IJDC | *Peer-Reviewed Paper*

The Location of the Citation: Changing Practices in How Publications Cite Original Data in the Dryad Digital Repository

Christine Mayo
School of Information and Library Science
University of North Carolina at Chapel Hill

Todd J. Vision
Department of Biology
University of North Carolina at Chapel Hill

Elizabeth A. Hull
Dryad Digital Repository

Abstract

While stakeholders in scholarly communication generally agree on the importance of data citation, there is not consensus on where those citations should be placed within the publication – particularly when the publication is citing original data. Recently, CrossRef and the Digital Curation Center (DCC) have recommended as a best practice that original data citations appear in the works cited sections of the article. In some fields, such as the life sciences, this contrasts with the common practice of only listing data identifier(s) within the article body (intratextually). We inquired whether data citation practice has been changing in light of the guidance from CrossRef and the DCC. We examined data citation practices from 2011 to 2014 in a corpus of 1,125 articles associated with original data in the Dryad Digital Repository. The percentage of articles that include no reference to the original data has declined each year, from 31% in 2011 to 15% in 2014. The percentage of articles that include data identifiers intratextually has grown from 69% to 83%, while the percentage that cite data in the works cited section has grown from 5% to 8%. If the proportions continue to grow at the current rate of 19-20% annually, the proportion of articles with data citations in the works cited section will not exceed 90% until 2030.

Received 20 October 2015 ~ *Accepted* 24 February 2016

Correspondence should be addressed to Todd J. Vision, Department of Biology, University of North Carolina, Chapel Hill, NC 27599-3280, USA. Email: tjv@bio.unc.edu

An earlier version of this paper was presented at the 11th International Digital Curation Conference.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

Introduction

Data citation is the practice of referencing within a scholarly publication data that is not reported in detail within that publication but is nonetheless integral to the findings therein (Altman and King, 2007). When the data have been made available in coordination with that publication, we refer to this an 'original data citation.' An original data citation is in contrast to a 'data reuse citation,' in which the data being referenced was produced by other researchers or reported in a prior publication. While we recognize that there are cases where the distinction may not be easy to make, it is nonetheless useful in discussing the practices of researchers archiving data, and citing data, associated with traditional scholarly publications. Both original and data reuse citations can serve to document the source of data used in a publication, provide recognition for the data creators, facilitate future access and enable tracking of data reuse (CODATA-ICSTI Task Group on Data Citation, 2013). As evidence of its perceived importance to the integrity of the scholarly record, data citation figures prominently within the Transparency and Openness Promotion (TOP) guidelines (Nosek et al., 2015).

Recently, several organizations came together to articulate a set of Joint Declaration of Data Citation Principles (Data Citation Synthesis Group, 2014) intended to govern implementation of data citation policy and practice. Among these principles are statements such as 'data citations should be accorded the same importance in the scholarly record as citations of other research objects, such as publications' and 'data citations should facilitate giving scholarly credit and normative and legal attribution to all contributors to the data, recognizing that a single style or mechanism of attribution may not be applicable to all data.' At the time of writing, these principles have been endorsed by over 90 organizations, including data repositories, scholarly societies, publishers, and organizations critical to the scholarly communications infrastructure, such as CrossRef and DataCite.

Notably absent from the Joint Declaration, and related guidance on its technical implementation (Starr et al., 2015), is a specific recommendation for where data should be cited. Data citation practices vary considerably in the literature. In some fields, providing the data identifier has historically sufficed to satisfy journal policy. For example, repository-specific identifiers known as 'accession numbers' have typically been reported somewhere within the main text of the publication for nucleic acid sequences deposited in the GenBank database, a practice we refer to as 'intratextual citation.' More recently, a number of influential organizations, such as the Digital Curation Center (Ball and Duke, 2011) and CrossRef (2012) have recommended that a more traditional scholarly citation be placed within the works cited section, or bibliography, of the publication. DataCite recommends that this citation include at least the names of the creator(s), the publication year, a title, the identity of the publisher and the identifier[1]. These recommendations have been taken up to varying degrees by publishers and may sometimes be seen in journals' instructions to authors, although there tends to be greater agreement that the works cited is appropriate for reuse data citations than for original data citations. To complicate matters, some publishers instruct authors to put original data citations in a dedicated data availability section, which may

1 Cite Your Data: https://www.datacite.org/services/cite-your-data.html, retrieved 15 September 2016.

reduce the likelihood that authors will feel it necessary to include original data citations with the works cited.

