlap of the three search engines databases.

4.1.1 Keyphrase sourcing and selection process

This section describes the first element of the methodology: selection of disciplinary keywords used to query the databases.

The keyphrases selection process had to ensure that queries based on the phrase return:

- a substantial number of records, but lower than the limit imposed by APIs (e.g. number of results of a Google Scholar query in Publish or Perish (Harzing, Available from 2007) is limited to 1000 records, and Scopus limits to 200 results),
- is discipline-specific to enable comparison of coverage of different fields of research,
- allows the representation of a variety of research in a given disciplinary field.

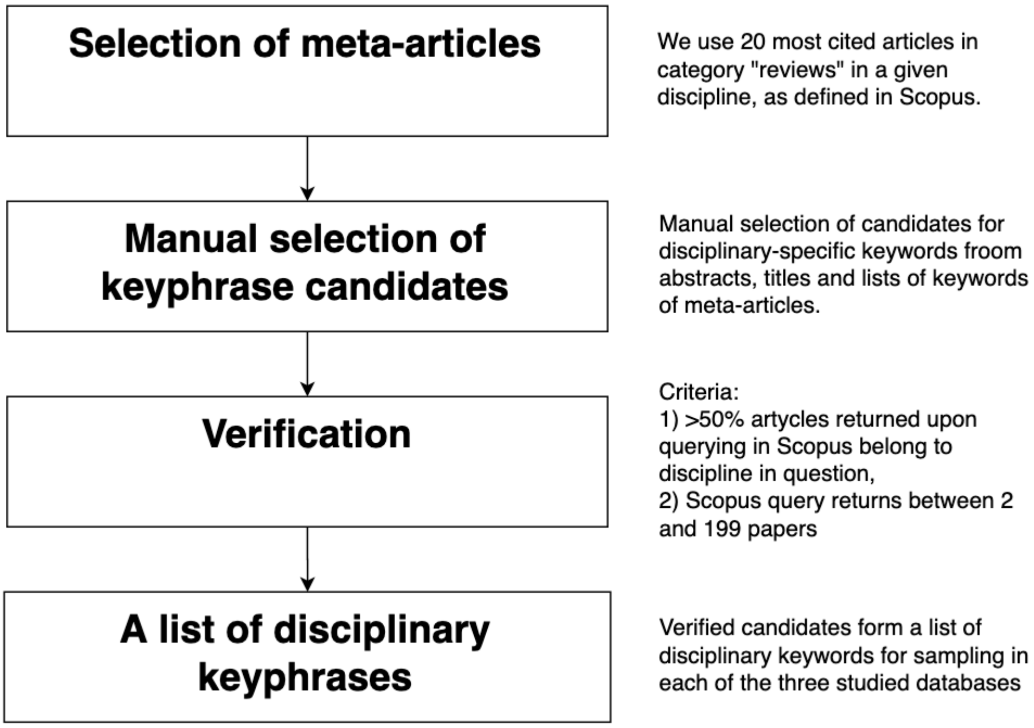

Selection of meta-articles
We use 20 most cited articles in category "reviews" in a given discipline, as defined in Scopus.
Manual selection of keyphrase candidates
Manual selection of candidates for disciplinary-specific keywords froom abstracts, titles and lists of keywords of meta-articles.
Verification
Criteria:
1) >50% artycles returned upon querying in Scopus belong to discipline in question,
2) Scopus query returns between 2 and 199 papers
A list of disciplinary keyphrases
Verified candidates form a list of disciplinary keywords for sampling in each of the three studied databases

FIGURE 4.2: Key-phrases sourcing, selection, and verification process

The process began with obtaining a list of 20 most cited review articles in each of the given six OECD disciplinary fields from Scopus. The OECD classification (Organisation for Economic Co-operation and Development (OECD), 2007) was used, as it is a widely accepted standard, which is independent of disciplinary classifications created within the studied databases and which can also be easily mapped onto the Scopus classification. A decision was made to use meta-review publications as a source of keywords, as such documents represent an overview of a live, developing topic of study and hence are a source of widely-used, but disciplinary-specific keywords. Furthermore, using the most highly cited publications helps to focus on the impactful areas of research within a discipline.

The list of reviews was cleansed to remove duplicates or wrongly classified documents. A further limit of one review per journal was introduced in order to broaden the coverage of sub-disciplines within each of the fields of study. Then, a manual search for keywords was conducted, using the reviews' titles, abstracts, and keywords lists. This was followed by the creation of a list of one to eight candidate keyphrases (individual keywords or longer phrases) per review publication. A Scopus search, looking for documents with a given keyword in the title, was then conducted for each candidate, where they had to pass all of the following criteria:

- Number of returned Scopus records – when querying for documents with the phrase in title – had to be lower than 200, the maximum number of records obtained in a single query in Scopus, as shown in Table 4.1
- More than half of the obtained records had to be classified by Scopus as belonging to sub-disciplines of the area of research in question

We have to note that the aim of the keyword selection is not to recreate real-world keywords used by researchers to describe fields of their research, as obtaining a set of those would demand the involvement of researchers from all studied areas. Instead, the method used aims to identify phrases that appear in the titles of research documents in particular disciplines in order to compare sets of results of identical queries performed on different databases.

This method of obtaining sample was, however, based on the methodology of conducting systematic reviews. In this type of meta-study, topic-related keywords are used to obtain all publications relevant to answering a given research question. It can be considered as a 'golden standard' of performing – and recording – a systematic review of literature, and hence methodology based on such standard should, in principle, enable us to obtain a representative sample of disciplinary publications.

As an example, in the field of Engineering and Technology, the most cited review found was "One-dimensional nanostructures: Synthesis, characterization, and applications" (Xia et al., 2003). A candidate keyphrase "1D Nanostructures" was selected. A Scopus query for this phrase in articles titles (in brackets) resulted in 71 records (fitting within the specified limit for that discipline, being 145 documents/query), with 53 of them classified as belonging to Materials Science, 33 as Engineering, 29 in Physics and Astronomy, 25 in Chemistry, 10 in Chemical Engineering, and fewer than 10 in 5 other disciplines. As we can see from these numbers, Scopus classification is non-unique (meaning that each record can be classified as belonging to more than one field of research – in this example, the average number of disciplines per paper is 2.42). In this case, as Materials Science, Engineering, and Chemical Engineering are sub-fields of the OECD discipline of Engineering and Technology, it has been concluded that indeed more than 50% of records obtained by querying for the keyphrase "1D Nanonstructures" belong to the field in question.

An important note regarding the latter criterion has to be made. The rather arbitrarily chosen 50% threshold may result in some sample noise; this is due to the presence of a

TABLE 4.1: Disciplinary keywords selection process

Field of study	Reviews	Max query results threshold	Keyword candidates	Keywords selected	Number of records
1. Natural Sciences	20	200	90	39	1918
2. Engineering and Technology	20	145	106	42	1806
3. Medical and Health Sciences	19	140	113	44	1968
4. Agricultural Sciences	19	125	113	46	1923
5. Social Sciences	19	125	111	53	1907
6. Humanities	21	180	119	38	1989

significant number of records from outside the field, hence hindering the opportunity of observing disciplinary differences in database coverage. However, we would argue that increasing the limit would also lead to bias. Because Scopus, as can be seen in the example above, often classifies multidisciplinary research outputs in more than one discipline, setting a high threshold for documents assigned only to one discipline would result in the rejection of keywords referring to interdisciplinary fields of research, e.g. on the intersection of engineering and medicine. The exclusion of such documents would prohibit us from obtaining a representative, diverse sample of publications from a broad area of study.

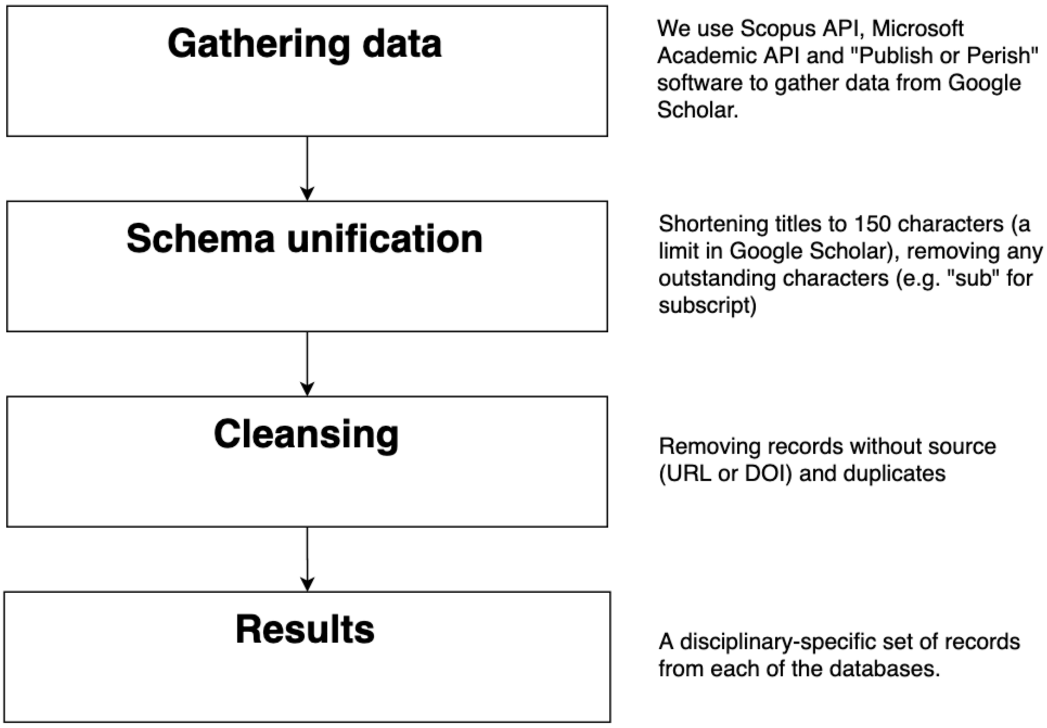
4.1.2 Data gathering

Lists of disciplinary keywords were then used in semi-automatic querying using Scopus and Microsoft Academic Application Programming Interfaces (APIs) and ”Publish or Perish” software (Harzing, Available from 2007) for Google Scholar access. Unfortunately, due to subscription limitations, it was not possible to obtain access to data that would include a number of citations of papers obtained from Clarivate Analytics’ Web of Science API, hence this database was not used in this comparison.

