UNIVERSITY OF SOUTHAMPTON

FACULTY OF PHYSICAL AND APPLIED SCIENCES

School of Electronics and Computer Science

Web Science Doctoral Training Centre

COMPARISON OF MICROSOFT ACADEMIC GRAPH WITH OTHER SCHOLARLY CITATION DATABASES

by

Bartosz Paszcza

Thesis for the degree of Masters of Science

September 2016

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF PHYSICAL AND APPLIED SCIENCES

School of Electronics and Computer Science

Web Science Doctoral Training Centre

Thesis for the degree of Masters of Science

COMPARISON OF MICROSOFT ACADEMIC GRAPH WITH OTHER SCHOLARLY CITATION DATABASES

Bartosz Paszcza

The project aims to study the Microsoft Academic Graph, a scholarly citation database, by comparison with three competitors in the field: Web of Science, Scopus, and Google Scholar. Openness, transparency of data gathering and processing, and completeness of data including the global unique identifiers has been researched in each of the four datasets. The analysis has been conducted using a set of 75 institutional affiliations, 6 randomly selected authors from the and 639 documents published by these authors. The coverage of total research output in MAG of the six selected authors had reached 76.0%, hence being on-par with coverage of Google Scholar (76.2%) and significantly better than that of Scopus (66.5%) and Web of Science (58.8%). The overall results indicate that Microsoft Academic Graph can be an interesting source of information for bibliometric or scientometric analysis. However, no definite conclusions regarding the scope of MAG can be drawn due to the small size of the sample. Furthermore, problems with affiliation and author disambiguation in MAG have been highlighted. Finally, studies focusing on the disciplinary coverage of the datasets in greater detail are proposed.

Table of Contents

List of Tables

List of Figures

List of Accompanying Materials

1. *Papers_by_affiliation_and_year.xslx* - spreadsheet documenting the data used in Sections 4.3.1 and 4.3.2
2. *Authors_total.xslx* – spreadsheet documenting data used in Sections 4.3.3 and 4.3.4

DECLARATION OF AUTHORSHIP

I, Bartosz Paszcza

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

"Comparison of Microsoft Academic Graph with other scholarly citation databases"

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

Signed: (-) Bartosz Paszcza .....

Date: 01/09/2016.....

Acknowledgements

I would like to thank supervisors overseeing this project: Leslie Carr, Jeremy Frey, and Stevan Harnad, for their continuous support and patient explanations during meetings, which greatly contributed to the outcome of the project. At all the stages: planning the research, designing methods, and reviewing the outcomes I was lucky to be able to receive their guidance and comments.

Finishing my thesis is as much of my accomplishment as it is of Dorota and Marek, my parents, and sister Agnieszka. Although I still struggle to explain to them what Web Science is about, without their help in the last (and only) twenty-three years of my life, I would not be in a position to even start this Master's project, not even mentioning finishing it.

Last, but not least, comes a large group of people contributed to this project indirectly – many of them probably did not even notice their contribution. Some of them provided the so-called social support, enabling me to enjoy those three months. Some motivated me to clarify what I am doing by asking irritating questions. Finally, some of them were taking over my extracurricular responsibilities in the crucial times I had to focus on the dissertation fully. This was the case of (take a deep breath): Mikołaj Buszko, Paweł Grzegorczyk, Piotr Kaszczyszyn, Ola Królik, Rafał and Asia Mostowy, Jacek Partridge, Jakub Słoń, Alek Smoczyński, all those whose company I enjoyed in Kraków and Southampton, and on goes the list. All I can say is: sorry for all grumbling!

Definitions and Abbreviations

API - Application Programming Interface

DOI - Digital Object Identifier

GS – Google Scholar

ISSN - International Serial Standard Number

MA - Microsoft Academic

MAG - Microsoft Academic Graph

UKPRN - UK Provider Reference Number

WoS - Web of Science

Chapter 1: Introduction

Throughout the last century, science has undergone rapid growth regarding a number of researchers, costs of conducting experiments, and resulting scientific output. Iconic science historian and one of the fathers of scientometrics, Derek de Solla Price (1962), estimated that "80% to 90% of scientists who ever lived are alive now". Public and private funds dedicated to research are – at least on average – constantly increasing (Stephan 2012). A symbol of the growth of costs of advance in human knowledge is possibly the Large Hadron Collider in CERN research laboratory. The costs of construction reached 13.25 billion dollars and the paper announcing a remarkable discovery made using the equipment in 2012, experimental confirmation of the existence of the Higgs boson, has a list of 5,154 authors (Aad et al. 2012). Notably, the scientific content of the paper and the list of authors occupy nearly an equal number of pages in the article.

The exponential growth in numbers of researchers and matching increase of science budgets pose new opportunities, but also create demands regarding the management. Effective allocation of resources – be it human or financial – is certainly one of the significant challenges. No surprise that a search for methods helping to assess the quality of scientist or institution work has been a continuous goal of science policy. In addition to an expert review, quantitative measures based on citation counts have been used in many cases, but the search for more robust and reliable methods continues (Wilsdon 2015; Mingers & Leydesdorff 2015). Another aim is to amend the existing scholarly communication system based on peer-reviewed articles published in journals in order to be able to facilitate effective knowledge transfer between the growing number of researchers (Byrnes et al. 2015).

In the meantime, the rise of the World Wide Web as a tool for knowledge exchange has been transforming scholarly communication since 1990. It comes as no surprise, as Berners-Lee idea for the WWW born in the above mentioned CERN laboratory had exactly that purpose. In his words, the system was "developed to be a pool of human knowledge" (Berners-Lee et al. 1994). The transformative power of the move from paper to a digital, online form of publication can be compared to the so-called first scientific revolution: the creation of the first open knowledge exchange system in the form of journals around 1665 (Jinha 2010).

The move to publication online does not yet, however, make full use of the opportunities posed by digitisation of knowledge and instant communication via the Internet. Scholarly publications are mostly still published in formats designed for printing (such as PDFs), and publishers keep many of the limitations initially caused by a paper form of journals, like word limits (Bartling & Friesike 2014, p.7). The move of scholarly communication to an online form also enables collection of data regarding the use of publications

by other researchers on an unprecedented scale. This data can be used for studies on the development of science, but also potentially in the evaluation of scientists and institutions.

1.1 The history and role of scholarly databases

Hence scholarly databases, tools indexing scholarly publications, citations, and other metrics, are of interest to multiple agents. Considerable attention is given to the analysis of the scope and depth of such datasets as information sources by the community of scientometricians and bibliometricians, who use them to study the scientific procedure. Policymakers' growing interest in scholarly metrics also highlights the importance of databases, as the backbones of the quantitative evaluation system. Finally, their contents are of interest to individual researchers or institutions who can use them to obtain an overview of work conducted and its reception by the community.

The attention to citations as links between publications, researchers, journals, institutions, and ideas has been first noticed by the 'father' of the bibliometrics, Eugene Garfield (1955). He created *Science Citation Index,* a first citation index (which belonged to his company - *Institute for Scientific Information,* and was later transformed into Web of Science). The aim of his activity was initially to improve researchers' ability to review literature – the citations were seen as an way to notice criticism or obsolescence of papers cited (Garfield 1955). Shortly afterward, the SCI was recognised as an source of information for studies regarding the scientific procedure. One of the first persons to use such information to analyse the networks created by researchers and their publications was Derek de Solla Price (1983). The interest in citation databases has gradually developed in the direction of creation of metrics: indicators of the impact of the publications. Hence, citation indices became of interest to higher education and research policy makers (Mingers & Leydesdorff 2015).

A rapid growth of Web of Science (WoS) has been taken place in the 1990s and 2000s, when the role of the Web as a medium of digitised knowledge exchange has substantially increased. Online publications enabled Web of Science to include more journals and expand the database to incorporate conference proceedings. In 2002 WoS, previously has been only distributed on CD-ROMs sent to institutions, for the first time become available via a web platform. The tipping point for the scholarly databases has been reached in the year 2004, when publishing company *Elsevier* has created a rival citation database – Scopus - and *Google* launched Google Scholar, a search engine dedicated for queries regarding scholarly literature (Hicks et al. 2015; Burnham 2006). A distinction between the two types of data gathering has to be drawn. While in Scopus and WoS the decision to index an article is based on whether the venue of its publication is on the list of manually approved journals, Google Scholar uses algorithms to crawl and automatically parse websites in search of scholarly publications (Harzing & van der Wal 2008).

Microsoft experimented with the creation of a robust citation database and scholarly search engine since 1996. The first efforts, a system called Windows Live Academic (later called Live Search Academic) has been called 'dishearting' by Peter Jacsó (2011) due to many critical flaws. A second attempt, firstly released under the name Libra, has been regarded as more successful. One of the keys to the creation of a more intelligent search platform was a focus on the literature regarding research in Computer Science. This field has been well covered by indexing systems and online libraries such as CiteSeerX, ACM Digital Library and IEEE (Giles et al. 1998; Caragea et al. 2014). The improvement in the quality of the portal and growth of the quantity of papers indexed has been noted. The service was soon renamed Microsoft Academic Search. A review by Jacsó (2011) declared the tool 'a project of great interest,' however, the coverage at the time of his publication has been still lagging behind Scopus or Web of Science. Unfortunately, an analysis conducted by Orduña-Malea three years later has shown that around 2011 first signs of discontinuation of development of the platform have been observed (Orduña-Malea et al. 2014). The same review established that since 2013, the service has ceased to be updated at all, and the indexing of new records has been proceeding at a 'minimal rate'. The third attempt by Microsoft has been made available to public in 2015 under the name Microsoft Academic (Sinha et al. 2015). This time, it has been based on both Bing search engine web crawlers and indexing information from online libraries, publishers, and other databases. Such design places it between GS and the two traditional citation databases regarding data gathering methods. Furthermore, the dataset behind the online search portal, containing papers (titles and abstracts), authors, affiliations, and citations has been openly published under the name Microsoft Academic Graph (MAG)[1] in the same year. An attempt on analysis of the MAG, relating its scope, openness, completeness of information and interoperability to the other three scholarly citation datasets (WoS, Scopus, GS) is the aim of this project.

In the meantime, relocation of the mainstream of scholarly publication and communication to the Internet has resulted in novel opportunities for observation of scholarly communication and development of metrics. Alternative metrics or *altmetrics* (Priem et al. 2010) are terms describing data collection and assessment tools using the web usage statistics to allow 'the impact of research to be measured more broadly than with citations' (Bornmann 2015), although the term is sometimes confusingly used referring to article-level metrics (Costas et al. 2015). Under this category, multiple databases and portals have been created to measure specific types of activity (Lin & Fenner 2013). Viewing statistics and number of PDF downloads are recorded by some publishers, such as the Public Library of Science (PLOS) or independent Altmetrics portals (Altmetric.com, ImpactStory). Reference managers, such as CiteULike or Mendeley, record the data on usage of papers by individuals saving them to their reference libraries. Discussions

[1] https://www.microsoft.com/en-us/research/project/microsoft-academic-graph/

around publications can be traced by counting the number of mentions of the URL on social media (Twitter), Wikipedia, or scholarly blogs (e.g. ResearchBlogging). Recommender systems, such as the F1000Prime have been created. F1000Prime is using a network of a few thousands of members to crowdsource the review, which in turns is used to decide whom to recommend it to, forming a post-publication peer review system (Bornmann 2014). The databases which are the backbones of such systems are becoming a promising source of information for scientometric studies. However, due to their focus on measurement of impact other than traditional citations, only some of the dimensions of analyses conducted in this project could be related to them (such as the openness and interoperability of datasets). Hence a decision has been made to focus on the comparison between the four above-mentioned citation databases.

Chapter 2: Related Literature

2.1 Requirements on scholarly databases

The role of metrics in the evaluation of scientists' work has been debated continuously over the recent years. Notably, the Higher Education Founding Council for England created a Steering Group with the aim of performing a study on the perspectives of use of quantitative metrics in research evaluation (Wilsdon 2015). The report is based on case studies performed as part of the Research Evaluation Framework 2014 (REF 2014). As a result, it indicates growing pressures for an audit of public spending on science, resulting in the adoption of metrics as a faster and less expensive alternative to traditional expert review. On the other hand, the researchers themselves contest usefulness of such quantitative indicators in the evaluation of work, highlighting the fact that misuse and narrowly designed metrics can have a detrimental effect on research (Wilsdon 2015).

The overall conclusion of the report states that metrics can provide support for qualitative evaluation based on peer-reviewed case studies, but cannot replace it. As one of the key elements of the call for 'responsible metrics', it has been indicated that it is necessary to base quantitative indicators on best, 'in terms of accuracy and scope', available data (Wilsdon 2015). Hence a call for 'open and interoperable data infrastructure' was devised, demanding a robust data infrastructure enabling to construct meaningful metrics. The report encourages a creator (or owner) of a database to openly present information on data collection and processing. Furthermore, the call asked for adoption of cross-database identifiers, data infrastructure standards, and semantics to improve the clarity of metrics. Finally, it has been highlighted that common semantics (including definitions of concepts, such as 'impact') and identifiers will increase the interoperability of sources of data, in turn increasing scope and robustness of metrics (Wilsdon 2015).

These recommendations come in line with two other documents concerned with the usage of quantitative indicators in the assessment of research. Created by the American Society for Cell Biology, the 'San Francisco Declaration on Research Assessment' (DORA) has, among other points, drawn attention towards transparency on underlying data and methods of processing. Additionally, attention has been drawn to making data available for unrestricted reuse (with computational access to it) (ASCB 2012). Similarly, the 'Leiden manifesto for research metrics' describes ten principles of responsible metrics creation and use. The fourth principle asks for the openness of data collection and processing (Hicks et al. 2015). As an example of such 'black-boxed' metric, the online scholarly social network ResearchGate's 'RG Score' may be used: it was found to be irreproducible and nontransparent (Jordan 2015; Kraker & Lex 2015). Additionally, the Manifesto highlights that the data collected should help metrics take into account the disciplinary

variations in publishing and citation practices (disciplinary normalisation) allow researchers to verify the data collected, and should support simplicity of metrics, which in turn helps to spread understanding and transparency of an indicator.

