PREPRINT. PLEASE CITE AS:
Orduna-Malea, E.; Martin-Martin, A.; Ayllón, Juan M.; Delgado López-Cózar, E. (2014). "The silent fading of an academic search engine: the case of Microsoft Academic Search". *Online information review*, 38(7), pp. 936-953.

The silent fading of an academic search engine: the case of Microsoft Academic Search

Enrique Orduña-Malea*
EC3 Research Group
Polytechnic University of Valencia
Valencia, Spain

Alberto Martín-Martín, Juan M. Ayllón and Emilio Delgado-López-Cózar
EC3 Research Group
University of Granada
Granada, Spain

Acknowledgements

This research was funded under Project HAR2011-30383-C02-02 from Dirección General de Investigación y Gestión del Plan Nacional de I+D+I (Ministry of Economy and Competitiveness) and Project APOSTD/2013/002 from the Regional Ministry of Education, Culture and Sport (Generalitat Valenciana) in Spain.

About the authors

*__Enrique Orduña-Malea__ holds a PhD in documentation from the Polytechnic University of Valencia where he currently works as a postdoctoral researcher and lecturer, particularly in the EC3 Research Group. He specialises in web metrics, particularly in the creation, diffusion and consumption of content and products on the web. Dr Orduña-Malea is the corresponding author and may be contacted at enorma@upv.es.

Alberto Martín-Martín is an FPU (university professor training) research fellow and PhD candidate in the field of bibliometrics and scientific communication at the University of Granada. His earlier degrees in information science and scientific information are from the same university. He is also a member of the EC3 Research Group.

Juan M. Ayllón is an FPI (predoctoral research grant) research fellow and a PhD candidate in the field of bibliometrics and scientific communication at the University of Granada. His earlier degrees in information science and scientific information are from the same university. He is also a member of the EC3 Research Group.

Emilio Delgado-López-Cózar is Professor of Research Methods at the University of Granada and founder of the EC3 Research Group (Science and Scientific Communication Evaluation), specialising in bibliometrics and research evaluation. He is a principal creator of numerous tools for scientific evaluation in the Spanish environment, including IN-RECS, IN-RECJ, IN-RECH, I-UGR Ranking of Spanish Universities, H Index Scholar, Integrated Classification of Scientific Journals or EC3 Meta-ranking of Spanish Universities, among others.

Paper received 24 July 2014
Second revision approved 21 September 2014

Abstract

Purpose – The main objective of this paper is to describe the obsolescence process of Microsoft Academic Search (MAS) as well as the effects of this decline on the coverage of fields and journals, and their influence on the representation of organisations.

Design/methodology/approach – The total number of records and those belonging to the most reputable journals (1,762) and organisations (346), according to the Field Rating indicator in each of the 15 fields and 204 sub-fields of MAS, were collected and statistically analysed in March 2014, by means of an automated querying process via http, covering academic publications from 1700 to the present.

Findings – Microsoft Academic Search has not been updated since 2013, although this phenomenon began to be glimpsed in 2011, when its coverage plummeted. Throughout 2014, indexing of new records is still ongoing, but at a minimal rate, without following any apparent pattern.

Research limitations/implications – There are also retrospective records being indexed at present. In this sense this research provides a picture of what MAS offered during March 2014 when queried directly via http.

Practical implications – The unnoticed obsolescence of MAS affects the quality of the service offered to its users (both those who engage in scientific information seeking and also those who use it for quantitative purposes).

Social implications – The predominance of Google Scholar as a monopoly in the academic search engines market as well as the prevalence of an open construction model versus a closed model (MAS).

Originality/value – A complete longitudinal analysis of fields, journals and organisations on MAS has been performed for the first time, identifying an unnoticed obsolescence. There has not been any public explanation or disclaimer note announced by the company responsible, which is incomprehensible given its implications for the reliability and validity of the bibliometric data provided on fields, journals, authors and conferences as well as their fair representation by the search engine.

Keywords Academic search engines, Microsoft Academic Search, Google Scholar, Scientific fields, Academic journals, Universities
Article classification Research paper

Introduction

At the beginning of the second decade of the twenty-first century the two major academic search engines with information about scientific citations are Google Scholar (GS) and Microsoft Academic Search (MAS), developed by two companies (Google and Microsoft). They compete not only in the design of these tools but in a wide range of products and web services, with the competition between their search engines (Google and Bing) being especially important for our research area.

Google Scholar, launched in 2004 (Jacsó, 2005; Mayr and Walter, 2005), constituted a great revolution in the retrieval of scientific literature, since for the first time bibliographic search was not limited to the library or to traditional bibliographic databases. Instead, because it was conceived as a simple and easy-to-use web service, Google Scholar enabled simple bibliographic search for everyone with access to the web. This was the birth of academic search engines (Ortega, 2014a), and secondarily, of academic search engine optimisation, which can be defined as "the creation, publication, and modification of scholarly literature in a way that makes it easier for academic search engines to both crawl it and index it" (Beel *et al*., 2010, p. 177). It has been noted, however, that such optimisation may occasionally be implemented for illegitimate purposes (Labbé, 2010; Delgado López-Cozar *et al.*, 2014), for example aiming to cheat academic search engines, something that is more difficult to achieve in traditional bibliometric databases.

Google Scholar, made in the image and likeness of its parent product (Google), started offering simple services to facilitate the search of academic papers (a search box and little else). It maintained a beta version from 2004 to 2011

(http://googlescholar.blogspot.com.es/2012/05/our-new-modern-look.html), with which GS started gaining users. According to Compete.com the number of unique users in May 2013 for the URL scholar.google.com amounted to 1,665,193 while in May 2014 it reached 2,427,903, considering only US data and apart from the web traffic generated by local versions of Google Scholar (https://siteanalytics.compete.com/scholar.google.com). The inclusion of the advanced search option (which enables users to search by author, publication year or limiting the search to the entire document or only the title) came later (Jacsó, 2008a; Beel and Gipp, 2009).

Microsoft's response came two years after, when the company announced in 2006 a new product, different and distant from Google Scholar's philosophy. It was named Windows Live Academic Search (Carlson, 2006), changing its name later to Live Search Academic (Jacsó, 2008b, 2010) and finally converting it late in 2009 into Microsoft Academic Search (Jacsó, 2011) after a complete redesign of the service carried out by its affiliate, the Microsoft Asia Research Group in China.

As Ortega and Aguillo (2014) state, the coverage of MAS at the beginning was limited to the computer science and technology fields, expanding in March 2011 to other categories thanks to agreements with different source providers, becoming a platform oriented to the identification of the top papers, authors, conferences and organisations (including universities, research institutes or companies) in 15 fields of knowledge and more than 200 sub-fields (http://social.microsoft.com/Forums/en-US/mas/thread/bf20d54a-ede2-48a9-8bbb-f6c1c1f30429). It provided both the bibliographic description of the publications and their citation counts. In short it offered everything needed to identify the most relevant research and to carry out comparative performance assessments.

Over two years Microsoft improved at a relentless pace not only the site navigation and browsing capabilities, but also the bibliometric performance indicators and especially the visualisation options (maps of publications, authorship, citation graphs, organisation comparisons, etc.). Google's response came in two stages. First, with the launch of Google Scholar Citations (Ortega and Aguillo, 2013), first restricted to a test group of users in July 2011 and then available to everyone in November 2011 (http://googlescholar.blogspot.com/2011/11/google-scholar-citations-open-to-all.html). This was a web service designed for authors to compute their citation metrics (h-index, the i-10 index and the total number of citations) and track them over time (each metric is computed over all citations and also over citations in papers published in the last five years). Second, they released Google Scholar Metrics

(http://scholar.google.com/citations?view_op=top_venues), available since April 2012 and focused on providing a means for identifying the most influential academic journals in different countries and scientific specialties (Jacsó, 2012; Delgado López-Cózar and Cabezas-Clavijo, 2013; Orduña-Malea and Delgado López-Cózar, 2014), thus setting a new competitive landscape, as was reported in the specialised press (Butler, 2011).