The Application Programming Interfaces are a standard data-gathering method in the Web, due to their convenience and opportunity for automation. Therefore, the lack of such interface in Google Scholar is a serious drawback of that platform. It has to be noted that Scopus and its API are available to paying subscribers, whilst MA at the time the study was conducted used to offer a free quota of 10,000 queries/month (Hermannova and Knoth, 2016; Paszcza, 2016), and a pay-per-query fee system above this limit. Furthermore, Scopus limits the scope of data gathering to a specific discipline, and imposes a limit of 200 documents per query and a quota of 20,000 queries per 7 days [1].

Furthermore, one of the significant drawbacks of Google Scholar is a very limited metadata structure (effectively limited to title, authors, URL, citations, and venue of publication – lacking DOI, for example), whilst Microsoft Academic supports a broader schema, shown in Figure 4.4.

[1] https://dev.elsevier.com/policy.html



We use Scopus API, Microsoft Academic API and "Publish or Perish" software to gather data from Google Scholar.

Shortening titles to 150 characters (a limit in Google Scholar), removing any outstanding characters (e.g. "sub" for subscript)

Removing records without source (URL or DOI) and duplicates

A disciplinary-specific set of records from each of the databases.

FIGURE 4.3: Data gathering and cleansing process overview

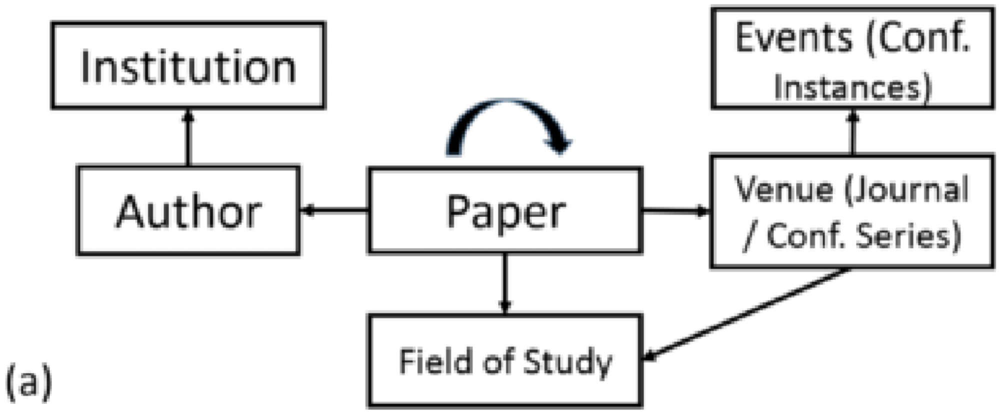


FIGURE 4.4: Microsoft Academic's data schema, from (Sinha et al., 2015)

One clear drawback of the Microsoft Academic API found during the study is the prohibition of use of stop-words (e.g. 'the') in query based on a number of words. The system in such cases returned zero results; such practice prohibits the automatic conversion of titles to words for usage in queries. Combined with the inability to query by part of the title (except for the ability to query with the beginning of the title) and lack of DOI-based query, this hampers methods of searching for documents with incorrectly parsed titles (regardless of whether the error in parsing is in MA or another dataset).

TABLE 4.2: Summary of data cleansing statistics

	Google Scholar		Microsoft Academic		Scopus	
Records	23,144	100.0%	14,102	100.0%	7,903	100.0%
* Without source (URL/DOI)	7,246	31.3%	388	2.8%	0	0.0%
* Duplicates - identical DOI/URL	149	0.6%	153	1.1%	105	1.3%
* Duplicates - identical metadata	392	1.7%	190	1.3%	107	1.4%
* Duplicates - PPM clustering	261	1.1%	153	1.1%	34	0.4%
Cleansed records	15,096	65.2%	13,218	93.7%	7,657	96.9%

Before cleansing, some unification of the data schema was conducted, as databases use different data formats for record storage. It was found during the study that for example, the publication's title length in Google Scholar is limited to 150 characters, imposing a similar limitation on results from other datasets in the study. Furthermore, the strings containing titles from Microsoft Academic contained indications of subscripts or superscripts in the form of letters ″sup″ or ″sub″ within the string, which had to be removed.

In order to unify the results, the search conducted on databases only looked for keyphrases in titles of documents. This type of query was the single common keyword-search method in all three of the studied datasets. As an example, Google Scholar enables search in the full body of document (without ability to search in titles and abstracts only), whilst both Scopus and MA offer opportunities for search in abstracts, without an option for a full-text query.

It has to be concluded that Microsoft Academic stands out in terms of its openness and methods of access (free-of-charge API quota) and robust metadata structure. On the other hand, however, some inefficiencies in methods of querying (e.g. necessary elimination of stop-words, lack of search for phrases in titles) demanded scrupulous input manipulation in order to retrieve the desired documents from the database.

4.1.3 Data cleansing

The data cleansing procedure aimed to remove unidentifiable, duplicate, or wrongly parsed records from the analysed set. A few methods were used to find and remove such documents. Results of these efforts are presented in Table 4.2.

First, all records without a link to full document were removed. In practice, in the cases of Microsoft Academic and Scopus, these consisted of records lacking both URL and DOI. Because Google Scholar does not specify records' DOIs, in this case all documents without URL were removed.

Secondly, within each database, a search for duplicate records, using URLs and DOIs was performed to remove repeated papers. Manual verification of other metadata

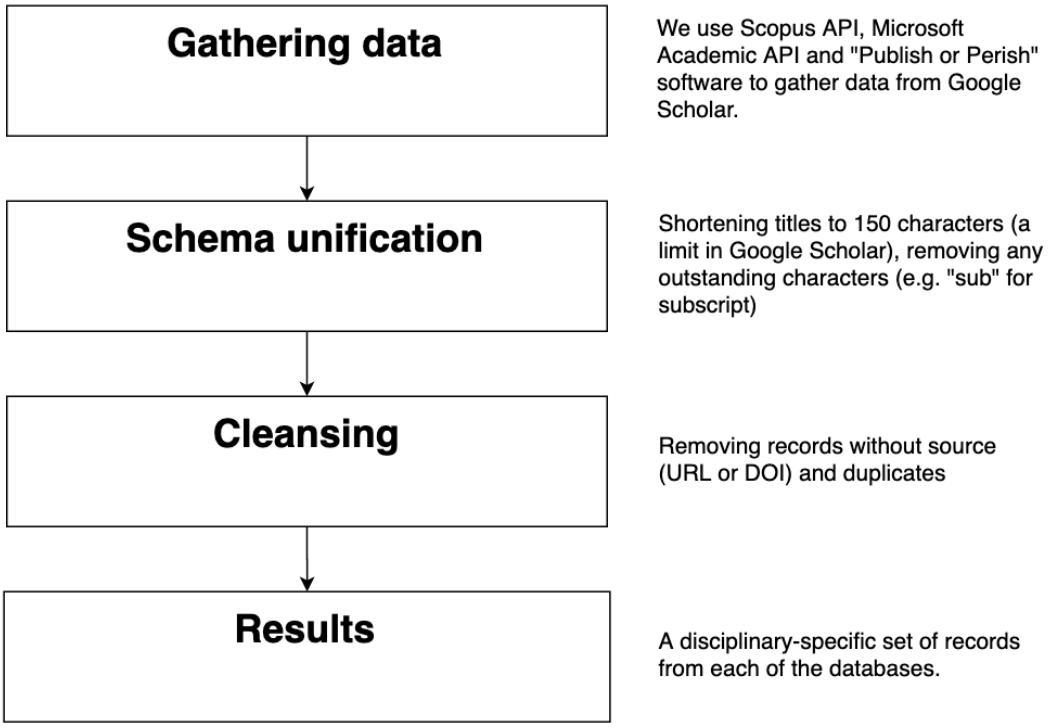

Gathering data
We use Scopus API, Microsoft Academic API and "Publish or Perish" software to gather data from Google Scholar.
Schema unification
Shortening titles to 150 characters (a limit in Google Scholar), removing any outstanding characters (e.g. "sub" for subscript)
Cleansing
Removing records without source (URL or DOI) and duplicates
Results
A disciplinary-specific set of records from each of the databases.

FIGURE 4.5: Data cleansing statistics

fields was also conducted to ensure that the records were indeed duplicate, but DOIs and URLs can be considered a reliable source of an article's unique identification (Thelwall, 2018).

To further investigate duplicate records within each of the three databases, title-based query was performed. In order to find those records with erroneous or no DOIs/URLs, the quest was to find records with the same titles. Manual verification followed, defining duplicates as articles having the same titles, at least one author in the authors list, and publication dates within a range of ±1 year of each other. An exception was made for serial publications under the same title (e.g. ”Computer Modelling Working Group Report”). It should also be noted that such methods tend to remove ’upgraded’ publications as duplicates, e.g. when a paper was first presented at a conference and then developed into a journal article. In such cases, the record with larger number of citations was kept in the dataset. It is believed by the authors that although this is a simplification, it is a necessary one – allowing to cleanse the dataset without the effort of analysing papers in detail to find the degree of similarity. Finally, all records containing the word ’patent’ in title or name of venue of publication were removed from datasets.

Finally, a manual review was performed on duplicates for which titles and other metadata are not identical, for example due to errors in parsing or difference in versions of the document. To find such cases, a clustering algorithm named prediction by partial matching (PPM) with default settings (of radius 1.0 and block consisting of 6 words), implemented in OpenRefine, was run on records' titles. Lists of clusters of potential duplicates were created as a result of running the algorithm. Clusters were then manually reviewed, using the criteria described in the paragraph above (of authors and publication dates), and identified duplicates were removed.

To summarise, it has to be noted that Microsoft Academic's data quality has been found to be only marginally lower than that of Scopus (93.7% to 96.9% of the input records passing the data cleaning process), whilst the set obtained from Google Scholar has seen removal of a third of the records (65.2%). It has to be noted, however, that this was mainly due to the number of records without URL in Google Scholar (31.3%) – and probably also a lack of DOIs in metadata – highlighting the limitations of the small amount of data describing each document that can be obtained from this source. The three methods of duplicates identification return fairly uniform results in each of the datasets, with the number of removed documents being in the range of 0.6% to 1.7%. The low number of duplicates due to repeated DOIs/URLs in Google Scholar (0.6%) compared with the results of 1.1% and 1.4% in MA and Scopus can also be partially attributed to the fact that lack of DOIs in GS limits the chances of identifying duplicates in the set.

