

September 2008

Appraisal of Citation Data Sources

**A report to HEFCE by the Centre for Science and Technology Studies,
Leiden University**

By Henk F. Moed and Martijn S. Visser
Centre for Science and Technology Studies (CWTS)
Leiden University
PO Box 9555
2300 RB Leiden, the Netherlands
Tel.: ++31 71 5273909
Fax: ++31 71 5273911
Email: moed@cwts.leidenuniv.nl

Contents

Summary	1
Section 1 Introduction	4
Section 2 Methodology	6
Section 3 Elsevier's Scopus	8
3.1 Scopus versus Web of Science	8
3.2 Scopus coverage of publications submitted to 2001 RAE	10
3.3 Scopus coverage of ACM and IEEE computer science articles	14
3.4 Other aspects	14
Section 4 Thomson's ISI Proceedings	16
Section 5 Discussion and conclusions	17
Acknowledgements	20
References	20

Summary

1. This report presents a study for the development of a new research assessment and funding system, commissioned by the Higher Education Funding Council for England (HEFCE).
2. During the past 50 years bibliometric indicators of research performance based on publication and citation counts were almost exclusively derived from the citation indexes (currently the Web of Science) invented by Eugene Garfield and published by Thomson Scientific (formerly the Institute for Scientific Information). The objective of the study presented in this report is to discuss the potential usefulness of citation databases other than the Web of Science (WoS) and to propose criteria for assessing such databases. It focuses on two citation databases:
 - a. Scopus, a new large, multidisciplinary citation index launched by Elsevier. Scopus claims to cover some 15,000 sources, including a number of proceedings volumes of important international conferences, especially in the applied and engineering sciences.
 - b. ISI Proceedings, published by Thomson Scientific. This covers numerous conference proceedings and journals not covered by the WoS. Use of this database is expected to be fruitful especially in Engineering and Computer Science, a field in which the coverage by the WoS has been found to be moderate.
3. In assessing the adequacy of a citation index for evaluative-bibliometric purposes, it is proposed to take into account the following three issues:
 - How adequate is the coverage of fields covered by the database
 - How accurate are the citation links established in the database?
 - Is the information on authors and their affiliations complete and well structured?
4. The core of the study presented in this report provides a systematic analysis of the *coverage* of Scopus in the various domains of science and scholarship. Since the WoS has become a standard in bibliometric research, this database will be used as a benchmark.
5. Any comparison of these two databases is hampered by the fact that both are in continuous development. Source coverage is being expanded, backlogs are being added and data capturing and standardization are being improved. Specific outcomes may therefore become quickly obsolete. In view of this, this report can only present general patterns and conclusions.
6. The Scopus database used in this analysis is based on raw data provided by the Scopus team on 1 August 2007. Additions and corrections made after this date are *not* included. The WoS database used in the analysis is based on annual deliveries of raw WoS data by Thomson Scientific up until and including 2006.

7. The outcome of the comparison of WoS and Scopus coverage at the level of individual articles indicates that in science-related fields the overwhelming part of articles and reviews in journals covered by the WoS is also included in Scopus. The overall percentage of WoS-covered, science-related papers found in Scopus increased over the years, from 89 per cent in 1996 to 97 per cent in 2005. In other words, in these fields and for published articles and reviews the WoS now constitutes almost a complete subset of Scopus.

8. Even the 2005 Scopus dataset used in this study contains 'gaps'; about 400 journals are not fully covered. This set includes 122 journals for which Scopus covers only 5 per cent or less of the number of articles and reviews indexed in the WoS. The Scopus team informed the authors of this report that during past months a major effort has been made to fill the gaps it had detected in its internal quality control processes. The results of this effort could not be evaluated in the current study.

9. The comparison of WoS and Scopus coverage of the 'best' publications submitted to the 2001 RAE showed that Scopus coverage is specifically better in the Subject Group Subjects Allied to Health and, to a lesser extent, also in Engineering & Computer Science and Health Sciences. In Clinical Medicine, Biological Sciences and Physical Sciences, however, Scopus coverage is slightly lower than WoS coverage.

10. As regards the accuracy of citation linking within Scopus, it must be underlined that this database includes, in principle, all information on a cited work that is given in a cited reference, including all authors and (if available) even the title of the cited work. The WoS includes only the first author, abbreviated source title (at most 20 characters), publication year, volume number and starting page number. The additional information in Scopus makes possible, in principle, the development of citation linking algorithms that are more accurate than those based on the more limited information on a cited work captured by Thomson Scientific.

11. With respect to information on authors and their institutional affiliations, it must be noted that the WoS preserves, in principle, only the link between the first or the reprint author and his or her address. For all other authors there is no link between author and address in the database. Scopus, however, keeps this link between author and affiliation for each author of a paper. This property is expected to make the process of assigning papers to individuals or departments much easier.

12. The findings presented in this report suggest that the criteria for selecting sources are rather different among the two databases. The WoS's coverage is primarily based on Eugene Garfield's concept of measuring the importance of journals on the basis of their citation impact, as well as including the most important ones as sources in the database. Scopus coverage is more comprehensive and the citation impact of journals is apparently less discriminative, although it includes the overwhelming part of WoS journals in science-related fields.

13. A further characterization of the coverage surplus of Scopus in qualitative terms deserves special attention. This study found that Scopus covers a number of proceedings

of international conferences published by the Association for Computing Machinery (ACM) and the Institute of Electrical and Electronic Engineers (IEEE). In view of the importance of these sources, their inclusion in a citation index not only enhances the quantity, but also the quality, of its coverage of the field Computer Science. But studies related to other research fields revealed that not all Scopus journals absent from the WoS are equally important.