As an example, BioMedCentral's editorial policy states that "all authors must include an 'Availability of Data and Materials' section in their manuscript detailing where the data supporting their findings can be found" and then goes on to require that "all publicly available datasets be fully referenced in the reference list with an accession number of unique identifier such as a digital object identifier (DOI)"[2]

Data repositories may also provide instructions to authors on how to formulate data citations, and in many cases these follow the CrossRef/DCC/DataCite recommendations above. Dryad, for instance, provides a data citation string on each data package page that includes the author(s)/creator(s), publication year, title, publisher (Dryad Digital Repository) and a Digital Object Identifier (DOI). Although it is not explicitly stated that the original data citation should be included within the works cited, it is implied by the nature of the bibliographic information provided.

Since much of Dryad's content is associated with publications in the life sciences, with its tradition of intratextual original data citations using accession numbers, we wished to know to what extent the recommendations to include full original data citations within the works cited were having an effect on practice. Are Dryad DOIs being treated like GenBank accession numbers, or are authors citing data as recommended by CrossRef, DCC and DataCite, and is there evidence of a temporal trend? Dryad is a convenient object for such a study because the repository only hosts data associated with publications, and records one primary publication for each data package. This makes it easy to identify a corpus of articles and to interpret the counts of articles in this corpus for which the original data citation occurs in the main text, within the works cited, both, or neither. Also, many of the publications associated with Dryad content are available from Europe PubMed Central (EuropePMC), which provides API access to download and parse the XML-encoded full text of the associated article. We focused our efforts on detecting occurrences of the Dryad DOI because this is the most critical element of the data citation for unambiguous linkage to the data package.

Methods

For all data packages published in Dryad between 2011 and 2014 inclusive, we recorded the DOI for the data packages and the DOI for the associated publication. We retrieved the XML for the full text of all articles in this set that could be found within the Open Access subset of EuropePMC. The XML tags were used to split the text into the body and works cited sections. We then searched for the Dryad DOI within each section and classified each article as containing a hit within the body (intratextual), within the works cited, within both sections, or within neither. Hits to data availability statements were counted as intratextual.

Results

The full text of 1,125 articles associated with Dryad data packages published from 2011 to 2014 could be found in the EuropePMC Open Access (OA) subset. This represents 16.5% of all 6,834 data packages published by Dryad during that interval. A further 604

2 Availability of Supporting Data: https://www.biomedcentral.com/getpublished/editorial-policies, retrieved 15 September 2016.

articles were present in EuropePMC but not within the OA subset and could not be automatically classified. There are more articles from recent years, reflecting the growth in usage of the repository as a whole. The sample sizes ranged from 42 articles in 2011 to 558 articles in 2014.

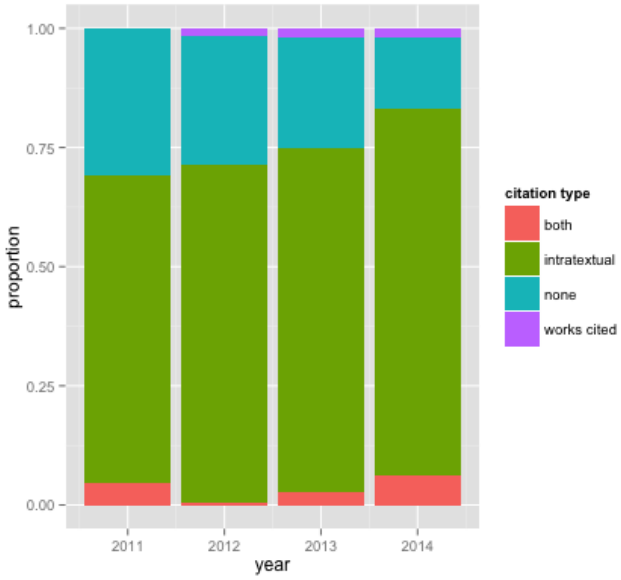

Figure 1. The proportion of articles by year in each citation category. The numbers of articles each year were 42 (2011), 175 (2012), 350 (2013) and 558 (2014). A 4x4 contingency test yields χ^2=30.14 with nine degrees of freedom. Thus, the observed values differ significantly (p<0.001) from what would be expected in the absence of a temporal trend.

Across all years, we found 833 articles with only an intratextual reference to the corresponding Dryad DOI, 21 in which the DOI appeared only within the works cited section, and 47 publications in which it was present in both. The DOI was not found in 224 (19.9%) of the articles. The proportions show a temporal trend (see Figure 1). The percentage of articles that did not include the DOI at all declined each year, from 31% in 2011 to 15% in 2014. The percentage of articles with Dryad DOIs placed either intratextually alone, or both intratextually and in the works cited, grew from 69% to 83%, an average of 20% annual growth. The proportion that included the DOI in the works cited (either alone, or with an intratextual citation) grew from 5% to 8%, an average of 19% annual growth.