4.1.4 Overlap analysis

To analyse overlap across databases, we need to choose a set of criteria and design an algorithm. Unfortunately, only a narrow set of articles mention the details of methods used to do so, along with statistics on effectiveness of this matching. As was mentioned in Section 3.2.1, (cleansed) article titles, authors are commonly used (Prins et al., 2016; Zuccala and Cornacchia, 2016), sometimes supplemented by less accurate metadata: year of publication and venue of publication (Bar-Ilan, 2008; de Winter et al., 2014).

A recent study on Microsoft Academic used three independent methods (the meaning each individual criterion was sufficient to form a match): DOIs, cleansed title, and bibliographical information (journal title, volume, issue) (Hug et al., 2017). The last of the three methods was deemed the least reliable. The same article mentioned that studies should elaborate on the methods used and present basic statistics (precision, recall) – a step that is often omitted in publications in the research field.

Therefore, the aim of the study was first to verify which metadata would yield best possible method of finding overlap by analysing the precision and recall.

Furthermore, Thelwall (2018) studied the issue of identifying matching journal articles in Microsoft Academic. This study focused on evaluating the queries to MA themselves, rather than finding overlap of two sets of records from different sources. The results largely confirm the results from the study by Hug et.al. (2017).

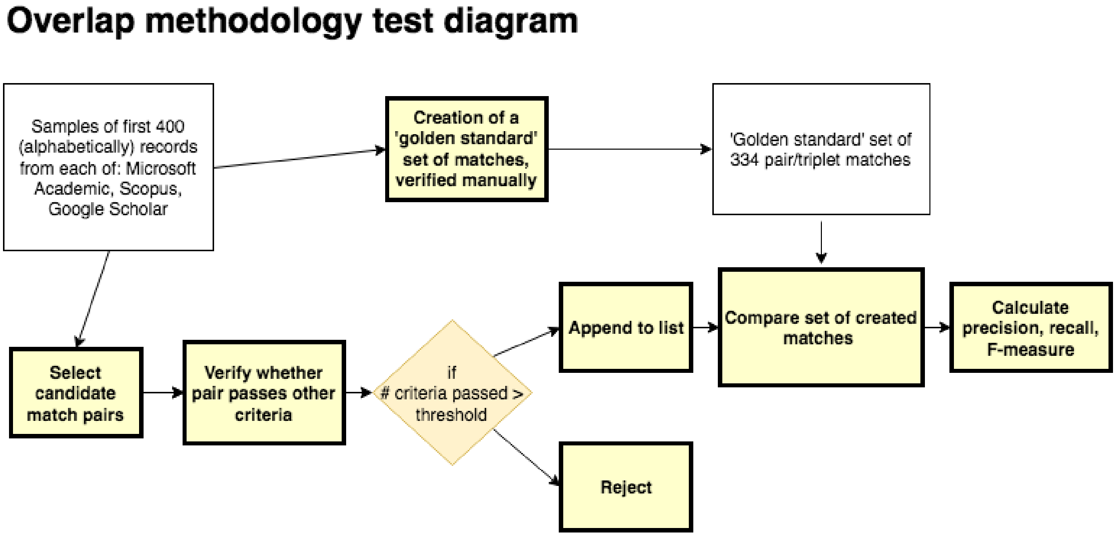

Overlap methodology test diagram
Samples of first 400 (alphabetically) records from each of: Microsoft Academic, Scopus, Google Scholar
Creation of a 'golden standard' set of matches, verified manually
'Golden standard' set of 334 pair/triplet matches
Select candidate match pairs
Verify whether pair passes other criteria
if # criteria passed > threshold
Append to list
Reject
Compare set of created matches
Calculate precision, recall, F-measure

FIGURE 4.6: Overlap identification test diagram

This study aimed to design and verify a methodology to analyse the overlap of the Microsoft Academic, Google Scholar and Scopus datasets. Our study aimed to test the criteria and choose the best performing ones. The task of finding overlap of three databases adds another layer of complication (as some metadata, e.g. DOI, was available only in two of the sets), so testing the selection of metadata used to identify match was deemed as necessary.

Firstly, we aimed to create ’a golden set’ - a sample of documents for which we semi-manually find their records in all three databases. The set was obtained by choosing the first 400 records in the field of Natural Sciences, alphabetically sorted, from our study samples in each of the three datasets. Duplicate records were then removed. Due to differences in authors’ names format recording (e.g.in Google Scholar author’s name would be presented as ”BM Paszcza”, in Scopus as ”Paszcza B.” and in Microsoft Academic can be presented as ”bartosz paszcza”, ”paszcza bartosz”, or ”b m paszcza”) unifications to surnames only had to be conducted. Titles were also unified: truncated to 150 characters (a maximum title length in Google Scholar), non-ASCII characters, diacritic signs removal. After investigation, a decision was made to also remove all spaces, convert double letters to single (”illiterate” to ”iliterate” and vowels from the titles (”iliterate” to ”ltrt”). This decision has been conducted after observing that the automated indexing was prone to make errors in such cases (e.g. either by omitting the repeated letter, or making a mistake in distinguishing simliar letters such as ”l” and ”i”). Then, a procedure to identify pairs and triples of matching records was conducted, with manual verification of the candidate matches in place.

Then, five criteria were tested separately and in combinations (involving 2 or more criteria needed to be fulfilled). The five criteria studied are: DOI/URL (DOIs are only provided in Scopus and MA, URLs were used for matching across GS and MA), 95%

similarity of cleansed titles, venue of publication name, at least one common author, year of publication. The algorithm then compared each of the three databases pairwise, and took into account already existing pairs. In case a record was matched across the three databases, it was stored as a triple match. Each two records compared were checked for fulfilment of each of the above-mentioned criteria.

Each of these criteria were programmed, and combinations from 2 to 5 criteria used simultaneously were automatically tested. The standard measures for comparison of methods in pattern recognition: precision, recall and F1 scores, were then calculated.

The results of the experiment, as presented in 5.3, show that the optimal performance was achieved by combining the DOI/URL criterion with title characters similarity larger than 95%.

Software used to select the 'golden set sample', cleanse the records, and match the cross-database data using each of the criteria presented above is made openly available on GitHub[2] for reproducibility.

[2] https://github.com/bpaszcza/MicrosoftAcademic_study

Chapter 5

Application of Completeness Methodology

5.1 Scope of the databases

Tables 5.1 and 5.2 present the data gathered using keyword-based queries during the data collection. We need to note that because of the methodology used, comparison between disciplines within a single database. In other words, the fact that Google Scholar dataset contained 2582 records from Natural Sciences and only 2005 from Engineering and Technology does not indicate that overall GS indexes less documents in the latter category, nor does it indicate that less articles are being published in the field of Engineering and Technology.

What the study intends to show, instead, is a comparison between databases within each of the disciplines. The methodology was designed to compare the coverage of field of Natural Sciences by Scopus, Google Scholar, and Microsoft Academic.

5.1.1 Microsoft Academic and Scopus

Number of MA records was higher than in Scopus in all six OECD fields of science, recorded number of citations to these documents was higher in five out of six disciplines (Engineering and Technology being the exception). Overall, MA returned 72.6% more documents and 12.6% more citations than Scopus. Most notable differences in favour of MA appear in Social Sciences and Humanities, where number of retrieved records in MA is more than double (Social Sciences) and almost triple (Humanities). This observation confirms earlier studies, highlighting ineffective indexing of research outputs in these disciplines by Scopus, mentioned for example by Harzing and Alkangas (2016a)). Another study conducted by Hug and Braendle (2017) also observed

TABLE 5.1: Number of documents per discipline

	Google Scholar	Microsoft Academic	Scopus
Natural Sciences	2582	2212	1493
Engineering and Technology	2005	1800	1325
Medical and Health Sciences	1651	1524	1201
Agricultural Sciences	2963	2359	1468
Social Sciences	3092	2578	1251
Humanities	2803	2745	919
TOTAL	15096	13218	7657

TABLE 5.2: Number of citations to documents found per discipline

	Google Scholar	Microsoft Academic	Scopus
Natural Sciences	186161	129601	126215
Engineering and Technology	37984	24487	25857
Medical and Health Sciences	80684	59050	55737
Agricultural Sciences	88078	59426	51126
Social Sciences	118235	59757	39679
Humanities	40158	13402	8498
TOTAL	551300	345723	307112

the lower number of MA citations records in Engineering and Technology. They, however, also observed similar result for Natural Sciences, where our study shows Scopus lagging, which may be explained by some uncertainty in the results or fast-paced development of the MA database. Our results confirm the ones of Alakangas and Harzing (Harzing and Alakangas, 2017), who also demonstrated significantly better performance of MA versus Scopus.

5.1.2 Microsoft Academic and Google Scholar

The number of cleansed records in GS is 14.2% higher than in MA, and the number of citations in GS exceeds the one recorded by Microsoft by 49.7%. However, as was denoted in 4.1.3, a third of GS papers were removed during cleansing process (compared to only 6.7% of MA's), which was not performed on papers citing the records in our sample. Hence the difference in citations may be in reality smaller due to lower quality of GS records - a hypothesis that needs further study.

Nevertheless, we need to conclude that despite improvement of the Microsoft Academic, it still lags in scope against the Google Scholar. Above mentioned study by Harzing and Alkangas (2017) also identified similar result, albeit being based on a smaller sample of research output of 145 researchers from a single academic institution. They also showed consistently higher numbers for GS when compared to Scopus, in all six OECD fields of research - which is also the case in our sample.

