Most commonly used multidisciplinary citation data sources

Citation databases include information on publications in the database and the citations between them. Citation information shows how often other publications in the database have cited the publication being examined. In addition to the citation numbers of an individual publication, citation databases typically allow users to review things such as how many times a certain author or the publications of a certain organisation have been cited. The number of citations received by a publication varies in different databases. Citation databases usually focus on scientific journal articles, but they can also include monographs, conference publications and reports. No citation database includes all publications. It should also be kept in mind that certain fields of sciences (such as medical and natural sciences) have better representation in databases than others (such as social sciences, humanities and arts). The number of citations in databases also depends on the length of the period from which the database has citation data and on how often the citation data is updated in the database. All databases also have some errors. The content of databases is constantly changing, because they add new materials to their collection and remove outdated content while also updating citation data of older publications.

The two of the most well-known multidisciplinary databases are Web of Science (WoS) by Clarivate Analytics and Scopus by Elsevier. Both WoS and Scopus contain curated materials, and they have certain quality criteria for their database content, for example, the journals must apply a peer review process and have an editorial board and an ISSN identifier. The title and abstract of the article must be written in English. Citation data can also be analysed in the publicly available Dimensions and Google Scholar services. Both Dimensions and Google Scholar index materials into their collections with automated methods, without a separate validation process. The basic content and functionalities of the Dimensions database are available openly and free of charge. Paying subscribers also have access to a version with more extensive content and more versatile analysis functionalities. Google Scholar is a free search engine that is specialised in finding scientific information, but it also holds non-scientific materials such as undergraduate theses. Its citation data also includes many citations from non-scientific publications, therefore their number of citations differs greatly from that of the WoS and Scopus databases. This guide presents Google Scholar along with citation databases, although it is not an actual database. All three databases, as well as Google Scholar, have been presented in more detail in their own chapters.

Comparing the sources of citation data

Because the data indexing methods and collection policies vary significantly in citation databases, so does the content of these databases. WoS and Scopus both apply a selective collection policy and have editorial boards with experts of different fields that help them select and index their content. Dimensions is heavily reliant on machine learning and data processing implemented with automated algorithms instead of manual curating done by experts. Dimensions mainly indexes all scientific publications and data sets with a DOI identifier. The automated methods of Google Scholar, on the other hand, crawl the internet and index all publications that they recognise as having an academic structure.

The publishing practices of different disciplines, such as the publication type and language, are reflected in the coverage of the databases. The traditional citation databases, such as WoS and Scopus, mainly contain scientific journals. Fields of sciences, where books are a typical type of publication, are poorly represented in these databases, in particular. The content also often lacks in terms of conference publications. Typically, social sciences and humanities have the poorest coverage. The fact that the databases mainly focus on literature written in English is another limiting factor of their coverage. Although the databases also accept publications in languages other than English, one of the selection criteria for the articles published on these publication channels is that their abstracts and titles are written in English. Dimensions and Google Scholar, which are based on automated indexing methods, reach a significantly higher degree of coverage in several disciplines. Dimensions is noticeably more extensive than WoS or Scopus, especially when it comes to edited books, their chapters or monographs.

Several studies have been carried out on the differences between databases and on the extent of their coverage of materials from different fields of science. Many of these studies show that the coverage of the WoS Core Collection database is lower within many fields of science than that of its competing databases (e.g. Mongeon, P. & Paul-Hus, A. 2016; Martín-Martín, A. et al. 2021; Singh, V.K. et al. 2021; Visser, M. et al. 2021). The citation data source that proved in many studies to have the best coverage in nearly all fields of science out of all sources presented in this guide is Google Scholar (e.g. Harzing, A.-W. 2019; Martín-Martín, A. et al. 2021). Some recent studies about the coverage of citation database materials have been listed at the end of this chapter. Additionally, a coverage comparison of databases carried out on the basis of publications produced by the University of Jyväskylä can be found as an attachment.

