Journal of Informetrics 9 (2015) 570–576

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

A systematic analysis of duplicate records in *Scopus*

Juan-Carlos Valderrama-Zurián[a], Remedios Aguilar-Moya[b], David Melero-Fuentes[c,*], Rafael Aleixandre-Benavent[d]

[a] *Departamento de Historia de la Ciencia y Documentación, Universidad de Valencia, Valencia, Spain*
[b] *Departamento de Ciencias de la Educación, Universidad Católica de Valencia "San Vicente Mártir", Spain*
[c] *Instituto de Documentación y Tecnologías de la Información, Universidad Católica de Valencia "San Vicente Mártir", Spain*
[d] *Ingenio (CSIC-UPV) UISYS-Universidad de Valencia, Spain*

ARTICLE INFO

Article history:
Received 23 December 2014
Received in revised form 11 May 2015
Accepted 11 May 2015
Available online 18 June 2015

Keywords:
Bibliographic control guidelines
Bibliometric indicators overdimensionalized
Duplicate records
Indexing errors
Scopus database

ABSTRACT

In recent years, the *Web of Science Core Collection* and *Scopus* databases have become primary sources for conducting studies that evaluate scientific investigations. Such studies require that duplicate records be excluded to avoid errors of overrepresentation. In this line, we identify duplicate records in *Scopus* and examine their origins. Identifying journals with duplicate records in *Scopus*, selecting and downloading bibliographic journal records, and identifying and analyzing the duplicate records is the methodology adopted. Duplicate records are found when articles published in a journal are incorrectly mapped by *Scopus* to this journal and to a different journal from the same publisher and when there are journal title changes, orthographic differences in the presentation of a journal name, and journal name variants. In these last three cases, one bibliographic record of each duplicate is mapped to *Medline* coverage of *Scopus*. Consequently, the identified duplicates and the significant differences in the number of citations received in duplicate articles may influence bibliometric studies. Thus, there is a need for rigorous quality control guidelines to govern database managers and editors to prevent the creation of duplicates.

© 2015 Elsevier Ltd. All rights reserved.

1. Introduction

Following two years of development by librarians and researchers worldwide, *Elsevier* launched its scientific literature database *Scopus*, officially named *SciVerse Scopus*, to the information market on November 4, 2004 (Bar-Ilan, 2008; Boyle & Sherman, 2005; Chadegani et al., 2013).

Scopus initially included approximately 14,000 research sources (journals, book series, conference proceedings, patents) from 4000 publishers, encompassing 27 million records (Bar-Ilan, 2008; Boyle & Sherman, 2005). Today, the database includes 21,915 sources from 5000 publishers, encompassing 54 million bibliographic records (Elsevier, 2014a, 2014b) from 1823 to the present (Elsevier, 2014c). The current thematic breakdown of the database is as follows: 32% Health Sciences, 30% Physical Sciences, 23% Social Sciences, and 15% Life Sciences (Elsevier, 2014a).

Scopus indexes the content included in *Embase*, *Compendex*, *World Textile Index*, *Fluidex*, *Geobase*, *Biobase*, and *Medline*. This bibliographic database combines characteristics of the *Medline* and *Web of Science Core Collection* databases (Falagas, Pitsouni, Malietzis, & Pappas, 2008). Consequently, one can search for *Medline* records via *Scopus* separately (using an advanced search

* Corresponding author.
E-mail address: davidmf@mail.ucv.es (D. Melero-Fuentes).

http://dx.doi.org/10.1016/j.joi.2015.05.002
1751-1577/© 2015 Elsevier Ltd. All rights reserved.

only) or through the entire database, and in addition to *MeSH* (Medical Subject Headings) terms, one can obtain access to citation data and cited references (De Granda-Orive et al., 2013).

Additionally, among other databases, *Scopus* also includes *Embase* records and corresponding *Emtree* terms (Falagas et al., 2008). In this way, descriptors from *Medline* and *Embase* can be within the same record in *Scopus*.