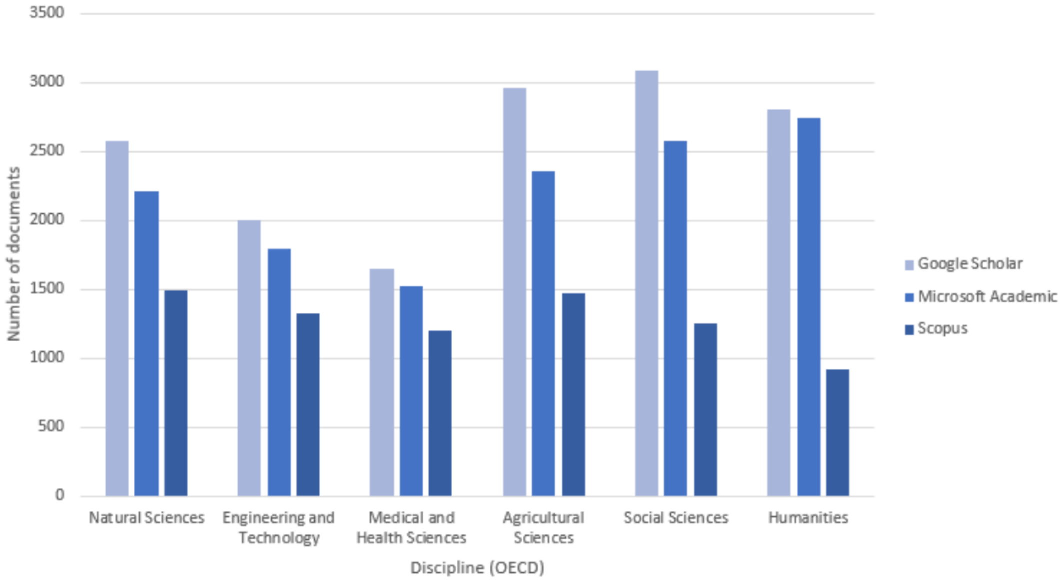


FIGURE 5.1: Number of documents found by discipline

5.1.3 Comparison of results to other studies

The results of this study have to be confirmed by comparison to other studies on MA, GS, and Scopus. They are presented in Tables 5.1, 5.2 as well as in Figures 5.1 and 5.2.

The number of documents found was 72.6% higher in MA when compared to Scopus in total, and individually higher in all of the six analysed disciplines. Citation counts were higher in five out of six fields, with the only exception being Engineering and Technology, where MA indexed 94.7% of the citations included in Scopus. However, overall MA indexed 12.6% more citations than Scopus.

Such results confirm those found by other researchers. In her original study, Harzing (2016b) has observed MA's superiority in terms of citation counts over Scopus only in some disciplines, with other comparison metrics indicating similar performance. In the study by Hug et al. (2017), it was found that in terms of citation scores of documents from a single institutional repository, MA compares on-par with or higher than Scopus in the fields of Humanities, Social Sciences, and Agricultural Sciences. This study also found that in the field of Engineering and Technology, MA's citation score lags behind Scopus, but by a very small margin. The greatest discrepancy between the results can be found in Natural Sciences, where Hug at al. found significantly fewer citations in MA than in Scopus.

It is also necessary to mention that some variance in the results is explainable by the constant development of MA in the most recent two years (Harzing and Alakangas, 2017), in terms of scope and quality, which results in a higher number of documents and in more indexed citations.

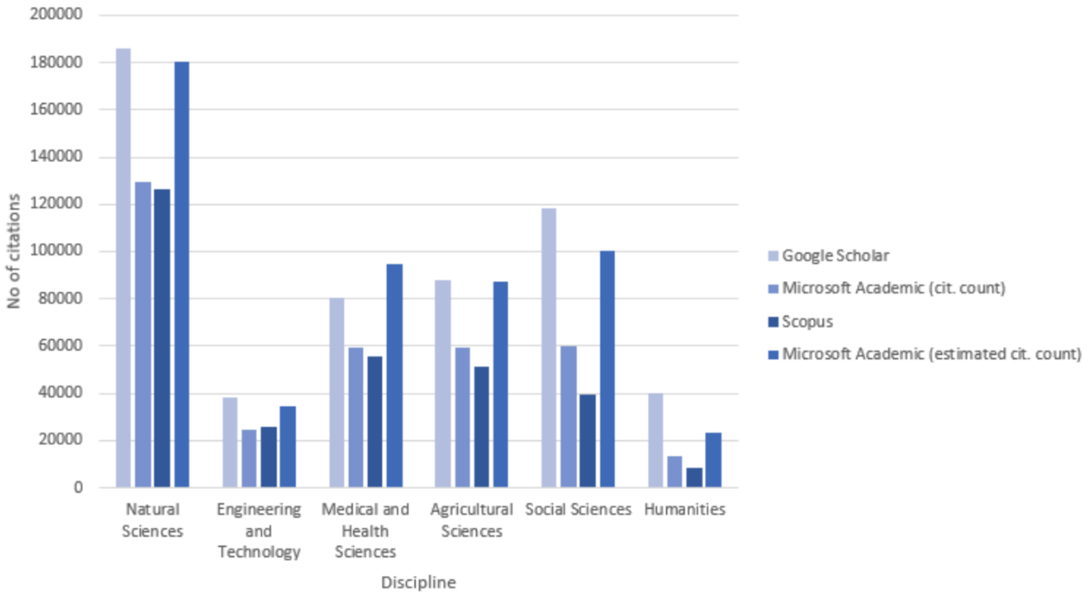


FIGURE 5.2: Number of citations to documents found by discipline

In another recent study, focusing on comparing MA with GS, Scopus, and WoS, Harzing (2017) has concluded that MA significantly outperforms Scopus in terms of number of indexed documents and citations, which seems to confirm our findings. In her study, GS outperformed MA in number of citations in all but one of the categories – Life Sciences – where the score was almost equal. In our study, this difference is confirmed, albeit to an even greater extent, as MA indexes 62.7% of citations indexed by GS. Similarity between the indexing of disciplines can be observed, with MA's poorest score being 33.4% in Humanities and Arts in our study, which is comparable to Harzing's 40.5%, for example.

However, we also need to remember that – contrary to sets of database records – citations to the documents have not been cleansed of duplicates and unidentifiable documents. As shown above, a third of the publications indexed by GS were removed in the cleansing process, compared to just 6.7% of MA's records. Hence, it should be further verified whether the observed differences in number of recorded citations hold when incorrect citing documents and links are removed from the study. The number of documents indexed (after cleansing) shows that the difference between databases can be significantly smaller, with MA indexing a number of documents that constitute 87.6% of GS's set.

TABLE 5.3: Summary of data cleansing statistics

	Google Scholar		Microsoft Academic		Scopus	
Records	23,144	100.0%	14,102	100.0%	7,903	100.0%
* Without source (URL/DOI)	7,246	31.3%	388	2.8%	0	0.0%
* Duplicates - identical DOI/URL	149	0.6%	153	1.1%	105	1.3%
* Duplicates - identical metadata	392	1.7%	190	1.3%	107	1.4%
* Duplicates - PPM clustering	261	1.1%	153	1.1%	34	0.4%
Cleansed records	15,096	65.2%	13,218	93.7%	7,657	96.9%

5.2 Data quality

The second comparison between the databases is concerned with the quality of data reviewed during the data cleansing process. Looking at Table 5.3, Scopus has demonstrated to be the most reliable database in terms of data quality (96.9%), unsurprisingly due to its method of data gathering and previous experiments. However, the rising contender - Microsoft Academic - has shown performance on only marginally lower level to Scopus (93.7% correct records), which is a pleasant surprise. Such a result, combined with significantly larger scope (13,218 records in MA, compared to 7,657 in Scopus), demonstrates that the Microsoft team has managed to remove many of the drawbacks of the original Microsoft Academic Search service.

Comparing MA to Google Scholar, one has to notice that the after-cleansing number of records is comparable, but Microsoft Academic still returns smaller number of records (87.6%). However, in terms of the quality of the data, the wider scope of metadata (especially availability of DOI, but also more than one - and quite often, a few - URLs per record) and higher overall quality leads to a significant difference (with only 65.2% of Google Scholar records passing the cleansing process).

5.3 Databases overlap

Thirdly, this study aimed to test various methods of automated cross-database overlap analysis against a manually verified 'golden sample' of matched records. The test sample was created by obtaining three sets of records from each of the three databases. The sets consisted of the first 400 alphabetically sorted records from the samples in the field of Natural Sciences.

Then, an automatic procedure with a manual review to find cross-database overlap (pairs/triples of records referring to the same research output) was performed. An algorithm proposed to link a pair of records across databases if any of the below described criteria matched:

- A matching DOI (Scopus - MA match) or URL (MA - GS)

- At least one common author in the authors' list and year of publication within ±1 year of each other (please note, Scopus metadata shows only the first author)
- The two titles were similar in more than 80% (calculated using Python's difflib.SequenceMatcher function, based on the Ratcliff-Obershelp algorithm[1]); titles have been truncated at the 150th character, as this is the maximum title length that Google Scholar provides

In the case that titles were identical, and at least one common author and publication date were found in both records, a match was created fully automatically. If, however, some of these three conditions had not been met, the proposed pair/triple was reviewed manually. Finally, after proceeding through the full set of records, another manual check was conducted to discard any duplicates or false positives by manual review of metadata of matched pairs/triples. This meticulous and time-consuming process was conducted to ensure that the test sample can be assumed to truthfully represent the overlap of test sets. A graph presenting an overview of the test sample creation process is presented in Figure 4.6.

As a result, 334 pairs and triples, representing records in the three databases that refer to the same publication, were identified.

Subsequently, a process of creating and testing an algorithm allowing automatic overlap test was initiated. The available metadata fields were tested to verify the precision and recall of matching using these criteria. Five separate metadata types were used:

- identical DOI (match between Scopus and Microsoft Academic) or URL (match between Microsoft Academic and Google Scholar)
- titles: at least 95% similarity (calculated using Python's difflib.SequenceMatcher function, based on the Ratcliff-Obershelp algorithm[2])
- the same venue of publication
- at least one common author's surname
- the same year of publication

Furthermore, multiple tests with an increasing threshold (defined as the minimal number of conditions met to form a match) were performed. The results are presented in Table 5.4.