Researcher and organisation profiles in citation databases

- Typically, researcher and organisation profiles are created with the help of the service providers' own algorithms for the organisations and authors represented in the publications.
- Researcher and organisation profiles of citation databases often have some errors. Some aspects that make identifying the authors of publications more difficult are common surnames, name changes and characters not used in the English language. Similarly, the names of organisations may change or the organisations may undergo structural changes. The algorithms of citation databases do not necessarily recognise changes such as this.
- The publication is linked to a certain researcher and organisation based on the information stated in the publications. Therefore, it is of utmost importance that the authors mark down their organisation correctly in all the publications they write. The names of organisations and their units and departments should be written down in full, avoiding any abbreviations.

When selecting a database, however, the coverage of its materials is not the only criteria. The search functions and the quality of the indexed materials also matter. Even though Google Scholar has greater coverage than its competitors, its search functions have some limitations, such as the limited use of Boolean operators and very limited filtering options for search results. The traditional WoS and Scopus databases, that have been on the market for a long time, have weaker coverage, the materials are indexed into the databases at a slower pace, and they are also paid services. On the other hand, they both have a wide range of search functions and the materials indexed by them have been validated. Table 1 presents the key functionalities and content of the Web of Science, Scopus and Dimensions databases as well as the Google Scholar service to allow for the easy comparisons of the citation databases. From the perspective of responsible publication metrics, the citation database used for an analysis should be selected, in addition to the usage requirements, by considering the publication practices of the scientific field being analysed. If possible, it is recommended that more than one citation database be used for any publication-based analyses.

- In the case of the Web of Science and Scopus databases, organisations and researchers can also themselves ensure that their profiles are up to date and accurate. Keeping the researcher and organisation profiles accurate is a continuous process. The researcher profiles may be broken despite the corrections, and the service provider can accidentally link corrected organisation profiles to publications that do not actually belong to it.
- Keeping the organisation profiles of citation databases up to date is an important task, as this information is used for various benchmarking and collaboration analyses. Many ranking organisations also utilise this data.

	Web of Science	Scopus	Dimensions	Google Scholar
Availability	Subscription needed	Subscription needed	Basic content and functionalities are publicly available, more advanced use and more extensive content are subject to a charge.	Free of charge
Number of journals **(Dimensions: Number of publication channels)**	~22 000	~24 000	~99 000 + ~1M books (own channel-level metadata for ~45,500 publication channels)	Not public
Other content	Publications: conference publications, monographs. Also small quantities of other publication types. Additionally, features information on funders.	Publications: conference publications, monographs, book chapters, professional journals and patents. Also small quantities of other publication types. Additionally, features information on funders.	Publications: book chapters, conference publications, preprints, monographs, patents and policy publications. Also includes information about the datasets, research funding and clinical trials. Scientific publications and information about the data sets are included in the free version. The rest of the content is subject to a charge.	Conference publications, monographs, pre-prints, theses, PowerPoint presentations, WWW pages
Fields of science	Natural sciences, medical science, technology, social sciences, arts and humanities	Natural sciences, technology, health sciences, social sciences, arts and humanities	Best coverage: natural sciences, medical science, technology and social sciences	Not public
Temporal coverage	Since 1900 (science), since 1956 (social sciences) and since 1975 (arts and humanities); the availability of materials depends on the licences acquired by the organisation.	The coverage varies, some journals back to the 1780s, citation data from the 1970s onwards.	No separate ground rule regarding the age of materials, mainly indexes all publications with a DOI.	Not public