14. In a study of a large medical field, oncology, it was found that the Scopus journals not indexed in the WoS tend to have much lower impact factors than WoS-covered journals. Moreover, compared to the oncological journals in both WoS and Scopus, the Scopus oncological journals not included in the WoS tend to be published in more diverse countries, be published in more non-English languages and be more recently founded (López-Illescas et al., 2008).

15. More research into the quality of the sources indexed by Scopus across research fields is needed in order to obtain a better understanding of its coverage and of its utility for evaluative-bibliometric purposes. Equally important, the implications of the use of a comprehensive citation index for the construction of citation-based indicators needs to be further explored. In particular, the potentialities of defining and applying within Scopus sub-universes of publications and citations deserve special attention, in which journals are selected or discarded according to their citation impact.

16. Such an analysis is based on the notion that, if one uses Scopus as the data source in a bibliometric study of research performance, it is not necessary to include all sources covered by the index in the analysis. Creating an 'off-line' bibliometric database of Scopus data enables one to mark a particular sub-universe of sources and to calculate bibliometric indicators within such a sub-universe.

17. Nevertheless, even at this stage of development the conclusion seems justified that Scopus is a genuine alternative to the WoS as a data source for bibliometric indicators of research performance in science-related fields, provided that the gaps ('missing' articles or issues of partly covered journals) from publication year 1996 onward are filled in the near future.

18. The fact that the Scopus database would then be a complete citation index only for sources published from 1996 does not constitute an obstacle in using it for calculating citation-based indicators in HEFCE's upcoming Research Excellence Framework, provided that the indicators relate to the past performance of research staff and departments during a time period of no longer than ten years.

19. At this moment no accurate data are available on the outcomes of paper-by-paper matching of ISI Proceedings against other databases. Nevertheless, since ISI Proceedings covers a large number of proceedings volumes, it is expected to be a valuable addition to the WoS. The two databases jointly can be assumed to cover the literature in many fields more completely, especially in Engineering and Computer Science, a field in which WoS coverage has proven to be moderate. Therefore, Thomson Scientific's plans to incorporate this database more fully into the WoS in 2008 are of great interest.

Section 1 Introduction

20. The Higher Education Funding Council for England (HEFCE) is currently in the process of developing a new framework for the assessment and funding of research which will make use of bibliometric indicators of research quality.

21. Bibliometric indicators of research performance based on publication and citation counts have become more and more important as tools in research evaluation and allocation of research funds. During the past 50 years such indicators were almost exclusively derived from the citation indexes (currently the Web of Science) invented by Eugene Garfield and published by Thomson Scientific (formerly the Institute for Scientific Information).

22. The objective of the study presented in this report is to discuss the potential usefulness of publication databases other than the Web of Science (WoS) and to propose criteria for assessing such databases. It focuses on two citation databases:

a. Scopus, a new large, multidisciplinary citation index launched by Elsevier. Scopus claims to cover some 15,000 sources, including a number of proceedings volumes of important international conferences, especially in the applied and engineering sciences (Scopus, 2008). Scopus supplies a broader coverage than the WoS at the journal level, but it tends to lack the depth of coverage in years, as a large part of its references do not go back further than 1996. Its broader coverage and the fact that it is a citation index make Scopus the most important candidate for further exploration, either as a supplementary or as an alternative data source.

b. ISI Proceedings, published by Thomson Scientific. This covers numerous conference proceedings and journals not covered by the WoS (Thomson Scientific, 2008). In addition, as from 1999 this database includes cited references in source articles and can be considered as a citation index. Use of this database is expected to be fruitful especially in Engineering and Computer Science, a field in which the coverage of the WoS is found to be moderate.

23. In assessing the adequacy of a citation index for evaluative-bibliometric purposes, the three issues presented in the box below are the most relevant.

a) *How adequate is coverage of the fields covered by the database?* It is important to test claims made by the producer of the database about coverage, both its *quantity* and its *quality*. This can be done only if one has a full bibliometric version of the entire database.

b) *How accurate are the citation links established in the database?* This can be assessed by checking whether links that are made in the database are correct and by analysing the extent to which links that are known to be present in the original documents are actually included in the database.

c) *Is the information on authors and their affiliations complete and well structured?* This is especially relevant in the process of assigning papers to individual researchers or research departments.

24. **Section 2** describes how Scopus and WoS were compared. The assessment focuses on their coverage in the various domains of science and scholarship. The WoS was selected as the benchmark because it is normally used in this type of study and has become a standard in bibliometric research.

25. **Section 3** of this report deals with Scopus. **Sections 3.1, 3.2** and **3.3** present a systematic analysis of the coverage of this database compared to the WoS. **Section 3.4** deals with the other two issues indicated in the textbox above: citation linking and author and address information. It highlights general points and preliminary findings. **Section 4** dedicates attention to the second database assessed in this report, Thomson Scientific's ISI Proceedings. Finally, **Section 5** presents a discussion and the main conclusions of the study.