Nonetheless, Google Scholar and Microsoft Academic Search differ greatly in their construction. Table 1 presents a comparison of the main characteristics of Google Scholar and Google Scholar citations (as personal profiles are directly related to the indexation of new records) and Microsoft Academic Search.

Table 1. Comparison of the different features of Google Scholar and Google Scholar citations and Microsoft Academic Search

Features	GS and GSC	MAS
Service integration	Different products for papers, authors and journals	The same product for papers, authors and journals
Inclusion guidelines	a) Trusted sources: Publishers that cooperate directly with GS, and publishers and webmasters who have requested that GS crawl their databases. b) Invited articles: papers cited by papers indexed from trusted sources	MAS gathers bibliographic information (metadata) from the principal scientific publishing (Elsevier, Springer) and bibliographic services (CrossRef).
Profile registration	The profile must be intentionally created by the user.	Profiles are automatically created from the signatures of each author that appear in a new paper indexed.
Author profile keywords	Directly supplied by the authors	Automatically added by MAS.
Subject classification	GS does not use any subject classification.	MAS uses its own classification scheme based on 15 disciplines and more than 200 sub-domains.
Document types	Uncontrolled: all rich files stored in the trusted source hosting (PDF,	Controlled: journal papers, conference proceedings, reports,

	DOC, PPT, PS, XLS, etc.).	white papers, and a variety of other content types
Multiple versions	GS groups multiple versions into one record, establishing one version as principal, which is not necessarily the publisher's version.	Multiple versions are grouped, selecting the published journal publication as definitive. The option "view publication" provides access to all versions.
Author data editing	Authors edit their personal data themselves after signing into Gmail.	Authors edit their personal data via a request after signing into MAS.
Publication information editing	The authors edit publication metadata (Add, Change, Merge, etc.) themselves after signing into Gmail.	The authors edit their publications metadata (Add, Change, Merge, etc.) themselves after signing into MAS.
Automated query	API is not allowed.	API is allowed.
No. of results retrieved	1,000 limit	No limit
Size	Officially: unknown	Officially: 39.85 million (Nov 2013)

The parameter of size (a Google trade secret) reflects the major difference between the two products: absolute transparency in MAS against almost nonexistent transparency in GS. Microsoft declares its sources (publishers, repositories, etc.) with great detail and absolute precision as well as all the benefits it provides. We know the size of the database and how it grows.

Despite this clarity and transparency, a routine check performed in March 2014 unexpectedly yielded confused and contradictory results about the real size of MAS. On the one hand the data displayed on the official website (http://academic.research.microsoft.com/About/Help.htm) declares as of September 2011 that "the number of publications increases to 35.3 million".

Throughout 2012 there is no further information, and finally as of January 2013 it is declared that "more than 10 million new publications from JSTOR, Nature, Public Library of Science (PLoS), SSRN, and others (23 publishers added)" had been added. Therefore we can

assume that there were at least 45.3 million publications at that moment. This figure is close to the one used by Khabsa and Giles (2014), which assumed the size of MAS as 48,774,763 documents as of January 2013.

However, on the other hand the data provided by Microsoft Azure Marketplace (http://datamarket.azure.com/dataset/mrc/microsoftacademic) as of March 2014 (see table name: paper) declares 39.85 million documents (only 4 million documents more than September 2013, according to the official website information). Finally, if a query is performed manually in the website platform (as of May 2014), we obtain 45.9 million documents.

This confusion in the official data depending on the source consulted made us notice that the number of new records indexed in 2014 was dramatically low (802 as of May 2014), which could be a sign of the demise of the system, although this has not been officially announced. These preliminary results were informally shared with the scientific community (Orduña-Malea *et al.*, 2014), reaching the specialised press (http://blogs.nature.com/news/2014/05/the-decline-and-fall-of-microsoft-academic-search.html). Nonetheless, although the MAS coverage downgrade was demonstrated, the process and details of this decline, as well as the effects of this obsolescence on the quality of the service offered to its users (both those who engage in scientific information seeking and also those who use it for quantitative purposes) have not been described to date, despite the relevance of this academic search engine.

Therefore the main objective of this paper is to describe the obsolescence process of Microsoft Academic Search in order to find out whether it was gradual or abrupt, and whether it followed some kind of order or was random. Similarly the effects of this decline in the coverage of journals, and in turn their influence on the fair representation of fields and organisations in the platform (that is, on the quality of the information offered) constitute secondary objectives.

Related research

Research literature has traditionally paid more attention to Google Scholar than Microsoft Academic Search due to its higher coverage and ease of use, among other things (Orduña-Malea *et al.*, 2014). Most of the works published about MAS study this product by comparing it with Google Scholar (Jacsó, 2011) whereas informetric analyses based on the data provided

by MAS are scarce but increasing (Gonçalves *et al.*, 2014; Li *et al*., 2014; Ortega, 2014b; Sarigöl *et al.* 2014).

Haley (2014) compares the bibliometric performance of 50 top economics and finance journals both in GS and MAS using the Publish or Perish (PoP) application (http://www.harzing.com/pop/htm). Two different timeframes were used: over their entire lifespan in the target databases, and 1993-2012. Data were collected in June 2013. The results obtained by the authors are clear and definite: GS doubled – and in some cases tripled – bibliometric values of all the indicators used to determine the impact of the 50 top economics and finance journals studied.

Nonetheless, this work suffers from two methodological weaknesses that may influence the results: 1) inaccurate search queries of journal titles due to not using either all possible variants of a journal name or the "exclusion operator" to remove irrelevant documents, and 2) the existing limitation in GS of showing only the top 1,000 results raises doubts about the validity of the results since most of the searches performed for the 50 journals analysed far exceed the threshold set by GS.

Moreover, Gardner and Inger (2013) seek to learn how readers discover, access and navigate the content of scholarly journals. These authors conducted a large scale survey of journal readers (n = 19,064) during May, June and July of 2012. All regions of the world and all professional sectors, especially academic researchers (50 percent of respondents) and students (20 percent of respondents) are well represented. This study concludes that when searching and following a citation, "academic search engines are the second most popular resource across the board. Instead, they are less important for people who want to discover [the] latest articles" (p. 17). Additionally the results show that Google and Google Scholar are always the first choice (especially for students) whereas Microsoft Academic Search is rarely used.

The number of author profiles has also been discussed in the literature, as this indicator reflects the use of these services by the research community, especially for MAS, as personal profiles are created automatically when a new paper is indexed containing a new author not covered previously.

Ortega and Aguillo (2014) offer a comparative analysis of the personal profiling capabilities of MAS and Google Scholar Citations (GSC). It should be specified to properly interpret the results that they do not offer a comparison between GS and MAS but between the author profiles provided by Google Scholar Citations and those offered by MAS. The results clearly show that the number of profiles in MAS is almost 200 times the number of

profiles in GSC. MAS contained 19 million author profiles in August 2012 whereas in the case of GSC this information is unknown, although the authors estimate 106,246 profiles in June 2012. The reason for this remarkable difference is the way in which the products are made (automatically in MAS, and manually in GSC).

Ortega and Aguillo (2014) additionally perform an analysis of 771 personal profiles appearing both in MAS and GSC. The results show that GSC gathers 158.3 percent more documents per profile than MAS, 327.4 percent more citations, and 155.8 percent higher h-index values. These differences occur in virtually every scientific field except for chemistry and medicine. However, it is striking that in these two fields MAS gathers more documents than GSC but at the same time recovers far fewer citations. This contradiction is surprising and it might have been caused by the samples taken in these areas not being big enough.