The least reliable conditions of selection are venue of publication, year of publication, and common authors. In a trial using only a single condition, low precision (defined as

[1] https://docs.python.org/2/library/difflib.html

[2] https://docs.python.org/2/library/difflib.html

TABLE 5.4: Results of conditions for cross-database record matching (overlap analysis)

A. Threshold: 1/5 (minimal number of criteria met)

Selection criterion:	DOI/URL	Title (>95% similarity)	Venue of publication	Common Author	Year of Publication	combined (DOI/URL + Title)
precision	0.99	0.89	0.01	0.09	0.01	—
recall	0.6	0.96	0.29	0.69	0.27	—
F	0.75	0.92	0.01	0.16	0.01	—

B. Threshold: 2/5

Selection criterion:	DOI/URL	Title (>95% similarity)	Venue of publication	Common Author	Year of Publication	combined (DOI/URL + Title)
precision	1	0.96	0.12	0.34	0.06	0.96
recall	0.6	0.96	0.32	0.9	0.75	0.97
F	0.75	0.96	0.18	0.49	0.11	0.97

C. Threshold: 3/5

Selection criterion:	DOI/URL	Title (>95% similarity)	Venue of publication	Common Author	Year of Publication	combined (DOI/URL + Title)
precision	0.98	0.95	0.61	0.86	0.83	0.96
recall	0.59	0.9	0.33	0.89	0.86	0.9
F	0.74	0.92	0.43	0.87	0.85	0.93

D. Threshold: 4/5

Selection criterion:	DOI/URL	Title (>95% similarity)	Venue of publication	Common Author	Year of Publication	combined (DOI/URL + Title)
precision	0.73	0.74	0.68	0.72	0.72	0.74
recall	0.41	0.5	0.31	0.49	0.49	0.5
F	0.53	0.6	0.42	0.58	0.58	0.6

E. Threshold: 5/5

Selection criterion:	DOI/URL	Title (>95% similarity)	Venue of publication	Common Author	Year of Publication	combined (DOI/URL + Title)
precision	0.53	0.53	0.53	0.53	0.53	0.53
recall	0.18	0.18	0.18	0.18	0.18	0.18
F	0.27	0.26	0.26	0.26	0.26	0.26

the ratio of true positives to the sum of true positives and false positives) of 0.01-0.09 suggests a very high (91%-99% of all matches found) number of cross-database matches that are not in fact referring to a single publication. The recall (ratio of true positives over the sum of true positives and false negatives) of these three methods was also relatively low, ranging from 0.27-0.69, meaning that of the matches that should have been found, only 27%-69% were indeed identified.

On the other side of the spectrum, DOI/URL condition has been shown to lead to a very high precision of 99%, indicating a low number of false positives, further confirming that identification of unique records using those two identifiers is indeed very reliable. Its recall, however, has been relatively low – reaching 60%. This effect is due to the fact that despite its popularity, DOIs are recorded in only approximately 70% of the records in the sample. Furthermore, whilst MA attempts to provide a list of all URLs where a given document can be found (e.g. institutional repositories, journals, personal web pages), GS only provides one, so that in some cases, the same document can be found at different addresses by crawlers of the two databases.

Finally, matching records using their title has also proved to be a very reliable method. Here, a condition of equal or more than 95% similarity of title strings was used to take into account common record parsing errors that could remain in the sample. This method has resulted in precision of 0.89, indicating some false positives have been selected, but also recall of 0.96. Such result proves that matching using titles helps to match records without recorded or common DOI/URL.

Hence, to improve the quality and reliability of the algorithm used, a further test was performed to determine how increase in threshold (minimal number of conditions met to record a match) reflects on the precision and recall scores. As an example, the column 'Venue of publication' in Table 5.4B. indicates records matches that have the same venue of publication and met at least any one of the other four conditions (DOI/URL, title, common author, or year of publication). We could expect the precision to increase and the recall to drop with the increasing threshold. Whilst the latter part of the statement seems to be generally true as recall is indeed decreasing gradually in parts B-E – the precision in some cases is fluctuating (e.g. 'Venue of publication' column, Table 5.4D.) or decreases when the threshold is larger than 3. This can be explained by taking into account the methodology of a study, where not only pairs of matched records were created, but also triples (three records, each from one of the databases, matched). The incomplete cases, where in the test sample we had a triple, and the algorithm only found a pair (e.g. found a match between a record in Scopus and one in Google Scholar, but missed a match to a record in MA) were counted as 'false positives' due to their incompleteness. As the threshold increased, the difficulty of creating a match resulted in more incomplete pairs instead of triples, leading to an increase in the number of false positives and hence a decrease in precision.

Furthermore, an additional column was added for a combined method that requires fulfilment of both DOI/URL and title criteria and any of the remaining three possible criteria, in cases the threshold threshold was set to a number higher than 2. This was an attempt to combine the two criteria with best performance, indicated by their individual a highest F-score. Indeed, the best result has been obtained by using both criteria simultaneously and discarding the other proposed criteria, reaching an F-score of 0.96.

To conclude, this study reaffirms the notion that unique identifiers such as DOIs and to some extent URLs are indeed the best method for finding the same records in multiple data sources. However, due to the fact that some records lack a DOI and have different URLs (meaning that they have been found by the databases in different online locations), comparison of titles is recommended as the next-best method. Similarly to (Hug et al., 2017), we found that attempting to find overlap based on other metadata fields e.g. authors, year of publication, venue of publication results in significantly higher error rate.

Chapter 6

Conclusions

6.1 Overall conclusions

The proposed methodology has been shown to enable a discipline-agnostic, large scale and significantly automated in-depth comparison of scholarly communication datasets available via APIs. Therefore, it can serve as a foundation for creation of a generalised, easily reproducible method for studying and verifying the databases which became an essential element of scientific literature reviews and policymaking, among other applications.

The proposed methodology was applied to study Microsoft Academic, Google Scholar, and Scopus. The results of this exercise, when compared with other subject literature, show high degree of similarity. Such is the case, especially in terms of scope and citation coverage differences between the databases. The proposed methodology also provides more in-depth results, e.g. regarding the disciplinary coverage based on document-level sampling, which in the future should be compared to other studies.

As bibliometric studies show, comparing the indexing of different disciplines is necessary, as the coverage of disciplines by different databases varies - a feature which the proposed methodology develops. With a significantly larger number of records amassed by Microsoft Academic and Google Scholar in the field of Humanities, for example, one has to conclude that those portals should serve as much better data sources for studies in this given area. Hence, methodology proposed in the study enables detailed comparison of databases, hence helping to gather an in-depth knowledge and understanding of scholarly information sources. This study has shown that it is possible to perform large-scale comparison aiming to cover all disciplines, based on individual research outputs instead of sets of publication by an author, an institution or journal. This important feature is a necessary to truthfully study the scholarly communication landscape and its indexing by the database in question.

6.2 Comparison to other studies of Microsoft Academic

Furthermore, our results confirm with only minor differences described in Section 5.1 the recent results of Microsoft Academic analyses. Hence, this study is highlighting the new portal as an interesting source of data. Microsoft Academic is combining robust metadata structure previously characteristic of Scopus and Web of science with scope and size comparable to that of Google Scholar. Characteristically, the source's openness (especially availability via an open API) also positively distinguishes it from its competitors. It remains to be seen, however, whether Microsoft maintains such approach towards openness of the dataset in the future.

Some expected disciplinary differences were indeed identified: MA indexed publications and - to a smaller extent - citations in areas of Social Sciences and Humanities two- to three-times better than Scopus, a dataset often criticised for poor coverage of these disciplines. The citation scores were almost uniformly higher in MA than in Scopus.

On the other hand, the number of citations in Google Scholar remained significantly higher than in MA. Future studies should examine the citing documents to identify, whether they indeed are academic papers. It may be that the GS quality problems (which were also found in the data cleansing process) are inflating the GS score. It is also possible - especially in the areas of Social Sciences and Humanities - the difference may come from better indexing of books and other non-traditional research outputs by Google. Inclusion of the ”estimated citation count”, a mathematical prediction of the ”real” number of citations to a paper is an attempt to work around the problem. However, it cannot be deemed sufficient, as the reliability of such prediction method is not yet verified.

6.3 Methodology

The data gathering and cleansing process was found to be significantly simpler and less time-consuming in Microsoft Academic than in Google Scholar, due to available API and developed metadata structure. The quality of the downloaded MA data was found to be almost on-par with Scopus' data. It is a significant achievement, bearing in mind that the data quality problems were plaguing earlier Microsoft Academic Search, but also Google Scholar - a service similarly using web crawlers to index content.

The newly designed, keyword-based methodology allows researchers to compare the sources performance within disciplines, whilst not being dependent on output of a single journal, institution or a group of individuals. The manual methodology for sourcing the disciplinary keyphrases used in this paper could also be further developed, by using expert reviews (getting experts from a given discipline to provide or verify keywords) or designing an automated script, based on Natural Language Processing techniques, that could verify the keywords – for example, checking whether it can find a Wikipedia entry (Łopuszyński and Bolikowski, 2014).

Furthermore, a review of metadata useful to perform an overlap analysis of the datasets may prove a foundation for standardisation of cross-dataset records matching. Our study largely confirmed earlier results in that matter, but has supplemented the results with in-depth statistics and verified the outcomes on a much more diverse sample consisting of six types of metadata drawn from three databases.

We have found that the best result for matching records accross databases in order to compare how much they overlap is obtained by using simultaneously the two criteria when comparing two records from different datasets:

- a matching DOI or URL (or both) and
- at least 95% similarity of titles.

This method obtained an F-score (harmonic mean of precision and recall) of 0.96. The 95% similarity criteria for the title of the publications allows for some minor and relatively common errors in the ways titles are recorded in the datasets. For example some titles were observed with erroneous letters or diacritic signs, an error found especially in databases using automated parsers of online content, such as Google Scholar and Microsoft Academic.

6.4 Further work

This study can be built on in order to create an easy-to-use and standardised method enabling researchers to gain an in-depth understanding of the advantages and disadvantages of the tools they used to conduct literature review and assess impact of individual publications, institutions or countries. Such knowledge could inform the usefulness of certain datasets for studies in a specific discipline (especially meta-reviews), evaluation of researchers (e.g. of the citation counts of often underrepresented non-STEM researchers) and applications in industry.

Furthermore, a keyword-based method of understanding the characteristics of databases could be useful for studies of other types of datasets and services. Examples include studies of scope of news datasources and news aggregators, insights into coverage of search engines or the characteristics of content selection algorithms in social media.

The study is proposed to be repeated on other datasets, with a more thorough evaluation and comparison of results. Since this study has been conducted on a relatively novel source of data, Microsoft Academic, studies comparing e.g. Web of Science and Scopus could provide more information into the differences between the new methodology and results of research conducted by others.

Furthermore, the data cleansing process should be automated to a larger extent. This problem may prove to be hard to generalise, as data quality problems may be specific to individual databases. However, data cleansing remains the most time-consuming element of the proposed procedure. On the other hand, we note that it is a standard, and often the most laborious, component of data science projects in general.