26. Three important, general comments must be made. Firstly, it needs emphasising that during the past 50 years many researchers have explored the use of the Thomson Scientific/ISI citation indexes (Science Citation Index, WoS) for evaluative-bibliometric purposes. CWTS has 25 years of experience with the use of these indexes and has created a bibliometric version of the WoS, denoted here as the CWTS-WoS database. As a result, CWTS has a detailed knowledge of the technical ins and outs of this database. For other databases, including ISI Proceedings and especially Scopus, such a detailed knowledge is as yet only partly developed.

27. Secondly, it needs emphasising that any comparison of these two databases is hampered by the fact that both are in continuous development. Source coverage is being expanded, backlogs are being added and data capturing and standardization procedures are being improved. Specific outcomes may therefore become quickly obsolete. In view of this, this report can only present general patterns and conclusions.

28. A third comment relates to the fact that both Thomson Scientific and Scopus make bibliometric indicators available through the Internet versions of their indexes or in special indicator information products. This report, however, does not discuss the validity, accuracy and usefulness of these indicators themselves, but rather focuses on the source coverage of the underlying databases in science fields, including physical and bio-medical sciences, mathematics and engineering.

Section 2 Methodology

29. In **Section 3** Scopus and WoS coverage are compared at the level of individual papers. To this end, a bibliometric version of the Scopus database was created, based on the raw Scopus data that the Scopus team provided to CWTS. This is denoted here as the CWTS-Scopus database. The raw data set was created on August 1, 2007 and covers the time period 1996-2006. Updates, additions and corrections added to Scopus after that date are not included in the CWTS-Scopus database.

30. In a first analysis, presented in **Section 3.1**, the two databases were matched against one another on a paper-by-paper basis. More specifically, for each individual article or review included in the CWTS-WoS database it was determined whether or not it is included in the CWTS-Scopus database. It is not relevant whether a document in Scopus matched to a WoS paper has the same document type as the corresponding paper in the WoS. If a WoS article or review was matched to a Scopus document, this match was accepted as a proper match, regardless of the document type of this Scopus document.

31. Linking the two databases on a paper-by-paper basis was carried out by defining a series of match-keys describing the base bibliographic information of an article in each database in the same way. A typical example of such a match-key is a string containing certain parts of the name of the first author of a particular article, a part of the journal title in which it was published, the publication year, volume number and the starting page number.

32. The matching process took into account a number of variations in the way the base bibliographic information is captured and formatted in each of the two databases, such as variations in the way author names are formatted. Although the outcomes are not 100 per cent accurate, they can be assumed to be sufficiently accurate to identify general patterns and the main differences in coverage between the two databases.

33. In a second analysis, presented in **Section 3.2**, Scopus and WoS were compared according to the extent to which they cover the 'best' publications submitted to the 2001 RAE. The WoS coverage of these RAE publications has already been analysed in the report "Development of Bibliometric Indicators of Research Quality" submitted to HEFCE on 20 March, 2008 (Moed, Visser and Buter, 2008). In the current report a similar analysis is presented with respect to Scopus.

34. The third analysis compared Scopus and WoS according to the extent to which they cover the publications in a database of Computer Science articles created at CWTS in an earlier study (Moed and Visser, 2007). It is presented in **Section 3.3**. CWTS carried out a pilot study in which research performance in Computer Science was assessed. In this study the CWTS-WoS database was expanded with the proceedings papers from the following two sources:

- The proceedings of over 200 recurring conferences that are made available as part of the ACM (Association for Computing Machinery) Digital Library.
- Data from over 400 recurring conferences accessible through the Digital Library of the Computer Society of the Institute of Electrical and Electronics Engineers (IEEE/CS).

Section 3 Elsevier's Scopus

3.1 Scopus versus Web of Science

35. As outlined in Section 2, the WoS and Scopus databases created at CWTS were matched against one another on a paper-by-paper basis. For each individual article or review included in the CWTS-WoS database it was determined whether or not it was included in the CWTS-Scopus database, regardless of the document type assigned to the corresponding paper in Scopus. The outcomes of this analysis are presented in Table 1 and Figure 1. They provide results for two publication years: 1996 and 2005. For an explanation of Figure 1 the reader is referred to the legend below this figure.

36. **Table 1** shows that 97 per cent of science related papers (articles and reviews only) published in WoS covered journals in 2005 were found in Scopus. For papers related to Social Sciences and Humanities this percentage amounts to 72. The overall percentage of WoS papers included in Scopus is 95. For the year 1996, these three percentages are lower (93, 57 and 89 per cent, respectively).

