STI 2018 Leiden

23rd International Conference on Science and Technology Indicators
"Science, Technology and Innovation Indicators in Transition"

## STI 2018 Conference Proceedings

*Proceedings of the 23rd International Conference on Science and Technology Indicators*

All papers published in this conference proceedings have been peer reviewed through a peer review process administered by the proceedings Editors. Reviews were conducted by expert referees to the professional and scientific standards expected of a conference proceedings.

**Chair of the Conference**

Paul Wouters

**Scientific Editors**

Rodrigo Costas
Thomas Franssen
Alfredo Yegros-Yegros

**Layout**

Andrea Reyes Elizondo
Suze van der Luijt-Jansen

The articles of this collection can be accessed at https://hdl.handle.net/1887/64521

ISBN: 978-90-9031204-0

23rd International Conference on Science and Technology Indicators (STI 2018)
"Science, Technology and Innovation indicators in transition"
12 - 14 September 2018 | Leiden, The Netherlands
#STI18LDN

# A criteria-based assessment of the coverage of Scopus and Web of Science

Dag W. Aksnes[*] and Gunnar Sivertsen[**]

[*] *dag.w.aksnes@nifu.no*
NIFU – Nordic Institute for Studies in Innovation, Research and Education, P.O. Box 2815 Tøyen, 0608 Oslo (Norway)

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

## Introduction

Although the providers of Scopus and Web of Science increasingly claim to cover the world's scientific and scholarly literature comprehensively, both are selective in practice as well as in principle. The products not only depend on the coverage, but also the quality and relevance of their contents, to have success on the market. The provider of Web of Science, Clarivate Analytics, in addition inherits a tradition in which Eugene Garfield demonstrated that information retrieval theory (Bradford's law of scattering) and citation analysis support the idea of indexing mainly the 'core journals'. For many decades, an in-house editorial team has been evaluating possible new source items for Web of Science according to a set of publicly available criteria and with the help of citation analysis.

Elsevier instead publicly states on the webpages of the product that "content included in Scopus is carefully curated and ultimately selected by the independent Scopus Content Selection and Advisory Board (CSAB), an international group of scientists, researchers and librarians who represent the major scientific disciplines." Although the coverage of Scopus is somewhat broader than that of Web of Science, all comparisons, including our own in this study, demonstrate a large overlap and indicate the same pattern of deficiencies when it comes to the social sciences and humanities, and the coverage of literatures in other languages than English. The business model and the criteria seem to be the same. Scopus is also selective in principle and practice.

The two products serve several purposes. Among them are information retrieval, science studies and research evaluation and funding. Here, we limit the perspective to *research evaluation and funding* as we ask two questions that normally must be answered all the time in this context: *How should research quality be assessed? And who should decide on the criteria?* With the use of Scopus and Web of Science for research evaluation and funding, the answers are already given above: The commercial providers decide how to select the information provided for the evaluation and who will be using the selection criteria. Even the 'independent' advisory board for Scopus is appointed by Elsevier. These procedures ensure the quality of the highly valued products that we use for information retrieval and science studies. Hence, it is easy to forget that the same procedures are less legitimate in research evaluation and funding.

In research evaluation, the procedures and criteria are normally developed and decided in the *public domain* and anchored in *representative bodies of the research communities*. In public *funding* of research, the procedures and criteria are normally decided by *democratically responsible authorities and policies* and made *public to society*.

We see a need for the international community of experts in bibliometrics and research evaluation to start discussing the use of Scopus and Web of Science from the perspective of *properly organized* research evaluation and funding. The two questions need to be renewed in this context: How should research quality be assessed? And who should decide on the criteria?

To initiate the discussion, we apply a *criteria-based* assessment of the coverage of Scopus and Web of Science in this study. The criteria have been developed by the *Norwegian Association of Higher Education Institutions (Universities Norway)* with the assistance of its underlying national disciplinary committees and in collaboration with the *Norwegian Ministry of Higher Education and Science* to support the latter's institutional funding model. The criteria are very similar to those applied for institutional funding purposes in three other countries: Belgium (Flanders), Denmark and Finland.