Language coverage and language requirements of the material being indexed	Mainly contains materials in English. Small amount of materials in other languages. Publication channels that have been accepted for indexing must have article titles and abstracts in English.	Mainly contains materials in English. Small amount of materials in other languages. Publication channels that have been accepted for indexing must have article titles and abstracts in English.	Mainly contains materials in English. Also contains materials in other languages if they have a DOI. Publication, patent and grant information is also available in other languages. No separate language requirements for indexed materials. The grant and patent information uses automated machine translation for abstracts and titles.	Not public, but also has materials in languages other than English.
Content policy	Public	Public	No separate content policy. Mainly indexes all publications with a DOI.	Not public, contract with most major publishers.
Citation analysis	Citation Report tool	Citation Overview tool	The general overview of the Analytical Views search results and separate views for disciplines, researchers and publication channels	Cited by link in connection to search results, allows for the publications citing the publication to be viewed and shows the number of citations per publication.
Temporal coverage of citation data	Since 1900 (science), since 1956 (social sciences) and since 1975 (arts and humanities)	Since 1970	Varies. In some places, the database has indexed citations more than a 100 years old, while some very recently indexed publications may have some of the cited sources missing.	Not public
Indicators	Number of citations, average number of citations per publication, average number of citations per publication year, annual numbers of citations, h-index, usage statistics of records (how many times an individual record has either been uploaded into the reference management system or how many times the publication's full text has been opened)	Number of citations, annual numbers of citations, h-index, record views, PlumX usage statistics, field-normalised citation impact indicator for articles and journals.	Number of citations, number of citations from past two years, average number of citations per publication, average number of citations per publication year, online attention to individual research result found by the Altmetric.com service, field-weighted and relative citation impact	Number of citations per publication The Google Scholar profile also provides the researcher-specific number of citations, the h-index and the i10 index from all years and last five years.
University rankings utilising data	Academic Ranking of World Universities (ARWU), i.e. Shanghai Ranking University Ranking by Academic Performance (URAP) U.S. News & World Report's Best Global Universities Rankings U-Multirank National Taiwan University (NTU) Ranking CWTS Leiden Ranking CWUR (Center for World University Ranking, United Arab Emirates) Round University Ranking (RUR) State of scientific research in Finland reports by the Academy of Finland	Times Higher Education World University Ranking Times Higher Education Impact Ranking QS Ranking State of scientific research in Finland reports by the Academy of Finland	No known rankings utilising Dimensions data	No known rankings utilising Google Scholar data
Researcher profiles	ResearcherID	Scopus Author Identifier	Dimensions researcher profile	Google Scholar Profile
Tools utilising data	InCites, Journal Citation Reports, Eigenfactor, ScienceWatch, Publish or Perish	SciVal, SCImago Journal and Country Rank, CWTS Journal Indicators, Publish or Perish	Dimensions Analytics, JYUcite	Publish or Perish

Table 1. Summary of the characteristics of the four key multidisciplinary citation data sources.

Sources

Clarivate Analytics (no date) *Web of Science LibGuides.* Available: https://clarivate.libguides.com/home (Accessed 4.2.2022)

Digital Science (no date) *Dimensions.* Available: https://www.dimensions.ai/products/free/ (Accessed 12.4.2022)

Google (no date) *About Google Scholar.* Available: https://scholar.google.com/intl/fi/scholar/about.html (Accessed 18.2.2022)

Harzing, A.-W. (2019) Two new kids on the block: How do Crossref and Dimensions compare with Google Scholar, Microsoft Academic, Scopus and the Web of Science? *Scientometrics,* 120(1), pp. 341–349. Available: https://doi.org/10.1007/s11192-019-03114-y

Martín-Martín, A., Thelwall, M., Orduna-Malea, E. and López-Cózar, E.D. (2021) Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and OpenCitations' COCI: a multidisciplinary comparison of coverage via citations. *Scientometrics,* 126: pp. 871–906. Available: https://doi.org/10.1007/s11192-020-03690-4

Mongeon, P. and Paul-Hus, A. (2016) The journal coverage of Web of Science and Scopus: a comparative analysis. *Scientometrics,* 106, pp. 213–228. Available: https://doi.org/10.1007/s11192-015-1765-5

Elsevier (no date) *Scopus.* Available: https://www.elsevier.com/solutions/scopus (Accessed 12.4.2022)

Singh, V.K., Singh, P., Karmakar, M. *et al.* (2021) The journal coverage of Web of Science, Scopus and Dimensions: A comparative analysis. *Scientometrics,* 126, pp. 5113–5142. Available: https://doi.org/10.1007/s11192-021-03948-5

Visser, M., van Eck, N.J. and Waltman, L. (2021) Large-scale comparison of bibliographic data sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic. *Quantitative Science Studies,* 2(1), pp. 20–41. Available: https://doi.org/10.1162/qss_a_00112