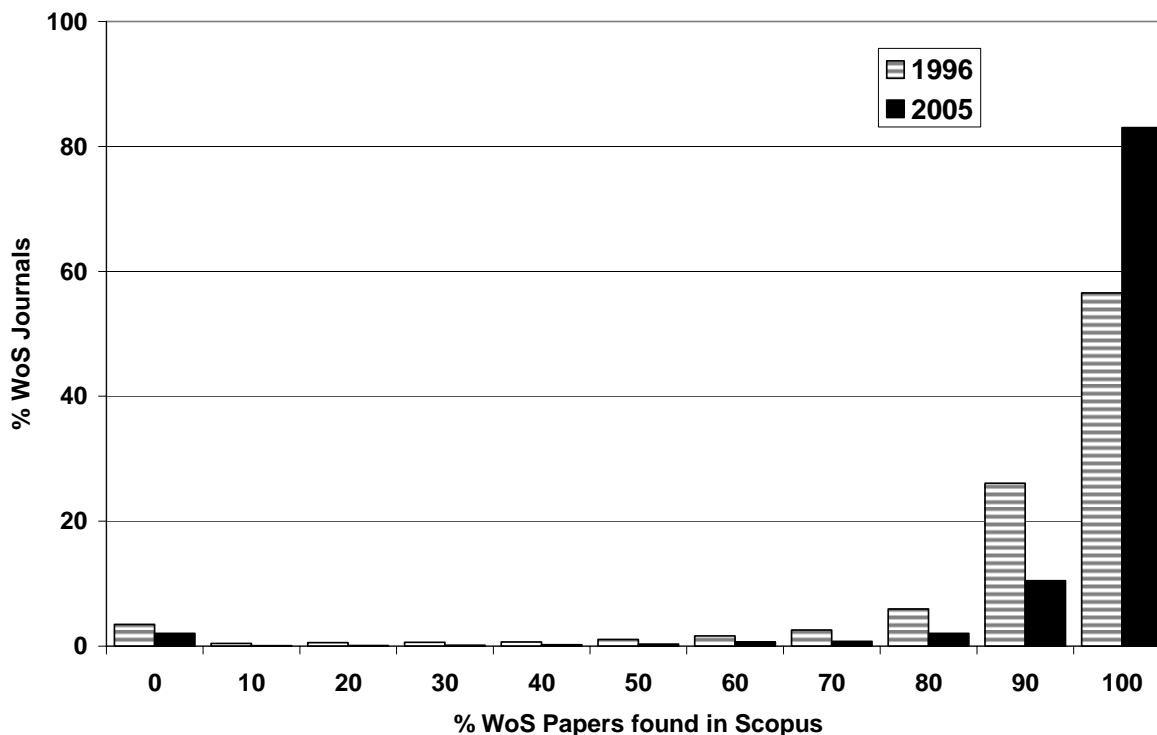
Table 1: Percentage of WoS papers found in Scopus

<i>Publication Year</i>	<i>WoS Database Segment</i>	<i>No. WoS-covered Journals</i>	<i>No. WoS Articles and Reviews</i>	<i>% WoS Articles and Reviews Found in Scopus</i>
1996	Science	5,320	673,271	93 %
1996	Social Sciences & Humanities	2,610	88,583	57 %
1996	Total	7,930	761,854	89 %
2005	Science	6,146	867,748	97 %
2005	Social Sciences & Humanities	2,719	94,629	72 %
2005	Total	8,865	962,377	95 %

37. **Figure 1** shows, for science-related articles and reviews published in 2005, that for 83 per cent of WoS-covered journals the percentage of papers found in Scopus is between 95 and 100 per cent. For 10 per cent of journals it is between 85 and 95 per cent. For almost 2 per cent of journals it is between 0 and 5 per cent. That is, these latter WoS journals are hardly covered or not covered at all by Scopus as of yet. Since the total number of science-related WoS journals in 2005 amounts to 6,146 (see Table 1), the absolute number of WoS journals hardly covered by Scopus or not covered at all amounts to 122. For the year 1996, the percentage of WoS journals not or hardly covered by Scopus amounts to 3.5, while the percentage of WoS journals almost fully covered is 57.

38. Figure 1 also shows that there are some WoS journals for which Scopus covers only between 15 and 85 per cent of papers. That is, these journals are only partly covered, at least in the raw Scopus dataset used in this study. Especially for the year 1996, the fractions of such journals can be substantial. For 17 per cent of WoS journals in 1996 less than 85 per cent of papers was included in Scopus. For the year 2005 this percentage declined to a level of 6 per cent. About 400 journals from this year are not fully covered in the dataset used in the study.

Figure 1: Distribution of % WoS papers found in Scopus across science-related WoS journals



39. Numbers on the horizontal axis indicate midpoints, so 0 means between 0 and 5%; 10 means between 5 and 15% and so on up to midpoint 90, which means between 85 and 95%; and midpoint 100, between 95 and 100%. The figure shows that in 2005 for 83 per cent of WoS-covered journals the percentage of WoS papers (articles and reviews) found in Scopus is between 95 and 100 per cent. For the year 1996 this percentage is 57. The Scopus database used in this analysis is based on raw data provided by the Scopus team on 1 August 2007. Additions and corrections made after this date are not included. The WoS database used in the analysis is based on annual deliveries of raw WoS data by Thomson Scientific up to and including 2006.

3.2 Scopus coverage of publications submitted to 2001 RAE

40. Sections 2.2 and 2.3 of the report by Moed, Visser and Buter (2008) submitted to HEFCE presented an analysis of the extent to which the WoS covers the 'best' publications submitted to the 2001 RAE. The current section presents a similar analysis of the extent to which Scopus covers these 'best' publications. Table 2 presents, per research field, the number of publications submitted to the 2001 RAE, the percentage of papers covered by the WoS and by Scopus, respectively, and in the last column the difference between these two percentages. The WoS coverage results are identical to those presented in Table 3 in Section 2.2 and Table 5 in Section 2.3 of the abovementioned report.

41. **Table 2** presents outcomes per discipline (Science, Mathematics and Social Sciences and Humanities) and per Unit of Assessment (UOA). In the results per UOA, the units in a discipline are ordered by descending difference in the coverage percentage between Scopus and WoS, given in the last column. Table 2 also presents results per Subject Group (SG). The categorisation of UOA into SGs is presented in **Table 3** below and is identical to the one applied in the earlier report to HEFCE. The names of the SGs are the same as those used in the 2008 RAE. However, it must be emphasised that their content is different for medical-biological SGs.