The inclusion criteria used in the 'Norwegian model' will be further described in the Methods section below, but essentially, peer-reviewed scientific and scholarly publications are defined and delimited in a way that is comparable to selecting only original research publications and reviews in Scopus and Web of Science. Source items are similarly selected one by one on the basis of a set of minimal criteria that are intended to promote proper peer review and research quality. In practice, these minimal criteria provide a wider selection of source items than in Scopus and Web of Science. We are thereby able to describe the differences between what the *academic communities of a country* regard should be included as original research publications for evaluation and funding and what the *commercial providers* of Scopus and Web of Science are able to provide within a similar limitation to publication type. The patterns of differences will be described both with regard to publication type (books, articles in books, articles in series and journals), field of research and language.

During recent years, several valuable studies have addressed how Web of Science, and more recently Scopus, cover the research literature of various fields and countries. Nevertheless, a criteria-based approach representing research evaluation standards has been absent. With a few examples in each category, these are the main types of approaches in earlier studies:

- The products have been compared to each other with no external reference data, usually confirming that both are suitable tools for evaluation (e.g. Archambault, Campbell, Gingras, & Lariviere, 2009)

- What is not covered has been determined by using citations to non-indexed items in the same products as data (e.g. Nederhof, 2006).
- The coverage of the products has been compared to Google Scholar in several studies with different conclusions regarding the usability of the latter (e.g. Harzing & Alakangas, 2016). None of the studies assert that Google Scholar represents inclusion criteria according to research evaluation standards.
- Ulrich's Periodicals Directory has also been used as an external reference, again with no assertion that it represents academic standards for evaluation (Mongeon & Paul-Hus, 2016).
- Closer to our approach are studies that base the comparison a wider dataset defined as the published research output of a discipline in a non-English speaking country, or area of research or a geographical region (Ossenblok, Engels, & Sivertsen, 2012; Mongeon & Paul-Hus, 2016). Particularly interesting among these is Chavarro (2017) with a critical discussion of the principles and practices of selectivity in the products, demonstrating how their alleged 'universalism' does not represent global research in practice.

Our study differs from such earlier studies by applying *an explicit set of general criteria developed by academic communities* with which we can observe what is included and excluded in the two products.

## Data and methods

The so-called 'Norwegian Model' (Sivertsen, 2016), which so far has been adopted at the national level by Belgium (Flanders), Denmark, Finland and Norway, has three components:

(A) A complete representation in a national database of structured, verifiable and validated bibliographical records of the peer-reviewed scholarly literature in all areas of research;
(B) A publication indicator with a system of weights that makes field-specific publishing traditions comparable across fields in the measurement of 'Publication points' at the level of institutions;
(C) A performance-based funding model which reallocates a small proportion of the annual direct institutional funding according the institutions' shares in the total of Publication points.

The experience is that even with only marginal influence on the total funding, component C will support the need for completeness and validation of the bibliographic data in component A. The data in component A are delimited by a definition, according to which a scholarly publication must 1) present new insight 2) in a scholarly format that allows the research findings to be verified and/or used in new research activity, and 3) in a publication channel (journal, series, book publisher) which represents authors from several institutions and organizes independent peer review of manuscripts before publication. While the first two requirements of the definition demand originality and scholarly format in the publication itself (this is checked locally by each institution), the third and fourth requirements are supported centrally by a dynamic register of approved scholarly publication channels.

Component A in our study is the bibliographic database Cristin (Current Research Information System in Norway), which covers almost all Norwegian higher education institutions, research institutes and hospitals. Only publications which have officially qualified as scientific or scholarly according to specific criteria given above are included in the study. We use simple counts of unique publications. A total of 128,872 scientific or scholarly publications are included from the period 2011-2016.