Since the databases were first launched, numerous studies have compared the coverage, characteristics, and citations of the *Web of Science Core Collection* to those of *Scopus* (Archambault, Campbell, Gingras, & Lariviere, 2009; Falagas et al., 2008; Torres-Salinas, Lopez-Cozar, & Jimenez-Contreras, 2009; Vieira & Gomes, 2009). These studies suggest that both databases include similar resources (Gavel & Iselid, 2008) and that, while *Scopus* offers greater coverage in various disciplines (Bartol, Budimir, Dekleva-Smrekar, Pusnik, & Juznic, 2014; De Granda-Orive et al., 2013; Jacso, 2005; LaGuardia, 2005; López Illescas, Moya Anegón, & Moed, 2008), the choice of one or both databases to perform bibliometric studies depends on the area of study (Chadegani et al., 2013; Baykoucheva, 2010; Falagas et al., 2008).

This aspect is related to *Scopus*' greater coverage of journals that do not publish papers in English (Bartol et al., 2014), thus causing *Scopus* to include more citations than the *Web of Science Core Collection* does (De Granda-Orive et al., 2013; Falagas et al., 2008; Kulkarni, Aziz, Shams, & Busse, 2009; Meho & Rogers, 2008; Meho & Sugimoto, 2009; Torres-Salinas et al., 2009); consequently, it is recommended that citation studies should be performed using *Scopus* (Gorraiz & Schloegl, 2008).

A study examining scientific studies on eating disorders[1] performed over the 2008–2013 period is currently being developed. Our work shows that a *Scopus* search retrieves duplicate articles from seven journals. Consequently, we identify duplicate entries from these journals in the *Scopus* database and their potential effects on bibliometric research. In this sense, the objective of this study is to analyze why there are duplicate records in the search and what their influence is on the scientific evaluation.

2. Methods

2.1. Identifying journals with duplicate records in Scopus

Currently, the authors of this paper are conducting a study that evaluates the scientific research activity on eating disorders through *Web of Science Core Collection* and *Scopus*. In the phase of identification of records covered by both databases, 11 pairs of identical studies (duplicates) from *Scopus* were found. All these duplicate records had differences in the field "journal title". Thus, we found duplicates in which one record was mapped to *Psychoterapia* and another record was mapped to *Archives of Psychiatry and Psychotherapy* (both journals are published by the *Polish Psychiatric Association*), in which one record was mapped to *Taehan Kanho Hakhoe chi* and another record was mapped to *Journal of Korean Academy of Nursing* (this journal is a continuation of *Taehan Kanho Hakhoe chi*), in which one record was mapped to *BMJ Case Reports* and another record was mapped to *BMJ case reports*, in which one record was mapped to *Current Topics Behavioral Neuroscience* and another record was mapped to *Current topics behavioral neuroscience*, in which one record was mapped to *Proceedings of the Royal Society B: Biological Sciences* and another was mapped to *Proceedings. Biological sciences/The Royal Society*, and in which one record was mapped to *Emergency Medical Journal* and another record was mapped to *Emergency medicine journal: EMJ*.

2.2. Selecting and downloading bibliographic journal records

With this discovery, a relational database was developed in August of 2014 that includes all articles published in these journals and indexed in *Scopus* to determine whether more duplicates may be found.

2.3. Identification and analysis of duplicate records

Duplicate records were identified by two biomedical index database experts[2] through an algorithm written with logical operators (IF, OR, AND, SAME) in a new field in each record. The algorithm compared the bibliographic information with that of the rest of the records and marked every pair of identical records. Manual revisions of the selected records were then performed. The sources (publisher or database) from which *Scopus* obtained duplicates, the number of duplicates for each journal, and the number of citations received for each duplicate article were determined.

3. Results

From 34,119 works, a total of 4237 duplicates (12.42%) are found in the seven journals studied in the 2004–2013 period (Table 1). The 99.17% of duplicate cases are articles. The causes of these duplications are as follows.

[1] Eating disorders are considered "A group of disorders characterized by physiological and psychological disturbances in appetite or food intake," as defined by *Medical Subject Headings* from *National Center for Biotechnology* Information. URL: http://www.ncbi.nlm.nih.gov/mesh/68001068

[2] The experts are Juan-Carlos Valderrama-Zurián, MD, PhD, and Rafael Aleixandre-Benavent, MD, PhD, who are both authors of this paper.