42. At the discipline level the Scopus coverage is slightly higher than the WoS coverage in Science and Social Sciences and Humanities, while it is lower in Mathematics. At the SG level Table 2 shows that Scopus coverage is substantially higher in Subjects allied to Health (the absolute difference between Scopus and WoS coverage amounts to +7.6 per cent) and to a lesser extent also in Engineering & Computer Science (+1.7 per cent) and Health Sciences (+0.6 per cent). In Clinical Medicine, Biological Sciences and Physical Sciences Scopus coverage is slightly lower than WoS coverage (around -2 per cent).

43. Table 2 reveals substantial differences between Scopus and WoS coverage among UOAs. In the Science fields of Nursing, Clinical Dentistry, Civil Engineering, other studies and professions allied to Medicine, Computer Science and Mineral and Mining Engineering the absolute difference between Scopus and WoS coverage is 3 per cent or more. Nursing ranks on top with +16 per cent. For Pharmacology, Food Science and Technology, Physics and Anatomy this absolute difference is -3 per cent or less. In fields related to Social Science Scopus coverage tends to be higher than WoS coverage, whereas in Humanities-related UOAs it is the other way around.

Table 2: WoS versus Scopus coverage of 2001 RAE 'best' publications

<i>Research Field</i>		<i>Total No. submitted publications</i>	<i>% in WoS</i>	<i>% in Scopus</i>	<i>% in Scopus – % in WoS</i>
Discipline					
Science		95,056	84.1	84.4	0.3
Mathematics		6,634	81.8	80.1	-1.7
Social Sciences and Humanities		91,324	24.9	25.9	1.0
Subject Group					
Clinical Medicine*		16,541	96.6	94.9	-1.7
Health Sciences*		10,621	85.0	85.6	+0.6
Subjects allied to Health*		10,203	73.7	81.3	+7.6
Biological Sciences*		16,694	93.5	91.2	-2.2
Physical Sciences		17,190	88.0	86.0	-1.9
Engineering & Computer Science		23,807	70.1	71.8	+1.7
Discipline	Unit of Assessment				
SCI	Nursing	2,509	53.3	69.3	+16.0
SCI	Clinical Dentistry	1,855	87.2	94.3	+7.1
SCI	Civil Engineering	2,162	67.1	73.1	+6.0
SCI	Other studies and professions allied to Medicine	4,146	73.9	79.1	+5.2
SCI	Computer Science	6,148	44.0	47.0	+3.0
SCI	Mineral and Mining Engineering	364	65.9	69.0	+3.0
SCI	Pharmacy	1,693	88.6	90.2	+1.6
SCI	General Engineering	4,245	72.1	73.7	+1.6
SCI	Community-based Clinical subjects	5,490	89.1	90.3	+1.3
SCI	Environmental Sciences	2,375	83.4	84.3	+1.0
SCI	Mechanical, Aeronautical and Manufacturing Engineering	4,351	82.0	82.9	+1.0
SCI	Electrical and Electronic Engineering	3,591	84.9	85.1	+0.2
SCI	Psychology	5,131	80.6	80.5	-0.2
SCI	Agriculture	2,125	86.9	86.6	-0.2
SCI	Chemistry	5,425	87.2	86.5	-0.7
SCI	Metallurgy and Materials	1,733	90.5	89.7	-0.8
SCI	Veterinary Science	1,274	96.5	95.6	-0.9
SCI	Earth Sciences	2,471	82.8	81.8	-1.0
SCI	Physiology	953	94.2	93.0	-1.3
SCI	Hospital-based Clinical subjects	11,586	96.5	94.9	-1.6
SCI	Clinical Laboratory Sciences	4,955	96.6	94.9	-1.7
SCI	Pre Clinical Studies	606	98.2	96.2	-2.0
SCI	Chemical Engineering	1,213	87.5	85.3	-2.1
SCI	Biological Sciences	9,921	94.5	91.9	-2.6
SCI	Pharmacology	745	94.1	90.9	-3.2
SCI	Food Science and Technology	476	89.7	86.1	-3.6
SCI	Physics	6,919	92.0	87.8	-4.2
SCI	Anatomy	594	89.6	84.2	-5.4
Math	Statistics and Operational Research	1,548	76.6	76.7	+0.1

