Formalised data citation practices would encourage more authors to make their data available for reuse


It is increasingly common for researchers to make their data freely available. This is often a requirement of funding agencies but also consistent with the principles of open science, according to which all research data should be shared and made available for reuse. Once data is reused, the researchers who have provided access to it should be acknowledged for their contributions, much as authors are recognised for their publications through citation. ***Hyoungjoo Park*** *and* ***Dietmar Wolfram*** *have studied characteristics of data sharing, reuse, and citation and found that current data citation practices do not yet benefit data sharers, with little or no consistency in their format. More formalised citation practices might encourage more authors to make their data available for reuse.*

Today's researchers work in a heavily data-intensive and collaborative environment in order to further scientific discovery across and within fields. It is becoming routine for researchers (i.e. authors and data publishers) to submit their research data, such as datasets, biological samples in biomedical fields, and computer code, as supplementary information in order to comply with data sharing requirements of major funding agencies, high-profile journals, and data journals. This is part of open science, where data and any publication products are expected to be made available to anyone interested.

Given that researchers benefit from publicly shared data through data reuse in their own research, researchers who provide access to data should be acknowledged for their contributions, much in the same way that authors are recognised for their research publications through citation. Researchers who use shared data or other shared research products (e.g. open access software, tissue cultures) should also acknowledge the providers of these resources through formal citation. At present, data citation is not widely practised in most disciplines and as an object of study remains largely overlooked.

Image credit: Data Security Breach by Blogtrepreneur. This work is licensed under a CC BY 2.0 license.

Our study examined characteristics of data sharing, reuse, and citation as documented in Clarivate Analytics' Data Citation index (DCI) and articles that cite authors whose datasets are indexed in the DCI. The DCI has only been available since 2012, whereas citation indexes to scholarly publications go back more than 50 years. We examined the following questions:

1. How prevalent is data reuse as measured by data citation?

Date originally posted: 2017-07-17

Permalink: http://blogs.lse.ac.uk/impactofsocialsciences/2017/07/17/formalised-data-citation-practices-would-encourage-more-authors-to-make-their-data-available-for-reuse/

Blog homepage: http://blogs.lse.ac.uk/impactofsocialsciences/

2. To what extent do authors formally and informally document data citation?
3. What are the ongoing challenges to studying data citation and reuse?

We initially identified which of the 156 fields listed in the DCI have been most active in formal data citation. Of these fields, the top ten account for the vast majority of citable sources indexed in the DCI. Genetics and Heredity is the top field with almost 2.3 million records (representing public datasets, software, data studies, and data repositories). Next, we identified 30 authors who were associated with the 15 most highly cited Genetics and Heredity records in the DCI. We then identified a sample of articles that cite these authors. We manually examined the 148 citing articles for evidence of data sharing and reuse in different areas of each article (e.g. the references, main text, acknowledgements, supplementary information, and author information) in order to identify formal (i.e. cited) and informal (i.e. mentioned in passing or implied) data sharing and reuse.

Findings

We found that data citations appear in the references section of an article less frequently than in the main text, making it difficult to identify the reward and credit for data authors (i.e. data sharers). Consistent data citation formats could not be found. Current data citation practices do not (yet) benefit data sharers. Also, data citation was sometimes located in the supplementary information, outside of the references. Data that had been reused was often not acknowledged in the reference lists, but was rather hidden in the representation of data (e.g. tables, figures, images, graphs, and other elements), which may be a consequence of the fact that data citation practices are not yet common in scholarly communications.

Ongoing challenges remain in identifying and documenting data citation. First, the practice of informal data citation presents a challenge for accurately documenting data citation. As we found, formal and informal data citation take place in different areas of articles. It would be reasonable to expect data citations to appear alongside standard bibliographic citations as acknowledgment of the author's use of the data.

Second, data recitation by one or more co-authors of earlier studies (i.e. self-citation) is common, which reduces the broader impact of data sharing by limiting much of the reuse to the original authors. This observation represents a key challenge to the identification of data reuse without analysing the content of the citing document to determine if data reuse actually took place. This finding demonstrates that an increase in citations does not necessarily indicate new and unique citers. Co-author self-citation needs to be studied in further detail in data citation.

Third, currently indexed data citations may not include rapidly advancing areas, such as in the hard sciences or computer engineering, because approximately 90% of indexed works were associated with journal articles. In a rapidly advancing area, conference proceedings can have greater importance than journal articles or books as research dissemination venues. Data citations included in conference proceeding papers are then less likely to be indexed.

Fourth, the number of authors associated with shared datasets raises questions of the ownership of and responsibility for a collective work, although some journals require one author to be responsible for the data used in the study. Large numbers of authors are common in some areas such as biomedical research because of the large research teams needed to carry out this research. Is this practical for data citation, particularly in determining the order of credit for data authors? Version control of datasets is also important because there may be multiple versions of publicly available data. This creates additional challenges for data citation.

The availability of data citation may encourage data authors to make their peer-reviewed data discoverable for reuse by others in order to increase data authors' recognition and rewards in scholarly communication. The frequency analysis of the subject categories in which data citation is taking place reveals that the formally recorded citations are largely concentrated in a small number of disciplines in the biomedical sciences and selected physical sciences. We cannot conclude from this that data citation is only predominant in these fields, but rather that these fields may have greater data repository representation in the DCI.

More detailed findings and conclusions can be found in our full paper. A pre-print version of the article is available here.

Date originally posted: 2017-07-17

Permalink: http://blogs.lse.ac.uk/impactofsocialsciences/2017/07/17/formalised-data-citation-practices-would-encourage-more-authors-to-make-their-data-available-for-reuse/

Blog homepage: http://blogs.lse.ac.uk/impactofsocialsciences/

This blog post is based on the authors' article, "An examination of research data sharing and re-use: implications for data citation practice", published in Scientometrics (DOI: 10.1007/s11192-017-2240-2).

Note: This article gives the views of the authors, and not the position of the LSE Impact Blog, nor of the London School of Economics. Please review our comments policy if you have any concerns on posting a comment below.

About the authors

***Hyoungjoo Park** is a PhD candidate in the School of Information Studies at the University of Wisconsin-Milwaukee. She holds her master's degree from the iSchool at Syracuse University. She has worked as a research engineer and a research analyst at LG CNS, a total IT service company at LG Corporation. Her research interests include data citation, data sharing and reuse, scholarly communication and linked data.*

***Dietmar Wolfram** is a Professor in the School of Information Studies at the University of Wisconsin-Milwaukee. He received his PhD from the University of Western Ontario. His research interests include applied informetrics, information retrieval system design and evaluation, scholarly communication and user studies. His ORCID iD is 0000-0002-4991-276X.*

Date originally posted: 2017-07-17

Permalink: http://blogs.lse.ac.uk/impactofsocialsciences/2017/07/17/formalised-data-citation-practices-would-encourage-more-authors-to-make-their-data-available-for-reuse/

Blog homepage: http://blogs.lse.ac.uk/impactofsocialsciences/