Haustein *et al.* (in press) determine the use and coverage of social media environments, examining both their own use of online platforms and the use of their papers on social reference managers. The survey was distributed among the 166 participants (71 returned the questionnaire) in the 17th International Conference on Science and Technology Indicators in Montréal from 5-8 September 2012.

As the authors state: “asked for personal publication profiles on Academia.edu, Google Scholar Citations, Mendeley, Microsoft Academic Search, ResearcherID (WoS), or ResearchGate, 32 participants listed their publications at least at one of these platforms. The most popular tool was Google Scholar Citations (22 respondents with profile; 68.8 percent of those with publication profiles)”; MAS was the second least used platform, at a considerable distance from Google Scholar Citations.

When bibliometricians were asked what they were doing with their publication profiles, the authors found that GSC was the most frequently used product in all the typical activities related to the maintenance of an author profile: adding missing publications, merging duplicate publications and, especially, checking citations. However, MAS was the least frequently used product for all these operations. It is also of note that people especially used GSC to delete misattributed publications from their profiles. This would confirm the technical problems of MAS (information editions are mediated via a request), subsequently detected by Ortega and Aguillo (2014). Recently Van Noorden (2014) obtained similar results in a survey of more than 3,000 researchers, finding that around 80 percent of respondents were not aware of the MAS website.

The lesser use of Microsoft Academic Search may be related to diverse technical problems, many of which have been previously identified by Jacsó (2011). On the one hand

there is a higher number of duplicate profiles. As illustrative examples in bibliometrics, we can find up to 14 different entries for Derek de Solla Price (http://academic.research.microsoft.com/Search?query=Solla%20Price) or 6 for Eugene Garfield (http://academic.research.microsoft.com/Search?query=Eugene%20Garfield). This issue is of special importance in languages with many possible name variants and different translations (such as Spanish, Portuguese, Chinese and Russian). Nonetheless, the considerable efforts made by Microsoft to avoid this problem should be mentioned, such as the Labelling Oriented Author Disambiguation approach (LOAD), which combines machine learning and human judgement to achieve author disambiguation (Qian *et al.*, 2011), as well as the ALIAS – identifying duplicate authors in Microsoft Academic Search – program (http://cwds.uw.edu/alias-identifying-duplicate-authors-microsoft-academic-search). On the other hand a lower updating rate (41 percent of the MAS profiles presented an outdated affiliation) was detected (Ortega and Aguillo, 2014).

As regards the total size of MAS and the coverage according to field, the work of Jacsó (2011) is at present the only remarkable contribution. He estimated the size of MAS at 27.2 million documents (as of September 2011), and calculated the size of the fields and sub-fields that MAS contained at the time, showing that clinical medicine, chemistry and computer science were, at that time, the most representative fields.

Notwithstanding, some important thematic areas such as social sciences, geosciences, arts and humanities, and especially multidisciplinary (labelled at the beginning as “other domains”) had not yet been added to the product. Likewise longitudinal studies describing the evolution of coverage per field have not been published. Therefore the analysis of data, not only at a thematic level but also at journal and organisation level, are of interest both for the years of decline, and for the years preceding the downfall.

Method

We compiled the following data from Microsoft Academic Search: total records, and records according to field, journal and organisation. The method consisted of a first phase based on a cross-sectional analysis (used to describe the cumulative coverage of MAS at present) and a second phase based on a longitudinal analysis (used to describe the obsolescence process annually). The procedure for collecting each of these indicators is explained below.

Total records: the total number of records indexed in MAS up to 2014 was collected. The procedure, for which we did not use the available API, consisted of querying the database

directly via http from the official website, filtering results by year and gathering the results manually. It must be noted that MAS, unlike GS, does not provide hit count estimates but the total number of records stored in the database. The query via http is possible due to the structured URL generated after any query performed on the search box. For example the URL http://academic.research.microsoft.com/Search?query=year%3d2013 retrieves all indexed records published in 2013, also providing the total figure (in this case 8,147 records) directly in the search results.

Records per field: we also collected the number of records indexed in each field, broken down by year of publication. For example the URL http://academic.research.microsoft.com/Search?query=year%3d2013&s=0&SearchDomain=4 provides the number of indexed records published in 2013 in the field of biology.

Records per journal: in this case we obtained a sample of the most representative journals in MAS by means of intentional sampling. For this purpose we used the "Top journals in" option, which identifies the better ranked journals in each field and sub-field by means of the Field Rating indicator. The Field Rating is similar to the h-index in that it calculates the number of publications by an author, journal or organisation, and the distribution of citations to the publications, except focusing within a specific field among the fields and sub-fields covered by MAS.

Thus we collected data for the top 10 journals in each sub-field in each of the 15 general fields of MAS, obtaining a total figure of 1,762 unique journals (it should be noted that there are sub-fields with fewer than 10 journals catalogued and that a journal can be classified in more than one field).

For each journal the total annual number of publications indexed in MAS and the Field Rating were directly collected. For example the URL http://academic.research.microsoft.com/Search?query=year%3d2012%20jour%3a%28PHYS%20REV%20LETT%29 retrieves the number of indexed records for a journal published in 2012, in this case *Physical Review Letters*.

Records per organisation: similar to journals, organisations are ranked by the Field Rating indicator, which assesses organisations by field and sub-field. In this case we proceeded to extract by means of intentional sampling the 10 highest ranked organisations in each sub-field of each of the fields, obtaining a total of 2,053 records, which correspond to 346 unique institutions (since an institution can be ranked in the top 10 in more than one sub-field).

For each record (corresponding to an institution in a sub-field), the total and annual number of publications and citations of the organisation, the total and annual number of publications and citations received in the corresponding area, and the Field Rating value were directly collected. For example the URL http://academic.research.microsoft.com/Search?query=org%3a%28Stanford%20University%29%20year%3d2012 retrieves all indexed records published in 2012 for an organisation, in this case Stanford University.

Finally, each query (for each journal, organisation, field and years considered) was matched to its corresponding URL. The querying process and the annotation of the resulting values were automated, and all data were entered into a spreadsheet for processing. Data collection was carried out in March 2014 (and updated in July 2014), while the analysis was conducted between April and July 2014.

Results

Total coverage of Microsoft Academic Search

The total number of documents indexed in MAS as of March 2014 was 45,997,996, although as noted above, this figure varies considerably depending on the query method and the source used. This value, as indicated in the previous section, was obtained after performing the summation of annual values up to 2014, from the official public website of MAS.

The retrospective coverage of the product is equally noteworthy. When we analysed the number of registered documents per century, we got the following distribution: up to 1800, there are 7,459 documents; from 1800 to 1899 we can find 456,038; from 1900 to 1999 the figure rises to 23.9 million; and finally, in the period 2000-2014 there are 21.5 million documents.

Figure 1 shows the annual evolution from 1995 to 2014. In 2010 (2,346,228) it reaches its highest point and after that the fall is abrupt: 1,393,964 collected in 2011 (a drop of about one million documents compared to the previous year) while in 2012 the figure is just 290,506 records (another drop of about one million documents). In March 2014 only 802 documents had been collected so far that year but, unexpectedly, in July 2014 this figure had risen to 1,856, of which 1,479 (79.7 percent) fall into the multidisciplinary field.