Math	Applied Mathematics	3,008	88.3	86.4	-1.9
Discipline	Unit of Assessment	Total No. submitted publications	% in WoS	% in Scopus	% in Scopus – % in WoS
Math	Pure Mathematics	2,078	76.4	73.5	-2.9
SSHU	Town and Country Planning	1,478	38.1	57.4	+19.4
SSHU	Social Work	1,642	22.9	36.7	+13.8
SSHU	Accounting and Finance	779	21.7	34.9	+13.2
SSHU	Built Environment	2,471	24.5	35.9	+11.4
SSHU	Social Policy and Administration	3,912	25.8	34.1	+8.3
SSHU	Politics and International Studies	4,382	26.4	34.6	+8.1
SSHU	Linguistics	830	26.1	34.2	+8.1
SSHU	Sociology	3,519	29.2	37.0	+7.8
SSHU	Business and Management Studies	9,746	37.9	45.5	+7.6
SSHU	Geography	4,890	61.6	68.6	+7.0
SSHU	Education	8,662	16.0	22.5	+6.5
SSHU	Economics and Econometrics	2,879	67.5	72.0	+4.5
SSHU	Sports-related subjects	1,301	60.5	63.4	+2.9
SSHU	Library and Information Management	1,259	31.7	34.4	+2.7
SSHU	Law	5,314	7.6	9.7	+2.2
SSHU	European Studies	2,210	14.7	15.3	+0.6
SSHU	Anthropology	1,180	27.5	28.0	+0.4
SSHU	Art and Design	2,788	8.4	8.1	-0.2
SSHU	Asian Studies	521	12.9	12.1	-0.8
SSHU	Communication, Cultural and Media Studies	1,360	13.8	12.8	-1.0
SSHU	Middle Eastern and African Studies	610	10.5	8.7	-1.8
SSHU	Archaeology	2,352	16.1	14.2	-2.0
SSHU	Celtic Studies	410	6.1	1.7	-4.4
SSHU	Italian	405	8.4	3.7	-4.7
SSHU	Theology, Divinity, Religious Studies	1,873	9.0	2.8	-6.2
SSHU	History	7,132	18.0	10.2	-7.8
SSHU	American Studies	475	19.8	9.7	-10.1
SSHU	French	1,741	14.9	4.8	-10.1
SSHU	English Language and Literature	5,882	14.1	3.7	-10.4
SSHU	Iberian and Latin American Languages	830	13.9	3.1	-10.7
SSHU	History of Art, Architecture and Design	1,424	12.9	2.2	-10.7
SSHU	Russian, Slavonic and East European Languages	344	13.1	2.0	-11.0
SSHU	Drama, Dance and Performing Arts	1,097	12.9	1.7	-11.2
SSHU	Classics, Ancient History, Byzantine and Modern Greek Studies	1,583	14.4	1.2	-13.2
SSHU	German, Dutch and Scandinavian Languages	1,057	15.6	2.1	-13.5
SSHU	Music	1,142	18.3	1.9	-16.4
SSHU	Philosophy	1,844	34.8	12.4	-22.4

* Categorisation of Units of Assessment into Subject Groups does not fully coincide with 2008 Subject Groups.

44. One comment should be made as regards the accuracy of the matches between 'best' RAE papers on the one hand and the WoS and Scopus databases on the other. The percentage of RAE papers not matched to an article in WoS or Scopus, but nevertheless published in a journal that is covered by one of these databases in the year of publication, is for Scopus 5.6 per cent and for the WoS only 0.6 per cent. This difference can be attributed to two factors. The first is the presence of 'gaps' in the Scopus database, especially in earlier years. The second factor is that the format of author names in Scopus deviates more strongly from that in the RAE database than the format of author names in the WoS does, while the match algorithm did not take into account the variations that emerge from this. As a result, the coverage percentages for the WoS are more accurate than those presented for Scopus.

Table 3: Preliminary categorisation of 2001 RAE Units of Assessment into Subject Groups

<i>Clinical Medicine*</i>	<i>Physical Sciences</i>
Clinical Laboratory Sciences	Chemistry
Hospital-based Clinical subjects	Earth Sciences
	Environmental Sciences
<i>Health Sciences*</i>	Physics
Community-based Clinical subjects	
Psychology	<i>Engineering & Computer Science</i>
	Chemical Engineering
<i>Subjects allied to Health*</i>	Civil Engineering
Clinical Dentistry	Computer Science
Nursing	Electrical and Electronic Engineering
Other studies and professions allied to Medicine	General Engineering
Pharmacy	Mechanical, Aeronautical and Manufacturing Engineering
	Metallurgy and Materials
<i>Biological Sciences*</i>	Mineral and Mining Engineering
Agriculture	
Anatomy	<i>Mathematics</i>
Biological Sciences	Applied Mathematics
Food Science and Technology	Pure Mathematics
Pharmacology	Statistics and Operational Research
Physiology	
Pre Clinical Studies	
Veterinary Science	

* Categorisation of Units of Assessment into Subject Groups does not fully coincide with 2008 Subject Groups.

3.3 Scopus coverage of ACM and IEEE computer science articles

45. Table 4 presents data on the Scopus coverage of articles published during the time period 1996-2005 and included in the ACM Digital Library and in the Digital Library of the Computer Society of IEEE. 64 Per cent of ACM papers and 56 per cent of IEEE/CS articles were found in Scopus. The articles included in these two digital libraries are not covered by the WoS.

Table 4: Scopus coverage of ACM and IEEE computer science articles

Source	No. papers	% papers in Scopus
ACM	36,280	64%
IEEE/CS	76,739	56%

ACM: The proceedings of over 200 recurring conferences are made available as part of the ACM (Association for Computing Machinery) Digital Library.

IEEE/CS: Data from over 400 recurring conferences accessible through the Digital Library of the Computer Society of the Institute of Electrical and Electronics Engineers.

Papers are those published during the time period 1996-2005.

46. It should be noted that an additional analysis showed that the Scopus database covered Springer's Lecture Notes in Computer Science as from 2003. No papers in this source from earlier years were found in the database. On the other hand, the WoS has selective coverage of this source, at least as from 1996.

3.4 Other aspects

How accurate are the citation links established in the database?

47. The accuracy of citation links can be assessed by checking whether links made in the database are correct and by analysing the extent to which links that are known to be present in the original documents are actually included in the database. In order to analyse this, one needs a full bibliometric database, with all relevant data structured in separate fields. Such a comprehensive database based on Scopus data is not (yet) available at CWTS. It is therefore impossible to present systematic empirical data on the accuracy of citation links within Scopus. A limited number of case studies revealed that the citation linking software used to create the Scopus database does take into account a number of errors or variations in author names, publication years, volume numbers and starting page numbers.