While Scopus is organised as one database, Web of Science consists of several individual databases. The core databases are included in the Web of Science Core Collection which are:

- Science Citation Index Expanded (SCIE)
- Social Sciences Citation Index (SSCI)
- Arts & Humanities Citation Index (AHCI)
- Conference Proceedings Citation Index (CPCI)
- Book Citation Index (BKCI)
- Emerging Sources Citation Index (ESCI)

Although these are the core databases of Web of Science, many bibliometric analyses and indicators are limited to the classical ("flagship") citation indexes, the SCIE, SSCI, and AHCI, which cover journal publishing, only. For example, this holds for the Leiden ranking (http://www.leidenranking.com/information/data). The CPCI and BKCI databases cover conference series and book publications, respectively. The ESCI database was launched in 2015 and contains journals with regional importance and journals under evaluation for being a part of SCIE/SSCI/AHCI (http://wokinfo.com/products_tools/multidisciplinary/esci/). In this study, we have analysed the various databases individually and provide figures for the entire Web of Science Core Collection and for the three classical journal indexes, SCIE/SSCI/AHCI. In some of the analyses, figures are also shown for individual databases.

The comparative analysis consists of several steps. For the journal articles indexed in Cristin, the analyses are based on the list of source journals for Scopus and Web of Science. For Scopus, the October 2016 source list was used, which was the most recent available when the study was carried out. For Web of Science, the 2017 journal source list has been applied. In order to map the journal records of Cristin indexed in Scopus and Web of Science (SCIE, SSCI, AHCI and ESCI), the journal name, ISSN-number and e-ISSN numbers were used as identifiers. Because both database produces apply a cover-to-cover indexing of the journal literature, and fully index all issues such a method is justified.[1]

The analysis of book publications is more complicated where information on the title/name of the monographs, edited books, book series, conferences, conference series, as well as ISBN-numbers in various ways were used as identifiers. The source lists of Scopus and Web of Science for book publications and proceedings were used as basis for comparison.

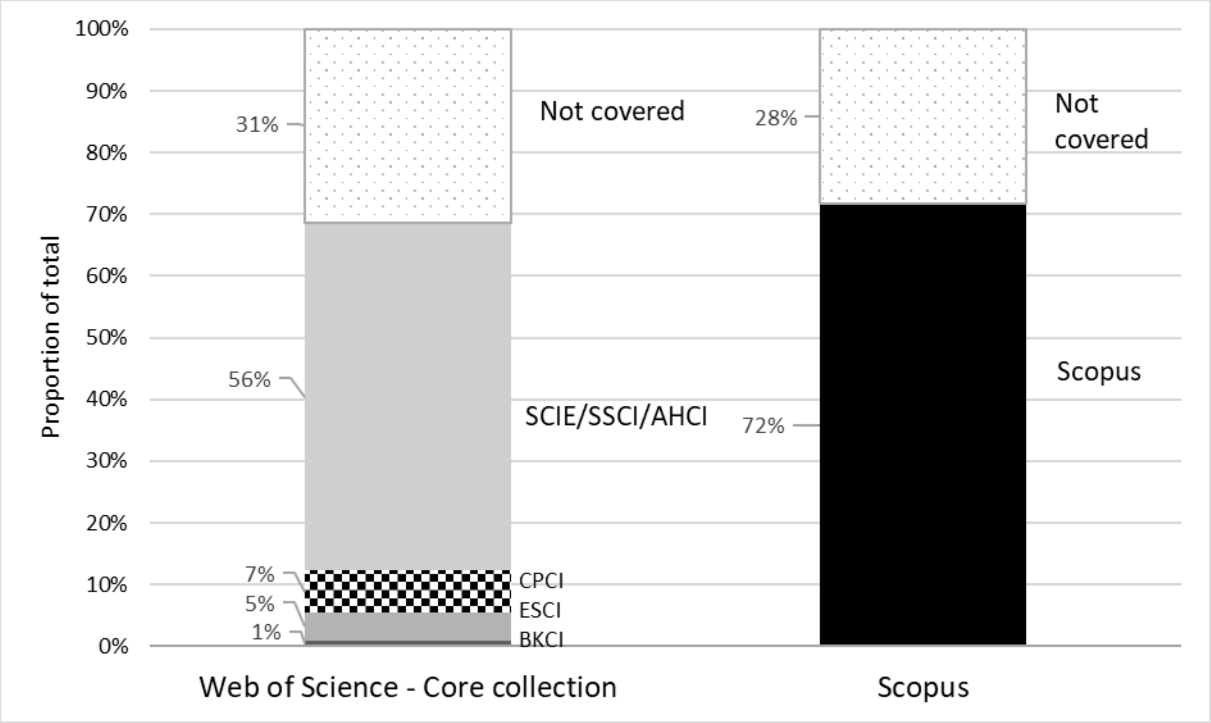
Although considerable efforts have been made to match the records as exact as possible, there inevitably will be cases where items mistakenly have been identified as being indexed or not. This is due to issues such as errors in core data, changes in the name of journals, or in the ISSN or ISBN numbers. Nevertheless, we believe that the sources of errors have rather minor importance when it comes to the overall findings of the study.