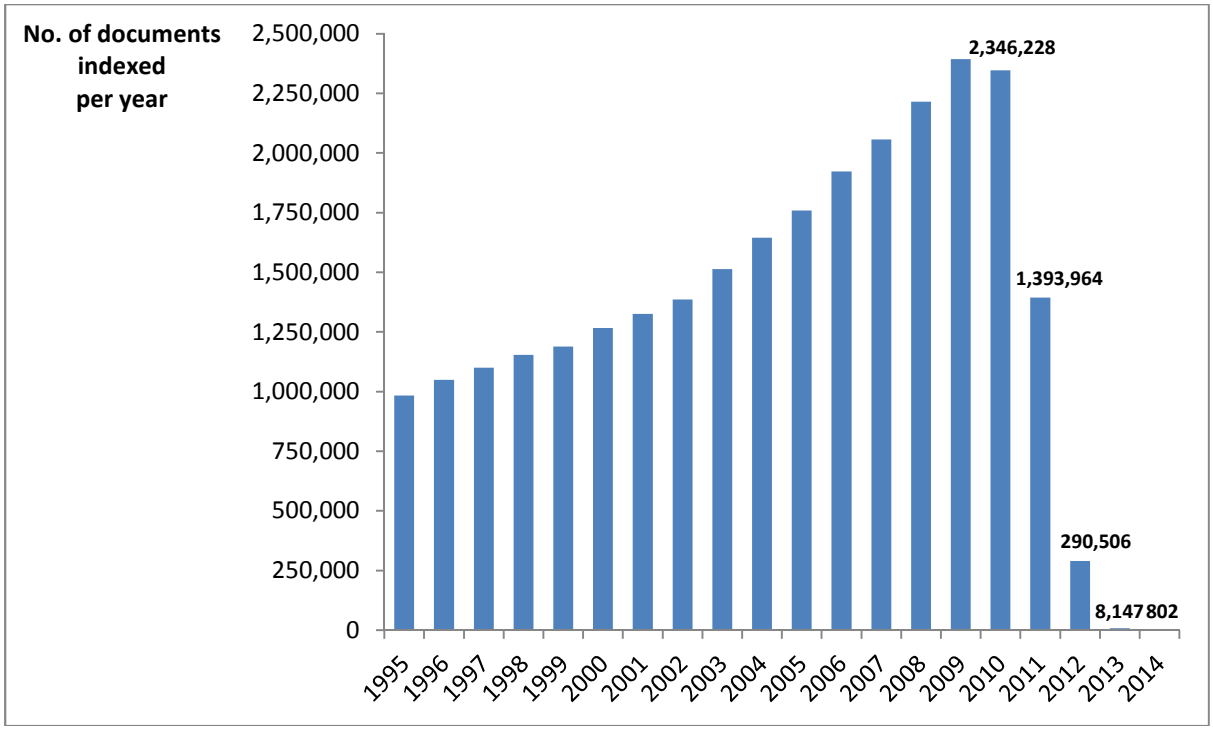

Figure 1. Number of documents indexed in MAS per year (1995- March 2014)

When the total number of historical records in each field were analysed, a clear preponderance of the medicine (23.95 percent), multidisciplinary (17.06 percent) and physics (10.28 percent) fields was observed. These data correspond to those records obtained for the recent period 2000-2014, where the contributions of computer science (passing from the historic 6.7 percent to 8.8 percent in the period 2000-2014) and engineering (passing from a historical 7.5 percent to 8.2 percent in the recent period) should be highlighted. These data can be observed in more detail in Table 2. It should also be indicated that the total values in the last row do not match the total number of records in the database, since obviously a record can be associated with more than one subject field.

Table 2. Number of documents according to fields

Discipline	Total		2000-2014		2013	
	N	%	N	%	N	%
Medicine	11,576,830	23.95	5,727,242	25.36	112	1.36
Multidisciplinary	8,248,315	17.06	3,513,039	15.55	6,513	79.28
Physics	4,967,621	10.28	1,991,266	8.82	40	0.49
Chemistry	4,380,349	9.06	1,846,333	8.17	32	0.39

Biology	3,954,030	8.18	1,886,939	8.35	60	0.73
Engineering	3,656,057	7.56	1,855,228	8.21	180	2.19
Computer science	3,229,591	6.68	1,991,996	8.82	685	8.34
Social Science	1,823,847	3.77	847,727	3.75	33	0.40
Arts and humanities	1,362,565	2.82	586,681	2.60	37	0.45
Geosciences	1,256,411	2.60	525,773	2.33	46	0.56
Mathematics	1,144,496	2.37	423,438	1.87	34	0.41
Economics and business	922,519	1.91	500,376	2.22	383	4.66
Material science	902,546	1.87	454,474	2.01	31	0.38
Agriculture science	463,559	0.96	190,664	0.84	9	0.11
Environmental sciences	449,363	0.93	245,325	1.09	20	0.24
Total	48,338,099	100	22,587,305	100	8,215	100

In addition Table 2 includes the total data for the last full year collected (2013), where the prevalence of multidisciplinary (79.3 percent), computer science (8.34 percent) and economics and business (4.66 percent) is observed, indicating that the fall in these disciplines has been milder than in medicine. To illustrate this behaviour Figure 2 includes time trends (2000 to 2014) of the fields that exceed one million records in that period, where a timely ascent in 2010 in chemistry and engineering are also noted.

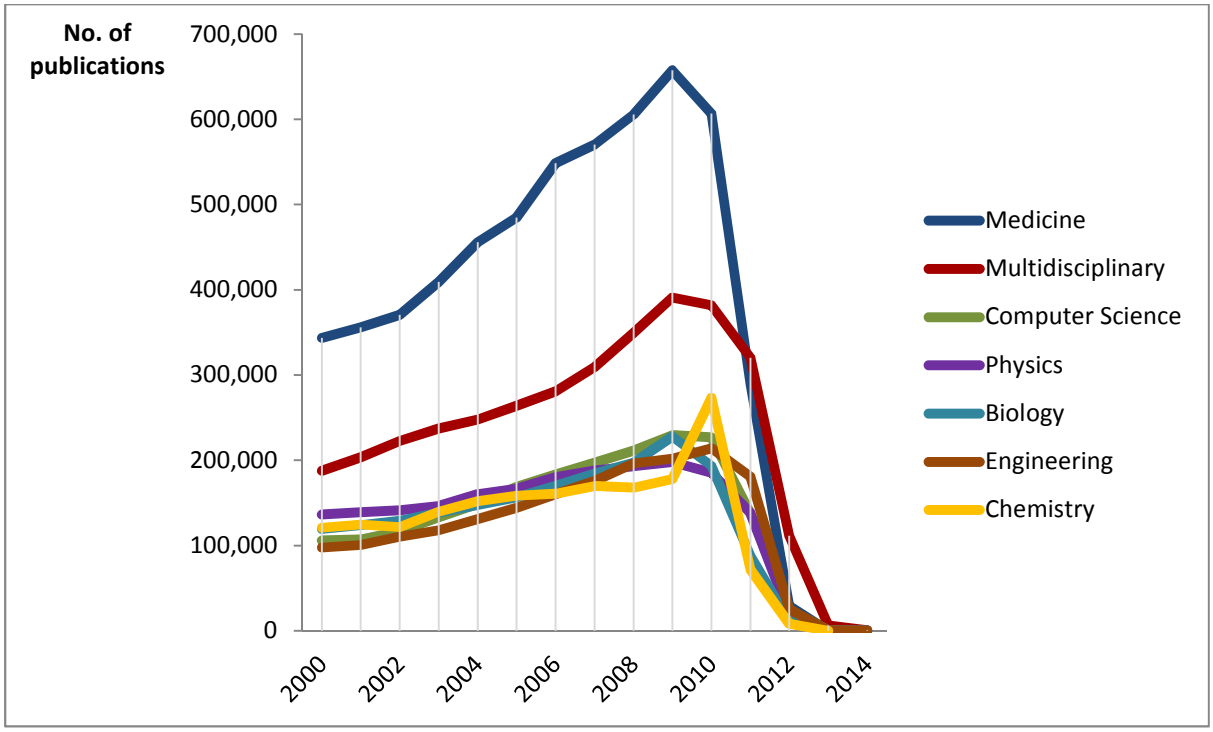

Figure 2. Number of publications according to fields (2000-2014)

Coverage according to journals

After checking the overall decline of records collected by MAS (both total and per field), in this section the data at journal level is analysed. Table 3 shows the 10 journals with the highest number of total records collected in MAS. Similarly the annual data from 2009 to 2013 is available, in order to observe the evolution over the years of decline.