Such recommendations seem to be shared by diverse groups of interests: researchers themselves, journal publishers, editors, scientometricians, and research evaluation bodies. The DORA has gained over 570 organisational and 12,300 individual signatories (Wilsdon 2015) since its creation by 'a group of editors and publishers of scholarly journals' (ASCB 2012), while the Leiden Manifesto has been created by academics working in scientometrics and bibliometrics areas. Finally, the HEFCE report presents recommendations based on the application of citation metrics in REF 2014. The above-mentioned proposed characteristics of ideal scholarly databases are similar in each those three documents. Other researchers have also mentioned some of the issues of these current state during their studies: lack of information regarding construction of disciplinary classification in WoS and Scopus (Wang & Waltman 2016), 'non-existent' transparency of sources of data in GS compared to Microsoft Academic Search (Orduña-Malea et al. 2014), or issues relating to interoperability of data (Zuccala & Cornacchia 2016). In light of the literature mentioned above, the dimensions of analysis of MAG in this report can be seen as themes of great importance to the scholarly community.

2.2 Citation Databases

2.2.1 Web of Science

The Web of Science (WoS) is a database created by the Institute for Scientific Information and then operated by Thomson Reuters. Recently, information has been published indicating that it is going to be sold to private equity funds Onex Corporation and Baring Private Equity Asia has been published. It is rumoured that the potential final buyer may be scholarly publishing corporation Nature Group, owner of the *Nature* journal series, among others[2]. On top of the database, multiple citation indexes have been created, including Science Citation Index Expanded, Arts&Humanities Citation Index, and the Social Sciences Citation Index. Recent addition to the index portfolio includes Conference Proceedings Citation Index and a Book Citation Index (Wouters et al. 2015). The database itself consists of the Core Collection, which includes the above-described indexes, and additionally incorporated databases, such as SciELO, a database based on open-access electronic publication model in Latin America and Caribbean journals (Lucio-arias et al. 2015). Access to the database is provided on a paid subscription basis.

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

The WoS portal enables search by publication title, author, topic (discipline or keyword), year of publishing, grant number, conference, affiliation, and DOI, among others (Falagas et al. 2008). There also exists an Application Programming Interface (API) provided to enable computational access to the data, but it demands an expanded subscription to make a full use of its capabilities.

2.2.2 Scopus

Scopus is a citation database launched in late 2004, owned by a Dutch publishing company Elsevier. Access is also provided on a subscription-based model. The database, apart from journals, is also covering books and conference proceedings. The web portal of the system allows for search based on title, abstract, keywords, author, affiliation conference and DOI. Similarly to the previously described dataset, an API service exists, although full access to it is limited to subscribers and only basic metadata can be obtained by the free user[3].

2.2.3 Google Scholar

Google Scholar (GS) was launched in 2004 by Google. This database is indexing scholarly literature available on the Web, using algorithms to search and parse them. Therefore, GS includes journal and conference proceedings, books, but also other types of research output: theses, preprints and technical reports that are not listed in Scopus or WoS (Wouters et al. 2015). Because the documents are retrieved and parsed automatically, no list of sources covered is available, and the quality of the indexed data remains an issue.

The GS website allows searching by keywords or phrases. An author, publication venue (journal) and date range can also be specified. However, because of lack of direct access to the GS database via an API, it is considerably harder to perform a bibliometric, large-scale analysis of the dataset. A program called *Publish or Perish* has been developed to help interested parties access information from the website (Harzing 2010) and it is used to retrieve information from Google Scholar in this study.

2.2.4 Microsoft Academic (Graph)

Microsoft Academic Graph (MAG) is a downloadable, free to use for academic applications dataset. The portal built on top of it, Microsoft Academic (MA), is a successor to the Microsoft Academic Search (MAS) project discontinued in 2012[4]. The new version of the portal is integrated with company's search engine,

[3] http://dev.elsevier.com/sc_apis.html
[4] https://microsoftacademic.uservoice.com/knowledgebase/articles/838965-microsoft-academic-faq

Bing. Confusingly, the official publication describing the dataset, published in 2015, still uses the term 'Microsoft Academic Search' when referring to the search portal (Sinha et al. 2015), probably as it is the term commonly adopted in publications about the service. The Microsoft Academic Graph is published as a set of tab-separated files of a total size of 28GB (compressed to a ZIP-format) and is also accessible via API. The downloadable versions of the database provide a snapshot at a given date, with the first version published on 5th June 2015 and the version used in this project created on 5th February 2016.

The Microsoft Academic portal allows search queries by keywords and has a menu consisting of disciplines and their subdisciplines. The search for article titles, keywords, disciplinary categories, affiliations or others can be performed using a unified search box. Additional statistics (e.g. top 10 institutions regarding a number of papers in given discipline) are also displayed alongside with a box presenting upcoming conferences in specified field of research.

2.3 Comparison between databases

The section below provides an overview of analyses performed on the four scholarly datasets. It has to be noted that a majority of studies focused on the Web of Science, Scopus, and Google Scholar. This situation may have arisen due to the discontinuation of the early version of Microsoft Academic in 2012 and the fact that the MAG dataset has been published only in 2015 (Harzing 2017). The criteria for comparison of the databases have been chosen based on the recommendations for scholarly databases, as highlighted in Section 2.1. It has to be mentioned that all four projects are in constant development, hence some presented analyses may already be dated**.**

2.3.1 Scope

Obtaining an accurate count of a total number of research publications is effectively impossible. An attempt to estimate this figure conducted by Jinha (2010) concluded that by the end of the year 2008, almost 50 million scholarly journal articles had been published. In an attempt to estimate the total number of scholarly documents available online, Khabsa and Giles (2014) found that GS covered at the time around 87% of all such publications – around 100 million. Hence the total number of English-language documents online was estimated to reach 114 million. The discrepancy between those findings and Jinha's estimate is most probably because Google Scholar also indexes non-traditional research output other than journal articles. In a study aiming to estimate the total number of documents in GS, Orduna-Malea et al. (2015) have used three independent methods to come to a conclusion that the size of the dataset in May 2014 was between 160 and 165 million documents. The same paper found out that the size of WoS at the time was 56.9 million and Scopus contained 53.4 million documents.

The scope of coverage of the databases has been analysed in multiple publications. One of the most detailed studies comparing the Scopus to Web of Science was conducted by Moed and Visser (2008) found that 97% of publications indexed by Scopus could also be found in WoS. The journals listed in the former, but not in the latter, have been found to have a lower number of citations and to be published in primarily nationally-oriented journals (López-Illescas et al. 2009). A study of Slovenian publications highlighted the superiority of coverage of Scopus versus WoS especially in the fields of social sciences, engineering and technology, humanities (Bartol et al. 2014).

A number of studies have reported that Google Scholar indexes larger number of publications than Scopus or WoS. Regarding publications in the fields of business and management, Mingers and Lipitakis (2010) concluded that GS has substantially better coverage than WoS and hence would form a better basis for research impact measurement in the area. At the same time, they highlighted that this opportunity is hampered by the unreliability of the GS data. Similar results were obtained in the fields of anthropology, education, and pedagogical sciences, where GS was shown to be superior to WoS regarding coverage, which may be due to the fact that these fields are characterised by more diverse types of output (Prins et al. 2016). To conclude, opportunities of use of GS in the evaluation of research in fields with moderate or low coverage in other databases have been highlighted before, but the need for a complex data cleansing and handling has to be taken into account because of the unreliability of the automatic scrapers collecting information for Google. Despite the problems of reliability, it has been found that 70% of citations indexed by Google Scholar, but not Web of Science, originate from full-text online documents of various types, thus enabling to measure the broader type of impact (Kousha & Thelwall 2008).

For some fields, however, the opposite trend has been identified. For example, a study on a set of Isreali researchers conducted by Bar-Ilan (2008) has shown that GS has worse coverage in the field of high energy physics than WoS or Scopus while Mikki (2010) concluded that neither WoS nor GS could be shown to be superior in the area of earth sciences. However, the improvement of the Google Scholar service, as reported by de Winter, Zadpoor, & Dodou (2014) or Harzing (2013, 2014) may imply that those results no longer hold true. It has also been noted, that the total number of citations to a specified set of 56 scholarly articles from diverse research fields has been higher in WoS than GS for 39 of the articles (de Winter et al. 2014). The repeating differences between databases demand further analyses to take into account possible differences in disciplinary coverage.

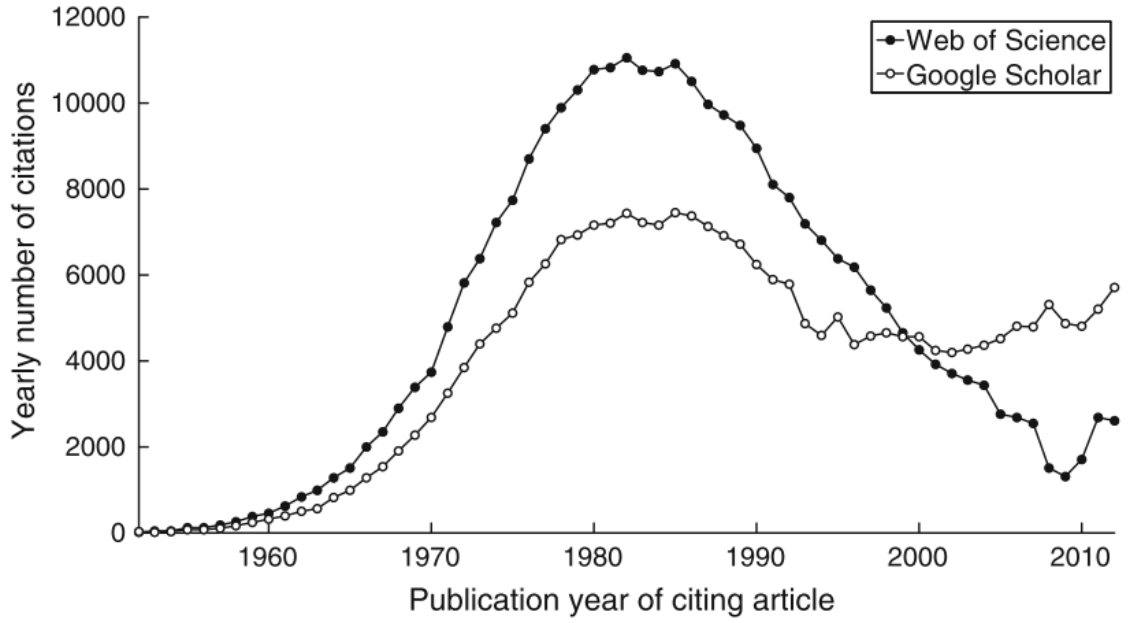

Figure 1 Comparison of number of publications indexed by Google Scholar and Web of Science by year of publication (de Winter et al. 2014)

As mentioned before, the Microsoft Academic database has not been a major point of interest for the community analysing scholarly databases. The initial description of the Microsoft Academic Search database performed by Jacsó (2011) concluded that it might be a 'promising' source of information for researchers interested in scientometrics. However, it took three years after publication of that paper for a new study concerned with the Microsoft Academic Search (MAS) to be conducted. The new study has shown evidence for a rapid decline in the number of papers indexed by MAS since 2012 to near-zero numbers (Orduña-Malea et al. 2014) when compared to WoS, even in fields in which MAS indexed more documents than WoS in 2011. Understandably, a comparison of 771 author profiles in MAS and GS performed the same year has shown that the former has a lower number of publications-per-author than the latter. However, the same study has also noted that the MAS has maintained more disciplinary balance than GS, which was found to index significantly more documents in the field of computer science (Ortega & Aguillo 2014). In a study aiming to estimate the total number of scholarly publications on the Web, MAS was used in comparison with GS to help estimate the total number of documents not indexed by the latter (Khabsa & Giles 2014). Finally, in their search for a method to reliably estimate the size of Google Scholar, counts of a number of publications indexed in MAS by year have been presented by Orduna-Malea et al., (2015). It has to be noted that the above-mentioned rapid decline of a number of indexed documents indexed by MAG can be noticed in results of the study, as shown in Figure 2.

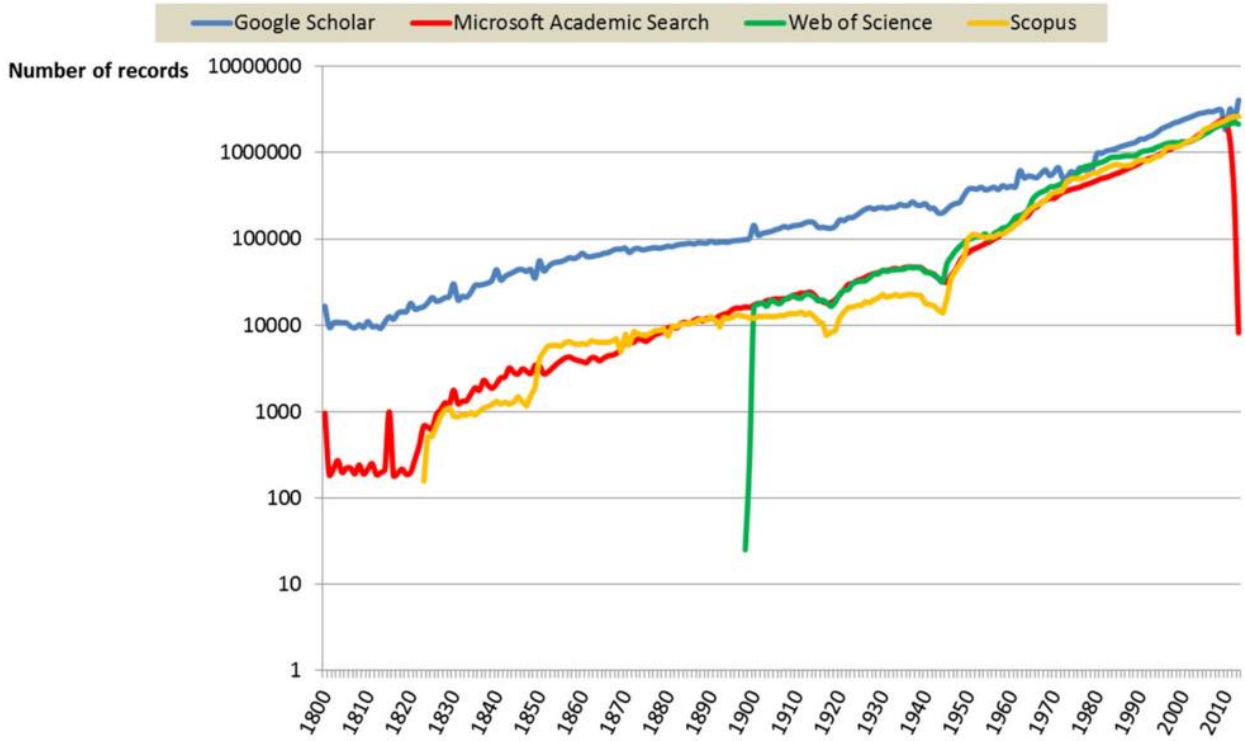

Figure 2 Comparison of WoS, Scopus, GS, and MAS by number of documents indexed (1800-2013); data for WoS available since 1900 (Orduna-Malea et al. 2015)

Since then, Harzing (2017) has been the first to research the new version of the Microsoft Academic (MA) platform, using the software *Publish or Perish* and a query of the author's own publications. The results, shown in Figure 3, have shown that although GS indexed 35 documents that were not indexed by MA ('A1: 35' in Figure 3), none of them were journal papers and the majority of them were book chapters, white papers, and conference papers. Furthermore, 17 of these publications were identified as 'citations' by GS, meaning they were documents identified only as references in other papers, without identified online presence themselves. On the other hand, MA indexed 43 publications unique when compared to WoS (out of which 20 were non-journal publications, 'B2: 43' in Figure 3). Most of the papers not found in WoS were either recently published, or circulated in journals which were not included in WoS at the time of their publishing but added to the database since then. Similar observations have been made regarding the 30 documents indexed by MA which were not found in Scopus. However, the number included only seven non-journal publications, indicating better coverage of diverse research outputs of Scopus when compared to WoS. Importantly, both Scopus and WoS had only a small number of publications that are not indexed by MA: two book chapters in the former case and just one in the latter. It has to be mentioned, however, that the study has been performed only on a single persons' scientific output and hence needs to be reproduced on a larger scale.