[1] Scopus does not cover book reviews and conference meeting abstracts. However, these publication types are not included in the study.

## Results

Figure 1 shows overall results for the 2015 and 2016 publications. Scopus covers 72 % of the total publication output, while the corresponding figure for Web of Science Core Collection is 69 %. Thus, the large majority of the Norwegian scientific and scholarly publication output is indexed in the two databases. Although Scopus has the highest coverage, the difference is not large. The three classical citation indexes, SCIE, SSCI, and AHCI, cover 56 % of the publication output, while the figures for the CPCI, ESCI and BKCI, are 7%, 5%, and 1% respectively.

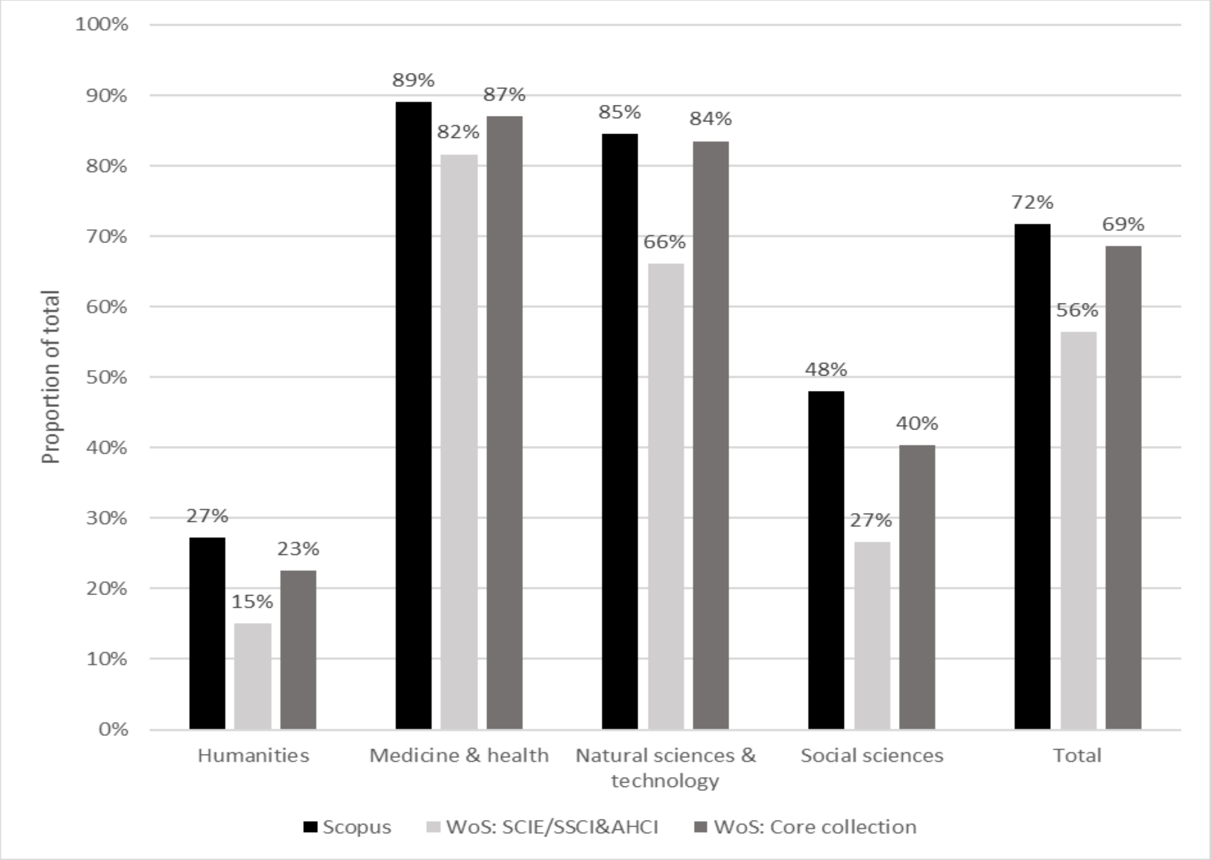
Figure 1: Coverage of 2015 and 2016 publications (n=45,972), total all fields and publication types, Scopus and Web of Science