Table 3. Top 10 journals according to documents indexed in MAS

Journal	Discipline	Total	2009	2010	2011	2012	2013
Nature	Multidisciplinary	480,580	4,861	4,200	3,783	900	3
Science	Multidisciplinary	290,006	4,470	3,827	2,265	54	1
Lancet	Medicine	233,248	1,788	1,935	5,175	1	0
Physical Review B	Physics	207,344	7,716	7,674	5,692	189	0
Journal of the Acoustical Society of America	Computer science	179,101	1,778	4,463	3,353	4	0
PNAS	Multidisciplinary	158,948	5,469	5,076	2,690	96	9
Journal of Geophysical Research	Geosciences	142,341	5,136	5,496	2,815	22	1
Astrophysical Journal	Physics	130,101	4,661	3,488	3,111	72	0
Physical Review D	Physics	102,232	6,626	6,590	4,177	80	0
Physical Review A	Physics	91,742	4,213	6,336	3,734	7	0

The data shown in Table 3 reflect the elevated representation of the multidisciplinary field, represented not only by *Nature*, *Science* and *PNAS* (*Proceedings of the National Academy of Sciences*) but also by *PLoS One*. Although the latter is not listed in the historical top 10 because it is of recent creation, it has the highest output data during the last decade (45,298 papers indexed from 2000 onwards). These results explain the lesser decline in this field, shown previously in Figure 2.

Similarly a disproportionate drop in the indexed journals was observed. The data obtained for these top 10 journals in 2012, dominated by *Nature* (900) and *Physical Review B* (189), bears little resemblance to the data obtained for the same journals in previous years. In fact the Pearson correlation between the number of papers indexed for these 10 journals between 2009 and 2010 is high ($r = 0.76$), but falls sharply thereafter (between 2010 and 2011 $r = 0.23$; between 2011 and 2012 $r = 0.11$).

If we concentrate only on the journals with the highest presence in the period 2000-2014, we obtain a total of 1,636 publications with at least 100 papers indexed in MAS. In Figure 3 we can observe the annual evolution (from 2008) for these top 10 journals.

We can notice the overall drop of records for all of these journals between 2010 and 2011, with two exceptions (*Gastroenterology* and *PLoS One*), for which MAS collected more records in 2011 than in the preceding year. From 2011 to 2012 the differences between the journals' indexed output are completely erased, with the exception of *Nature* (900) and again *PLoS One* (6,028).

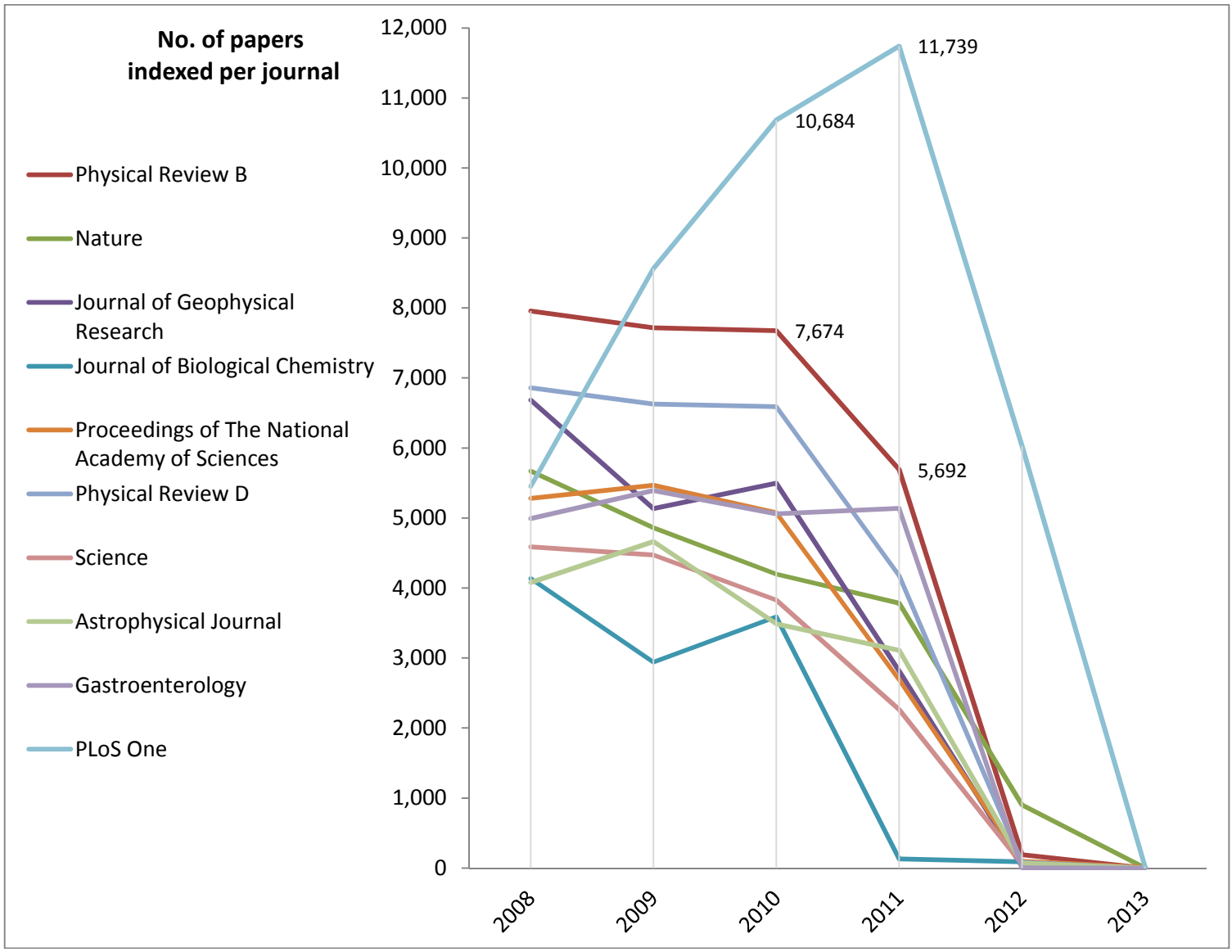

Figure 3. Papers indexed per year for the top 10 journals with more publications in the period 2000-2014 (2008-2013)

Obviously journals may vary from year to year as the number of published papers may fluctuate, although this variation, except in the case of *PLoS One*, should move between bounded values. In any case items indexed by MAS should be compared with the number of papers published by these journals. These values – extracted from Web of Science (WoS) for the top 10 journals shown in Figure 3 – are presented in Table 4, considering data for 2011 and 2012 (the first two years of decline).

Table 4. Papers indexed in MAS and Web of Science for the 10 journals with more publications indexed in MAS in the period 2000-2014

Journal	2011			2012		
	MAS	WoS	%	MAS	WoS	%
Physical Review B	5,692	6,307	90.25	189	5,816	3.25
Nature	3,783	2,591	146.01	900	2,651	33.95
*Journal of Geophysical Research	2,815	2,718	103.57	22	2,805	0.78
Journal of Biological Chemistry	130	4,501	2.89	88	4,165	2.11
**PNAS	2,690	4,127	65.18	96	4,360	2.20
Physical Review D	4,177	3,079	135.66	80	3,435	2.33
Science	2,265	2,750	82.36	54	2,760	1.96
Astrophysical Journal	3,111	2,508	124.04	72	3,115	2.31
Gastroenterology	5,139	5,044	101.88	3	4,883	0.06
PLoS One	11,739	13,786	85.15	6,028	23,452	25.70

* Not indexed by WoS: measured from Scopus. A problem related to the title was detected as well.

