QSS

an open access journal

Check for updates

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

DOI:
https://doi.org/10.1162/qss_e_00026

Corresponding Author:
Ludo Waltman
waltmanlr@cwts.leidenuniv.nl

The MIT Press

EDITORIAL

Special issue on bibliographic data sources

Ludo Waltman[1] and Vincent Larivière[2,3]

[1]Centre for Science and Technology Studies (CWTS), Leiden University, Leiden, the Netherlands
[2]École de bibliothéconomie et des sciences de l'information, Université de Montréal, Montréal, Quebec, Canada
[3]Observatoire des Sciences et des Technologies (OST), Centre Interuniversitaire de Recherche sur la Science et la Technologie (CIRST), Université du Québec à Montréal, Montréal, Quebec, Canada

1. INTRODUCTION

Research in quantitative science studies relies on an increasingly broad range of data sources, providing data on scholarly publications, social media activity, peer review, research funding, the scholarly workforce, scientific prizes, and so on. However, there is one type of data source that remains at the heart of research in quantitative science studies: bibliographic databases. These data sources have increasingly diversified over the last decade. Several organizations now provide large-scale databases of metadata on scholarly publications. For this special issue of *Quantitative Science Studies*, we invited the providers of major bibliographic data sources to provide insights on how their data can be used to support research in quantitative science studies.

This special issue comprises six papers. Three papers cover the most important commercial bibliographic data sources: Web of Science (Clarivate Analytics), Scopus (Elsevier), and Dimensions (Digital Science). Three other papers cover open data sources: Microsoft Academic, Crossref and OpenCitations.[1] There are of course many other bibliographic data sources. However, for this special issue, we have chosen to consider only data sources that cover publications from all fields of science and from all parts of the world. Data sources that focus on specific scientific fields (e.g., PubMed from the US National Library of Medicine) or specific countries (e.g., national databases of scholarly publications) are therefore not included.

We hope that this special issue will help authors of submissions to *Quantitative Science Studies* to choose the most suitable bibliographic data source for their research. In the past, Web of Science and Scopus often were the only data sources between which researchers could choose. Researchers typically used the data source to which their institution happened to have a subscription. In recent years, however, the number of options has increased considerably. This special issue aims to characterize the most important data sources currently available and to show how they differ in various dimensions, for instance in the data they provide, their level of openness, and their support for making research reproducible. As editors of *Quantitative Science Studies*, we consider openness and reproducibility to be of major importance. Research published in *Quantitative Science Studies* is expected to be as reproducible as possible, and reproducibility can be promoted by making use of open data.

Below we provide a brief overview of some of the key differences between the bibliographic data sources considered in this special issue, focusing on three questions: What does the data source provide? How to get access to the data? And how can the data be used?

[1] The team of Google Scholar was also invited to contribute to this special issue, but they did not respond to our invitation.

2. WHAT DOES THE DATA SOURCE PROVIDE?

Web of Science and Scopus are selective data sources. Both the Web of Science Core Collection and Scopus aim to cover only content that meets certain standards. These data sources for instance try to make sure they do not cover journals that adopt questionable publication practices ('predatory journals'). In the case of Scopus, content selection is carried out by an external Content Selection and Advisory Board consisting of independent researchers.

Rather than being selective, the other data sources aim to be comprehensive. Microsoft Academic obtains most of its data by crawling the Web, although it also makes use of data provided by publishers. Microsoft decides which content retrieved from the Web is considered to be of a scholarly nature and deserves to be included in Microsoft Academic. Data curation is performed using artificial intelligence techniques. Human intervention is minimized as much as possible. Crossref obtains its data from publishers that work with Crossref to obtain digital object identifiers (DOIs) for their content. Publishers decide whether they want to work with Crossref and what data they want to make available through Crossref. Dimensions builds on data from sources such as Crossref and PubMed, and complements this data with data received from publishers. Finally, OpenCitations also obtains its data from other sources, such as Crossref and PubMed Central. It does not receive data from publishers. In addition to other formats, OpenCitations makes its data available in RDF format as linked open data, using semantic web technologies.