For both databases, there are large variations in coverage across different domains. This is shown in Figure 2. In medicine and health, the coverage is not far from complete, with proportions of 89 % for Scopus and 87 % for Web of Science Core Collection. The three journal indexes of Web of Science, SCIE, SSCI, and AHCI capture 82 % of the production. The coverage is also very high for the natural sciences and technology, although for SCIE, SSCI, and AHCI the coverage is reduced (in particular due to the importance of proceeding papers in technology).

For the social sciences the coverage is significantly lower. Here, 48 % of the publications is indexed in Scopus and 40 % in Web of Science Core Collection, while 27 % appear in the SCIE, SSCI, and AHCI subset. Only a minor part of the publication output in humanities is indexed. Here the proportions are 27 % and 23 % for Scopus and Web of Science Core Collection.

Figure 2: Coverage of 2015 and 2016 publications (n=45,972) by domain, total all publication types, Scopus and Web of Science



Further details on the coverage by domains are provided in Table 1.

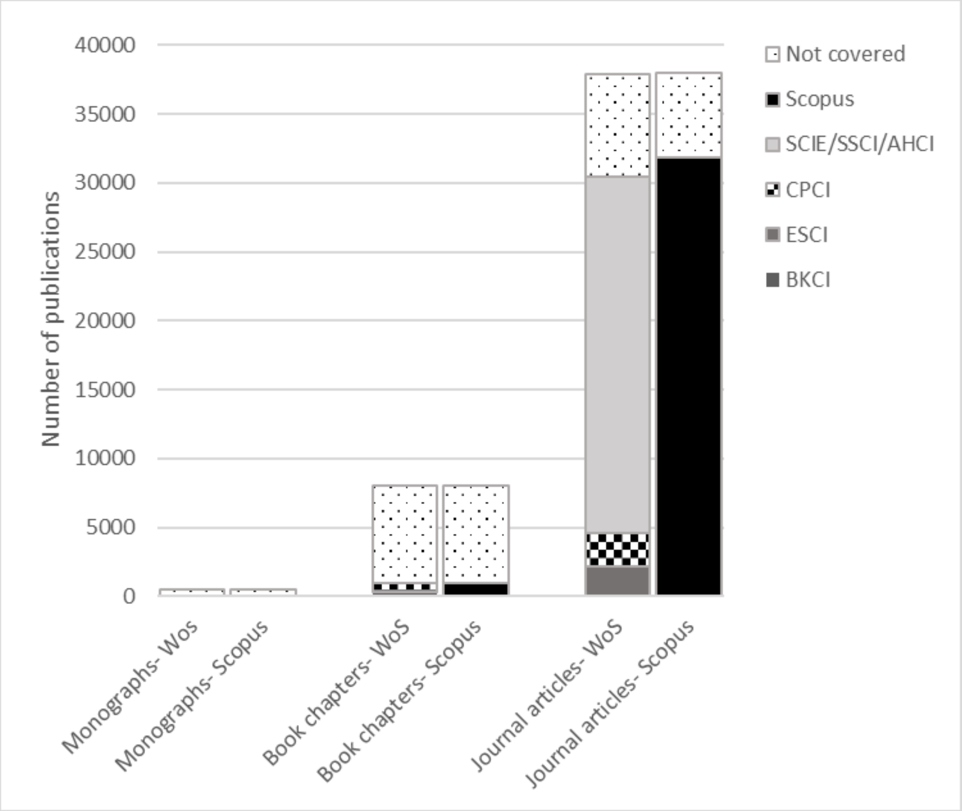
Table 1: Coverage of 2015 and 2016 publications (n=45,972) by domain, total all publication types, Scopus and Web of Science