Discussion

Our results show that intratextual references to the Dryad DOI currently make up the majority of original data citations in recent years. In the most recent year, such references were found in over 80% of articles, while present in the works cited section in only 8% of articles. The rate of growth is similar (19-20%) for use of the Dryad DOI intratextually and in the works cited, but articles in which the DOI is present in both sections still represents only about 6% of articles. Encouragingly, the total percentage of articles in which the DOI was present in any section has been rising steadily, from 69%

 doi:10.2218/ijdc.v11i1.400

of articles in 2011 to 85% in 2014. The continuing rise in references to data of any type would suggest that authors and journals at least appreciate the importance of original data citations, even if there is still not consensus about where they belong.

Less encouragingly, at the current rate of growth, the proportion of articles with data citations in the works cited section is not expected to exceed 90% until 2031. Thus, current efforts to promulgate best practice, at least as articulated by CrossRef and the DCC are working, but only very slowly. It is worth considering whether, if these policy recommendations are not adopted, there will be consequences for scholarly communications infrastructure to have original data citations occurring predominantly within the main body of text. Will this negatively effect the incentives for data publication among researchers, the discoverability of linkages between articles and data, and other intended benefits of including data citations within formal bibliographies?

Our results contrast interestingly with those of Kratz and Strasser (2015) on self-reported data citation behavior. In their study, 63% of scientists reported to have used formal data reuse citations (i.e., appearing in the works cited section) and none reported having used informal (i.e., intratextual) data reuse citations, although in their free-text response, 16% said that 'it was appropriate to cite data informally in the body of the text.' One explanation for the discrepancy may be that there are different attitudes toward and practices for original data citation versus data reuse citation. Another explanation, not mutually exclusive, is that some researchers do not clearly distinguish between the act of citing the article reporting original data and citing the original data itself. It is rare to find guidelines for authors provided by journals to be clear on these distinctions.

This study provides a limited view of data citation practices in the literature, examining only the Open Access subset of articles in EuropePMC with data deposited in Dryad. It would be of interest to contrast these results with those for other disciplines and repositories, particularly those without a strong tradition of intratextual citations of accession numbers. Also of value would be to do a more thorough text analysis in order to determine how many data citations exist that are lacking DOIs. Finally, it would be useful to distinguish original data citations that are in dedicated data availability sections from those occurring in other parts of the text. While it is clear that use of data availability sections is growing, the inconsistent way in which data availability sections are tagged within the documents makes it difficult to reliably identify them across many different journals.

Conflicts of Interest

Christine Mayo is employed as a curator for the Dryad Digital Repository, Elizabeth A. Hull is employed as Operations Manager by Dryad, and Todd J. Vision is a member of Dryad's Board of Directors.

Acknowledgements

We would like thank Ryan Scherle of Dryad for assistance with data collection, and Martin Fenner of DataCite and Joanna McEntyre of EuropePMC for methodological advice. This work was supported by NSF grants EF-0905606 to the National

Evolutionary Synthesis Center and DBI-1147166 to TJV. The data is available from the Dryad Digital Repository[3].

References

Altman, M., King, G. (2007). A proposed standard for the scholarly citation of quantitative data. *D-Lib Magazine, 13*(3/4). doi:10.1045/march2007-altman

Ball, A. & Duke, M. (2015). How to cite datasets and link to publications. DCC How-to Guides. Edinburgh: Digital Curation Centre. Retrieved from http://www.dcc.ac.uk/resources/how-guides

CODATA-ICSTI Task Group on Data Citation. (2013). Out of cite, out of mind: The current state of practice, policy, and technology for the citation of data. *Data Science Journal, 12*. doi:10.2481/dsj.OSOM13-043

CrossRef. (2012). DOIs in use: DataCite. *CrossRef Quarterly,* January, 2012. Retrieved from http://www.crossref.org/10quarterly/quarterly_jan12.html

Data Citation Synthesis Group. (2014). *Joint declaration of data citation principles*. Retrieved from https://www.force11.org/datacitation

Mayo, C., Vision T.J., & Hull E.A. (2016). Data from: The location of the citation: Changing practices in how publications cite original data in the Dryad Digital Repository. Dryad Digital Repository. doi:10.5061/dryad.8q931

Kratz, J.E., Strasser, C. (2015). Researcher perspectives on publication and peer review of data. *PLoS ONE, 10*(2): e0117619. doi:10.1371/journal.pone.0117619

Nosek, B.A., Alter, G., Banks, G.C., Borsboom, D., Bowman, S.D., Breckler, S.J., Buck, S., … Yarkoni, T. (2015). Promoting an open research culture. *Science, 348*(6242): 1422-5. doi:10.1126/science.aab2374

Starr J., Castro E., Crosas M., Dumontier M., Downs R.R., Duerr R., Haak L.L., ... Clark T. (2015) Achieving human and machine accessibility of cited data in scholarly publications. *PeerJ Computer Scien*ce 1:e1. doi:10.7717/peerj-cs.1

3 See: http://doi.org/10.5061/dryad.8q931