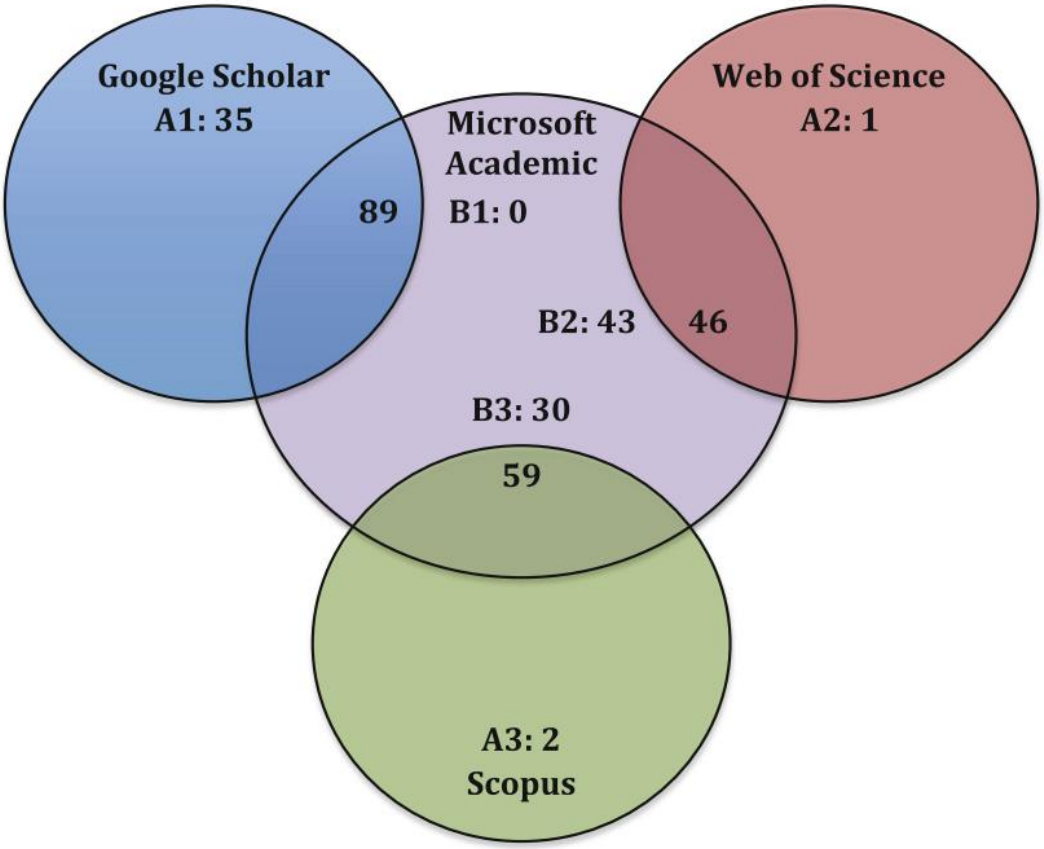

Figure 3 Comparison of coverage of new MAG dataset with WoS, Scopus, and GS (Harzing 2017)

2.3.2 Interoperability

The interoperability of the databases is defined here as the availability of an application programming interface (API) allowing to scrape the information from the databases and use of global unique identifiers for scholarly papers, authors, citations, and institutions.

The unique identifiers play a major role in disambiguation of entities in databases. One of the common problems with scholarly metadata is authors name disambiguation. Names are not unique, hence considerable effort has to be taken to link the correct author with paper. The problem arises due to a popularity of some surnames (e.g. Smith, Li), but also a translation of the name to a different alphabet (e.g. Chinese surnames) (Tang & Walsh 2010). Due to the sheer volume of indexed publications, manual disambiguation may be inefficient and prone to error. Although some progress has been made on the problem using machine learning and natural language processing methods (Treeratpituk & Giles 2009), the solution already available is to create and use unique identifiers (Wilsdon 2015).

The problem presented above also concerns papers (recognition of multiple versions of the same document), citations (as references are commonly mistyped or recorded with errors), and affiliations (similar to authors, the names of institutions can be presented in various formats and languages). Hence a set of identifiers is needed for each of types of entities. Table 1 describes the common identifiers used by the scientific community (Wilsdon 2015).

Table 1: Global unique identifiers in scholarly databases

Type of Entity	**Identifier**	**Degree of adoption**
Journals	International Serial Standard Number (ISSN)[5], with ISSN-L link as master identifier for both print and online edition	Widespread, with exceptions
Publishers and institutions	Multiple identifiers, although International Standard Name Identifiers (ISNI, worldwide)[6] and UK Provider Reference Number (UKPRN, more UK-centric and excluding funders)[7]	
Authors	Although multiple standards exist, ORCID[8] is regarded as superior	ORCID adoption growing in the UK and worldwide; endorsed by major science institutions, such as HEFCE, Jisc, Wellcome Trust
Papers	Digital Object Identifier (DOI)[9]	Commonly adopted, also issued to other forms of research output, such as conference papers or datasets

[5] http://www.issn.org/understanding-the-issn/the-issn-international-register/
[6] http://www.isni.org/
[7] https://www.ukrlp.co.uk/
[8] http://orcid.org/
[9] https://www.doi.org/

Chapter 3: Methods

The Microsoft Academic Graph database is available for download as a collection of tab-separated files. The individual files have been imported into a MySQL datastore, with the original structure of files preserved. Then, in order to answer the research question, the MySQL database was queried, with output saved as comma separated files. Python module Matplotlib was used to make visualisations of the output data, including histograms (which were created using the built-in 'hist()' function). As mentioned before, the data used for analysis is a snapshot of Microsoft Academic service published as MAG at 5th February 2016.

An alternative access to the dataset is provided via Microsoft's API service called Academic Knowledge API[10]. As was shown in Section 4.3.3, querying via the API has proved to result in richer responses in terms of scope, and hence has been used as a source of information in part of the study. The retrieval of information from the API proceeded using a Python script (with necessary user-key included), fetching the response in the form of a JSON file. Since the final comparison between databases was conducted in Excel, the output has been then converted to a CSV format.

3.1 Schema of the Microsoft Academic Graph

Figure 4 presents the entities in MAG, alongside with relationships between them. The complete list of files in the original MAG dataset, along with titles of columns in those, is attached to the downloadable dataset and is presented in Appendix A.

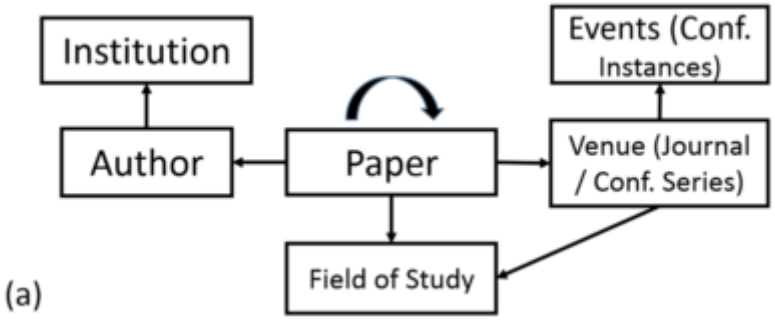

Figure 4 MAG entity relationship graph (Sinha et al. 2015)

Four independent entity types for initial analysis of the dataset have been identified: affiliations (institutions), authors, papers and fields of study. All of them are given an 8-symbols long unique ID,

[10] https://www.microsoft.com/cognitive-services/en-us/academic-knowledge-api

consisting of letters (A-Z) and digits (0-9). The details on information about them and directions of inquiry are presented below.

3.2 Openness

There are three dimensions of openness regarding the databases: licensing, access (including programmable access via API), and transparency of data sources and processing. Analysis of the approach of each of the four database owners is conducted based on information found on the official websites and previous studies.

3.3 Completeness of metadata

This degree of analysis focuses on the breadth of metadata available via each of the portals and MAG database. The richer the data surrounding authors or publications, the more options for analysis for scientometricians and bibliometricians exist. Therefore, a table containing the categories of information available in a local copy of MAG and via API is constructed for comparison with other databases. Furthermore, a review of the globally unique identifiers mentioned in Section 2.3.2 is presented to estimate opportunities for cross-database data use.

3.4 Scope

The primary direction of analysis focuses on a comparison of the scope of the Microsoft Academic Graph with the three other competitors. The low number of papers indexed by Microsoft Academic Search has been a repeating problem mentioned by previous reviewers (Jacsó 2011; Orduña-Malea et al. 2014), thus making the database a less reliable source for inquiry. With the new edition of the system (Microsoft Academic) and the newly published database (MAG), an attempt to estimate the scope of the dataset has to be repeated. The first attempt to analyse the new information source concluded that it might become an ‘excellent alternative for citation analysis’ if some of the identified problems are going to be resolved (Harzing 2017). The study, however, focused on a set of publications by a single author and hence shall be repeated on a larger scale. Four key entities in the dataset are analysed: affiliations (number of research papers and authors registered under a single chosen institution), authors (number of papers and citations), and publications (along with citation scores).

3.4.1 Affiliations

Graphs presenting the distribution of affiliated papers against institutions are presented in order to compare the databases. Based on the obtained data, a set of affiliations is selected and the resulting number of papers and authors for this institution compared with the other three databases. The range of selected affiliations for further analysis is designed to include those with high and low numbers of papers and authors in MAG and those coming from non-English countries.

To specify the set of institutions for comparison, the Webometrics Ranking[11] was used (Aguillo et al. 2008). A pseudo-random numbers generator has been employed with the job of selecting twenty-five numbers in three ranges: 0-200, 200-1000, and 1000-12000 (lowest available position on Webometrics Ranking website)**.** These ranges were arbitrarily chosen to obtain samples of top, average and low-ranked institutions, as ranked by the Webometrics Ranking. After identification of the names of the Higher Education Institutions (HEIs) and their country of origin, manual search has been performed in the Web of Science Core Collection, Scopus, and local instance of Microsoft Academic Graph. In the case of WoS and Scopus, features allowing enhanced search of the organisation has been used: the institutions have been firstly identified among the list of WoS or Scopus institutions and then the number of documents affiliated has been retrieved. Querying in MAG consisted of a search of string (or parts of it) among 'affiliation name', with the total counts of the number of authors and papers per each institution obtained from the database earlier. The operation was conducted using filtering in Microsoft Excel, on a set of institutions along with paper and author counts retrieved from MySQL.

3.4.2 Authors

A study regarding papers and citations of six selected authors is conducted in depth. In order to ensure fairness of judgment, a random selection of authors from the MAG database is made using the MySQL ORDER BY RAND() function performed on table 'authors'. Since author names disambiguation remains an issue in scholarly databases (e.g. Tang & Walsh 2010; Treeratpituk & Giles 2009), from the obtained set only people whose surnames and initials enable to uniquely identify a single person are selected. Such decision was made after the observation that the incoherence of author and his/her affiliation queries among databases did not allow for comparison of the results. The study is not aiming to uncover authors who are not represented in any of the datasets, hence a decision to randomly select profiles from one of the compared datasets does not bias the results in favor of MAG. However, the limitation of such design is

[11] http://www.webometrics.info/en

that the question of authors profiles missing from MAG or any other databases is not addressed by this study.

3.4.3 Papers

Set of publications authored by a given person is then retrieved from each of the four databases. Microsoft Excel is then used to process those sets and highlight the publications that are unique to MAG with respect to the three other databases (compared individually) and *vice versa*. The sets for Scopus and Web of Science are obtained using their web portals and author search capabilities, using initials and surname as a query. The set of documents from Google Scholar is collected via the *Publish or Perish* software, indicating initials and surname in a query field (in brackets, to ensure whole phrase search).

Papers stored in MAG can be divided into 'primary documents', which have complete (or almost complete) metadata present in the database (including authors, venue of publication, date of publication, references, and URL) and 'secondary documents', existing only as IDs. A similar division is observable in Google Scholar, where articles are divided into those parsed by algorithms and ones found only as references in other publications (marked '[citation]' [12]). The latter type is removed from the retrieved set before analysis.

A decision to exclude 'secondary papers' (meaning publications not directly indexed by the databases) from comparison was made after a careful inquiry into a set of papers for one of the authors, where out of 38 Google Scholar documents marked as '[citation]', only seven were identified to exist in both GS and one of the other datasets. However, Microsoft Academic Graph provided links to full-text documents in six cases and to the abstract in the seventh case. Interestingly, GS also included links to at least abstracts of the articles, but the marker '[citation]' remained. One of the possible reasons for such behavior is that the '[citation]' marker is updated independently from sources of entries in GS and simply is not up-to-date. Hence a decision has been made to follow the GS document type classification and to exclude documents marked as 'citations'.