References

ACRL Scholarly Communications Committee. Principles and strategies for the reform of scholarly communication. 2003. URL http://www.ala.org/acrl/publications/whitepapers/principlesstrategiess.

B. Amaro, P. A. Azrilevich, A. D. Babini, A. G. Beasley, N. Lossau, K. Mueller, K. Repanas, A. . Star, S. O. Rieger, E. Rodrigues, C. . Openaire, J. Ruttenberg, M. Viragos, and X. Zhang. Promoting Open Knowledge and Open Science: Report of the Current State of Repositories. 2015. URL https://www.coar-repositories.org/news-updates/promoting-open-science-and-open-knowledge-current-state-of-repositories/.

ASCB. San Francisco Declaration on Research Assessment. *Annual Meeting of The American Society for Cell Biology*, pages 1–10, 2012. ISSN 2046-6390. . URL papers3://publication/uuid/1AEB2F37-D0EA-4653-9E41-FBA2CD42E70B.

J. Bar-Ilan. Which h-index? A comparison of WoS, Scopus and Google Scholar. *Scientometrics*, 74(2):257–271, 2008. ISSN 01389130. . URL http://www.zalf.de/de/institute{_}einrichtungen/bib/Documents/BibliometrischeIndizes/Bar-Ilan{_}2008{_}h-factor.pdf.

S. Bartling and S. Friesike. *Opening Science*. Springer Open, 2014. ISBN 978-3-319-00025-1. . URL http://book.openingscience.org/.

T. Berners-Lee. Information Management: A Proposal. 2016, 1989. URL https://www.w3.org/History/1989/proposal.html.

C. L. Borgman. *Scholarship in the digital age: Information, infrastructure, and the Internet*. MIT Press, 2010. ISBN 9780262026192. URL https://mitpress.mit.edu/books/scholarship-digital-age.

L. Bornmann. Validity of altmetrics data for measuring societal impact: A study using data from Altmetric and F1000Prime. *Journal of Informetrics*, 8(4):935–950, 2014. ISSN 17511577. . URL http://dx.doi.org/10.1016/j.joi.2014.09.007.

L. Bornmann. Usefulness of altmetrics for measuring the broader impact of research. *Aslib Journal of Information Management*, 67(3):305–319, 2015a. ISSN 2050-3806. . URL http://www.emeraldinsight.com/doi/10.1108/AJIM-09-2014-0115.

L. Bornmann. Alternative metrics in scientometrics: A meta-analysis of research into three altmetrics. *Scientometrics*, 103(3):1123–1144, 2015b. ISSN 01389130. . URL http://dx.doi.org/10.1007/s11192-015-1565-y.

L. Bornmann, W. Marx, H. Schier, E. Rahm, A. Thor, and H. D. Daniel. Convergent validity of bibliometric Google Scholar data in the field of Chemistry. *Journal of Informetrics*, 3(1):27–35, 2009. ISSN 17511577. . URL https://www.sciencedirect.com/science/article/abs/pii/S1751157708000667.

J. Bosman and B. Kramer. Innovations in scholarly communication - data of the global 2015-2016 survey [Data set]. *Zenodo*, 2016. . URL http://doi.org/10.5281/zenodo.49583.

T. Brody. *Evaluating research impact through open access to scholarly communication*. PhD thesis, University of Southampton, 2006. URL http://eprints.soton.ac.uk/id/eprint/263313.

J. F. Burnham. Scopus database: a review. *Biomedical digital libraries*, 3:1, 2006. ISSN 1742-5581. . URL http://www.ncbi.nlm.nih.gov/pubmed/16522216.

C. Caragea, J. Wu, A. Ciobanu, K. Williams, and J. Fern. CiteSeerX: A Scholarly Big Dataset. *Advances in Information Retrieval*, 8416:311–322, 2014. . URL www.cse.unt.edu/{~}ccaragea/papers/ecir14.pdf.

A. Cavacini. What is the best database for computer science journal articles? *Scientometrics*, 102(3):2059–2071, 2015. ISSN 15882861. . URL https://link.springer.com/article/10.1007/s11192-014-1506-1.

D. de Solla Price. *Little Science, Big Science ...and Beyond*. Columbia University Press, 1983. ISBN 0231049560. URL http://www.garfield.library.upenn.edu/lilscibi.html.

D. J. de Solla Price. *Little Science, Big Science*, volume 5. Columbia University Press, 1963. ISBN 978-0231085625. URL https://www.degruyter.com/view/product/522665?format=EBOK.

J. C. de Winter, A. A. Zadpoor, and D. Dodou. The expansion of Google Scholar versus Web of Science: A longitudinal study. *Scientometrics*, 98(2):1547–1565, 2014. ISSN 01389130. . URL https://link.springer.com/article/10.1007/s11192-013-1089-2.

J. C. Fagan. An Evidence-Based Review of Academic Web Search Engines, 2014-2016: Implications for librarians' practice and research agenda. *Information Technology and*

Libraries, 36(2):7–47, 2017. ISSN 2163-5226. . URL https://ejournals.bc.edu/ojs/index.php/ital/article/view/9718.

M. E. Falagas, E. I. Pitsouni, G. A. Malietzis, and G. Pappas. Comparison of PubMed, Scopus, Web of Science , and Google Scholar: strengths and weaknesses. *The FASEB Journal*, 22(2):338–342, 2008. . URL http://www.ncbi.nlm.nih.gov/pubmed/17884971.

K. Flegal, K. Bk, H. Orpana, and B. Graubard. Association of All-Cause Mortality With Overweight and Obesity Using Standard Body Mass Index Categories: A Systematic Review and Meta-analysis. *Jama*, 309(1):71–82, 2013. ISSN 1538-3598. . URL http://jama.jamanetwork.com/article.aspx?articleid=1555137.

E. Garfield. Citation indexes for science: a new dimension in documentatio through association of ideas. *Science*, 122(July):108–11, 1955. URL http://science.sciencemag.org/content/122/3159/108.

E. Garfield. Citation Indexing for Studying Science. *Nature*, 227:669–671, 1970. ISSN 0028-0836. . URL http://garfield.library.upenn.edu/essays/V1p133y1962-73.pdf.

E. Garfield. Journal impact factor: a brief review. *Canadian Medical Association Journal*, 1999. URL https://www.cmaj.ca/content/161/8/979.

E. Garfield and I. H. Sher. New factors in the evaluation of scientific literature through citation indexing. *American Documentation*, 14(3):195–201, 1963. URL http://garfield.library.upenn.edu/essays/v6p492y1983.pdf.

T. Greenhalgh, G. Robert, F. MacFarlane, P. Bate, O. Kyriakidou, and R. Peacock. Storylines of research in diffusion of innovation: A meta-narrative approach to systematic review. *Social Science and Medicine*, 61(2):417–430, 2005. ISSN 02779536. . URL https://www.ncbi.nlm.nih.gov/pubmed/15893056.

C. Gumpenberger, W. Glänzel, and J. Gorraiz. The ecstasy and the agony of the altmetric score. *Scientometrics*, 2016. ISSN 0138-9130. . URL http://link.springer.com/10.1007/s11192-016-1991-5.

M. Gusenbauer. Google scholar to overshadow them all? comparing the sizes of 12 academic search engines and bibliographic databases. *Scientometrics*, 118(1):177–214, 2019. URL https://link.springer.com/article/10.1007%2Fs11192-018-2958-5.

S. Harnad. Publicly retrievable FTP archives for esoteric science and scholarship: The Subversive Proposal. 1995. URL https://groups.google.com/forum/?hl=en#!topic/bit.listserv.vpiej-l/BoKENhKO_00.

A.-W. Harzing. Microsoft Academic (Search): a Phoenix arisen from the ashes? *Scientometrics*, 108(3):1637–1647, 2016. ISSN 15882861. . URL http://www.harzing.com/blog/2016/06/microsoft-academic-search-a-phoenix-arisen-from-the-ashes.

A.-W. Harzing. Publish or perish [computer software]. Available from 2007. URL https://harzing.com/resources/publish-or-perish.

A.-W. Harzing and S. Alakangas. Google Scholar, Scopus and the Web of Science: a longitudinal and cross-disciplinary comparison. *Scientometrics*, 106(2):787–804, 2016a. ISSN 15882861. . URL http://eprints.mdx.ac.uk/18511/.

A.-W. Harzing and S. Alakangas. Microsoft Academic: is the phoenix getting wings? *Scientometrics*, 110(1):1–13, 2016b. ISSN 15882861. . URL https://link.springer.com/article/10.1007/s11192-016-2185-x.

A.-W. Harzing and S. Alakangas. Microsoft Academic is one year old: the Phoenix is ready to leave the nest. *Scientometrics*, 112(3):1887–1894, 2017. ISSN 15882861. . URL https://link.springer.com/article/10.1007%2Fs11192-017-2454-3.

A.-W. Harzing and R. van der Wal. Google Scholar as a new source for citation analysis. *Ethics in Science and Environmental Politics*, 8:61–73, 2008. ISSN 18635415. . URL http://www.int-res.com/articles/esep2008/8/e008pp5.pdf.

S.-U. Hassan, A. Visvizi, and H. Waheed. The 'who'and the 'what' in international migration research: data-driven analysis of Scopus-indexed scientific literature. *Behaviour & Information Technology*, 38(9):924–939, 2019. URL https://www.tandfonline.com/doi/abs/10.1080/0144929X.2019.1583282.

R. Haunschild, S. E. Hug, M. P. Brändle, and L. Bornmann. The number of linked references of publications in Microsoft Academic in comparison with the Web of Science. *Scientometrics*, pages 1–4, 2017. ISSN 15882861. .

D. Herrmannova and P. Knoth. An Analysis of the Microsoft Academic Graph. *D-Lib Magazine*, 22(9-10):1–16, 2016. ISSN 10829873. .

D. Hicks and P. Wouters. The Leiden Manifesto for research metrics. *Nature*, 520(7548):9–11, 2015. ISSN 0028-0836. . URL https://www.researchgate.net/publication/275335177_The_Leiden_Manifesto_for_research_metrics.

J. Higgins and S. Green. Cochrane Handbook for Systematic Reviews of Interventions, 2008. URL http://handbook.cochrane.org/.

J. E. Hirsch. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of USA*, 102(46):16569–16572, 2005. ISSN 0027-8424. . URL http://www.ncbi.nlm.nih.gov/pubmed/16275915.