48. It must be emphasised that there is an important difference between Scopus and the WoS as regards the way in which these databases capture cited references. While the WoS includes only the first author, abbreviated source title (at most 20 characters), publication year, volume number and starting page number, Scopus includes, in principle, all information on a cited work that is given in a cited reference, including all authors, and (if available) even the title of the cited work. This additional information enables one, in

principle, to develop citation linking algorithms that are more accurate than those based on the more limited information on a cited work captured by Thomson Scientific.

Is the information on authors and their affiliations complete and well structured?

49. This aspect is especially relevant in the process of assigning papers to individual researchers or research departments. CWTS has not yet carried out a detailed assessment of the way this information is structured in Scopus, but one relevant aspect can be underlined in this stage. Scientific papers tend to have multiple authors. An original source paper normally gives a list of authors, followed by a list of their institutional affiliations. In addition, it gives for each author the name of the institution to which he or she is affiliated, often by means of special symbols. The WoS preserves, in principle, only the link between the first or (if available) reprint author and his or her address. For all other authors there is no link between author and address in the database. Scopus, however, keeps this link between author and his affiliation for each author of a paper. This property is expected to make the process of assigning papers to individuals or departments easier. It must be noted that, as from July 2008, Thomson Scientific will start linking each author of a paper included in the WoS to its corresponding address.

Section 4 Thomson's ISI Proceedings

50. ISI Proceedings provides Web access to bibliographic information and author abstracts from papers delivered at prestigious international conferences, symposia, seminars, colloquia, workshops and conventions in a wide range of disciplines (see Thomson Scientific, 2008). Some 385,000 records are added each year. ISI Proceedings is available in two editions, the Science & Technology edition and the Social Sciences & Humanities edition. Together, they claim to cover about 5.2 million papers from over 60,000 conferences. As from 1999 the database includes cited references. All addresses of publishing authors are included as well. Thomson Scientific has plans to incorporate this database more fully into the WoS in 2008.

51. CWTS has no experience as of yet with the use of this database in bibliometric assessments of research performance. Visual inspection of the list of sources covered by ISI Proceedings indicates that it covers a number of proceedings included in ACM Digital Library and IEEE/CS. It must be noted that ISI Proceedings also covers a substantial number of articles included in the WoS. These are mainly articles published in proceedings volumes or issues of 'regular', WoS-covered journals.

Section 5 Discussion and conclusions

52. The comparison of WoS and Scopus coverage at the level of individual articles, as presented in Section 3, indicates that in science-related fields the overwhelming part of articles and reviews in journals covered by the WoS are also included in Scopus. The overall percentage of WoS-covered, science-related papers found in Scopus increased over the years, from 89 per cent in 1996 to 97 per cent in 2005. In other words, in these fields and for published articles and reviews, the WoS constitutes almost a genuine subset of Scopus.

53. The percentage of WoS-covered, science-related journals for which at least 85 per cent of all papers are included in Scopus increased from 83 per cent in 1996 to 94 per cent in 2005. Taking the WoS database as a benchmark, the percentage of journals for which Scopus covered 85 per cent or less of papers (articles and reviews) declined from 17 per cent in the year 1996 to 6 per cent in 2005. Therefore, the Scopus dataset used in this study, which was created on 1 August 2007, contains 'gaps' even in 2005: about 400 journals are not fully covered. This set includes 122 journals for which Scopus covers only 5 per cent or less of the number of articles and reviews indexed in the WoS.

54. While Thomson Scientific processes each journal 'cover to cover' and includes, in principle, all types of documents in the WoS, Scopus is somewhat more restrictive in this respect. Meeting abstracts are normally not included in Scopus. Theoretically it is possible that documents categorized in the WoS as articles or reviews are assigned in the Scopus production process to a document type that Scopus does not include. This factor could lead to a lower coverage of WoS articles and reviews in Scopus. A follow-up study should examine in more detail the differences between WoS and Scopus in the assignment of document types. However, it is unlikely that this factor has a major influence on the results of the analyses presented above.

55. The Scopus team informed the report authors that, during the past few months, a major effort has been made to fill the gaps it had detected in its internal quality control processes. The results of this process of filling the gaps cannot be evaluated in the current study. It can only be done in a follow-up study, based on the most recent Scopus data.

56. The comparison of WoS and Scopus coverage of the 'best' publications submitted to the 2001 RAE showed that Scopus coverage is especially better in the Subject Group Subjects allied to Health, as well as to a lesser extent also in Engineering & Computer Science and Health Sciences. In Clinical Medicine, Biological Sciences and Physical Sciences, however, Scopus coverage is slightly lower than WoS coverage. It must be noted that the 'best' publications submitted to the 2001 RAE were published during the years 1996-2000. In these years Scopus coverage was lower than in later years.

57. The finding that Scopus covers 64 per cent of papers published during the time period 1996-2005 and included in the ACM Digital Library, and 56 per cent of articles in the Digital Library of the Computer Society of IEEE illustrates that Scopus has an added value in Computer Science. On the other hand, it must be noted that the version of the

Scopus database used in this study covers Springer's Lecture Notes in Computer Science as from the year 2003.

