Data quality and consistency in Scopus and Web of Science in their indexing of Czech Journals

Pavel Mika,* Jakub Szarzec** and Gunnar Sivertsen***

* mika@knav.cz
Library of the Czech Academy of Sciences,
Národní 1009/3, 115 22 Praha 1, Czech Republic

** jakub.szarzec@techlib.cz
National Library of Technology,
Technická 6/2710, 160 80 Praha 6 – Dejvice, Czech Republic

*** gunnar.sivertsen@nifu.no
Nordic Institute for Studies in Innovation, Research and Education (NIFU)
P.O. Box 2815 Tøyen, N-0608 Oslo, Norway

ABSTRACT
This study addresses the discussion of "quality versus coverage" that often arises if a choice is needed between Scopus and Web of Science (WoS). We present a new methodology to detect problems in the quality of indexing procedures. Our preliminary findings indicate the same degree and types of errors in Scopus and WoS. The more serious errors seem to occur in the indexing of cited references, not in the recording of traditional metadata.

Keywords
Scopus; Web of Science; data quality; journal coverage; citation indexing; references; Levenshtein distance.

Submission type: Research in progress paper.
Relevant track: Data infrastructure for research metrics.

INTRODUCTION
This study addresses the discussion of "quality versus coverage" that often arises if a choice is needed between Scopus and Web of Science (WoS). With regard to coverage of source documents and citations, there are large differences in favour of Scopus, although there is not full overlap with WoS content (Gavel & Iselid 2008). The consequences of different coverage depend on the purpose of a particular usage. The two data sources need to supplement each other from an information retrieval perspective (Bar-Ilan 2010). They can, however, replace each other as the basis for indicators of scientific production and citations at the country level (Archambault et al. 2009), but less so at the level of institutions (Vieira & Gomes 2009) or in fields of research that tend to be marginally covered in both sources (Bartol et al. 2014; Haddow & Genoni 2010; Sivertsen 2014).

The quality and consistency of citation indexing procedures are important for all purposes, however. Franceschini et al. (2015) recently published indications of serious types of errors in Scopus that WoS is not free from either. Our study aims at resolving the same question of

data quality. We present a new methodology to reach this aim. Our preliminary findings indicate fewer errors and less difference in this respect between Scopus and WoS than we expected from the earlier study. More serious errors seem to occur in the indexing of cited references, not in the recording of traditional metadata. Our further research – also to be presented at the conference – will clarify the extent of this problem.

METHODS

We chose to study journals published by organizations or publishers in the Czech Republic. The reason for this choice is that we wanted to compare Scopus and WoS mainly where they differ: in coverage of the "periphery" of the international core journals. We chose the Czech Republic because the printed versions of the indexed journals are easily available to us. There are 49 Czech journals in the 2014 edition of the Journal Citation Report (WoS) and 159 Czech journals in the 2014 Scopus Journal Title List. Among these, 46 journals are indexed in both databases. They cover Agriculture, Chemistry, Business Economics, Engineering, Plant Sciences, Food Science Technology, Veterinary Sciences, Entomology, Physiology and Microbiology. Most of them (84 per cent) are published in the English language; some are bilingual; the remaining few publish in the Czech language only.

We downloaded the data manually in early December 2015 using the web interface of each database. The queries were limited by ISSN for five years, 2010-2014. We retrieved 13,281 records from Scopus and 13,947 records from WoS in the same 46 journals. The completeness of both downloads was checked against the online versions of the databases after download.

Matching supposedly identical records was crucial in the preparation of data for further analysis. We used an iterative process in several phases where we combined manual and automatic methods based on the Levenshtein distance metric. We were able to match a total of 12,494 records. The matched records thereby constituted 94 percent of the records retrieved from Scopus and 90 percent of the records retrieved from WoS.

The quality and consistency of the data in the two databases was studied by making two types of systematic comparisons. First, the matched records were compared to each other to study possible differences in indexing between the two databases. Second, all records, including those that could not be matched, were compared to the electronic archives of the indexed journals. In addition, two of the journals were analysed using their printed versions. In both types of comparisons, the official indexing policies of the two databases (Scopus Elsevier 2016; Thomson Reuters 2016), which are not identical, provided important guidelines with regard to expected outcomes.

RESULTS

The results of the comparison of the 12,494 matched records are shown in Table 1.