S. E. Hug and M. P. Brändle. The coverage of Microsoft Academic: Analyzing the publication output of a university. *Scientometrics*, 113(3):1551–1571, 2017. URL https://arxiv.org/pdf/1703.05539.

S. E. Hug, M. Ochsner, and M. P. Brändle. Citation analysis with Microsoft Academic. *Scientometrics*, 111(1):371–378, 2017. ISSN 15882861. . URL https://arxiv.org/abs/1609.05354.

N. Jacobs. *Open Access: Key strategic, technical and economic aspects*. Elsevier, 2006. URL https://www.elsevier.com/books/open-access/jacobs/978-1-84334-203-8.

P. Jacsó. Metadata mega mess in Google Scholar. *Online Information Review*, 34(1):175–191, 2010. ISSN 1468-4527. . URL http://www.emeraldinsight.com/doi/10.1108/14684521011024191.

P. Jacsó. The pros and cons of Microsoft Academic Search from a bibliometric perspective. *Online Information Review*, 35(6):983–997, 2011. ISSN 1468-4527. . URL https://www.emerald.com/insight/content/doi/10.1108/14684521111210788/full/html.

A. Jinha. Article 50 million: An estimate of the number of scholarly articles in existence. *Learned Publishing*, 23(3):258–263, 2010. ISSN 09531513. . URL http://onlinelibrary.wiley.com/doi/10.1087/20100308/abstract.

JISC. Researchers of Tomorrow: the research behaviour of Generation Y doctoral students. Technical report, 2012. URL https://www.fosteropenscience.eu/index.php/content/researchers-tomorrow-research-behaviour-generation-y-doctoral-students.

M. Khabsa, Z. Wu, and C. L. Giles. Towards Better Understanding of Academic Search. *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries - JCDL '16*, pages 111–114, 2016. ISSN 15525996. . URL http://dl.acm.org/citation.cfm?doid=2910896.2910922.

K. S. Khan, R. Kunz, J. Kleijnen, and G. Antes. Five steps to conducting a systematic review. *Journal of the Royal Society of Medicine*, 96(3):118–121, 2003. ISSN 0141-0768. . URL https://www.ncbi.nlm.nih.gov/pubmed/12612111.

A. Kodakateri, S. Gauch, and J. Eno. Conceptual Recommender System for CiteSeer x. *RecSys CBRecSys 2015: New Trends on Content-Based Recommender Systems*, pages 50–53, 2015. ISSN 00349887. .

K. Kousha and M. Thelwall. Sources of Google Scholar citations outside the Science Citation Index: A comparison between four science disciplines. *Scientometrics*, 74 (2):273–294, 2008. URL https://ideas.repec.org/a/spr/scient/v74y2008i2d10.1007_s11192-008-0217-x.html.

T. S. Kuhn. *The Structure of Scientific Revolutions*. University of Chicago Press, Chicago, 1st edition, 1970. ISBN 0-226-45808-3. .

E. Landhuis. Scientific literature: Information overload. *Nature*, 535:457, 2016. URL https://www.nature.com/nature/journal/v535/n7612/full/nj7612-457a.html.

M. Levine-Clark and E. L. Gil. A comparative citation analysis of Web of Science, Scopus, and Google Scholar. *Journal of Business and Finance Librarianship*, 14(1):32–46, 2009. ISSN 08963568. . URL https://www.tandfonline.com/doi/abs/10.1080/08963560802176348.

L. Leydesdorff and I. Rafols. A global map of science based on the ISI subject categories. *Journal of the American Society for Information Science and Technology*, 60(2):348–362, 2009. ISSN 15322882. . URL http://www.leydesdorff.net/map06/texts/map06.pdf.

L. Leydesdorff, L. Bornmann, W. Marx, and S. Milojević. Referenced Publication Years Spectroscopy applied to iMetrics: Scientometrics, Journal of Informetrics, and a relevant subset of JASIST. *Journal of Informetrics*, 8(1):162–174, 2014. URL https://www.sciencedirect.com/science/article/pii/S1751157713001077.

A. Liberati, D. G. Altman, J. Tetzlaff, C. Mulrow, P. C. Gøtzsche, J. P. A. Ioannidis, M. Clarke, P. J. Devereaux, J. Kleijnen, and D. Moher. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *Journal of clinical epidemiology*, 62(10): e1–34, 2009. ISSN 18785921. . URL https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1000100.

A. H. Lichtenstein, E. A. Yetley, and J. Lau. Application of systematic review methodology to the field of nutrition. *The Journal of nutrition*, 138(12):2297–306, 2008. ISSN 1541-6100. . URL http://www.ncbi.nlm.nih.gov/pubmed/19022948{%}5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3415860.

M. Łopuszyński and Ł. Bolikowski. Towards robust tags for scientific publications from natural language processing tools and Wikipedia. *International Journal on Digital Libraries*, oct 2014. ISSN 1432-5012. . URL http://link.springer.com/10.1007/s00799-014-0132-0.

A. J. Lotka. The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences*, 16(12):317–323, 1926. URL https://www.worldcat.org/title/frequency-distribution-of-scientific-productivity/oclc/51555046.

M. Mabe, D. Price, and M. Cole. Gold or green: which is the best shade of open access?, 2012. URL https://www.timeshighereducation.com/news/gold-or-green-which-is-the-best-shade-of-open-access/420454.article.

M. P. Major, P. W. Major, and C. Flores-Mir. Benchmarking of reported search and selection methods of systematic reviews by dental speciality. *Evidence-based dentistry*, 8(3):66–70, 2007. ISSN 1462-0049. . URL http://www.nature.com/articles/doi:10.1038%2Fsj.ebd.6400504.

A. Martín-Martín, M. Thelwall, E. Orduna-Malea, and E. D. López-Cózar. Google scholar, microsoft academic, scopus, dimensions, web of science, and opencitations' coci: a multidisciplinary comparison of coverage via citations. *Scientometrics*, 126(1):871–906, 2021.

L. I. Meho and K. Yang. Impact of data sources on citation counts and rankings of LIS faculty: Web of Science versus Scopus and Google Scholar. *Journal of the American Society for Information Science and Technology*, 58(13):2105–2125, 2007. ISSN 15322882. . URL http://eprints.rclis.org/8876/1/meho-yang-03.pdf.

S. Mikki. Comparing Google Scholar and ISI Web of Science for earth sciences. *Scientometrics*, 82(2):321–331, 2010. ISSN 01389130. .

J. Mingers and L. Leydesdorff. A Review of Theory and Practice in Scientometrics. *European Journal of Operational Research*, 241(1):1–19, 2015. URL http://arxiv.org/vc/arxiv/papers/1501/1501.05462v2.pdf.

S. Mnookin. *The panic virus: The true story behind the vaccine-autism controversy*. Simon and Schuster, 2012. URL https://www.simonandschuster.com/books/The-Panic-Virus/Seth-Mnookin/9781439158654.

H. F. Moed. *Citation Analysis in Research Evaluation*, volume 9. 2005. ISBN 1402037139. . URL http://link.springer.com/book/10.1007/1-4020-3714-7/page/1.

H. F. Moed and M. Visser. Appraisal of Citation Data Sources: A report to HEFCE by the Centre for Science and Technology Studies, Leiden University. Technical Report September, Centre for Sciene and Technology Studies, Leiden University, 2008. URL http://www.hefce.ac.uk/pubs/rdreports/2008/rd17{_}08/.

D. Moher, L. Shamseer, M. Clarke, D. Ghersi, A. Liberati, M. Petticrew, P. Shekelle, L. L. A. Stewart, H. Bastian, P. Glasziou, I. Chalmers, A. Chan, A. Hróbjartsson, M. Haahr, P. P. Gøtzsche, D. Altman, J. Kirkham, D. Altman, P. Williamson, J. Kirkham, K. Dwan, D. Altman, C. Gamble, S. Dodd, R. Smyth, P. Williamson, K. Dwan, C. Gamble, P. Williamson, J. Kirkham, S. Norris, H. Holmer, L. Ogden, R. Fu, A. Abou-Setta, M. Viswanathan, M. McPheeters, B. Ma, J. Guo, G. Qi, H. Li, J. Peng, Y. Zhang, Y. Ding, K. Yang, D. Moher, J. Tetzlaff, A. Tricco, M. Sampson, D. Altman, A. Liberati, D. Altman, J. Tetzlaff, C. Mulrow, P. Gotzsche, J. Ioannidis, M. Clarke, P. Devereaux, J. Kleijnen, D. Moher, D. Moher, A. Liberati, J. Tetzlaff, D. Altman, S. Straus, D. Moher, D. Moher, A. Booth, L. L. A. Stewart, A. Booth, M. Clarke, D. Ghersi, D. Moher, M. Petticrew, L. L. A. Stewart, A. Booth, M. Clarke,

G. Dooley, D. Ghersi, D. Moher, M. Petticrew, L. L. A. Stewart, L. Turner, L. Shamseer, D. Altman, K. Schulz, D. Moher, N. Smidt, A. Rutjes, D. V. der Windt, R. Ostelo, P. Bossuyt, J. Reitsma, L. Bouter, H. de Vet, S. Prady, S. Richmond, V. Morton, H. MacPherson, H. Williams, S. Green, J. Higgins, P. Alderson, M. Clarke, C. Mulrow, A. Oxman, A. Chan, J. Tetzlaff, D. Altman, A. Laupacis, P. P. Gøtzsche, K. Krleža-Jerić, A. Hróbjartsson, H. Mann, K. Dickersin, J. Berlin, C. Doré, W. Parulekar, W. Summerskill, T. Groves, K. Schulz, H. Sox, F. Rockhold, D. Rennie, D. Moher, E. Antman, J. Lau, B. Kupelnick, F. Mosteller, T. Chalmers, A. Oxman, G. Guyatt, D. Moher, K. Schulz, I. Simera, D. Altman, A. Booth, M. Clarke, D. Ghersi, D. Moher, M. Petticrew, L. L. A. Stewart, A. Chan, J. Tetzlaff, P. P. Gøtzsche, D. Altman, H. Mann, J. Berlin, K. Dickersin, A. Hróbjartsson, K. Schulz, W. Paruleka, K. Krleža-Jerić, A. Laupaucis, D. Moher, L. Shamseer, D. Moher, M. Clarke, D. Ghersi, A. Liberati, M. Petticrew, P. Shekelle, L. L. A. Stewart, A. Stevens, L. Shamseer, E. Weinstein, F. Yazdi, L. Turner, J. Thielman, D. Altman, A. Hirst, J. Hoey, A. Palepu, K. Schulz, D. Moher, S. Hopewell, D. Altman, D. Moher, K. Schulz, A. Hirst, D. Altman, E. Mills, P. Wu, J. Gagnier, D. Heels-Ansdell, V. Montori, P. Craig, P. Dieppe, S. Macintyre, S. Michie, I. Nazareth, M. Petticrew, P. Davies, A. Walker, J. Grimshaw, B. Carlsen, C. Glenton, C. Pope, K. Dwan, D. Altman, L. Cresswell, M. Blundell, C. Gamble, P. Williamson, D. Moher, L. L. A. Stewart, and P. Shekelle. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Systematic Reviews*, 4(1):1, 2015. ISSN 2046-4053. . URL http://www.systematicreviewsjournal.com/content/4/1/1.