** Not indexed in WoS, measured from Scopus.

Table 4 provides conflicting data. According to the data for 2011 there are many cases where MAS indexes a much higher number of records than WoS does for the same journal (*Nature*, *Journal of Geophysical Research*, *Physical Review D*, *Astrophysical Journal*, *Gastroenterology*). This pattern (not detected in 2012), might be explained by the differences in the policies each of the products follow regarding the indexing of the different document types that may be found in a journal (papers, letters, reviews, editorials, etc.), although these differences seem to be too high to be fully explained by these policies.

In any case the data of 2012 also contain unexpected figures. Of the 23,462 papers published by *PLoS One* (considering the data from WoS) only 6,028 (25.7 percent) are indexed in MAS, and from the 2,651 papers published by *Nature* MAS only collected 900 (33.9 percent). Randomness in these percentages for all the journals analysed was detected as well.

Furthermore, as was previously identified, there is an over-representation of certain fields in 2013 (multidisciplinary, computer science and economics and business) as well as an important growth in the number of records indexed in chemistry in 2010. The analysis performed at the journal level in this section explains the causes behind these values.

First, we performed manual queries in MAS, filtering by year (2010) and by field (chemistry). Despite locating certain journals whose indexed production significantly increased in 2010 (e.g. *Journal of Biological Chemistry*) there were others whose presence was reduced (e.g. *International Journal of Quantum Chemistry* and *Journal of Organic Chemistry*).

However, for the journal *Nachrichten aus und Chemie Technik Laboratorium*, MAS collected 3,260 items in 2010, the only year in which this journal is indexed. There are several other examples of this phenomenon, which is probably the cause of the growth of its chemistry field in 2010.

Second, of the 685 records assigned to computer science in 2013, we detected that 258 belong to the *International Journal of Advanced Computer Science and Applications*, and 162 to the *International Journal of Advanced Research in Artificial Intelligence*. In the case of economics and business, the journal *Applied Economics* published 234 of the 383 publications indexed in 2013.

These data confirm that the exaggerated presence – in terms of indexed records – of these two fields in 2013 is due to the indexing of a few random journals that have retained a high degree of indexation in 2013, but they were not representative of the total size of these fields in previous years, nor in their assessment (none of them belong to the original journal sample, which consisted of the 10 journals with the highest Field Rates by sub-field).

However, the multidisciplinary field presents a different situation. The journal that provides the greatest number of publications in this area in 2013 is *PLoS One*, with 21 papers, followed by *PNAS*, with 9, far from the 6,513 total records gathered in 2013 for this field, as shown above (Table 2). In order to find the location of this number of records, a general search filtered by year 2013 and the multidisciplinary area was performed. We noticed the existence of a large amount of catalogued records which were not assigned to any journal.

This circumstance is unexpected since one of the essential characteristics of MAS is precisely that it obtains the metadata of papers directly from publishers with which it has entered into a commercial agreement.

Coverage according to organisations

Finally, in this section the coverage of organisations in MAS is described. The top 10 institutions according to the total historical number of indexed documents in MAS are shown in Table 5. Additionally, the annual data from 2009 to 2014 are presented in Figure 5.

Table 5. Organisation rankings according to the number of documents indexed in MAS

University	T	2009	2010	2011	2012	2013	2014
Harvard University	598,929	34,437	36,145	17,563	2,308	72	13
Chinese Academy of Sciences	492,400	54,222	57,665	36,903	4,334	85	18
Stanford University	463,878	21,131	21,219	11,773	2,276	49	7
U. of California Berkeley	426,973	18,351	18,116	10,611	1,892	38	1
U. of California Los Angeles	400,837	20,414	21,301	11,015	1,826	71	8
U. of Michigan	356,453	18,885	18,915	10,310	1,885	48	3
U. of Tokyo	353,190	19,856	20,128	10,175	1,135	47	1
U. of Oxford	351,259	17,977	18,639	10,060	1,910	58	3
MIT	350,469	15,865	16,306	9,752	1,713	51	3
U. of Cambridge	349,649	17,502	17,473	9,468	1,591	39	4

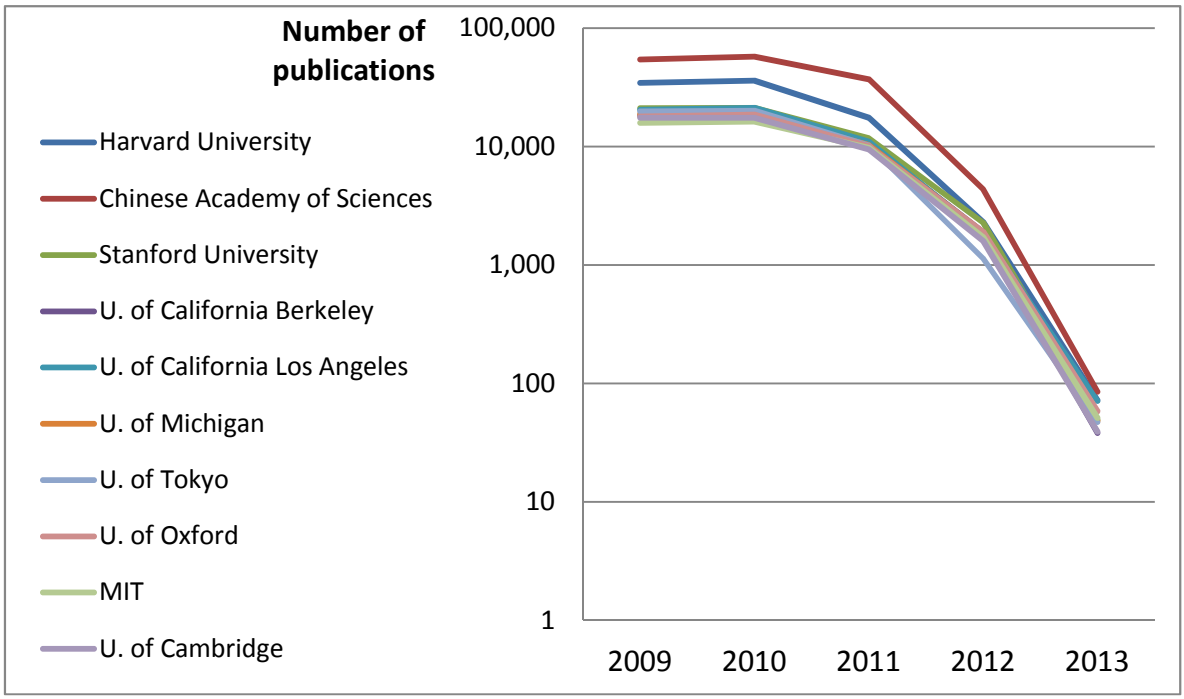

Figure 5. Organisation ranking annual data from 2009 to 2014

A general decline for all universities was detected (however, the predominance of the Chinese Academy of Sciences in recent years should be noted). Unlike the fall in the indexing of journal content, in this case a fairly uniform decrease in these institutions' publications (those with the highest coverage in the system) is observed. It is probable, however, that universities with more specialised profiles (especially in medicine) would present more abrupt falls in relation to their corresponding overall sizes.

In order to determine whether this decline has been uneven across the various fields (directly dependent a priori on the indexing ratio of journals), we proceeded to analyse the annual trend for each institution for each of the 15 fields covered by MAS. The thematic profile for the two largest universities in terms of the number of indexed documents (Harvard University and Chinese Academy of Sciences) is offered in Figure 6 for the years 2010, 2011 and 2012, which correspond to the last year of growth (2010) and the first two years of decay.

In this case we double-checked that the downfall in document indexing by field was, in general and with minor exceptions (mathematics and environmental sciences for Harvard, and materials sciences for the Chinese Academy of Sciences), quite similar during these years. That is to say the thematic profile of these institutions remains stable during the years of decline, although the coverage of journals falls unevenly.