Furthermore, more general statistics regarding the total number of papers in the database are to be produced. Such activity contrasted with appropriate numbers for WoS, Scopus, and GS is to help verify whether Microsoft Academic Graph can be taken as a reliable source of publication and citation information, covering diverse fields of study and an appropriate number of records.

[12] https://scholar.google.com/intl/en/scholar/help.html#general

3.4.4 Citations and References

The number of citations recorded by databases is compared for each of the six authors studied. The citation score is important for two reasons. Firstly, it is a major point of interest to users of the datasets – be it scientometricians, researchers, or policymakers. Therefore, a consistent and reliable citation indexing is needed to declare a dataset to be of interest to the researchers. Secondly, citations themselves can be regarded as indicators of 'depth' of the database: they provide information on the scope of citing publications that the database creators or algorithms have indexed.

3.4.5 Disciplinary classification

The problem with comparing disciplinary classifications encountered in WoS, Scopus, and MAG is that each of these has been independently defined by the owner of a given dataset and is characterised by a different total number of disciplines and sub-disciplines. This study provides only a general description of the disciplinary classification in Microsoft Academic Graph, compared with the two other classification schemes and proposes further work using this classification to compare disciplinary coverage of the four datasets.

Chapter 4: Results

4.1 Openness

As has been noted above, Google Scholar is a free service, but the data itself is only available via the search portal, with no direct access to the database itself. Hence, for example, it is not possible to obtain the number of documents and author profiles in the service and estimates need to be used (Orduna-Malea et al. 2015). Web of Science and Scopus are restricting access to the dataset to subscribers. Both Scopus and WoS web page interfaces can be used for a limited scale bibliographic analyses, however for a large scale queries a direct access to the database is needed (Waltman 2016). The direct access is included in a more expensive, hence only a limited number of institutions can perform such experiments. Microsoft Academic is hence the only one of the four to have made the complete dataset freely available to download and reuse for 'any non-revenue/no-fee academic purpose'[13]. The Microsoft Academic API is also open to the public, limited to 10,000 transactions per month, with no limit on the depth of information retrievable by free users[14]. Scopus API restricts access for non-subscribers only to 'basic metadata for most citation records'[15], while Web of Science API is open only to subscribers[16]. Therefore, MAG may be considered the most open dataset of the four analysed from the perspective of a free user with regards to both licensing and options for accessing the data.

Indexing of an article in WoS and Scopus can be predicted based on whether the venue of its publication is itself chosen to be indexed by owners of these services. The Web of Science provides three criteria for inclusion of a new journal: 1) three consecutive editions published on time, 2) reaching a threshold in a number of citations in journals already included in WoS, 3) special factors – such as the inclusion of a journal appealing to policymakers. The last category has arisen because the policy documents are commonly referred to as 'grey literature' – although having a real-world impact, they do not commonly produce direct references to scientific literature (Leydesdorff 2008). Additionally, peer-review, specific data formats (XML, PDF) compatible with WoS, as well as providing full text in English or at least bibliographic information in that language is demanded from the journal[17]. Similar criteria regarding journal inclusion have been created by curators of Scopus. They also use a range of qualitative (editorial policy review, peer

[13] http://academic.research.microsoft.com/about/Microsoft%20Academic%20Search%20API%20User%20Manual.pdf
[14] https://www.microsoft.com/cognitive-services/en-us/academic-knowledge-api
[15] http://dev.elsevier.com/sc_apis.html
[16] http://ip-science.interest.thomsonreuters.com/data-integration?utm_source=false&utm_medium=false&utm_campaign=false
[17] http://wokinfo.com/essays/journal-selection-process/

review, diversity in the geographical distribution of authors and editors) and quantitative principles ('citedness of journal articles in Scopus'). It is worth noting that only serial titles, such as journals, book series or conference series can be included in Scopus. Contrary to WoS, Scopus does not focus on publications in English but demands journal's home website availability in English and the full content of the journal to be published online [18].

Table 2 Criteria for inclusion of journal in Scopus

Category	Criteria
Journal Policy	Convincing editorial policy Type of peer review Diversity in geographical distribution of editors Diversity in geographical distribution of authors
Content	Academic contribution to the field Clarity of abstracts Quality of and conformity to the stated aims and scope of the journal Readability of articles
Journal Standing	Citedness of journal articles in Scopus Editor standing
Publishing Regularity	No delays or interruptions in the publication schedule
Online Availability	Full journal content available online English language journal home page available Quality of journal home page

It should be taken into account that both the Web of Science [19]and Scopus[20] publish a complete list of all journals indexed in the databases.

The two databases underlying Web search engines Google Scholar and Microsoft Academic are much less specific about criteria of inclusion of publications. The nature of platforms based on algorithms indexing

[18] https://www.elsevier.com/solutions/scopus/content/content-policy-and-selection
[19] http://ip-science.thomsonreuters.com/mjl/
[20] https://blog.scopus.com/posts/titles-indexed-in-scopus-check-before-you-publish

documents found on the Web prohibits from providing a complete list of sources of publications as the decision on inclusion of paper is made on a case-by-case scenario. However, URLs of the documents found are provided in both GS and Microsoft Academic, but the complete dataset (including URLs) is downloadable only in the latter case.

The official description of the Microsoft Academic service highlights that both partners' content and algorithmically found content are used as information sources: '(1) feeds from publishers (e.g. ACM and IEEE), and (2) webpages indexed by Bing' (Sinha et al. 2015). Authors of the paper drew attention to the fact that the majority of input comes from the search engine parsers, but it is the publishers' data that is of better quality and hence presumably contains richer metadata. Microsoft Academic Search curators published a list of content providers participating in the creation of their platform, the header of which declares the list to show the state as of 'early 2013'. Interestingly, partners in the project range from pre-print repositories such as arXiv, other scholarly publication databases (CiteSeerX, DBLP) and publishers themselves, such as the Public Library of Science (PLOS) or Elsevier (owner of Scopus). Notably, Thomson Reuters was not part of the project as of early 2013[21]. There are no criteria for inclusion specified on site, nor it has been stated that all of the partners' publications are included. The only statement found on site defines the Microsoft Academic as a portal including 'journal publications, conference proceedings, reports, white papers, and a variety of other content types.'[11]

It has to be concluded that apart from the 'almost nonexistent' (Orduña-Malea et al. 2014) transparency of Google Scholar, the three other databases openly publish the sources of content. However, neither the rather general description of criteria for inclusion in the case of Scopus and Web of Science, nor the lack of knowledge regarding Bing parsers used for Microsoft Academic allows to predict whether a journal or publication will be automatically included in the datasets.

4.2 Completeness of metadata

This section focuses on the types of metadata regarding papers, citations, authors and affiliations that are available in the databases. All of the databases include author list, year of publication, venue of publication and number of citations. Google Scholar provides the most limited metadata on the publication, where only the author (with a link to author's profile in Google Scholar available, providing the profile exists), date, venue of publication, and a number of citations is provided. Interestingly, not only the number of citations as indexed by Google Scholar is shown, but also the number of citing papers in WoS is displayed. GS also

[21] http://academic.research.microsoft.com/About/Help.htm#5

lists versions of an article, which enables to access the document from multiple sources. This feature is especially important in the case of articles published in journals which are not Open Access (OA), where the second or third version may lead to an institutional repository where the document is freely available (Jamali & Nabavi 2015). A similar mechanism of versions clustering is implemented in MA, where a list of sources of publication is presented, alongside with formats available at a given URL (PDF, HTML, other), as shown in Table 3 and Table 4.

Table 3 Breakdown of MAG tables and information contained in them

Table	**Information**
Affiliations	Affiliation ID, Name
Authors	Author ID, Name
Fields of Study	Field of Study ID, Name
Fields of Study Hierarchy	Child FOS ID and level (L3-L0), Parent FOS ID and level (L3-L0), Confidence level (0-100%)
Journals	Journal ID, Name
Papers	Paper ID, Title, Publish date, DOI, Publication Venue, Journal ID mapped to venue, Paper Rank
Paper-Author-Affiliations	Paper ID, Author ID, Affiliation ID, Affiliation name, Author sequence number (place on lists of authors)
Paper Keywords	Paper ID, Keyword, Field of Study ID
Paper References	Paper ID, Referencing paper ID
Paper URLs	Paper ID, URL

The Web of Science by default provides abstract of the publication, information on venue of publication, the DOI of the paper, extracted information on author, date of publishing, paper keywords, details of funding of the research, publisher, Web of Science disciplinary classification, number of citations and other information on document type and ISSN identifier. Such information can also be obtained from Scopus.

Table 4 Comparison of types of metadata available in GS, WoS, Scopus and MAG

Information published	**Google Scholar**	**Web of Science**	**Scopus**	**Microsoft Academic Graph**
Author list for a document	+	+	+	+
Abstract	-	+	+	- (available in Microsoft Academic, but not in MAG)
Date of publication	+	+	+	+
Venue (e.g. journal)	+	+	+	+
Affiliation	+ (if Author's profile created)	+	+	+
URLs	+	+	+	+
Citations	+	+	+	+
References	-	+	+	+
Database Keywords	-	+	+	+
Funding	-	+	-	-
Disciplinary classification	-	+ (WoS classification)	+ (Scopus classification)	+ (MA classification)
Document type	-	+	+	-
Language	-	+	+	-

In terms of global unique identifiers available, Table 5 provides results of an inquiry into information obtained from the datasets. It is worth noting that identifiers issued by the database owners themselves were not included in the comparison, as only those fostered as open, independent standards (or proposed standards) give hope of wider adoption by the community of researchers and publishers.

Table 5 Overview of usage of independent, unique identifiers in databases

Identifier	**Google Scholar**	**Web of Science**	**Scopus**	**Microsoft Academic**
Journals (ISSN)	-	+	+	-
Publishers and institutions (ISNI)	-	-	-	-
Authors (ORCID)	-	+	- (enables search by ORCID id)	-
Papers (DOI)	-	+	+	+

Unfortunately, only the already widely adopted DOI is commonly included in the output of queries. Web of Science and Scopus stand out as services also providing ISNI - an identifier for series of publications (journals, books), information which is unavailable in GS or MAG. Furthermore, both of these services provide an option to include PubMedID (alternative document identifier, issued by PubMed[22]) in the results. ORCID ID can be retrieved only from Web of Science, but Scopus enables search for a person based on this identifier. Hence it is assumed that Scopus also stores that information. Neither GS nor MAG allows retrieving global unique identifiers regarding series of publications, publishers, and institutions or authors.

[22] http://asklib.hsl.unc.edu/a.php?qid=37565

4.3 Scope

4.3.1 Basic Statistics

Table 6 Counts of types of entries in MAG

Type of entity	Count
Affiliations	19,843
Authors	114,698,044
Fields of Study (disciplinary classification)	53,834
Journals	23,404
Documents (titles of documents)	126,909,021
Documents URLs	454,070,767
Paper-Reference pairs	528,682,710

Table 6 presents the number of entities of each type, counted as the number of rows in corresponding documents.

The number of affiliations in MAG can be compared to the Webometrics Ranking of World Universities. The ranking is based on a number of webpages of an institution, how well they are interlinked and how many rich documents do the pages contain, and the number of publications affiliated found in Google Scholar (Aguillo et al. 2008). Being based on an online presence of an institution and supported by the GS database, the authors claim that Webometrics Ranking 'is probably a complete directory of universities having independent web domains' and their current count of affiliations reaches 21,000 Higher Education Institutions (HEIs)[23], which is close to the count of affiliations in MAG. This would suggest that MAG covers universities and research institutions well, however, it has to be recognised that while Webometrics Ranking counts HEI only, the affiliations in MAG are of more diverse nature. For example, private companies (such as 'Microsoft' itself) or government ministries (e.g. 'Brazilian Ministry of Finance') can be found in MAG. Furthermore, authors of the Webometric Ranking estimate the total number of HEIs to be around 40,000, showing that both MAG and Webometric Ranking do not cover a complete list of such institutions, but merely around half of them.

The number of author profiles (individual IDs and names) in MAG reaches 114 million. This is a considerable improvement compared to a report from 2012, where the number of authors in Microsoft Academic Search

[23] http://www.webometrics.info/en/node/24

(the initial version of the project) was estimated to be 19 million (Ortega & Aguillo 2014). Unfortunately, no estimates regarding the number of authors in Google Scholar are available. The fundamental difference is that within GS, an author profile has to be manually created by an author himself, whilst in Microsoft Academic service an automatic profile is set up with at least a list of co-authored affiliations and fields of study that the researcher has authored papers in, as shown in Figure 5.

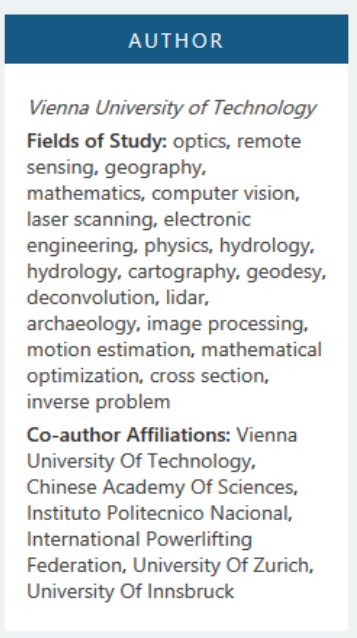

Figure 5 An example of an author profile in MAG

The three-level classification schema consists of 53,834 Fields of Study (FOS). The classification provides an opportunity for the much more detailed location of paper among a variety of disciplines, comparing to between 200 and 300 disciplines and sub-disciplines available in Scopus and WoS and no classification system in GS (Paragraph 4.3.5). FOS in MAG are mapped to papers based on keywords. Relations between the categories are stored in a separate file, consisting with tuples of child category and parent category, levels of both FOS and a probability of such relation, allowing to infer the top-level disciplines having assigned a few third-level Fields of Study. Unfortunately, no information on the calculation of the likelihood of child-parent relationship among categories has been found. The total number of paper-keyword-FOS triples in the database is 158,280,968, and the number of tuples describing child-parent relationships between FOS is 182,103.

The number of papers with at least some of the metadata available in the database (title, year of publication) reached 126 million. Hence a considerable improvement has been observed, as of 2012 the number of such entities in Microsoft Academic Search was estimated to be around a third of that number (40 million, Ortega and Aguillo 2014). The total number of documents in Google Scholar is estimated to be between 100 million (English-only) and 160-165 million (Khabsa & Giles 2014; Orduna-Malea et al. 2015). Additionally, Khabsa et al. in their study estimated the number of scholarly papers written in English and available online to be 114 million as of 2014. Orduna-Malea at al. also found out that the size of WoS is 56.9 million and Scopus is 53.4 million documents, as presented in Figure 6.