J. Niyibizi, N. Zanré, M.-H. Mayrand, and H. Trottier. The association between adverse pregnancy outcomes and maternal human papillomavirus infection: a systematic review protocol. *Systematic reviews*, 6(1):53, 2017. ISSN 2046-4053. . URL http://systematicreviewsjournal.biomedcentral.com/articles/10.1186/s13643-017-0443-5{%}0Ahttp://www.ncbi.nlm.nih.gov/pubmed/28284227{%}0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5346269.

E. Orduña-Malea, A. Martín-Martín, J. M. Ayllon, and E. Delgado López-Cózar. The silent fading of an academic search engine: the case of Microsoft Academic Search. *Online Information Review*, 38(7):936–953, 2014. ISSN 1468-4527. . URL http://arxiv.org/abs/1404.7045.

E. Orduna-Malea, J. M. Ayllon, A. Martin-Martin, and E. Delgado Lipez-Cozar. Methods for estimating the size of Google Scholar. *Scientometrics*, 104(3):931–949, 2015. ISSN 01389130. . URL http://arxiv.org/abs/1506.03009.

Organisation for Economic Co-operation and Development (OECD). Revised field of science and technology (fos) classification in the frascati manual. 2007. URL https://www.oecd.org/science/inno/38235147.pdf.

J. L. Ortega. *Academic Search Engines: A Quantitative Outlook*. 2014. ISBN 9781780634722. . URL https://www.elsevier.com/books/academic-search-engines/ortega/978-1-84334-791-0.

S. Parekh-Bhurke, C. S. Kwok, C. Pang, L. Hooper, Y. K. Loke, J. J. Ryder, A. J. Sutton, C. B. Hing, I. Harvey, and F. Song. Uptake of methods to deal with publication bias in systematic reviews has increased over time, but there is still much scope for improvement. *Journal of Clinical Epidemiology*, 64(4):349 – 357, 2011. ISSN 0895-4356. . URL http://www.sciencedirect.com/science/article/pii/S0895435610001976.

B. Paszcza. Comparison of the Microsoft Academic Graph with other scholarly citation databases. Master's thesis, University of Southampton, 2016. URL http://eprints.soton.ac.uk/id/eprint/408647.

D. Pieper and F. Summann. Bielefeld Academic Search Engine (BASE): An end-user oriented institutional repository search service. *Library Hi Tech*, 24(4):614–619, 2006. URL http://eprints.rclis.org/9207/1/pieper_summann_final_web.pdf.

H. Piwowar, J. Priem, V. Larivière, J. P. Alperin, L. Matthias, B. Norlander, A. Farley, J. West, and S. Haustein. The state of oa: a large-scale analysis of the prevalence and impact of open access articles. *PeerJ*, 6:e4375, 2018. URL https://peerj.com/articles/4375/.

J. Priem, D. Taraborelli, P. Groth, and C. Neylon. Altmetrics: A manifesto, 2010. URL http://altmetrics.org/manifesto/.

A. A. M. Prins, R. Costas, T. N. van Leeuwen, and P. F. Wouters. Using Google Scholar in research evaluation of humanities and social science programs: A comparison with Web of Science data. *Research Evaluation*, (February), 2016. ISSN 0958-2029, 1471-5449. . URL http://rev.oxfordjournals.org/content/early/2016/02/02/reseval.rvv049.

M. Reyhani Hamedani, S.-W. Kim, and D.-J. Kim. SimCC: A novel method to consider both content and citations for computing similarity of scientific papers. *Information Sciences*, 334-335:273–292, 2016. ISSN 00200255. . URL http://www.sciencedirect.com/science/article/pii/S0020025515008828.

P. O. Seglen. Why the impact factor of journals should not be used for evaluating research. *British Medical Journal*, 314(7079):497, 1997. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2126010/.

A. Sinha, Z. Shen, Y. Song, H. Ma, D. Eide, B.-j. P. Hsu, and K. Wang. An Overview of Microsoft Academic Service (MAS) and Applications. *Proceedings of the 24th International Conference on World Wide Web Companion (WWW 2015 Companion)*, pages 243–246, 2015. . URL http://research.microsoft.com/apps/pubs/default.aspx?id=246609.

E. Sponsler and E. F. Van de Velde. Eprints.org software: A review. 2001. URL https://core.ac.uk/download/pdf/4890905.pdf.

P. Sud and M. Thelwall. Evaluating altmetrics. *Scientometrics*, 98(2):1131–1143, 2014. URL https://link.springer.com/article/10.1007%2Fs11192-013-1117-2.

J. Tang. AMiner. *Proceedings of the 25th International Conference Companion on World Wide Web - WWW '16 Companion*, pages 373–373, 2016. . URL http://dl.acm.org/citation.cfm?doid=2872518.2890513.

J. Tang, J. Zhang, D. Zhang, L. Yao, C. Zhu, and J. Li. ArnetMiner: An expertise oriented search system for web community. *CEUR Workshop Proceedings*, 295:990, 2007. ISSN 16130073. . URL http://portal.acm.org/citation.cfm?id=1402008.

M. Thelwall. Microsoft Academic: A multidisciplinary comparison of citation counts with Scopus and Mendeley for 29 journals. *Journal of Informetrics*, 11(4):1201–1212, 2017a. ISSN 17511577. . URL http://linkinghub.elsevier.com/retrieve/pii/S1751157717302900.

M. Thelwall. Does Microsoft Academic find early citations? *Scientometrics*, (August): 1–10, 2017b. ISSN 15882861. . URL http://www.scit.wlv.ac.uk/~cm1993/papers/DoesMicrosoftAcademicFindEarlyCitations_Preprint.pdf.

M. Thelwall. Microsoft Academic automatic document searches: Accuracy for journal articles and suitability for citation analysis. *Journal of Informetrics*, 12(1):1–9, 2018. ISSN 17511577. . URL http://linkinghub.elsevier.com/retrieve/pii/S1751157717303346.

D. Tranfield, D. Denyer, and P. Smart. Towards a methodology for developing evidence-informed management knowledge by means of systematic review. *British Journal of Management*, 14:207–222, 2003. ISSN 1045-3172. . URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1467-8551.00375.

N. J. van Eck and L. Waltman. Accuracy of citation data in Web of Science and Scopus. *arXiv preprint arXiv:1906.07011*, 2019. URL https://arxiv.org/pdf/1906.07011.pdf.

E. S. Vieira and J. A. Gomes. A comparison of Scopus and Web of Science for a typical university. *Scientometrics*, 81(2):587–600, 2009. ISSN 01389130. . URL https://ideas.repec.org/a/spr/scient/v81y2009i2d10.1007_s11192-009-2178-0.html.

M. Visser, N. J. van Eck, and L. Waltman. Large-scale comparison of bibliographic data sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic. *Quantitative Science Studies*, 2(1):20–41, 04 2021. ISSN 2641-3337. . URL https://doi.org/10.1162/qss_a_00112.

W. H. Walters. Google Scholar coverage of a multidisciplinary field. *Information Processing and Management*, 43(4):1121–1132, 2007. ISSN 03064573. . URL https://www.sciencedirect.com/science/article/pii/S0306457306001439.

L. Waltman. A review of the literature on citation impact indicators. *Journal of Informetrics*, 10(2):365–391, 2016a. ISSN 18755879. . URL http://dx.doi.org/10.1016/j.joi.2016.02.007.

L. Waltman. Special section on size-independent indicators in citation analysis. *Journal of Informetrics*, 10(2):645, 2016b. ISSN 17511577. . URL http://dx.doi.org/10.1016/j.joi.2016.04.001http://linkinghub.elsevier.com/retrieve/pii/S1751157716300864.

L. Waltman and V. Larivière. Special issue on bibliographic data sources, 2020.

K. Wang, Z. Shen, C.-Y. Huang, C.-H. Wu, D. Eide, Y. Dong, J. Qian, A. Kanakia, A. Chen, and R. Rogahn. A review of microsoft academic services for science of science studies. *Frontiers in Big Data*, 2:45, 2019. URL https://www.frontiersin.org/articles/10.3389/fdata.2019.00045/full.

R. Whalen, Y. Huang, A. Sawant, B. Uzzi, and N. Contractor. Natural Language Processing, Article Content & Bibliometrics: Predicting High Impact Science. In *ASCW'15 Workshop at Web Science 2015*, pages 6–8, 2015. URL http://ascw.know-center.tugraz.at/wp-content/uploads/2015/05/ASCW15_whalen-etal-nlp-article-content-and-bibliometrics.pdf.

J. Wilsdon. The Metric Tide: Report of the Independent Review of the Role of Metrics in Research Assesment and Management. Technical report, HEFCE, 2015. URL doi:10.13140/RG.2.1.4929.1363.

Y. Xia, P. Yang, Y. Sun, Y. Wu, B. Mayers, B. Gates, Y. Yin, F. Kim, and H. Yan. One-dimensional Nanostructures: synthesis, characterization, and applications. *Advanced materials*, 15(5):353–389, 2003. URL https://www.onlinelibrary.wiley.com/doi/abs/10.1002/adma.200390087.

A. Zuccala and R. Cornacchia. Data matching, integration, and interoperability for a metric assessment of monographs. *Scientometrics*, 108(465), 2016. ISSN 0138-9130. . URL http://link.springer.com/10.1007/s11192-016-1911-8.