Table 1
Journals and type of documents: duplicate cases per year in *Scopus*.

Journal title per duplicate case	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	Percentage
BMJ Case Reports – BMJ case reports					17	5	337	816	1052	1509	88.18
Current Topics in Behavioral Neurosciences – Current topics in behavioral neurosciences						1	6	13			0.47
Emergency Medicine Journal – Emergency medicine journal: EMJ	1	2	1								0.09
Journal of Korean Academy of Nursing – Taehan Kanho Hakhoe chi					62						1.46
Proceedings of the Royal Society B: Biological Sciences – Proceedings. Biological sciences/The Royal Society			207	122		49					8.92
Psychoterapia – Archives of Psychiatry and Psychotherapy						6	8	18	5		0.87
Percentage of duplicate cases	0.02	0.05	4.91	2.88	1.86	1.44	8.28	19.99	24.95	35.61	100
Articles	1	2	204	116	79	56	345	843	1053	1503	99.17
Other documents (reviews; notes; short surveys)			4	6		5	6	4	4	6	0.83

3.1. Articles published in a journal are incorrectly mapped by Scopus to this journal and to a different journal from the same publisher

A total of 37 duplicate studies from the *Psychoterapia* and *Archives of Psychiatry and Psychotherapy* journals (Table 2, Cause A) are included in *Scopus*. It was confirmed that these studies are published in *Archives of Psychiatry and Psychotherapy* (http://www.archivespp.pl/) and that *Scopus* has erroneously mapped them to *Psychoterapia* (http://www.psychoterapiaptp.pl/). Additionally, it was observed that five of the duplicated studies have received either one or two citations.

3.2. Duplicates resulting from journal name changes

A total of 62 duplicates (published in 2008) were identified for *Taehan Kanho Hakhoe chi*, which is now entitled the *Journal of Korean Academy of Nursing* (Table 2, Cause B); the articles identified as having been published in *Taehan Kanho Hakhoe chi* were generated through *Medline* and indexed in *Scopus* (Fig. 1). There was an average difference of two citations between duplicates, with a standard deviation (SD) of 1.79.

3.3. Duplicates due to orthographic differences in journal titles

A total of 3736 pairs of duplicate records were identified in *BMJ Case Reports* (Table 2, Cause C1). In each case, one of the duplicated records lists the name of the journal with uppercase letters (*BMJ Case Reports*), while the other displays this name in lowercase letters (*BMJ case reports*). The latter are indexed via *Medline* in *Scopus*. Neither of these duplicates has been cited.

Another similar case was observed for *Current Topics Behavioral Neurosciences* (Table 2, Cause C2), for which 20 duplicated records with differences in the capitalization of the last three words (*topics*, *behavioral*, and *neurosciences*) of the journal title were found; those that listed the title in lowercase letters were indexed in *Medline* (Fig. 2). Of these 40 records, 33 were cited, with an average difference of 6.2 citations between duplicates (SD = 6.83).

Table 2
Duplicates and citation of different journals by the same publisher, renamed journal, journals with title spelling differences and journals with title variations.

Cause	Journal title	Period indexed in *Scopus*	*N* of papers indexed in *Scopus*	*N* (%) of duplicates	*N* of duplicates with citations	Mean ± SD of the difference in citations per duplicate
A	Psychoterapia[a]	2001–2014	393	37 (9.41%)	5	1.29 ± 0.49
	Archives of Psychiatry and Psychotherapy[a]	2001–2014	394	37 (9.39%)	4	
B	Journal of Korean Academy of Nursing[a]	2008–2014	587	62 (10.56%)	35	2 ± 1.79
	Taehan Kanho Hakhoe chi[b]	2004–2008	624	62 (9.94%)	41	
C1	BMJ Case Reports[a]	2008–2014	13,785	3736 (27.1%)	0	–
	BMJ case reports[b,c]				0	
C2	Current Topics Behavioral Neuroscience[a]	2009–2014	244	20 (8.2%)	20	6.2 ± 6.83
	Current topics behavioral neuroscience[b,c]				13	
D1	Emergency Medical Journal[a]	1996–2014	5217	4 (0.08%)	1	4.5 ± 7
	Emergency medicine journal: EMJ[b,c]				2	
D2	Proceedings of the Royal Society B: Biological Sciences[a]	1946–2014	12,875	378 (2.94%)	373	17.96 ± 25.13
	Proceedings. Biological sciences/The Royal Society[b,c]				373	

Abbreviations: %, percentage; *N*, number; SD, standard deviation.
[a] All *Scopus* coverage except for *Medline* coverage in *Scopus*.
[b] *Medline* coverage in *Scopus* (generated through *Medline* and indexed in *Scopus*).
[c] *Medline* journal title variation.