58. Section 3.4 underlined two important features of Scopus. Firstly, it includes, in principle, all information on a cited work that is given in a cited reference, including all authors, and (if available) even the title of the cited work. The WoS includes only the first author, abbreviated source title (at most 20 characters), publication year, volume number and starting page number. The additional information in Scopus makes it, in principle, possible to develop citation linking algorithms that are more accurate than those based on the more limited information on a cited work captured by Thomson Scientific.

59. Secondly, the WoS preserves, in principle, only the link between the first or the reprint author and his or her address. For all other authors there is no link between author and address in the database. Scopus, however, keeps this link between author and affiliation for each author of a paper. This property is expected to make the process of assigning papers to individuals or departments much easier.

60. The findings presented in this report suggest that the criteria for selecting sources are rather different among the two databases. The WoS's coverage is primarily based on Eugene Garfield's concept of measuring the importance of journals on the basis of their citation impact and including the most important ones as sources in the database (Garfield, 1964; 1979). Scopus coverage is more comprehensive and the citation impact of journals is apparently less discriminative, although it includes the overwhelming part of WoS journals in science-related fields.

61. How can this surplus of Scopus coverage be characterized in qualitative terms? The abovementioned inclusion in Scopus of a substantial number of proceedings of important international conferences published by ACM and IEEE does not only enhance the quantity, but also the quality, of its coverage of the field Computer Science. But other case studies related to other research fields found that not all Scopus sources that are not included in the WoS are equally important.

62. In a study of a large medical field, oncology, it was found that the Scopus journals not indexed in the WoS tend to have much lower impact factors than WoS-covered journals. In addition, compared to the oncological journals both in WoS and Scopus, the Scopus oncological journals not included in the WoS tend to be published in more diverse countries, be published in more non-English languages and be more recently founded (López-Illescas et al., 2008).

63. The same study also found that countries that have published a large share of papers in the Scopus journals that are not included in the WoS tend to have a strongly reduced average citation rate calculated in Scopus compared to the outcomes based on WoS data. This pattern can be expected to occur also at the level of institutions or even individual authors active within such countries.

64. More research into the quality of the sources indexed by Scopus is needed in order to obtain a more complete insight. It is also important to further explore the implications of the use of a comprehensive citation index for the construction of citation-based

indicators. In particular, the potentialities of defining and applying within Scopus sub-universes of publications and citations in which journals are selected or discarded according to their citation impact deserve special attention. Such an analysis is based on the idea that, if one uses Scopus as the data source in a bibliometric study of research performance, it is not necessary to include in the analysis all sources covered by the index. Creating an 'off-line', bibliometric database of Scopus data enables one to mark a particular sub-universe of sources and to calculate bibliometric indicators within such a sub-universe.

65. Nevertheless, even at this stage of development, the conclusion seems justified that Scopus is a genuine alternative to the WoS as a data source for bibliometric indicators of research performance in science-related fields, provided that the gaps ('missing' articles or issues of covered journals) are filled.

66. The fact that the Scopus database would then be a complete citation index only for sources published from 1996 onward is certainly a limitation in longitudinal bibliometric analyses covering long time periods. However, it does not constitute an obstacle in using it in the construction of citation-based indicators for HEFCE's upcoming Research Excellence Framework, provided that they relate to the past performance of research staff and departments over a time period no longer than ten years. It must be noted that the Scopus team informed the report authors that Scopus has recently carried out backfills of the content of several important publishers, including Springer and the American Chemical and Physical Societies, and of the journals Science, Nature and The Lancet.

67. At this moment no accurate data are available on the outcomes of paper-by-paper matching of ISI Proceedings against other databases. Therefore, the current study could not generate data on the degree of overlap between ISI Proceedings on the one hand and the WoS, Scopus or the ACM and IEEE/CS digital libraries on the other. Nevertheless, since ISI Proceedings covers numerous proceedings, it is expected to be a valuable addition to the WoS. The two databases jointly can be assumed to cover the literature in many fields more completely, especially in Engineering and Computer Science, a field in which WoS coverage has proven to be moderate. Thomson Scientific's plans to incorporate this database more fully into the WoS in 2008 are certainly of interest.

Acknowledgements

The authors are grateful to Mrs. Nancy Bayers and her colleagues at Thomson Scientific and to Mrs. Helen de Mooij and her colleagues at Elsevier-*Scopus* for their comments on an earlier draft of this report.

References

Garfield, E. (1964). The Citation Index – A new dimension in indexing. *Science*, 144, 649–654.

Garfield, E. (1979). *Citation Indexing. Its theory and application in science, technology and humanities*. New York: Wiley.

López-Illescas, C., de Moya-Anegón, F. and Moed, H.F. (2008). Coverage and citation impact of oncological journals in the *Web of Science* and *Scopus*. Submitted to *Journal of Informetrics*.

Moed, H.F. and Visser, M.S. (2007). Developing Bibliometric Indicators of Research Performance in Computer Science: An Exploratory Study. Research Report to the Council for Physical Sciences of the Netherlands Organisation for Scientific Research (NWO). Leiden (the Netherlands): Centre for Science and Technology Studies (CWTS). CWTS Report 2007-01.

www.cwts.nl/cwts/NWO_Inf_Final_Report_V_210207.pdf

Moed, H.F., Visser, M.S. and Buter, R.S. (2008). Development of Bibliometric Indicators of Research Quality. Report submitted to the Higher Education Funding Council for England (HEFCE), 20 March 2008.

Scopus (2008). www.Scopus.com.

Thomson Scientific (2008). ISI Proceedings.

www.thomson.com/content/scientific/brand_overviews/proceedings