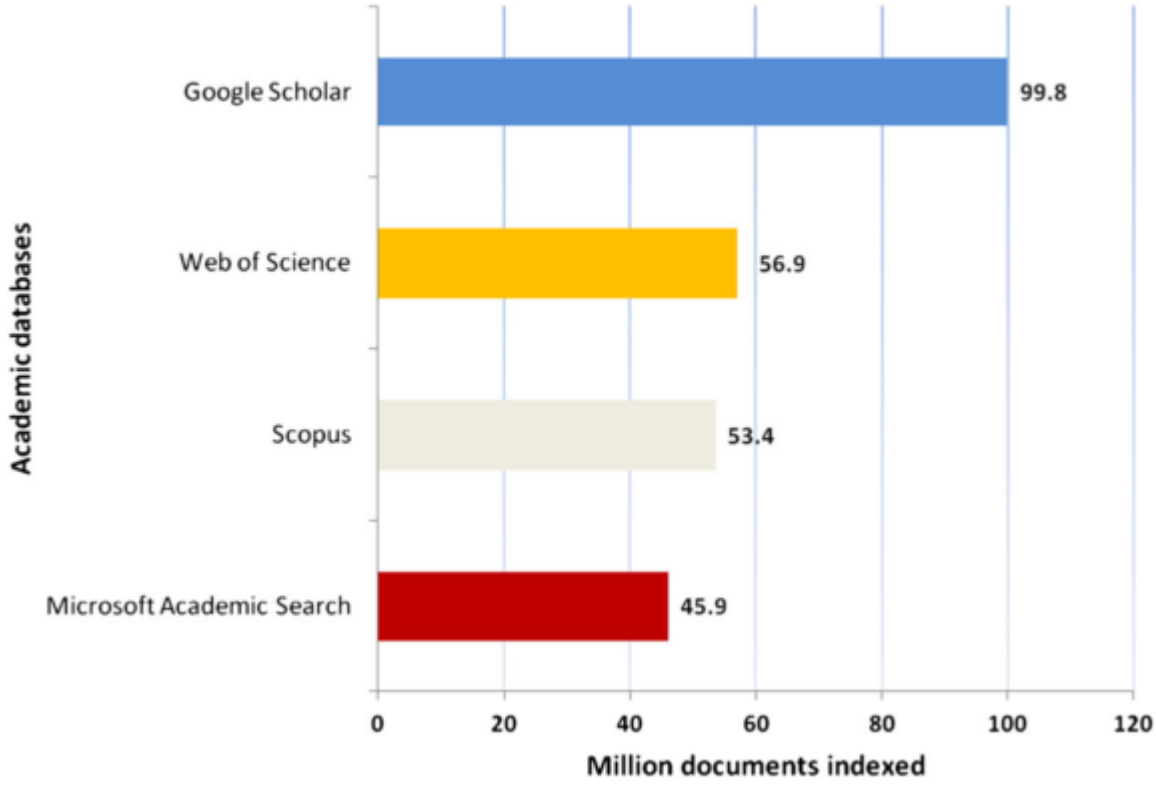

Figure 6 Comparison of number of documents in GS, WoS, Scopus, and the discontinued MAS service (Orduna-Malea 2015)

Another way to compare the sheer size of the databases via the number of journals indexed. These figures are officially presented by WoS and Scopus. The latter of the two claims to index 21,500 journals[24], while the former contains 12,665[25]. Judging by the number of journals in MAG database (23,404) and the total number of documents in it, the Microsoft dataset should show a better coverage of publications.

To conclude, the dataset used in the study is almost certainly smaller than Google Scholar, but greatly exceeds the sizes of the two traditional scholarly citation databases in terms of sheer size, showing potential for coverage of more diverse research outputs and/or balanced coverage in various disciplines. However, the next step of the comparison is designed to focus on the reliability of those numbers, by a careful examination of entities listed as affiliations, authors and documents in the Microsoft Academic Graph.

[24] https://www.elsevier.com/solutions/scopus/content
[25] http://wokinfo.com/products_tools/products/related/webservices/

Access to the whole dataset of papers and affiliations allows for large-scale queries which cannot be performed (case of GS) or at least access to them is heavily restricted for most researchers by subscription mechanism (WoS, Scopus). As an example, histograms of a number of papers and authors per MAG affiliation have been created.

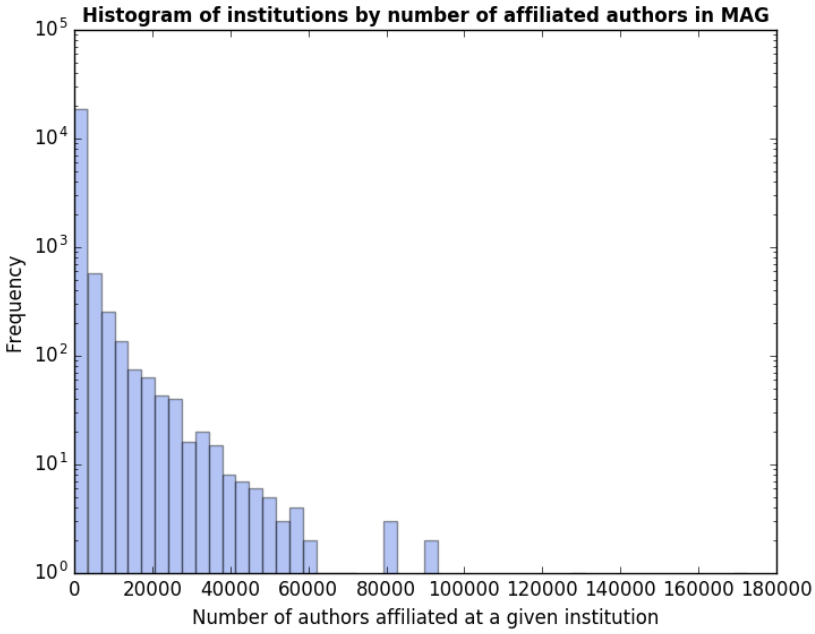

Figure 7 Frequency graph of authors per institution

As shown in Figure 7 and Figure 8**,** both distributions resemble the power-law distribution. Such feature could help generalise the long-existing Lotka's Law, which describes the frequency of publication by authors in a discipline of scientific inquiry to follow a specific power law distribution (Friedman 2015). Using MAG, a similar relationship could be studied among frequency of publications affiliated with an institution or frequency of authors affiliated with an institution. However, due to the low quality of affiliation recognition – and hence the input data – established quantitative relationship (mathematical formulas) would probably be unreliable using the currently available snapshot of the database.

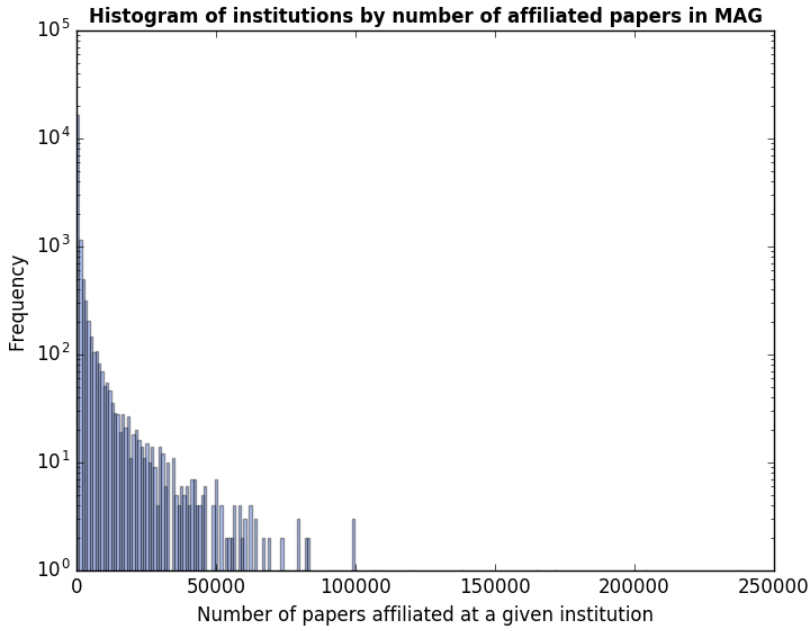

Figure 8 Frequency of papers per institutional affiliation

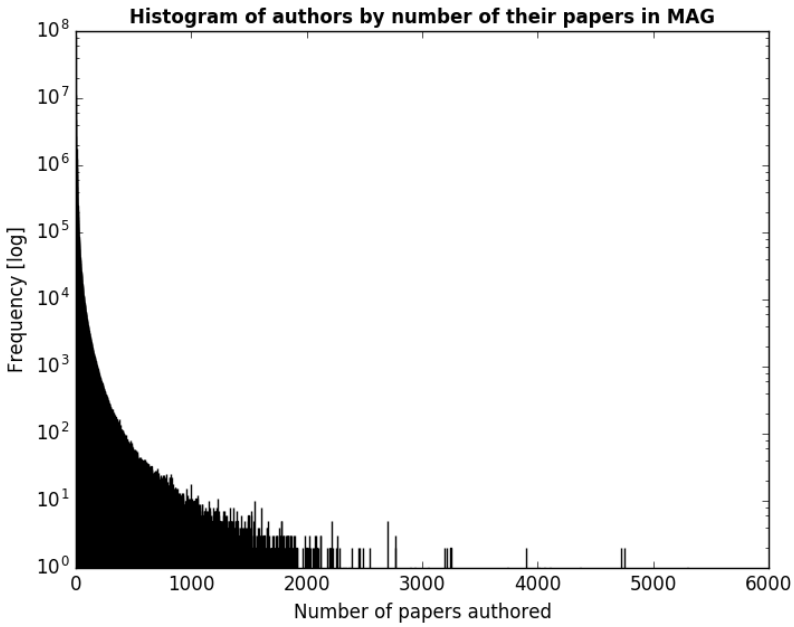

Figure 9 Frequency of papers per author

Figure 9, showing frequency of papers by the author again displays a trend resembling power-law distribution, which is an understandable generalisation considering that Lotka's Law applies to such frequency counts for individual disciplines (Mingers & Leydesdorff 2015). However, the possibility of a more detailed study of such relation is currently hindered by the problems with name and affiliation disambiguation, which are mentioned below.

The comparison of a number of papers indexed by each database by year of publication of the document has been conducted. Search by year has been carried out on Scopus and WoS web portals. Retrieval of reliable figures for Google Scholar has been shown to be complicated in previous studies. Here, an absurd query (search for articles not containing a long string of random characters) limited to each single year has been conducted. It has to be mentioned that the number of results presented by GS is only approximate (Orduna-Malea et al. 2015). Hence although they are included in Figure 10, the value of comparison of the other databases with GS is very limited. Locally-maintained Microsoft Academic Graph data has been obtained using the COUNT() function in MySQL.

The comparison presented in Figure 10 shows that the number of documents indexed by MAG has been consistently higher than numbers for Scopus and MAG. For most of the spectrum (1970-2002) the numbers have been similar to those in GS, and the discrepancy arising between 2002-2016 is most probably due to the above-mentioned unreliability in data gathering from Google Scholar, judging by visual comparison with Figure 2.

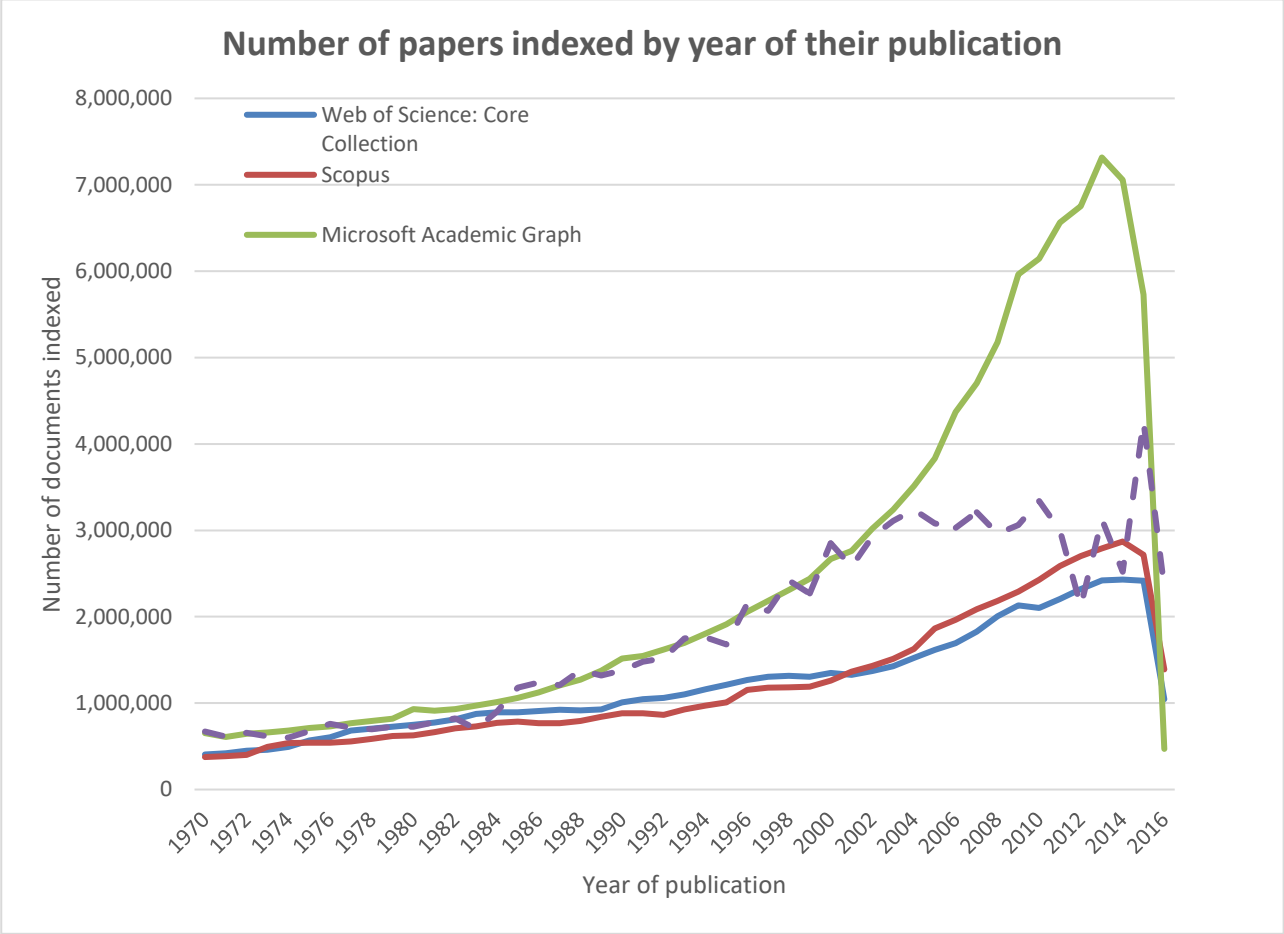

Figure 10 Number of papers indexed by databases by year of publication (1970-2016)

The rapid decrease in a number of indexed by MAG documents published in 2016 has to be highlighted. Since the version of the database published on 5th February has been used for comparison, while retrieval of data from other databases was conducted on 31st August, it should not be surprising that the number of articles in MAG is significantly lower than in other datasets. Therefore, more studies regarding the delay, defined as the time between publication of document and indexing of it by the database, needs to be conducted.