| | **Scopus** | **WoS Core Collection** | | | | | **N (total number of publications)** |
|---|---|---|---|---|---|---|---|
| | | **SCIE/SSCI/AHCI** | **CPCI** | **BKCI** | **ESCI** | **Total** | |
| Humanities | 27% | 15% | 1% | 2% | 5% | 23% | 5,067 |
| Medicine & health | 89% | 82% | 0% | 0% | 5% | 87% | 12,879 |
| Natural sci & tech | 85% | 66% | 15% | 0% | 2% | 84% | 18,223 |
| Social sciences | 48% | 27% | 3% | 2% | 9% | 40% | 9,803 |
| Total | 72% | 56% | 7% | 1% | 5% | 69% | 45,972 |

The Norwegian publication data are classified into three publication types: monographs, book chapters (articles/chapters in anthologies) and articles in journals/series. The latter category accounts for the large majority of the publications (81%), while 17% appear as book chapters and 1% as monographs.

Figure 3 shows how the coverage of publications varies according to publication type. In total, 84 % of the journal articles are indexed in Scopus, 80% in Web of Science Core Collection, while 68 % appear in the SCIE, SSCI, and AHCI subset. The coverage of the book chapters is much lower, 14 % for both Scopus and Web of Science Core Collection.

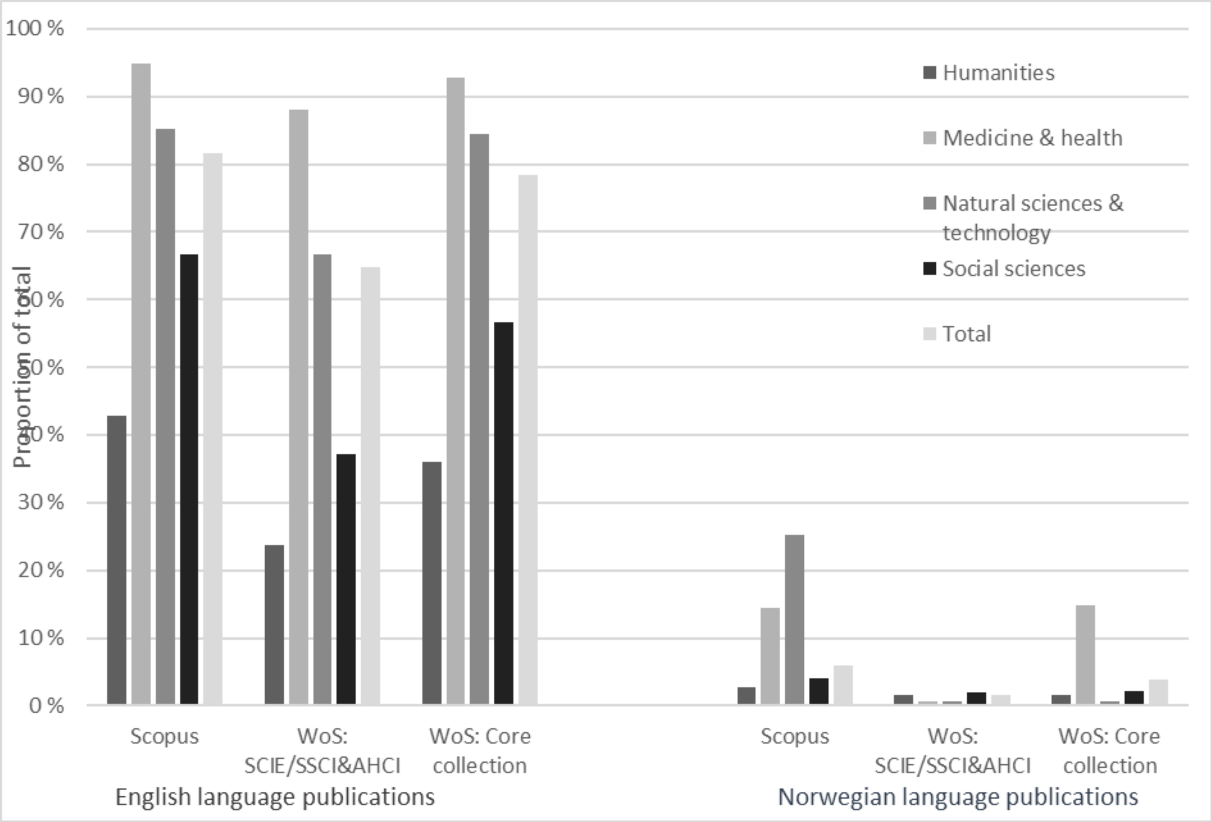
Figure 3: Coverage of 2015 and 2016 publications by publication types, Scopus and Web of Science