Recent studies on the coverage and characteristics of citation data sources

Baas, J., Schotten, M., Plume, A., Côté, G. and Karimi, R. (2020) Scopus as a curated, high-quality bibliometric data source for academic research in quantitative science studies. *Quantitative Science Studies,* 1(1), pp. 377–386. Available: https://doi.org/10.1162/qss_a_00019

Birkle, C., Pendlebury, D. A., Schnell, J. and Adams, J. (2020) Web of Science as a data source for research on scientific and scholarly activity. *Quantitative Science Studies,* 1(1), pp. 363–376. Available: https://doi.org/10.1162/qss_a_00018

Gusenbauer, M. (2022) Search where you will fnd most: Comparing the disciplinary coverage of 56 bibliographic databases. *Scientometrics,* 127, pp. 2683–2745. Available: https://doi.org/10.1007/s11192-022-04289-7

Harzing, A.-W. (2019) Two new kids on the block: How do Crossref and Dimensions compare with Google Scholar, Microsoft Academic, Scopus and the Web of Science? *Scientometrics,* 120(1), pp. 341–349. Available: https://doi.org/10.1007/s11192-019-03114-y

Herzog, C., Hook, D. and Konkiel, S. (2020) Dimensions: Bringing down barriers between scientometricians and data. *Quantitative Science Studies,* 1 (1), pp. 387–395. Available: https://doi.org/10.1162/qss_a_00020

Huang, C.-K., Neylon, C., Brookes-Kenworthy, C., Hosking, R., Montgomery, L., Wilson, K. and Ozaygen, A. (2020) Comparison of bibliographic data sources: Implications for the robustness of university rankings. *Quantitative Science Studies,* 1(2), pp. 445–478. Available: https://doi.org/10.1162/qss_a_00031

Martín-Martín, A., Orduna-Malea, E. and López-Cózar, E.D. (2018) Coverage of highly-cited documents in Google Scholar, Web of Science, and Scopus: A multidisciplinary comparison. *Scientometrics,* 116(3), pp. 2175–2188. Available: https://doi.org/10.1007/s11192-018-2820-9

Martín-Martín, A., Orduna-Malea, E., Thelwall, M. and López-Cózar, E.D. (2018) Google Scholar, Web of Science, and Scopus: A systematic comparison of citations in 252 subject categories. *Journal of Informetrics,* 12(4), pp. 1160–1177. Available: https://doi.org/10.1016/j.joi.2018.09.002

Martín-Martín, A., Thelwall, M., Orduna-Malea, E. and López-Cózar, E.D. (2021) Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and OpenCitations' COCI: A multidisciplinary comparison of coverage via citations. *Scientometrics,* 126, pp. 871–906 (2021). Available: https://doi.org/10.1007/s11192-020-03690-4

Mongeon, P. and Paul-Hus, A. (2016) The journal coverage of Web of Science and Scopus: a comparative analysis. *Scientometrics,* 106, pp. 213–228. Available: https://doi.org/10.1007/s11192-015-1765-5

Singh, V.K., Singh, P., Karmakar, M. et al. (2021) The journal coverage of Web of Science, Scopus and Dimensions: A comparative analysis. *Scientometrics,* 126, pp. 5113–5142. Available: https://doi.org/10.1007/s11192-021-03948-5

Visser, M., van Eck, N.J. and Waltman, L. (2021) Large-scale comparison of bibliographic data sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic. *Quantitative Science Studies,* 2(1): pp. 20–41. Available: https://doi.org/10.1162/qss_a_00112

Waltman, L. and Larivière, V. (2020) Special issue on bibliographic data sources. *Quantitative Science Studies,* 1(1), pp. 360–362. Available: https://doi.org/10.1162/qss_e_00026

Coverage comparison of databases carried out on the basis of publications produced by the University of Jyväskylä

Seppänen, J-T. (no date) *Comparing citation database coverage: University of Jyväskylä research publications in Dimensions, Scopus, Web of Science and PubMed.* This work has not yet been published. Comparing citatation database coverage _draft.pdf