4.3.2 Affiliations

As mentioned above, the sheer number of affiliations indexed by MAG (19,843) and diversity of the included institutions (from governmental, via private companies, to institutes and universities) indicates a decent coverage of affiliations. However, this picture should be contrasted with the results comparing MAG with WoS and Scopus regarding the number of papers per institution presented in Figure 11.

In all three databases, there were cases where an institution existing as a single HEI in the Webometrics Ranking has been recorded in multiple entries in WoS, Scopus, or MAG. Such was a common case with

institutes or departments of health. Therefore, only records being clearly affiliated with the institution in question have been included in the total count of papers of that institution. This uniform policy should not have discriminated against any of the databases.

The results are presented in Figure 11**.** The number of papers affiliated with institutions in the selected sample of 75 in MAG has been lower than in WoS and Scopus almost in every case. This discovery seems to be consistent with the problems experienced during querying. In MAG, affiliations were commonly dispersed, e.g. with departments existing as affiliations independent from the main university database record. Such case has been identified for example in the case of University of Cambridge, where many of the 31 university colleges, some of the departments (e.g. Department of Engineering or Department of Geography) or institutes (Cambridge Institute of Criminology) had been counted independently. In the case of Cambridge, the number of individual records of institutions being part of University of Cambridge (46) is significantly larger than such number in Scopus (19) or WoS (1). Even in a case of Ohio State University, where a corresponding number of records in MAG (8) was lower than in Scopus (13), the disambiguation has been considerably harder, as the Scopus portal presents information regarding the higher-level institution that a record is a feature lacking in MAG. It is worth noting that in both cases WoS has automatically grouped multiple sub-institutional profiles into a single profile of the searched affiliation. The lack of any hierarchy of the institutional entries in MAG makes it a harder to use source of information than both Scopus (where, although multiple entries exist and have to be manually selected and information regarding higher-level affiliation of those is presented) and Web of Science (where commonly sub-affiliations are grouped into a single institutional profile automatically).

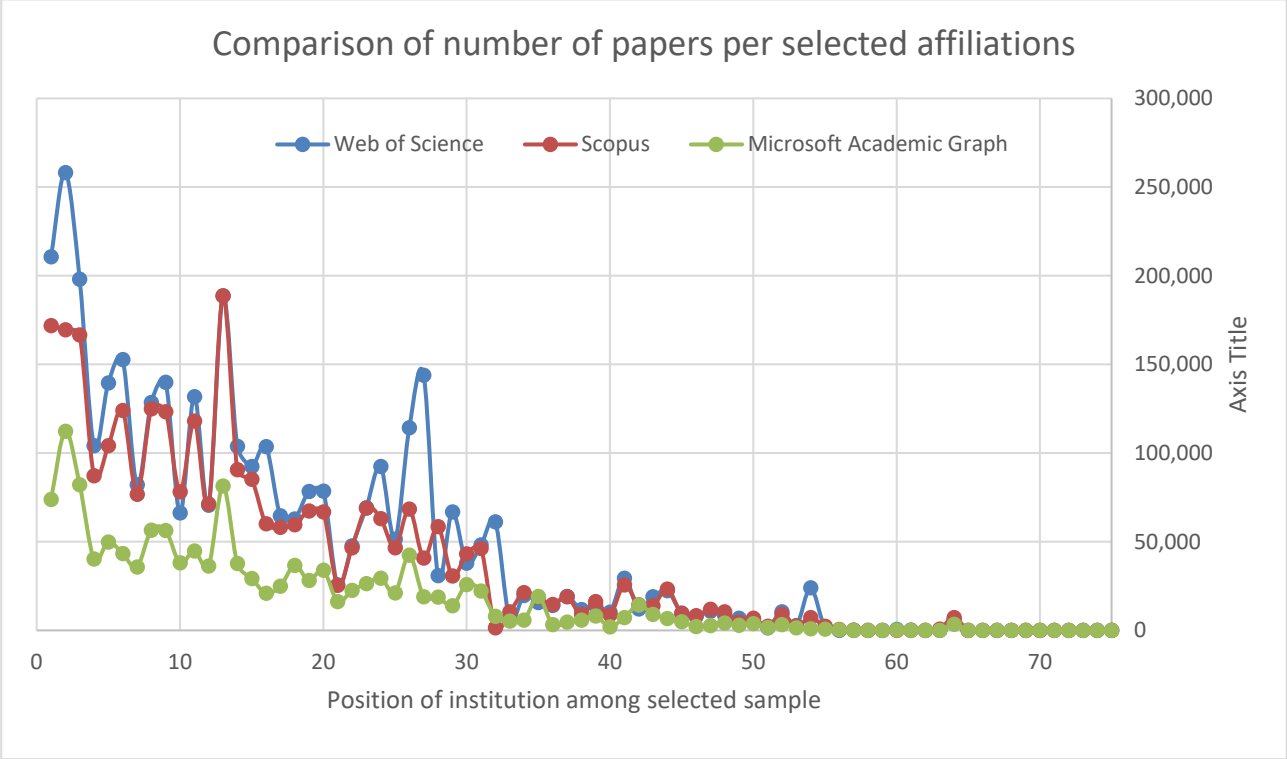

Figure 11 Comparison of number of papers per selected affiliations in databases

It has also been verified that the 'master' profile of an institution does not include the sub-institutional papers in the total count in Microsoft Academic portal. A joint query of 'Ohio State University' and 'Ohio State University College of Medicine' returns only 276 papers, authored by at least two people of which one is affiliated at the former and second at the latter. The problem with name disambiguation in MAG has been identified in 20 cases out of the studied 75 and has been particularly visible among the top-tier HEIs according to Webometrics Ranking (19 out of the sample of 25). It remains to be verified, however, whether querying the database via the available API improves the quality of results. Unfortunately, the fact that individual sub-institutional profiles are not hierarchically structured not only in Microsoft Academic Graph but also in the Microsoft Academic portal, suggests that the disambiguation of institutional profiles remains an issue in the service even when more advanced natural-language processing algorithms are used.

It is also worth noting that MAG superseded both Scopus and WoS in the identification of low-ranked institutions (between a 1000th and 12,000th place in Webometrics Ranking). Among the sample of 25 such HEIs, only five institutions have not been found in MAG, a number lower than the alternative datasets (10 affiliations missing Scopus and 16 from WoS). This effect can be observed as the non-existence of some data-points (due to a log-scale presentation of results) in Figure 12. A possible reason for the effect is that Microsoft Academic service is based on Bing algorithms parsing websites, which enables a broader

discoverability of various research outputs than the manually-curated Scopus and WoS, which are based on publication series (most commonly – journals) (Sinha 2015).

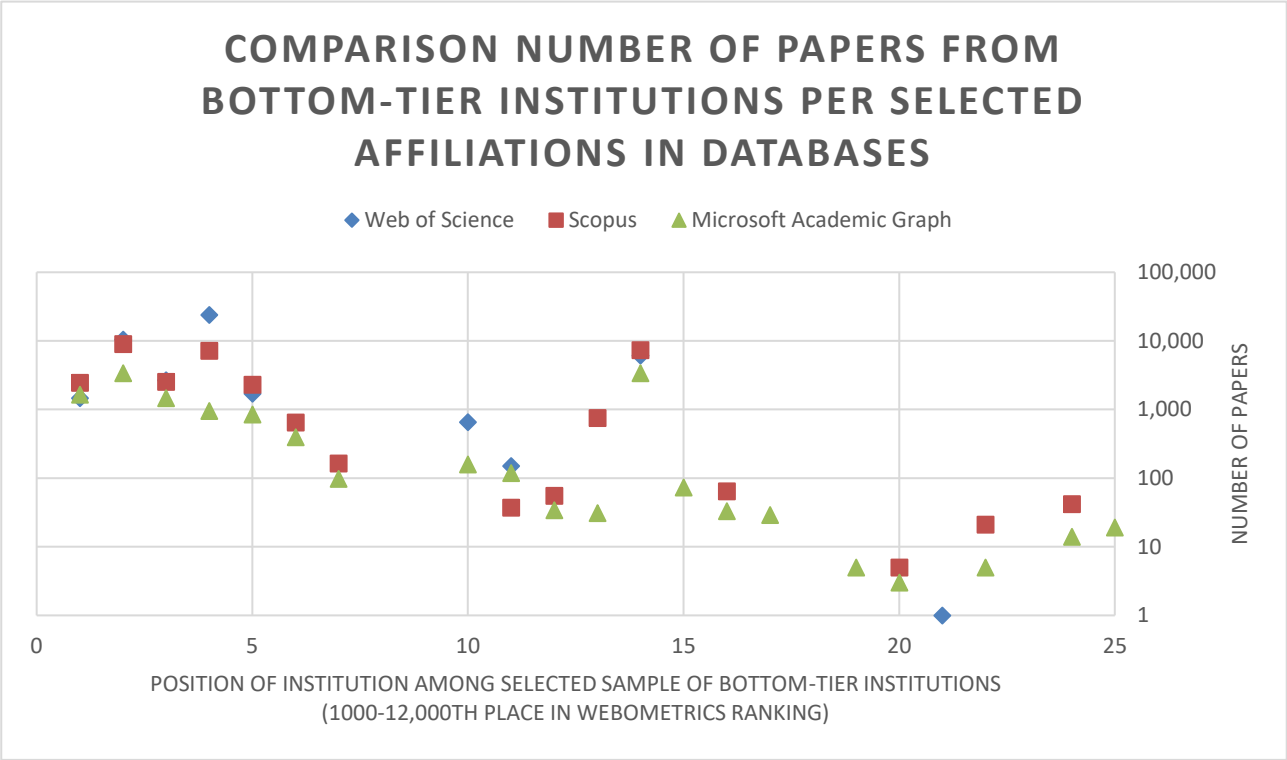

Figure 12 Number of papers of the twenty-five selected bottom-tier institutions, missing data points indicating lack of institutional profile in the given database

4.3.3 Authors and Citations

From the set of randomly selected authors, six author profiles were chosen due to the observable uniqueness of the combination of their initials and surname, enabling to retrieve papers authored by a single person. Regarding their fields of research, two of them were recognised as biologists, two as medics, one as a computer scientist and one physicist. Although their interests are not confined to a single discipline boundary, it has to be noted that the set is biased in the direction of natural sciences and completely excludes the fields of social sciences and humanities. Then, complete lists of documents authored by each of them have been retrieved from each of the databases. The set of articles has been manually compared, and a list of all papers authored by a given person has been constructed. The need for manual cleansing of data arises due to errors and differences in title parsing between the four sources and inclusion of duplicates, which had to be removed. Each questionable entry in the gathered data was manually verified online to see whether the title and authorship have been correctly recorded in the databases.

The first comparison focused on two options for retrieval of information from the Microsoft Academic Graph dataset: a local copy and usage of the API. The description of the API includes information that a natural language interpreter is included in its design[26], which may suggest better quality results on querying. Indeed, as presented in Table 7, the analysis conducted on a set of 164 documents of a single author, the performance of the API exceeds that of manually created MySQL database. Such state may find its' roots in errors of data pre-processing by the author of the thesis, but it can be at least partly blamed on the construction of the author profiles in MAG. Even though the author's name and surname had been carefully verified in order to make sure that it represents a single person, six profiles with the same name and surname combination have been found in a locally maintained copy of MAG.

Table 7 Comparison of Microsoft Academic data retrieved from a downloaded, local copy and information available from the API

	Downloaded MAG	***%***	**MAG accessed via API**	***%***
Total number of documents retrieved	99		132	
Unique documents in set w.r.t all other three datasets	9		7	
Total number of missing documents	64	*39.3%*	32	*19.5%*
Missing documents when compared to Web of Science	44	*50.0%*	9	*10.3%*
Missing documents when compared to Scopus	47	*37.0%*	16	*12.6%*
Missing documents when compared to Google Scholar	46	*39.3%*	25	*19.2%*
TOTAL NUMBER OF DOCUMENTS	163		164	

The multiplication of profiles can be explained as each profile in MAG is allowed only to have a single affiliation, hence after changing institution, each author 'earns' a new profile. As it was also later found, some problems with recognition and interpretation of language-specific characters (such as the Turkish g-brave, 'ğ', in the analysed case) has been experienced, although even in the case of API querying individual cases of wrongly disambiguated names remained. The number of documents missing from MAG is compared to a complete list of publications by analysed author constructed using all four datasets and individually to each dataset. In every case, the set of documents retrieved via API is more complete: the total number of missing papers is reduced by a half and better performance is observed in each of the

[26] https://www.microsoft.com/cognitive-services/en-us/academic-knowledge-api

individual comparisons. Although the analysis was conducted on a small set of 164 articles and one author, the visible superior performance of the Application Programming Interface led to choosing it as a source of information for further comparisons.

In order to study coverage of MAG when compared to the three competitors on a larger scale, a set of six uniquely-identifiable authors has been chosen. For each of the authors, a set of between 51 and 164 documents has been identified.