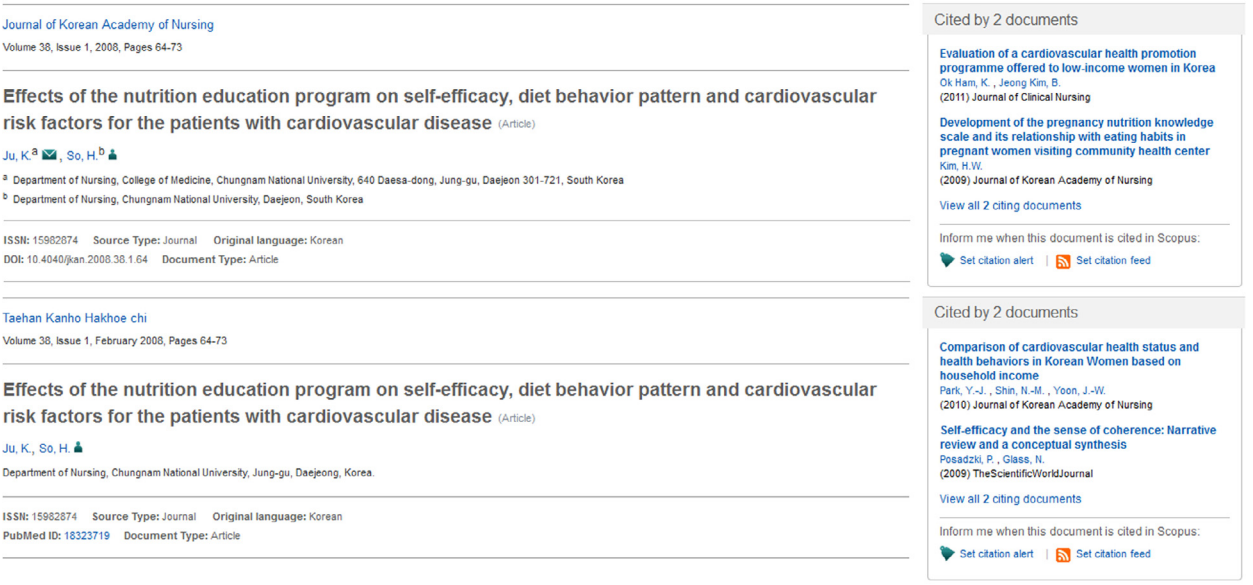

Fig. 1. Example of duplicate records in *Scopus* resulting from journal name changes.