In the Cristin database, all publications are classified according to publication language. Overall, 87 % of the Norwegian publications are written in English (2015-2016). Of the remaining publications, most of them are written in Norwegian and a small minority in other languages. However, Norwegian accounts for a much higher share of the publications in humanities and social sciences than in the other domains.

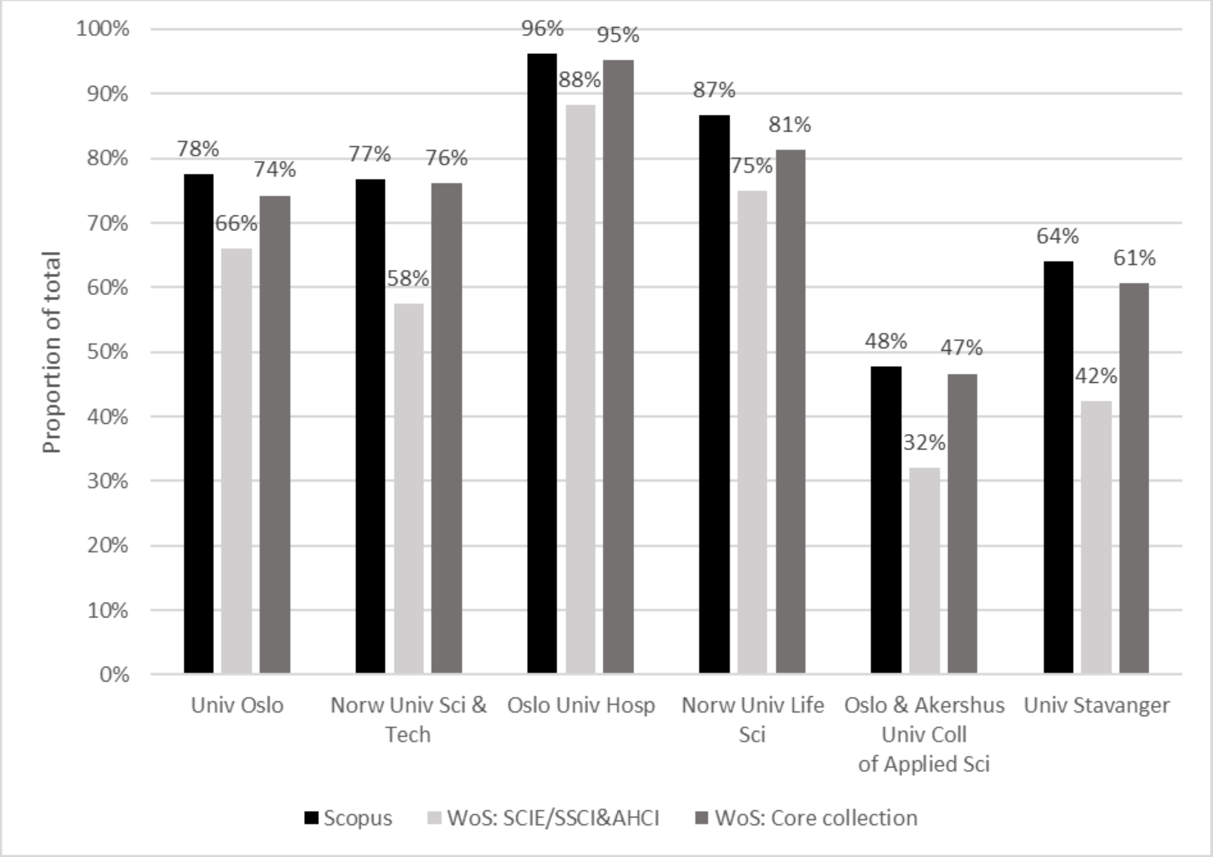
Figure 4 shows that both databases have a poor coverage of the Norwegian-language literature. This is an important reason why the databases cover humanities and social sciences less well than what is the case for the other domains. However, also the English language publications of these domains are less well covered. For the humanities, Scopus covers 43% of this literature, while the corresponding figure for Web of Science Core Collection is 36 %. The English language publications of the social sciences are better covered with 67% and 57% indexed in Scopus and Web of Science Core Collection, respectively.