Table 8 Comparison of MAG to other databases using author query

	Web of Science	**Scopus**	**Google Scholar (excl. Citation-type)**	**Microsoft Academic Graph**
Number of documents	376	425	487	486
%	*58.8%*	*66.5%*	*76.2%*	*76.0%*
Missing documents	263	214	152	153
Unique documents	21	7	53	49
Citation count	7,180	9,566	13,244	6,474
Citation count (excluding best and worst case)	3,466	4,177	6,898	3,786
Found in [x], but missing from MAG	77	71	114	
%	*20.5%*	*16.7%*	*23.4%*	
Found in MAG, but missing from [x]	187	132	113	
%	*38.5%*	*27.2%*	*23.3%*	
Found in [x] and MAG	301	355	388	
TOTAL NUMBER OF PAPERS	639			

As presented in Table 8, the total number of documents identified in MAG has been almost equal to Google Scholar (486 vs. 487) and higher than for both Web of Science (376) and Scopus (425). It is worth mentioning that the numbers in the table relate to a list of consistent, cleaned documents (with multiple versions of a single publication merged into one). The striking similarity of breadth of Microsoft Academic dataset to the one found in Google Scholar is also observable in the number of records missing from the duo (153 and 152, respectively), number of papers unique (meaning a document only retrieved from the single source and lacking in all three remaining databases; 49 to 53), and those articles found in Google Scholar, but missing from MAG (114; 23.4% of GS set) and vice versa (113; 23.3% of MAG set). However, in terms of a total number of citations, Google managed to identify more than twice the number of those indexed by MAG. The number of overall unique documents is of particular importance, since it indicates the

volume of ‘new information’ brought to the table by using the given dataset as information source, hence the fact that numbers for Google Scholar (53 such documents) and MAG (49) are significantly higher than for Web of Science (21) or Scopus (7) indicates that MAG indeed can be a promising source of information. However, similarly to the concerns regarding the quality of data in Google Scholar, the sheer number of documents retrieved cannot be easily taken as a positive sign, as it may include wrongly parsed or non-scholarly publications.

Furthermore, in five out of six profiles analysed, Microsoft Academic enabled to identify more documents than either Web of Science or Scopus. Unsurprisingly, the total number of records unique in MAG compared to Web of Science (187) or Scopus (132) individually has been higher than the opposite (77 and 71, respectively). Again, this statistic reinforces the argument that MAG can be an interesting source of bibliometric information. However, relatively high numbers of unique documents on a comparison of pairs of databases indicates also that MAG cannot replace any of the datasets, as neglecting WoS or Scopus as information source would lead to the exclusion of a large set of author’s documents.

Finally, total citation counts in each of the databases are compared. Again, in five out of six cases (the sixth author being different to the sixth author in the paragraph above) author’s citation count retrieved from MAG was comparable to the other databases and was higher than at least one of the respective numbers in three cases. However, because the relatively lowest number of citations (less than a half of the next-worst citation count for the author) has been identified for a person with the highest number of total papers (164), the total summed citation score of MAG was significantly lower than any of the three other datasets. This may indicate the overall worse performance of MAG in terms of citation counting, but exclusion of both the worst and best cases (measured as relation of number of citations in MAG with respect to lowest citation count for the author) shows a different picture: the number of citations indexed by Microsoft Academic (equalling 3,786) is between than the one based on Web of Science (3,466) and Scopus (4,177). In both cases, the champion of is Google Scholar (6,898 in the latter analysis). Hence it may be that the performance of MAG in that matter is not as bad as it would look at first glance, but the tiny sample size prohibits definitive conclusions. The ‘wrongly-indexed’ profile is an interesting case itself: although the citation score in MAG was more than two times lower compared to the next-worst and three times lower than the count in top dataset, the total number of documents for this author found in MAG has been highest (132; Web of Science 87; Scopus 127; Google Scholar 130) and the number of papers not included in MAG the lowest (MAG – 32, WoS – 77, Scopus – 37, Google Scholar - 34).

However, the existence of such cases allows us to conclude that inconsistency in the number of registered citations is a problem in Microsoft Academic. Therefore, a broader study is advised to be able to conclude

whether MAG performance in the matter is on-par with Scopus or WoS. The study should also take into account profiles from a wider set of disciplines.

4.3.4 Papers

Finally, the papers missing from Microsoft Academic dataset but available in other databases were scrutinised in detail.

Table 9 Documents missing from MAG after performing an author query

	Web of Science	Scopus	Google Scholar (excl. 'citation'-type)
Found in [x]	77	71	114
Available in MAG when searched by title	56	62	75
without queried author information	54	56	66
with alternative queried author name	2	6	9
Missing from MAG	21	9	39
% of articles in [x]	5.59%	2.12%	8.01%
TOTAL documents missing from MAG	57		
% of total number of documents	8.92%		

Table 9 is used to present the information found. The majority of the documents which were not retrieved using author query were found using search by title. In those cases, a substantial majority of documents were found to be missing the queried author from the authors list, while a small minority included an alternative notation of author's name. It is worth noting that the proportion of papers not indexed at all by MAG relative to a total number of documents found in each of the three databases individually ranged between 2.12% and 8.01%, which indicates a high (91.99% to 97.88%) overlap between MAG and any of the other retrieved sets. Such a study shall be repeated on a larger scale to confirm such a good coverage of other datasets by MAG. At this point it should be reminded that publications marked as 'citations' (not directly indexed by Google, but found as references in other publications) have been excluded from the Google Scholar set, possibly lowering the Google Scholar performance. However, this should not have affected the overall evaluation of the Microsoft Academic Graph, since two other independent sources of information (Scopus and WoS) were used for comparison.

Table 10 Breakdown of types of documents missing from MAG

Type of research output	Count
Journal papers (in English, peer-reviewed)	20
Journal papers (non-English, peer-reviewed)	5
Preprints (from repositories)	8
Books & book chapters	2
Conference papers	2
Meeting abstracts	15
Other (erratum, book review)	5
TOTAL MISSING	**57**

The form considered to be traditional research output is a peer-reviewed paper published in journals. Twenty-five of this type of documents were not found in MAG, of which five were published in a language other than English. Other missing documents included eight repository papers and 15 meeting abstracts (which were mainly indexed by Web of Science), as shown in Table 10. The data further confirms that MAG cannot be considered a complete replacement of any of the other datasets, even if only traditional forms of research output (journal papers) are considered.

4.3.5 Disciplinary classification

The Web of Science classification consists of around 250 categories in sciences, social sciences, and humanities. It is based on the so-called Hayne-Coulson algorithm, details of which have never been published. Besides, the classification system is constructed based on citation patterns, journal titles and expert review (Leydesdorff & Rafols 2009). Scopus classification – All Science Journal Classification – consists of two levels. The lower one, containing 304 categories, and a higher level composed of 27 fields of study. No details regarding the backbones of the Scopus classification are published, but a study comparing the above mentioned two systems found that WoS is significantly more accurate (Wang & Waltman 2016).

The MAG classification consists of four levels of categories, with the top level supposedly corresponding to the 19 disciplines listed in the Microsoft Academic portal. In their description of the database, Sinha et al. (2015) described the process of assigning fields of study (FOS) to individual papers. Before the operation, only around 5% papers have their FOS assigned. To assign the category to other papers, a 'seeding' method is followed based on the papers mentioned above with assigned discipline and ones with specified keywords, which can be mapped to a FOS. Then, using the already amassed knowledge on the relationship between categories and algorithms FOS candidates are assigned. Sinha et al. claim the process enables

them to classify all papers with a 98% accuracy (Sinha et al. 2015), a claim which is to be verified during this study.

Chapter 5: Conclusions

The aim of the project is to provide an insight into the internal structure and scope of the Microsoft Academic Graph database. The analysis has been conducted using a set of 75 institutional affiliations, 6 authors, and their 639 documents. Additionally, a basic large-scale analysis regarding the relationships between types of entities inside the database has been conducted. The overall results indicate that Microsoft Academic Graph can be an interesting source of information for bibliometric or scientometric analyses. The total number of indexed documents (126 million) is lower than the estimated number for Google Scholar (160-165 million) but considerably higher than relative numbers for Web of Science (57 million) and Scopus (53 million). The coverage of total research output (including not only peer-reviewed journal papers, but also other types of documents, such as books, reviews, letters) of the six selected authors had reached 76.0%, hence being on-par with coverage of Google Scholar (76.2%) and significantly better than that of Scopus (66.5%) and Web of Science (58.8%). However, the performance regarding disambiguation of affiliations or authors has been shown to be at least inconsistent.

5.1 Openness, transparency, and interoperability

The openness of databases has been scrutinised. Web of Science, Scopus, and Microsoft Academic are openly publishing either the lists of included journals or participants in data gathering. Algorithmic gathering of publications by MA can also be traced, as URLs are provided for the indexed papers in the complete downloadable Microsoft Academic Graph. Google Scholar's transparency has to be described as poor because no analysis of sources of included publications on a large scale can be conducted.

Furthermore, the Microsoft Academic ranks highest in terms of open access to the data. While Scopus and Web of Science at least partially restrict access to their search portals and APIs to paying subscribers of their products, Microsoft Academic provides a free downloadable version of the dataset and unrestricted access to search portal. MAG API is also open to the public, although it contains a limit of 10,000 queries monthly and statistics regarding the whole dataset cannot be obtained this way.

Regarding interoperability, MAG is still lagging behind Scopus or Web of Science, as it only provides article unique identifier (DOI), without providing publication series identifiers (such as ISSN) or authors identifiers (ORCID). None of the datasets enables to retrieve institutional id (e.g. ISNI). The lack of interoperability hinders the prospects of usage of the MAG dataset in conjunction with other databases, including services focusing on alternative metrics, forcing researchers to compare lists of retrieved authors, publications or institutions manually.

5.2 Affiliation search

Regarding the affiliation classification, it has been observed that the numbers of papers linked with the institution of the author were lower in MAG than in Scopus or Web of Science, despite the manual joining of sub-affiliations, which are not connected to their mother institutional profile. The difference between traditional citation databases and MAG hampers the opportunity for use in bibliometric or scientometric studies. Construction of a hierarchy of affiliations stored in MAG, which would enable automatic linking of sub-institutions (departments and institutes) to top-level institutions (mother institutions, such as universities) would probably prove useful. At the same time, the total number of papers affiliated with any institution, indexed with a name in MAG, is under 29 million – a significantly smaller number than the total of around 126 million papers indexed by the database. Such discrepancy implies problems with retrieval of metadata from the documents stored.

5.3 Author search

The performance of author search has been found to be of higher quality, although not completely clear of issues similar to the ones described above. Overall, the number of papers retrieved from MAG matched the numbers for Google Scholar (considered the biggest of the databases), with a similar number of relatively unique documents. The numbers of publications and unique documents were also higher in MAG than in Web of Science and Scopus. Therefore, it can be concluded that we are observing formation of two separate groups of scholarly publications and citations datasets: traditional (based on manual verification of quality of journals and conditional approval for inclusion of its publications, such as Scopus or WoS) and web-based (with extensive use of automatic algorithms for identification of scholarly content on the Web, such as Google Scholar and MAG). The latter group is showing a potential for a broader measurement of scientific impact by the inclusion of more diverse research output content (such as theses, conference papers, see Harzing & van der Wal 2008) and incoming citations from a wider variety of sources. On the other hand, automatic recognition and parsing of scholarly content can bring reliability issues concerning duplicate entries or wrongly identified metadata.

These were to some extent the problems in the performance of author queries. In the initial attempt to search by author name, the local version of MAG performed poorly. As an example, one of the authors queried was found to have six independent profiles in the database, which were not linked to each other. The existence of some the profiles can be explained by the fact that only a single affiliation per profile is allowed, but the lack of a hierarchical structure or even links between profile makes author search in MAG considerably more challenging than in Scopus or WoS. Much better results were obtained using the API,

probably due to capabilities of the natural language parsers possibly being able to link independent profiles and take into account names variants.

Finally, out of the 153 documents missing from the set obtained via API (out of the total of 639), 96 were found to exist in Microsoft Academic. However, the initial query did not retrieve them because they were missing the sought author from authors list or author's name spelling has differed. Therefore, improvement the performance of either the parsing or query processing algorithms is sought to allow reliable querying by affiliations or authors. An addition of tables enabling to link affiliations or authors profiles to the downloadable version of MAG would help address the situation, even if the relations can be established only with certain probability. A similar kind of table is already provided and describes the Fields of Study (disciplinary classification) hierarchy, enabling to link sub-disciplines to major fields.

5.4 Papers and citation count

Regarding total coverage of documents, MAG was found to include 91.99% of Google Scholar, 94.01% of Web of Science, and 97.88% of Scopus documents, indicating that in terms of a total number of publications Microsoft Academic can be considered a reliable source of information, comparable to the other three datasets. However, out of the total of 57 of documents missing from MAG, as much as 25 are peer-reviewed journal publication – type of a document which cannot be omitted if a complete analysis of publications is to be reliable. Therefore, although the performance of MAG has exceeded that of Scopus or WoS, the database itself cannot be considered a replacement for them.

Regardless of the higher number of documents indexed compared to WoS or Scopus, the total number of citations of those documents indexed by MAG remained low. Even after exclusion of the best and the worst cases from the dataset, the number of citations (3,786) stayed between the score of WoS (3,466) and Scopus (4,177). The number was still lower than the respective number for Google Scholar (6,898) despite a similar number of covered documents in both datasets. Furthermore, in one of the six author cases the MAG citation score had been between two and three times lower than the other scores, indicating inconsistency in MAG performance. It is possible that the underperformance of MAG is due to it being a new project (launched a year ago) and may be quickly improved. Nevertheless, the current low citation scores should be considered the most serious drawback of Microsoft Academic from the perspective of usage of its data for studies.

The Microsoft Academic API service has performed better than the downloaded version of a database loaded to MySQL. However, it does not allow to carry out some large-scale analyses. As an example, the

free version of an API did not allow to gather data for a histogram of papers per author or publications by year on the whole dataset.

5.5 Limitations of the study and further research

The numbers of documents indexed by MAG and other databases do not tell the whole story. Bearing in mind problems with the quality of indexed content identified in Google Scholar (e.g. Aguillo 2012; de Winter et al. 2014), a careful analysis of the research output uniquely found in MAG, or MAG and GS, has to follow. Such analysis has to focus on identification of documents lacking the rigor demanded by science, as those would have to be removed from any bibliometric or scientometric study using these datasets. A detailed description of the documents indexed by MAG, but missing in WoS or Scopus would also help in addressing the question of usability of that database in scientometric studies or research evaluation.