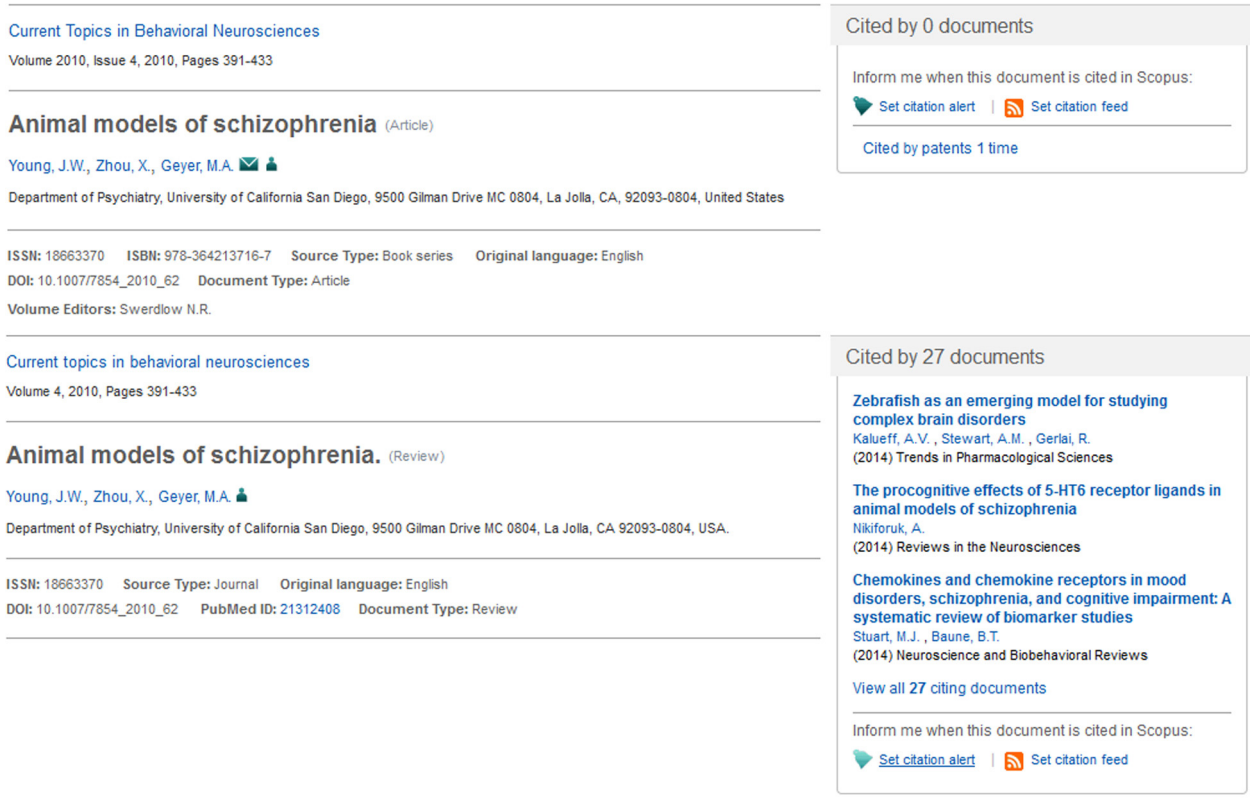

Fig. 2. Example of duplicate records in *Scopus* due to orthographic differences in journal title.

3.4. Duplicates due to journal title variations

Four cases of duplicate records were identified in the *Emergency Medicine Journal* (Table 2, Cause D1) where the duplicate entry was listed under the title *Emergency medicine journal: EMJ*. The duplicate records for the journals with title variations were generated through *Medline* and were indexed in *Scopus*. Three of the records were cited, with an average difference of 4.5 citations between duplicates (SD = 7).

A total of 378 duplicates were identified in the *Proceedings of the Royal Society B: Biological Sciences* (Table 2, Cause D2) when also considering the alternate title *Proceedings. Biological sciences/The Royal Society*, which is found through *Medline* via *Scopus*. One study that was cited 495 times not only was the most cited record in our examination but also exhibited the greatest difference in the number of citations between duplicates, with the other duplicate being cited 128 times (Fig. 3). The second highest citation difference between duplicates was 144, and the third highest was 125 citations. A total of 15