3. HOW TO GET ACCESS TO THE DATA?

Access to Web of Science and Scopus normally requires a payment. Dimensions has a free and a paid version. The paid version offers additional features not available in the free version. It for instance provides access to data that is not accessible in the free version. For research purposes, it is possible to apply for no-cost access to the full version of Dimensions, including access through an API. Likewise, Elsevier's International Center for the Study of Research plans to create a 'virtual laboratory' that provides free access to Scopus data for research purposes. While Dimensions data has already been made available for bibliometric research, the details of Elsevier's plan are not yet clear at the moment.

Microsoft Academic, Crossref, and OpenCitations make all their data openly available. Microsoft Academic can be queried through an API. Up to a certain limit, this API can be used free of charge. Crossref and OpenCitations can also be queried through APIs. Their APIs can be freely used without any limit. Crossref also offers a paid service called Metadata Plus, which provides improved API access and the possibility to download a snapshot of the full Crossref database. A snapshot of the full Microsoft Academic database can be downloaded through Microsoft's Azure platform. A small fee may be required to cover the costs associated with the use of this platform. OpenCitations also releases snapshots of its databases. These snapshots can be freely downloaded.

4. HOW CAN THE DATA BE USED?

All data sources allow their data to be used for research purposes. For some of the data sources, in particular Web of Science and Dimensions, the providers ask researchers that use their data to share their results and to report problems identified in the data.

Ideally, bibliographic data used in research projects is made openly available. The commercial data sources (Web of Science, Scopus, and Dimensions) do not allow their data to be made openly available. They impose restrictions on the sharing or redistribution of their data.

Such restrictions make it more difficult to reproduce research based on data from these sources. Even if different research teams all have access to a specific data source, it may be hard for one team to reproduce the work of another team, as the data is a moving target because of continuous updates of the data source. To address this problem, data providers would need to provide access to archived time-stamped versions of their data.

Microsoft Academic, Crossref, and OpenCitations make their data openly available. Researchers are therefore allowed to share or redistribute their data, which makes research easier to reproduce. Microsoft Academic releases its data under an ODC-BY license. This license requires Microsoft to be acknowledged when Microsoft Academic data is used. Crossref considers its data to be facts. The data cannot be owned and is therefore made available without a license. Finally, OpenCitations makes its data available under a CC0 license, releasing the data into the public domain and minimizing restrictions on the use of the data.

5. CONCLUSION

We hope that this special issue will help researchers working with bibliographic data to better understand the characteristics of different data sources and to choose the most suitable data source for their research. There are advantages and disadvantages to each data source. While the selectivity of Web of Science and Scopus may for instance be beneficial for some studies, it may be problematic for others. Likewise, some studies may use an open data source because reproducibility is considered essential, while other studies may have to rely on a closed data source because they require data that is not openly available.

This special issue does not provide comparisons of the different data sources in terms of their coverage, completeness, and data quality. Such comparisons can best be performed by independent research groups rather than by the data providers themselves. Submissions of papers presenting comparisons of the data provided by different bibliographic data sources are very much welcomed at *Quantitative Science Studies*. We hope to publish such papers in the near future.

Finally, we would like to express our gratitude to Clarivate Analytics, Elsevier, Digital Science, Microsoft, Crossref, and OpenCitations for working together with us in making this special issue possible.

COMPETING INTERESTS

Ludo Waltman is deputy director of the Centre for Science and Technology Studies (CWTS) at Leiden University. CWTS has commercial relationships with Clarivate Analytics, Elsevier, and Digital Science. CWTS and Digital Science also work together as founding partners of the Research on Research Institute (RoRI; http://researchonresearch.org). Waltman works together with OpenCitations in his capacity as chair of the Advisory Board of the Research Centre for Open Scholarly Metadata.

Vincent Larivière is associate scientific director of the Observatoire des sciences et des technologies at Université du Québec à Montréal (OST-UQAM). OST-UQAM has a commercial relationship with Clarivate Analytics. Larivière also sits on the Advisory Board of the Research Centre for Open Scholarly Metadata.