It has to be taken into account that the study focused on a minuscule set of only six authors and 639 documents. Repetition of such analysis on a larger scale is needed before definitive conclusions can be drawn. However, the extent of this study was limited due to the fact that each entity had to be manually verified. The following questions had to be answered in each case: whether an independent institutional profile should be merged with the main profile of an institution, whether author's name is unique, and whether each found document was indeed authored by the sought person. Therefore, other approaches may be less time-consuming and allow for a large-scale study. Possibly, using a set of documents obtained by a discipline-specific keyword search would enable not only a broader study but also analysis of disciplinary coverage of the databases.

Future analysis should also look into further detail on the disciplinary classification, systematically dividing the sets of authors and/or publications selected according to disciplines. Comparison of a number of papers found using a disciplinary-specific keyword search in each of the four databases should be made to study differences in coverage of the dataset depending on the field of research. Also, an analysis of the number of publications and citations per author shall be repeated including authors from a more diverse range of disciplines (especially from outside the natural sciences, as mentioned in section 3.4.2). Such approaches are interesting areas of study, as the performance of the other databases had been previously found to significantly vary between disciplines, as was mentioned in Section 2.3.1.

Bibliography

Aad, G. et al., 2012. Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. *Physics Letters, Section B: Nuclear, Elementary Particle and High-Energy Physics*, 716(1), pp.1–29. Available at: https://arxiv.org/abs/1207.7214.

Aguillo, I.F., 2012. Is Google Scholar useful for bibliometrics? A webometric analysis. *Scientometrics*, 91(2), pp.343–351. Available at: http://link.springer.com/article/10.1007/s11192-011-0582-8.

Aguillo, I.F., Ortega, J.L. & Fernández, M., 2008. Webometric Ranking of World Universities: Introduction, Methodology, and Future Developments. *Higher Education in Europe*, 33(2-3), pp.233–244. Available at: http://www.tandfonline.com/doi/abs/10.1080/03797720802254031?journalCode=chee20.

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

Bar-Ilan, J., 2008. Which h-index? - A comparison of WoS, Scopus and Google Scholar. *Scientometrics*, 74(2), pp.257–271. Available at: http://www.zalf.de/de/institute_einrichtungen/bib/Documents/BibliometrischeIndizes/Bar-Ilan_2008_h-factor.pdf.

Bartling, S. & Friesike, S., 2014. *Opening Science* S. Bartling & S. Friesike, eds., Available at: http://book.openingscience.org/.

Bartol, T. et al., 2014. Assessment of research fields in Scopus and Web of Science in the view of national research evaluation in Slovenia. *Scientometrics*, 98(2), pp.1491–1504. Available at: http://link.springer.com/article/10.1007/s11192-013-1148-8.

Berners-Lee, T. et al., 1994. The World-Wide Web. *Communications of the ACM*, 37(8), pp.76–82. Available at: http://www.sciencedirect.com/science/article/pii/016975529290039S.

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

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

Burnham, J.F., 2006. Scopus database: a review. *Biomedical digital libraries*, 3, p.1. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16522216.

Byrnes, J. et al., 2015. The Four Pillars of Scholarly Publishing: The Future and a Foundation. *PeerJ PrePrints*, pp.1–17. Available at: https://peerj.com/preprints/11/.

Caragea, C. et al., 2014. CiteSeerX: A Scholarly Big Dataset. *Advances in Information Retrieval*, 8416, pp.311–322. Available at: www.cse.unt.edu/~ccaragea/papers/ecir14.pdf.

Costas, R., Zoreh, Z. & Wouters, P., 2015. Do 'Altmetrics' Correlate With Citations? Extensive Comparison of Altmetric Indicators With Citations From a Multidisciplinary Perspective. *Journal of the Association for Information Science and Technology*, 66(10), pp.2003–19. Available at: https://arxiv.org/abs/1401.4321.

Falagas, M.E. et al., 2008. Comparison of PubMed, Scopus, Web of Science , and Google Scholar: strengths and weaknesses. *The FASEB Journal*, 22(2), pp.338–342. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17884971.

Friedman, A., 2015. The Power of Lotka's Law Through the Eyes of R. *Romanian Statistical Review*, 63(2), pp.69–77. Available at: http://scholarcommons.usf.edu/si_facpub/135/.

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

Giles, C.L., Bollacker, K.D. & Lawrence, S., 1998. CiteSeer: An Automatic Citation Indexing System. *ACM Conference on Digital Libraries*, pp.89–98. Available at: https://clgiles.ist.psu.edu/papers/DL-1998-citeseer.pdf.

Harzing, A., 2014. A longitudinal study of Google Scholar coverage. *Scientometrics*, 94(3), pp.1057–1075. Available at: http://link.springer.com/article/10.1007/s11192-013-0975-y.

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

Harzing, A.W., 2013. A preliminary test of Google Scholar as a source for citation data: A longitudinal study of Nobel prize winners. *Scientometrics*, 94(3), pp.1057–1075. Available at: http://link.springer.com/article/10.1007/s11192-012-0777-7.

Harzing, A.-W., 2017. Microsoft Academic (Search): a Phoenix arisen from the ashes? *in press for Scientometrics, available online*. Available at: http://www.harzing.com/blog/2016/06/microsoft-academic-search-a-phoenix-arisen-from-the-ashes.

Harzing, A.-W., 2010. The Publish or Perish Book. *Publish*, 7(7), p.250. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0960982206003861.

Hicks, D. et al., 2015. Bibliometrics: The Leiden Manifesto for research metrics. *Nature*, 520(7548), pp.429–431. Available at: http://www.nature.com/news/bibliometrics-the-leiden-manifesto-for-research-metrics-1.17351.

Jacsó, P., 2011. The pros and cons of Microsoft Academic Search from a bibliometric perspective. *Online Information Review*, 35(6), pp.983–997. Available at: https://www.researchgate.net/publication/241699155_The_pros_and_cons_of_Microsoft_Academic_Search_from_bibliometric_perspective.

Jamali, H.R. & Nabavi, M., 2015. Open access and sources of full-text articles in Google Scholar in different subject fields. *Scientometrics*, 105(3), pp.1635–1651. Available at: http://dx.doi.org/10.1007/s11192-015-1642-2.

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

Jordan, K., 2015. Exploring the ResearchGate score as an academic metric: Reflections and implications for practice. *Quantifying and Analysing Scholarly Communication on the Web (ASCW'15)*, pp.1–3. Available at: http://ascw.know-center.tugraz.at/wp-content/uploads/2015/06/ASCW15.

Khabsa, M. & Giles, C.L., 2014. The number of scholarly documents on the public web. *PLoS ONE*, 9(5).

Available at: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0093949.

Kousha, K. & Thelwall, M., 2008. Sources of Google Scholar citations outside the Science Citation Index: A comparison between four science disciplines. *Scientometrics*, 74(2), pp.273–294. Available at: http://link.springer.com/article/10.1007/s11192-008-0217-x.

Kraker, P. & Lex, E., 2015. A Critical Look at the ResearchGate Score as a Measure of Scientific Reputation. *ASCW'15 Workshop at Web Science 2015*, (May), pp.7–9. Available at: http://ascw.know-center.tugraz.at/2015/05/26/kraker-lex-a-critical-look-at-the-researchgate-score/.

Leydesdorff, L., 2008. Caveats for the use of citation indicators in research and journal evaluations. *Journal of the American Society for Information Science and Technology*, 59(2), pp.278–287. Available at: https://arxiv.org/abs/0911.1440.

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

Lin, J. & Fenner, M., 2013. Altmetrics in Evolution: Defining and Redefining the Ontology of Article-Level Metrics. *Information Standards Quarterly*, 25(2), pp.20 – 26. Available at: http://www.niso.org/apps/group_public/download.php/11273/IP_Lin_Fenner_PLOS_altmetrics_isqv25no2.pdf.

López-Illescas, C., Moya-Anegón, F. & Moed, H.F., 2009. Comparing bibliometric country-by-country rankings derived from the Web of Science and Scopus: the effect of poorly cited journals in oncology. *Journal of Information Science*, 35(2), pp.244–256. Available at: http://jis.sagepub.com/content/35/2/244.

Lucio-arias, D., Velez-cuartas, G. & Leydesdorff, L., 2015. SciELO Citation Index and Web of Science: Distinctions in the Visibility of Regional Science. *Proceedings of ISSI 2015*, (JUNE), pp.1152–1160. Available at: http://www.issi2015.org/files/downloads/all-papers/1152.pdf.

Mikki, S., 2010. Comparing Google Scholar and ISI Web of Science for earth sciences. *Scientometrics*, 82(2), pp.321–331. Available at: https://www.researchgate.net/publication/220365311_Comparing_Google_Scholar_and_ISI_Web_of_Science_for_Earth_Sciences.

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

Mingers, J. & Lipitakis, E.A.E.C.G., 2010. Counting the citations: A comparison of Web of Science and Google Scholar in the field of business and management. *Scientometrics*, 85(2), pp.613–625. Available at: http://link.springer.com/article/10.1007/s11192-010-0270-0.

Moed, H.F. & Visser, M., 2008. *Appraisal of Citation Data Sources: A report to HEFCE by the Centre for Science and Technology Studies, Leiden University*, Available at: http://www.hefce.ac.uk/pubs/rdreports/2008/rd17_08/.

Orduna-Malea, E. et al., 2015. Methods for estimating the size of Google Scholar. *Scientometrics*, 104(3), pp.931–949. Available at: http://arxiv.org/abs/1506.03009.

Orduña-Malea, E. et al., 2014. The silent fading of an academic search engine: the case of Microsoft Academic Search. *Online Information Review*, 38(7), pp.936–953. Available at: http://arxiv.org/abs/1404.7045.

Ortega, J.L. & Aguillo, I.F., 2014. Microsoft Academic Search and Google Scholar citations: comparative analysis of author profiles. *Journal of the Association for Information Science and Technology*, 65(6), pp.1149–1156. Available at: http://onlinelibrary.wiley.com/doi/10.1002/asi.23036/abstract.

Priem, J. et al., 2010. Altmetrics: A manifesto. Available at: http://altmetrics.org/manifesto/ [Accessed August 8, 2016].

Prins, A.A.M. et al., 2016. Using Google Scholar in research evaluation of humanities and social science programs: A comparison with Web of Science data. *Research Evaluation*, (February). Available at: http://rev.oxfordjournals.org/content/early/2016/02/02/reseval.rvv049.

Sinha, A. et al., 2015. An Overview of Microsoft Academic Service (MAS) and Applications. *Proceedings of the 24th International Conference on World Wide Web Companion (WWW 2015 Companion)*, pp.243–246. Available at: http://research.microsoft.com/apps/pubs/default.aspx?id=246609.

de Solla Price, D., 1983. *Little Science, Big Science ...and Beyond*, Columbia University Press; Revised edition edition (17 Sept. 1986). Available at: http://www.garfield.library.upenn.edu/lilscibi.html.

de Solla Price, D.J., 1962. Science since Babylon. *Technology and Culture*, 3(2), p.175. Available at: http://www.jstor.org/stable/3101441?origin=crossref.

Stephan, P., 2012. *How Economics Shapes Science* 1st ed., Harvard University Press. Available at: http://www.hup.harvard.edu/catalog.php?isbn=9780674088160.

Tang, L. & Walsh, J.P., 2010. Bibliometric fingerprints: Name disambiguation based on approximate structure equivalence of cognitive maps. *Scientometrics*, 84(3), pp.763–784. Available at: http://link.springer.com/article/10.1007/s11192-010-0196-6.

Treeratpituk, P. & Giles, C.L., 2009. Disambiguating authors in academic publications using random forests. *Proceedings of the 2009 joint international conference on Digital libraries - JCDL '09*, pp.39–48. Available at: http://portal.acm.org/citation.cfm?doid=1555400.1555408.

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

Wang, Q. & Waltman, L., 2016. Large-scale analysis of the accuracy of the journal classification systems of Web of Science and Scopus. *Journal of Informetrics*, 10(2), pp.347–364. Available at: http://dx.doi.org/10.1016/j.joi.2016.02.003.

Wilsdon, J., 2015. *The Metric Tide: Report of the Independent Review of the Role of Metrics in Research Assesment and Management*, Available at: DOI: 10.13140/RG.2.1.4929.1363.

de Winter, J.C.F., Zadpoor, A.A. & Dodou, D., 2014. The expansion of Google Scholar versus Web of Science: A longitudinal study. *Scientometrics*, 98(2), pp.1547–1565. Available at: http://link.springer.com/article/10.1007/s11192-013-1089-2.

Wouters, P. et al., 2015. *The Metric Tide: Literature Review (Supplementary Report I to the Independent Review of the Role of Metrics in Research Assessment and Management)*, Available at: www.nationalarchives.gov.uk/doc/open-government-licence/version/2.

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

Appendix A Breakdown of files and columns available in the downloadable version of MAG

#File

#Column number Column description

2016 KDD Cup Selected Affiliations

1	Affiliation ID
2	Affiliation name

2016 KDD Cup Selected Papers

1	Paper ID
2	Original paper title
3	Paper publish year
4	Conference series ID mapped to venue name
5	Conference series short name (abbreviation)

Affiliations

1	Affiliation ID
2	Affiliation name

Authors

1	Author ID
2	Author name

Conference Series

1	Conference series ID
2	Short name (abbreviation)
3	Full name

Conference Instances

1	Conference series ID
2	Conference instance ID
3	Short name (abbreviation)
4	Full name
5	Location
6	Official conference URL
7	Conference start date
8	Conference end date
9	Conference abstract registration date
10	Conference submission deadline date
11	Conference notification due date
12	Conference final version due date

Fields Of Study

1	Field of study ID
2	Field of study name

Field Of Study Hierarchy

1	Child field of study ID
2	Child field of study level
3	Parent field of study ID
4	Parent field of study level
5	Confidence

Journals

1 Journal ID

2 Journal name

Papers

1 Paper ID

2 Original paper title

3 Normalized paper title

4 Paper publish year

5 Paper publish date

6 Paper Document Object Identifier (DOI)

7 Original venue name

8 Normalized venue name

9 Journal ID mapped to venue name

10 Conference series ID mapped to venue name

11 Paper rank

PaperAuthorAffiliations

1 Paper ID

2 Author ID

3 Affiliation ID

4 Original affiliation name

5 Normalized affiliation name

6 Author sequence number

PaperKeywords

1 Paper ID

2 Keyword name

3 Field of study ID mapped to keyword

Paper References

1 Paper ID

2 Paper reference ID

Paper URLs

1 Paper ID

2 URL