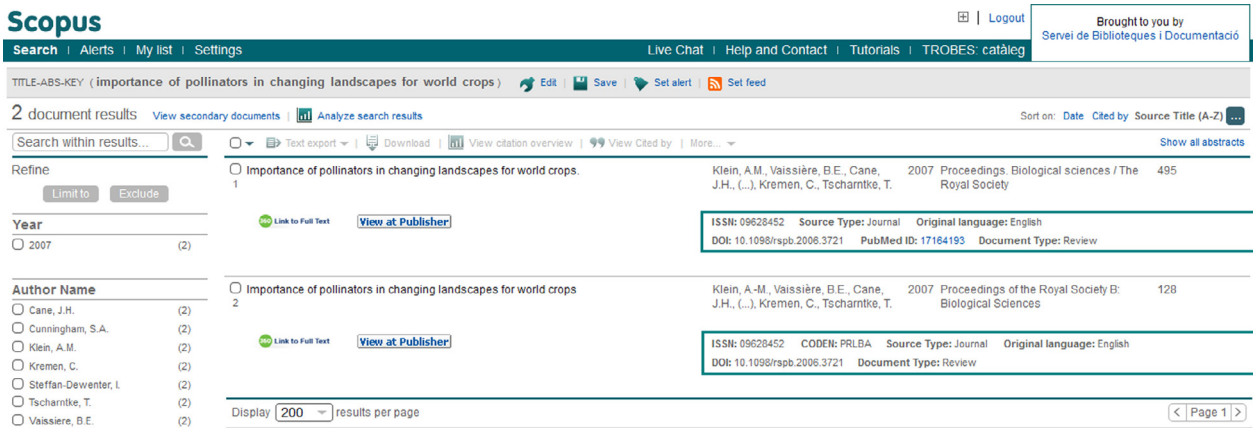

Fig. 3. Example of duplicate records in *Scopus* due to journal title variation.

duplicate results yielded between 93 and 54 citation differences; 200 duplicate results yielded between 49 and 10 citation differences, and 153 duplicate results yielded between nine and one citation differences. The overall average difference in citations between duplicates was 17.96 (SD = 25.13).

4. Discussion and conclusion

Duplicates accounting for 0.08% to 27.1% of the indexed articles were detected. Duplicate records were found to result when articles were mapped to different journals from the same publisher *Polish Psychiatric Association*, as a result of journal title changes, orthographic differences in the presentations of journal names, and journal name variants. In these last three cases, one bibliographic record of each duplicate is mapped to *Medline* coverage of *Scopus*.

Duplicate records by *Scopus* contribute to increasing the number of published works by authors and the institutions participating in them. In the case of having received citations, the authors and their institutional affiliations can have some citations distributed in two records of *Scopus* when the citations actually belong to a single work. Both pairs of duplicates can have equal number of citations (Fig. 1), different number of citations (Fig. 3), or one of the duplicates can be cited while the other is not (Fig. 2). For example: suppose that papers A1 and A2 are duplicates. Suppose further that work B cites paper A. It is possible that the citation from work B to paper A is counted twice, once for publication A1 and once for publication A2; or it is also possible that the citation is counted only once, either for publication A1 or for publication A2.

Significant differences in the number of citations in duplicate articles are observable. There are significant variations in the number of times that duplicates are cited where the minimum mean of the difference of citations per duplicate is 1.29 and the maximum is 17.96, with one case involving *Proceedings of the Royal Society B: Biological Sciences* and *Proceedings. Biological sciences/The Royal Society* showing a difference of 367 citations.

In addition to these findings, studies from García-Pérez (2010) about the quality of citation links in the *Web of Science Core Collection* and from Harzing (2012) about the quality of the document type classification in the *Web of Science Core Collection* highlight the need for database managers to implement rigorous quality-control guidelines to prevent indexing errors.

Journal publishers may consider systematically ascribing DOIs to articles, and database editors should more systematically harvest the existing DOIs from articles (when DOIs exist); a DOI uniquely identifies a record, which makes it easy to detect DOI duplicates. This step may be complemented with classic identifiers such as the International Standard Serial Number (ISSN) and International Standard Book Number (ISBN) or used in tandem with new identifiers, such as the Open Researcher and Contributor ID (http://orcid.org/).

Based on samples of duplicates identified in our study and in a recent study by Franceschini, Maisano, and Mastrogiacomo (2015), who found non-duplicate records with the same DOI in *Scopus*, it seems possible that *Scopus* is failing to check DOIs (see Figs. 2 and 3). There may also be a failure in ISSN checking between journals with title variations included in *Medline* (Fig. 2).

In conclusion, we strongly recommend that studies that evaluate scientific activity systematically confirm that duplicate records are not counted twice, even when records are obtained exclusively from one database. Indeed, a failure to do so could result in errors in author, institution, country, and journal productivity rankings and in the determination of citations.