Table 1. Comparison of selected fields

WoS abrevation/name of field		Scopus name of field	Number of identical (provided) values	Base for percentage	% of identical (provided)	Comparison method
AU	Authors	Authors	12,405	12,494	99.3	Number
TI	Document Title	Title	8,394	12,383	67.8	Levensthein
DT	Document Type	Document Type	11,424	12,318	92.7	Identical YES/NO
TC	Times Cited	Cited by	3,713	12,494	29.7	Number
PY	Year Published	Year	12,452	12,494	99.7	Identical YES/NO
VL	Volume	Volume	12,325	12,494	98.4	Identical YES/NO
IS	Issue	Issue	11,766	12,494	94.1	Identical YES/NO
BP	Beginning Page	Page start	12,302	12,494	98.4	Identical YES/NO
EP	Ending Page	Page end	11,944	12,494	95.6	Identical YES/NO
DI	Digital Object Identifier (DOI)	DOI	2,235	2,296	97.3	Identical YES/NO
LA	Language	Language of Original Doc.	11,186	12,494	89.5	Identical YES/NO
DE	Author Keywords	Author Keywords	12,015	12,494	96.1	Number
AB	Abstract	Abstract	11,901	12,494	95.3	Provided YES/NO
NR	Cited Reference Count	Reference count	3,376	4,445	76.0	Number

Generally, we find a high degree of consistency in indexing between the two databases, measured as the percentage identical data in each field, with one important exception, the number of references. All smaller or larger differences between the two databases can be technically explained without altering the general impression that the metadata are of relatively high quality in both databases. Here are several explanations before we turn a discussion of the exception:

- A higher rate of identical titles (68%) could not be expected, because 20 percent of the Scopus titles are multilingual. Other differences were caused mainly by the transcription of technical terms using the Greek alphabet into Latin, for Scopus titles.
- The number of times cited is expected to be different because the two databases cover different numbers of source journals.
- The differences in document type classification are mainly explained because the two resources use different classification schemes. The differences are small. The most common differences are shown in Table 2.

Table 2. Document type differences

WoS doc. type	Scopus doc. type	Number of docs.	% of explored dataset
Article	Review	208	1.7
Review	Article	205	1.7
Proceedings Paper	Article	171	1.4
Editorial Material	Article	139	1.1
Article	Proceedings Paper	107	0.9
Editorial Material	Review	17	0.1

Document type information is important in bibliometric analysis in order to normalize citation indicators. Our results indicate that this type of information is relatively reliable. However, even more important is the indexing of the reference lists in each document. An exception to the finding that metadata are of high quality is the indication we get as we see that 24 percent of the matched records have different reference counts in Scopus and WoS. This is a clear indication that the reference lists in the source documents are not appropriately or fully indexed.

We found 222 WoS records with more references than in Scopus and 847 Scopus records with more references than in WoS. The number of missing references for each comparison is shown in Table 3. The most common difference (12%) was caused by one missing reference in WoS records.

Table 3. Differences in number of references

Reference difference	Number of records	% of records	Number of missing references
WoS>SC	222	5	-1,913
SC>WoS	847	19	2,005
SUM	1,069	24	92

This observation of differences was the starting point for further research when we tried to compare all references from observed records. Unfortunately we still weren´t able to match all the references to find out any pattern in missing (or excess) references.

In the second part of the study, we compared matched as well as unmatched records (Scopus versus WoS) to the electronic archives of the 46 indexed journals. A total of 17,759 records could be used for the study of how and to what extent the journals are indexed. A quantitative overview is given for each of the journals in Table 4 (Appendix). Here, we compare the

number of records in the original source journal to the number of records indexed in WoS and Scopus and the number of records that could be matched between them. No numbers are the same for any of the journals and there are wide differences for some journals. The right column in Table 4 (A-C) refers to the following explanations for the differences:

A. There are only small differences for nine journals. The differences can mainly be explained because of differently defined document types used for indexing hybrid journals with a large array of document types.
B. There are larger differences between Scopus and WoS for 25 journals; however, the number in one of the databases resembles the number of records in the original source. The differences between the two indexing databases can be explained by differing indexing policies, with the exceptions below.
C. There are large differences between the original sources and the two indexing databases for nine journals. In these cases, we found that the electronic archive of the journal does not cover the journal completely or the archive includes supplemental items not published in the regular journal.

An example of C is *Chemicke listy* (0009-2788), where the archive includes supplementary material such as conference abstracts of plenary lectures, oral sessions and posters.

Differences of type B were examined by inspecting the printed versions of two journals. In *Folia Biologica* (ISSN 0015-5500), we discovered that the larger number of records in Scopus was caused by an error in which 71 records from a Polish journal with the same name but different ISSN (0015-5497) were included. We also found two instances of duplicate records in Scopus. All in all, we found 14 cases of the duplicate Scopus records in the whole dataset, which is less than expected from earlier studies of the same error (Valderrama-Zurián et al. 2015).

Inspecting *Československá psychologie* (0009-062X) in the same way, we found that neither Scopus nor WoS covered this journal completely. In spite of the indexing policy, 12 items were not indexed in WoS – mostly news, errata, and discussions. Of 214 items not indexed by Scopus, 51 were classified as research articles in WoS. If this classification is correct, they should have been indexed in Scopus according to its policy. The other missing items in Scopus can be explained by the policy of not indexing such items.

DISCUSSION AND FOCUS FOR FURTHER RESEARCH

We have established a methodology for two types of comparisons that aim to test the quality and consistency of the data and indexing in Scopus and WoS, by:

a. Matching and measuring the degree of similarity in supposedly identical records in both databases.
b. Comparing data from both databases to the sources that were indexed.