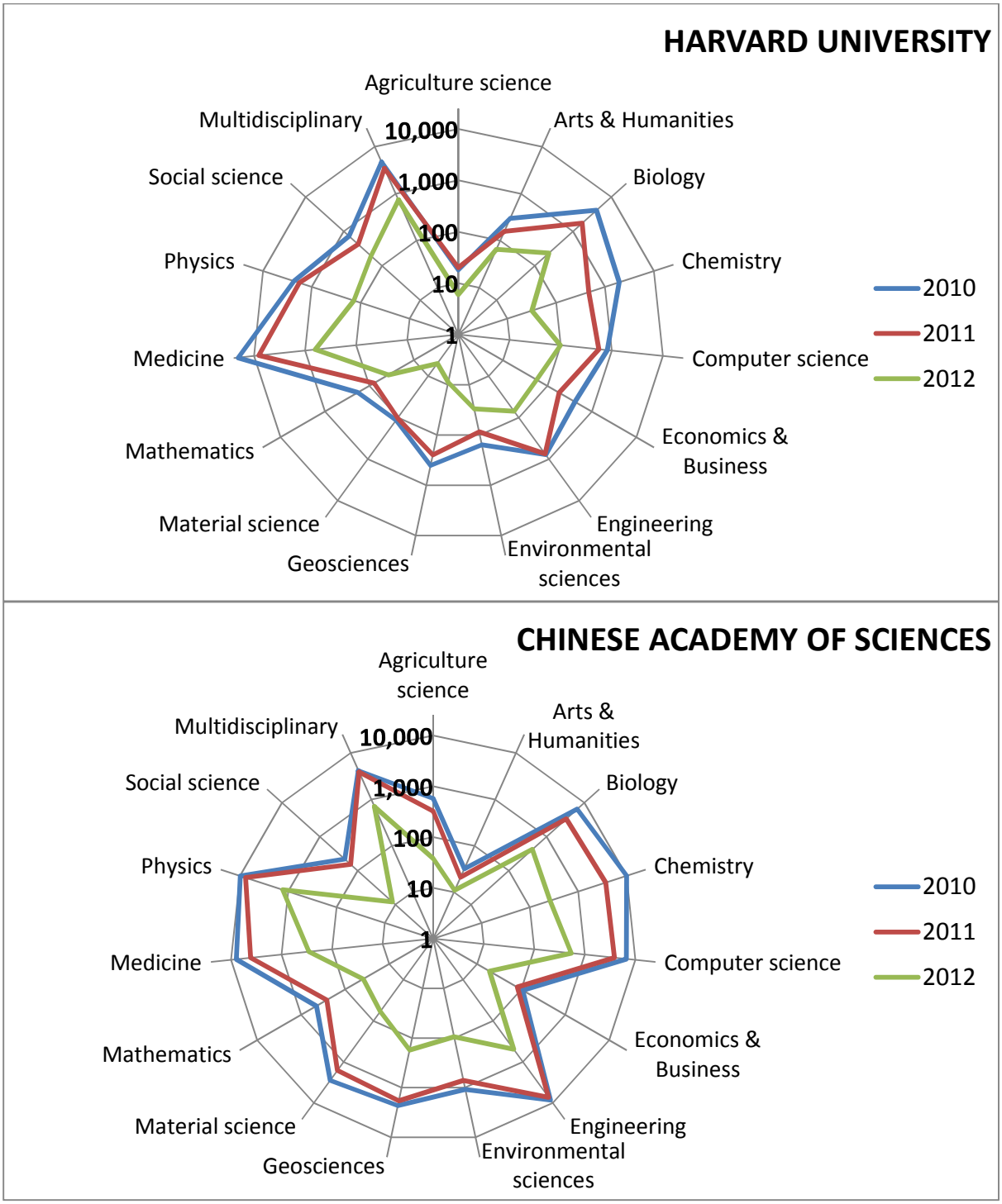

Figure 6. Thematic profiles of Harvard University and Chinese Academy of Science in MAS (2010-2013)

Discussion and conclusions

Microsoft Academic Search has not been updated since 2013, although this phenomenon began to be glimpsed in 2011, when its coverage plummeted. Throughout 2014 indexing of new records is still ongoing, but at a minimal rate, without following any apparent pattern.

Some journals of paramount importance in the scientific information market (such as *Nature*, *Science*, *PNAS*, *Lancet*, etc.) are indexed only partially, which suggests that currently the indexing of new records is working automatically, without supervision, in a way that seems almost random.

This situation is what has disturbed the balance in the sizes of the fields (which ultimately depends on the subject categorisation of the journals indexed) exaggerating during 2012 and 2013 the sizes of various areas of knowledge (for example computer science, economics) because of the existence of some journals (with a moderate journal rating in the field at best) that have had more records indexed as a result of an automatic procedure. Moreover, the lack of allocation of new records to journals generates significant inaccuracies in other areas (in this case multidisciplinary).

Thus the results obtained prevent us from using this tool in its present state. In any case the data must be used with some caution, because the total number of records varies according to the source, as indicated in the introduction. In this case the results offered are those obtained from a direct query to the database.

Since the obsolescence process has been demonstrated to be caused by the incomplete indexation of all papers belonging to the indexed journals, the effects of this phenomenon on users can be summarised at different levels:

- Final users have had access to misleading information since 2010. This is especially critical for the comparison tools offered by the platform, since the performance differences among authors and organisations are false (if we consider the coverage of journals indexed).
- For editors the incomplete indexation of their journals and publications undermines their diffusion.
- For organisations a distorted academic performance is shown. In any case, this effect has not directly affected the thematic profile of the organisations (which lose coverage in the database in a proportional manner, at least when we consider the top 10 organisations with a higher historical number of indexed records).
- For authors incomplete author profiles are built.
- For researchers using MAS as a data source for quantitative analysis the downgrade of the database may affect the validity of their results, especially when using data from 2010 to the present.

Moreover, this issue has gone unnoticed, as far as we know, in the bibliometric and webometric arena. In view of these problems it seems logical not only that MAS was hardly ever used to search for papers by academics and students (who mostly use Google or Google Scholar), as recently noted by Van Noorden (2014), and virtually ignored by bibliometricians. Even its disappearance has been ignored, although the activity of official forums and inclusion of new journals in 2014 should be further analysed in order to better explain what is really happening with the product.

In any case although the platform has only served as a technological testing bench and/or the downgrade is due to strategic business issues (the causes of the downgrade are out of the scope of this research), keeping a tool purposefully out of date without giving any sort of public explanation or disclaimer note implies irresponsibility, given its implications for the representation, reliability and validity of the bibliometric data provided on fields, journals, authors, and conferences.

Although both Google Scholar and Microsoft Academic Search are academic search engines which share a common goal (finding academic information) and work on the same data source (the academic web) these systems have different architecture, design and features. Therefore although technical problems (such as name disambiguation) may have influenced their obsolescence, the different philosophy for product development of each company (open and uncontrolled by Google Inc., while closed and controlled by Microsoft) could be the ultimate key that explains the success of one product and the failure of the other.

These different philosophies are specifically reflected in the different processes for profile creation, an aspect already pointed out by Ortega and Aguillo (2014): Google Scholar relies on self-edited personal profiles while MAS adopts a restricted model in which the researchers can only suggest changes to their automatically supplied profiles.

This circumstance implies not only less user interaction in MAS, but a limitation in detection and correction of errors in large quantities, impossible to identify in most cases without direct user involvement, creating a vicious circle that provides a competitive advantage to its competitor in the market, which also has intense activity.

In fact in December 2011, just a month after the public launch of Google Scholar Citations, MAS announced an update feature, but it ceased its activity (during 2012 it did not publish any issues); and the next update, in January 2013, would be the last to date. Therefore it is likely that the expansion of Google Scholar profiles was decisive in the obsolescence of MAS.