Future research will calculate the indicators of *SCImago Journal Rank* (http://www.scimagojr.com/) and *Source Normalized Impact per Paper* (http://www.journalindicators.com/) using the productivity and citation data from each journal analyzed in this study. Additionally, we will systematically study the locations from which the received citations in the duplicates derive, quantifying them when a bibliographic record cites both pairs of duplicate records or only one (see Fig. 1), and we will quantify other deficiencies related to incorrect mapping of the type of document and to incorrect data in the volume and issue fields (Fig. 2).

References

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

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

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

Baykoucheva, S. (2010). Selecting a database for drug literature retrieval: A comparison of Medline, Scopus, and Web of Science. *Science & Technology Libraries*, *29*, 276–288.

Boyle, F., & Sherman, D. (2005). Scopus™: The product and its development. *Serials Librarian*, *49*, 147–153.

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

De Granda-Orive, J. I., Alonso-Arroyo, A., García-Río, F., Solano-Reina, S., Jiménez-Ruiz, C. A., & Aleixandre-Benavent, R. (2013). Ciertas ventajas de Scopus sobre Web of Science en un análisis bibliométrico sobre tabaquismo. *Revista Española de Documentación Científica*, *36*, e011.

Elsevier. (2014a). *Scopus Content Overview.* , <http://www.elsevier.com/online-tools/scopus/content-overview> Accessed on September 2014

Elsevier. (2014b). *Scopus.* , <http://www.elsevier.com/online-tools/scopus> Accessed on September 2014

Elsevier. (2014c). *Scopus Timeline.* , <http://10.scopus.com/timeline/> Accessed on September 2014

Falagas, M. E., Pitsouni, E. I., Malietzis, G. A., & Pappas, G. (2008). Comparison of PubMed, Scopus, Web of Science, and Google Scholar: Strengths and weaknesses. *The FASEB Journal*, *22*, 338–342.

Franceschini, F., Maisano, D., & Mastrogiacomo, L. (2015). Errors in DOI indexing by bibliometrics databases. *Scientometrics*, *102*, 2181–2186.

García-Pérez, M. A. (2010). Strange attractors in the Web of Science database. *Journal of Informetrics*, *5*, 214–218.

Gavel, Y., & Iselid, L. (2008). Web of Science and Scopus: A journal title overlap study. *Online Information Review*, *32*, 8–21.

Gorraiz, J., & Schloegl, C. (2008). A bibliometric analysis of pharmacology and pharmacy journals: Scopus versus Web of Science. *Journal of Information Science*, *34*, 715–725.
Harzing, A. W. (2012). Document categories in the ISI Web of Knowledge: Misunderstanding the Social Sciences? *Scientometrics*, *94*, 23–34.
Jacso, P. (2005). As we may search: Comparison of major features of the Web of Science, Scopus, and Google Scholar citation-based and citation-enhanced databases. *Current Science*, *89*, 1537–1547.
Kulkarni, A. V., Aziz, B., Shams, I., & Busse, J. W. (2009). Comparisons of citations in Web of Science, Scopus, and Google Scholar for articles published in general medical journals. *JAMA*, *302*, 1092–1096.
LaGuardia, C. (2005). E-views and reviews: Scopus vs. Web of Science. *Library Journal*, *130*, 40–42.
López Illescas, C., Moya Anegón, F., & Moed, H. F. (2008). Coverage and citation impact of oncological journals in the Web of Science and Scopus. *Journal of Informetrics*, *2*, 304–316.
Meho, L. I., & Rogers, Y. (2008). Citation counting, citation ranking, and h-index of human–computer interaction researchers: A comparison of Scopus and Web of Science. *Journal of the American Society for Information Science and Technology*, *59*, 1711–1726.
Meho, L. I., & Sugimoto, C. R. (2009). Assessing the scholarly impact of information studies: A tale of two citation databases – Scopus and Web of Science. *Journal of the American Society for Information Science and Technology*, *60*, 2499–2508.
Torres-Salinas, D., Lopez-Cozar, E. D., & Jimenez-Contreras, E. (2009). Ranking of departments and researchers within a university using two different databases: Web of Science versus Scopus. *Scientometrics*, *80*, 761–774.
Vieira, E. S., & Gomes, J. A. N. F. (2009). A comparison of Scopus and Web of Science for a typical university. *Scientometrics*, *81*, 587–600.