Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Regular article

DataCite as a novel bibliometric source: Coverage, strengths and limitations☆

Nicolas Robinson-Garcia[a,*], Philippe Mongeon[b], Wei Jeng[c], Rodrigo Costas[d,e]

[a] *INGENIO (CSIC-UPV), Universitat Politècnica de València, Spain*
[b] *École de bibliothéconomie et des sciences de l'information, Université de Montréal, Canada*
[c] *Department of Library and Information Science, National Taiwan University, Taiwan*
[d] *CWTS, Leiden University, The Netherlands*
[e] *Centre for Research on Evaluation, Science and Technology (CREST), Stellenbosch University, Private Bag X1, Matieland 7602, South Africa*

ARTICLE INFO

Article history:
Received 6 March 2017
Received in revised form 17 July 2017
Accepted 17 July 2017
Available online 30 August 2017

Keywords:
Data sharing
Data citations
Bibliometric sources
Open data
Data infrastructure
Data metrics
DataCite

ABSTRACT

This paper explores the characteristics of DataCite to determine its possibilities and potential as a new bibliometric data source to analyze the scholarly production of open data. Open science and the increasing data sharing requirements from governments, funding bodies, institutions and scientific journals has led to a pressing demand for the development of data metrics. As a very first step towards reliable data metrics, we need to better comprehend the limitations and caveats of the information provided by sources of open data. In this paper, we critically examine records downloaded from the DataCite's OAI API and elaborate a series of recommendations regarding the use of this source for bibliometric analyses of open data. We highlight issues related to metadata incompleteness, lack of standardization, and ambiguous definitions of several fields. Despite these limitations, we emphasize DataCite's value and potential to become one of the main sources for data metrics development.

1. Introduction

Calls for data availability and sharing can be traced back to the beginning of the 20th century when Galton stated: "I have begun to think that no one ought to publish biometric results, without lodging a well arranged and well bound manuscript copy of all his data, in some place where it should be accessible, under reasonable restrictions, to those who desire to verify his work" (Galton, 1901, as cited in Perneger, 2011). However, it has been just a few decades since technology has made possible the development of the necessary infrastructure to make this happen (Peng, 2011). In the last decade, public funding agencies, publishers and institutions have directed their efforts towards developing such infrastructure as well as to incentivizing data sharing and reuse within the scientific community by promoting data citations (Robinson-García et al., 2015).

☆ "The peer review process of this paper was handled by Staša Milojević, Associate Editor of Journal of Informetrics."
[*] Corresponding author.
E-mail addresses: elrobin@ingenio.upv.es (N. Robinson-Garcia), philippe.mongeon@umontreal.ca (P. Mongeon), wjeng@ntu.edu.tw (W. Jeng), rcostas@cwts.leidenuniv.nl (R. Costas).

http://dx.doi.org/10.1016/j.joi.2017.07.003
1751-1577/© 2017 Elsevier Ltd. All rights reserved.

Data sharing and reuse practices have been adopted at a different pace by the different scientific communities. For instance, data infrastructure is widely developed within the crystallography community, dating back to the early 1970s (Torres-Salinas, Robinson-García, & Cabezas-Clavijo, 2012). A similar expansion can be observed in Genomics or Astronomy (Borgman, 2012). On the other hand, social sciences and the humanities have thus adopted these new practices at a slower pace than STEM fields (Doorn, Dillo, & van Horik, 2013; Kim & Adler, 2015).

Infrastructure design is a key factor towards fostering data sharing and reuse. Piwowar, Becich, Bilofsky, and Crowley (2008) analyzed how certain elements of data sharing frameworks may influence the usability, discoverability, and data reuse for different stakeholders.

Although measuring the impact of data is a highly relevant element in the research policy agenda, a direct measure of data reuse is very difficult to achieve (Missier, 2016). Attempts of metrics such as downloads of datasets or data citations have been proposed to track data reuse (Konkiel, 2013). While the former seem to be problematic on capturing different dimensions of usage (Mayernik, Hart, Maull, & Weber, 2016), –e.g., data might be downloaded for research validating purposes, –- more effort has been put into the call of movement of "data citations" (Costas, Meijer, Zahedi, & Wouters, 2013; Piwowar, Day, & Fridsma, 2007).

For data citations to become a valid indicator on data reuse, a shift is needed on the communication behavior of researchers when citing sources, as well as on the meaning they attach to their references (Mayernik, 2012; Parsons & Fox, 2013). Initiatives such as the launch of the Data Citation Index and the DataCite consortium are examples of efforts directed at promoting data citations. However, little is known about the production of data, field-specific practices, and other basic requirements such as the format a data record should have to facilitate information retrieval and bibliometric analyses. Previous studies focusing on Thomson Reuters' Data Citation Index (now Clarivate Analytics) have explored disciplinary biases and data types included (Torres-Salinas, Martín-Martín, & Fuente-Gutiérrez, 2014), data citation practices between fields (Robinson-García et al., 2015), and the relation between data citations and data mentions in social media (Peters, Kraker, Lex, Gumpenberger, & Gorraiz, 2016).

In a recent report, Costas et al. (2013) highlighted the need for developing data publication standards, reducing the dispersion of data repositories, and facilitating the traceability, citation and measurement of data records. The most comprehensive source for open data currently available is DataCite, which contains more than 7 million freely accessible records, almost doubling the figures last reported for the Data Citation Index (Peters et al., 2016).

In line with the open science movement and calls for increased data sharing and reuse, we highlight the importance of data publications and citations. This paper analyzes the structure and type of metadata offered by DataCite to assess its potential to become an important source for developing data-level metrics. DataCite is an international non-profit organization formed in 2009. It is a consortium of public research institutions, funding bodies and publishers worldwide whose mission is to promote open research data accessibility and tracking. For the latter, DataCite advocates for the use of Digital Object Identifiers (DOI) by assigning DOIs to their records (DataCite Metadata Working Group, 2015).

2. Objectives

This paper aims to explore the characteristics of the data collected by DataCite to determine its potential as a new source of bibliometric data for the study of open data production. Specifically, we examine the database structure and the level of standardization of the information provided in each field, to assess the usability of the data for bibliometric purposes. The paper is structured as follows. Firstly, we present the metadata scheme of DataCite records (2015). Then we assess the completeness of the data in each specific field and give an overview of the database coverage. Finally, we discuss the potential of DataCite as a source for tracking open data production, and we provide some recommendations for its use as tool for studying data production and citation patterns.

3. Data and methods

This section is structured in three parts. The first one describes the different points of access available by DataCite and advantages and limitations of using one or the other. Second, we recollect and describe the information provided by DataCite as to its structure, definition of data record fields, and information requested to each repository. The aim is to give the reader a full account as to what DataCite expects to receive from each data repository and how this information is expected to be presented to the final user. The last part describes the dataset downloaded from DataCite's public OAI API. The information retrieved and its structure is compared with the information provided in the first subsection.

3.1. Points of access to DataCite

DataCite provides two APIs to the public for downloading records indexed in its database. These two points of access contain the same number of records but differ in the structure in which they are presented as well as in the detail of information provided.

DataCite Metadata Store (https://oai.datacite.org/). The DataCite Metadata Store is a service to manage activities related to Digital Object Identifier (DOI) registration at DataCite. The MDS is used to create, register, store and manage DOIs and