As the strength of a house depends on its foundation, the pillars of a bibliographic database are the number of documents that it can identify and update, as well as the quality of this data. The rest of the features are interesting, but if the pillars fail, everything fails. That is precisely what happened to MAS. Their data, obtained from the largest source of information available today (the web) is lower (45.9 million documents) than the collections of other traditional databases (WoS and Scopus possessing more than 50 million). Even worse, it has not been updated since 2013, and had started showing severe drops in the number of records indexed from 2011.

The brilliant visualisation tools for thematic domains and fields, documents, authors and organisations that MAS has deployed (of a portentous quality), and which appeal so much to specialists, are worthless if the underlying data are insufficient, not updated and/or dirty.

In the meantime Google focused on recovering more documents and citations and cleaning up full bibliographic records and has beaten Microsoft. Offering speed and exhaustiveness in searches, and providing four popular bibliometric indicators (number of papers, citations, h-index, median h-index, number of documents with at least 10 citations) as well as basic visualisations (histograms), Google Scholar has come to stay and predominate, which in turn implies the prevalence of an open construction model (GS) versus a closed model (MAS). This situation is not positive in the sense that market competition is always desirable. In any case we must still wait to evaluate the future evolution of Microsoft Academic Search.

References

Beel, J. and Gipp, B. (2009), "Google Scholar's ranking algorithm: an introductory overview", in *Proceedings of the 12th International Conference on Scientometrics and Informetrics (ISSI'09)*, BIREME/PAHO/WHO and Federal University of Rio de Janeiro, Vol. 1, pp. 230-41.

Beel, J., Gipp, B. and Wilde, E. (2010), "Academic search engine optimization (ASEO)", *Journal of Scholarly Publishing*, Vol. 41 No. 2, pp. 176-90.

Butler, D. (2011), "Computing giants launch free science metrics: new Google and Microsoft services promise to democratize citation data", *Nature*, Vol. 476 No. 7358, p. 18.

Carlson, S. (2006), "Challenging Google, Microsoft unveils a search tool for scholarly articles", *Chronicle of Higher Education*, Vol. 52 No. 33, p. A43.

Delgado López-Cózar, E. and Cabezas-Clavijo, A. (2013), "Ranking journals: could Google Scholar Metrics be an alternative to Journal Citation Reports and Scimago Journal Rank?", *Learned Publishing*, Vol. 26 No. 2, pp. 101-13.

Delgado López-Cózar, E., Robinson-García, N. and Torres-Salinas, D. (2014), "The Google Scholar Experiment: how to index false papers and manipulate bibliometric indicators", *Journal of the Association for Information Science and Technology*, Vol. 65 No. 3, pp. 446-54.

Gardner, T. and Inger, S. (2013), *How Readers Discover Content in Scholarly Journals. Comparing the Changing User Behaviour between 2005 and 2012 and its Impact on Publisher Web Site Design and Function*, Renew Training, Abingdon.

Gonçalves, G.D., Figueiredo, F., Almeida, J.M. and Gonçalves, M.A. (2014), "Characterizing scholar popularity: a case study in the computer science research community", in *Joint Conference on Digital Libraries (JCDL 2014)*, [online], available at http://homepages.dcc.ufmg.br/~flaviov/papers/goncalves2014-dl.pdf (accessed 19 August 2014).

Haley, M.R. (2014), "Ranking top economics and finance journals using Microsoft academic search versus Google scholar: how does the new publish or perish option compare?", *Journal of the Association for Information Science and Technology*, Vol. 65 No. 5, pp. 1079-84.

Haustein, S., Peters, I., Bar-Ilan, J., Priem, J., Shema, H. and Terliesner, J. (in press), "Coverage and adoption of altmetrics sources in the bibliometric community", *Scientometrics*, DOI: http//dx.doi.org/10.1007/s11192-013-1221-3.

Jacsó, P. (2005), "As we may search – comparison of major features of the Web of Science, Scopus, and Google Scholar", *Current Science*, Vol. 89 No. 9, pp. 1537-47.

Jacsó, P. (2008a), "Google scholar revisited", *Online Information Review*, Vol. 32 No. 1, pp. 102-14.

Jacsó, P. (2008b), "Live Search Academic, Gale: Péter's Digital Reference Shelf", [online], available at: web.archive.org/web/20070207104154/http://projects.ics.hawaii.edu/~jacso/gale/windows-live-acad/windows-live-acad.htm (accessed 24 September 2014).

Jacsó, P. (2010), "Microsoft Academic Search, Gale: Péter's Digital Reference Shelf", [online], available at: http://www.jacso.info/PDFs/jacso-microsoft-academic-search-2010-Gale.doc (accessed 24 September 2014).

Jacsó, P. (2011), “The pros and cons of Microsoft Academic Search from a bibliometric perspective”, *Online Information Review*, Vol. 35 No. 6, pp. 983-97.

Jacso, P. (2012), “Google Scholar Metrics for publications: the software and content features of a new open access bibliometric service”, *Online Information Review*, Vol. 36 No. 4, pp. 604-19.

Khabsa, M. and Giles, C.L. (2014), “The number of scholarly documents on the public web”, *PloS One*, Vol. 9 No. 5, e93949.

Labbé, C. (2010), “Ike Antkare one of the greatest stars in the scientific firmament”, *ISSI Newsletter*, Vol. 6 No. 1, pp. 48-52.

Li, L., Wang, X., Zhang, Q., Lei, P., Ma, M. and Chen, X. (2014), “A quick and effective method for ranking authors in academic social network”, in Park, J.J., Chen, S-C, Gil, J-M and Yen, N.Y. Ed.), *Multimedia and Ubiquitous Engineering*, Lecture Notes in Electrical Engineering, Vol. 308, Springer-Verlag, Berlin, pp 179-85.

Mayr, P. and Walter, A.K. (2005), “Google Scholar – wie tief gräbt diese Suchmaschine?”, in *Die Zukunft Publizieren: Herausforderungen an das Publizieren und die Informationsversorgung in den Wissenschaften*, 9-11 May, Bonn, [online], available at: http://www.ib.hu-berlin.de/~mayr/arbeiten/Mayr_Walter05-preprint.pdf (accessed 24 September 2014).

Orduña-Malea, E. and Delgado López-Cózar, E. (2014), “Google Scholar Metrics’ evolution: an analysis according to languages”, *Scientometrics*, Vol. 98 No. 3, pp. 2353-67.

Orduña-Malea, E., Ayllón, J.M., Martín-Martín, A. and Delgado López-Cózar, E. (2014), “Empirical evidences in citation-based search engines: is Microsoft Academic Search dead?”, Working paper, No. 16, EC3 Research Group, Universidad de Granada, 21 May, [online], available at: http://arxiv.org/abs/1404.7045 (accessed 24 September 2014).

Ortega, J.L. (2014a), *Academic Search Engines: A Quantitative Outlook*, Elsevier, Netherlands.

Ortega, J.L. (2014b), “Influence of co-authorship networks in the research impact: ego network analyses from Microsoft Academic Search”, *Journal of Informetrics*, Vol. 8 No. 3, pp. 728-37.

Ortega, J.L. and Aguillo, I.F. (2013), “Institutional and country collaboration in an online service of scientific profiles: Google Scholar Citations”, *Journal of Informetrics*, Vol. 7 No. 2, pp. 394-403.

Ortega, J.L. and 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*, Vol. 65 No. 6, pp. 1149-56.

Qian, Y., Hu, Y., Cui, J., Zheng, Q. and Nie, Z. (2011), “Combining machine learning and human judgment in author disambiguation”, in *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, ACM, New York, pp. 1241-6.

Sarigöl, E., Pfitzner, R., Scholtes, I., Garas, A. and Schweitzer, F. (2014), “Predicting scientific success based on coauthorship networks”, *ArXiv*, [online], available at: http://arxiv.org/pdf/1402.7268.pdf (accessed 19 August 2014).

Van Noorden, R. (2014), “Scientists and the social network”, *Nature*, Vol. 512 No. 7513, pp. 126-9.