With both methods, most of the differences we observed could be explained according to differing methods and policies for indexing in Scopus and WoS or the specific publishing policies of journals.

There are two major exceptions, however, that will be the focus of our further studies:

a. Differences in the number of cited references in a record may be an indication that reference lists in the source documents are not appropriately or fully indexed.
b. Differences between the number of records in the archive of the source journal and the databases can be an indication that the contents are not appropriately or fully indexed.

REFERENCES

Archambault, É, Campbell, D., Gingras, Y., & Larivière, V. (2009). Comparing bibliometric statistics obtained from the Web of Science and Scopus. Journal of the American Society for Information Science and Technology, 60(7), 1320-1326.

Bar-Ilan, J. (2010). Citations to the "Introduction to informetrics" indexed by WOS, Scopus and Google Scholar. Scientometrics, 82(3), 495-506.

Bartol, T., Budimir, G., Dekleva-Smrekar, D., Pusnik, M., & Juznic, P. (2013). Assessment of research fields in Scopus and Web of Science in the view of national research evaluation in Slovenia. Scientometrics, 98(2), 1491-1504.

Franceschini, F., Maisano, D., & Mastrogiacomo, L. (2016). The museum of errors/horrors in Scopus. Journal of Informetrics, 10(1), 174-182.

Gavel, Y., & Iselid, L. (2008). Web of Science and Scopus: A journal title overlap study. Online Information Review, 32(1), 8-21.

Haddow, G., & Genoni, P. (2010). Citation analysis and peer ranking of Australian social science journals. Scientometrics, 85(2), 471-487.

Scopus Elsevier. (2016). Content Policy and Selection. (https://www.elsevier.com/solutions/scopus/content/content-policy-and-selection) [21 March 2016].

Sivertsen, G. (2014). Scholarly publication patterns in the social sciences and humanities and their coverage in Scopus and Web of Science. In E. Noyons (Ed.), Proceedings of the science and technology indicators conference 2014 Leiden (pp. 598-604). Leiden: CWTS.

Thomson Reuters. (2016). The Thomson Reuters Journal Selection Process. (http://wokinfo.com/essays/journal-selection-process/) [21 March 2016].

Valderrama-Zurián, J., Aguilar-Moya, R., Melero-Fuentes, D., & Aleixandre-Benavent, R. (2015). A systematic analysis of duplicate records in Scopus. Journal of Informetrics, 9(3), 570-576.

Vieira, E. S., & Gomes, J. A. (2009). A comparison of Scopus and Web of Science for a typical university. Scientometrics, 81(2), 587-600.

Table 4. Number of records in the original source journal compared to the number of records indexed in WoS and Scopus, and the number of records that could be matched between them. The right column (A-D) refers to explanations for the differences given in in the text.

Journal title abbrev.	Source N	WoS N	SC N	Matched N	Differences
Acta Ent Mus Nat Pra	293	291	241	239	B
Acta Geodyn Geomater	216	218	217	217	A
Acta Chir Orthop Tr	393	346	360	340	C
Acta Vet Brno	388	398	388	388	B
Agr Econ-Czech	304	302	292	292	B
Appl Math-Czech	176	175	169	169	B
Biomed Pap	728	346	344	317	C
B Geosci	234	233	245	220	B
Cent Eur J Publ Heal	233	224	286	219	B
Ceram-Silikaty	278	280	278	278	A
Cesk Slov Neurol N	668	594	574	536	C
Cesk Psychol	223	378	176	171	B
Czech J Anim Sci	331	320	317	317	C
Czech J Food Sci	408	406	386	386	B
Czech J Genet Plant	208	207	192	191	B
E M Ekon Manag	237	286	234	232	B
Epidemiol Mikrobi Im	193	171	165	132	C
Eur J Entomol	445	409	403	398	C
Financ Uver	136	137	136	130	A
Folia Biol-Prague	199	198	272	198	B
Folia Geobot	144	136	137	135	C
Folia Microbiol	439	439	457	437	B
Folia Parasit	237	232	211	211	B
Folia Zool	198	193	194	191	A
Fottea	106	108	103	103	A
Hortic Sci	128	128	126	126	A
Chem Listy	4,160	1,290	1,254	1,033	C
J Appl Biomed	133	129	106	103	B
J Geosci-Czech	127	126	111	110	B
Kybernetika	344	343	337	335	B
Listy Cukrov Repar	603	416	456	390	C
Morav Geogr Rep	109	86	108	84	B
Neural Netw World	218	212	205	198	A
Photosynthetica	379	373	401	364	B
Physiol Res	631	634	619	610	B
Plant Protect Sci	134	97	132	88	B
Plant Soil Environ	445	437	437	437	A
Polit Ekon	266	279	218	214	B
Prague Econ Pap	136	138	128	127	B
Preslia	127	127	127	120	A
Radioengineering	728	735	725	724	A
Slovo Slovesnost	177	160	73	66	B
Sociol Cas	760	471	204	197	C
Soil Water Res	111	110	111	110	A
Stud Geophys Geod	224	224	232	223	B
Vet Med-Czech	404	405	394	387	B