Figure 4: Coverage of 2015 and 2016 publications by publication language and domain, Scopus and Web of Science



At the level of individual institutions there are quite large differences in how well Scopus and Web of Science cover the publication output. For the largest hospital in Norway, Oslo University Hospital, almost all publications are indexed in Scopus and Web of Science Core Collection (96% and 95%), cf. Figure 5. On the other hand, Oslo and Akershus University College of Applied Sciences (now OsloMet) has less than half of their publications indexed. These differences reflect the field and publication profile of the institutions.

Figure 5: Coverage of 2015 and 2016 publications for selected institutions, Scopus and Web of Science



## Discussions

Our study differs from earlier studies by applying *an explicit set of general criteria developed by academic communities* with which we can observe what is included and excluded in the two products. After decades of letting commercial providers act as the 'neutral guarantors of quality', we wish to empower the academic communities to take back responsibility for criteria and procedures also in the domain of bibliometrics for research evaluation and funding.

Within the scope of this short paper, there is not space to provide a throughout discussion of our results and relate them to previous studies (to be added the final version). We note that Scopus appears to have the largest coverage, but the difference compared with the entire Web of Science Core Collection is minor.  The study shows, in correspondence with several previous studies, that both databases have the same problems in terms of coverage of the social sciences and humanities literature and with coverage of non-English languages (Sivertsen & Larsen, 2012). Moreover, although the number of indexed books has been increasing in both databases, the coverage of book publications is still very limited. This publication type accounts for a small share of the indexed publications of both Scopus and Web of Science.[2] While the coverage of the English language journal publications is almost

[2] By 2017, 150,000 books were indexed in Scopus, while the total number of indexed items was 69 million.

complete, this does not hold for the corresponding book publications. Apparently, many important publishers of scholarly books, particularly in the social sciences and humanities, are not covered by the databases (Sivertsen, 2014).

## References

Archambault, E., Campbell, D., Gingras, Y., & Lariviere, V. (2009). Comparing of Science Bibliometric Statistics Obtained From the Web and Scopus. *Journal of the American Society for Information Science and Technology, 60*(7), 1320-1326.

Chavarro, D. (2017). *Universalism and Particularism: Explaining the Emergence and Development of Regional Indexing Systems* (doctoral thesis), University of Sussex, Brighton.

Harzing, A. W., & Alakangas, S. (2016). Google Scholar, Scopus and the Web of Science: a longitudinal and cross-disciplinary comparison. *Scientometrics, 106*(2), 787-804.

Mongeon, P., & Paul-Hus, A. (2016). The journal coverage of Web of Science and Scopus: a comparative analysis. *Scientometrics, 106*(1), 213-228.

Nederhof, A. J. (2006). Bibliometric monitoring of research performance in the social sciences and the humanities: A review. *Scientometrics, 66*(1), 81-100.

Ossenblok, T. L. B., Engels, T. C. E., & Sivertsen, G. (2012). The representation of the social sciences and humanities in the Web of Science-a comparison of publication patterns and incentive structures in Flanders and Norway (2005-9). *Research Evaluation, 21*(4), 280-290.

Sivertsen, G., & Larsen, B. (2012). Comprehensive bibliographic coverage of the social sciences and humanities in a citation index: an empirical analysis of the potential. *Scientometrics, 91*(2), 567-575.

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

Sivertsen, G. (2016). Publication-Based Funding: The Norwegian Model. In M. Ochsner, S.E. Hug, H.D. Daniel (Eds.), *Research Assessment in the Humanities. Towards Criteria and Procedures* (pp. 79-90). Zürich: Springer Open.