The grand challenge of data citation: a critical look at the Making Data Count Data Citation Corpus

Roderic D. M. Page[1]
[1]SBOHVM, MVLS, University of Glasgow, Glasgow G12 8QQ, UK

Data citation

Bibliographic citation - where one publication cites another - is standard practice in academia. Citations between publications generate a graph or network of citations. This graph enables us to track the provenance of statements and compute metrics to assess the impact of an individual article, its authors, and the journal where it was published. The value of the citation graph is reflected in the fact that it was, until recently, only accessible as a commercial product provided by a small number of vendors, such as Elsevier and Clarivate. The Open Citation Index now provides free access to a large part of this graph (Peroni and Shotton, 2020).

Yet bibliographic citation only captures relationships between publications. Much of modern science, especially in molecular biology, is data-driven. Researchers use (and reuse) data, most notably biomolecular sequences (e.g., DNA and protein). Early efforts to track citations of sequence data discovered that publications may use sequence data without citing the original publication of those sequences (Page 2010). Without a mechanism to keep track of data citation we lose the ability both to track the provenance of data and to give credit to those who generated that data.

Recently, the Make Data Count initiative (https://makedatacount.org) has released the Data Citation Corpus (DataCite & Make Data Count, 2024a), providing millions of citation links between publications and datasets. While this development could greatly enhance our ability to document data use and reuse, a critical look at the corpus reveals numerous errors, including citations of fictitious data, and a lack of clarity over what constitutes a “citation” (Page, 2024a, 2024b).

The corpus and its problems

Version 1 of the corpus was released in early 2024 and contained 10 million data citations (DataCite & Make Data Count, 2024a). However, this release contained millions of duplicates and has been superseded by version 2 (DataCite & Make Data Count, 2024b) containing 5.2 million citations. Citations in the corpus comprise links between a publication Digital Object identifier (DOI) and an identifier for a data item, which may be a dataset, or in the case of molecular data the accession number for an individual item such as a nucleotide sequence. The corpus combines information from two sources, DataCite (1.2 million citations), and the Chan Zuckerberg Initiative (CZI) (4 million citations). The citations provided by the CZI were obtained by data mining the scientific literature using a language model derived from BERT (Devlin et al., 2019), but full details of the method used have not been published, nor is the code or model available.

I downloaded version 2 from Zenodo. The data is in JavaScript Object Notation (JSON) format, which I then loaded into an Apache CouchDB database. CouchDB stores JSON documents natively and provides a "map-reduce" framework to summarise those documents.

In a listing of repository by number of citations the largest repository in the corpus is the European Nucleotide Archive (a member of the International Nucleotide Sequence Database Collaboration or INSDC, Karsch-Mizrachi et al. 2012). There are some obvious data quality issues, for example, Figshare appears twice, as "Figshare" and as "figshare". Combining those two entries still underestimates the role FigShare plays as the repository "Taylor & Francis" is a branded instance of Figshare: https://tandf.figshare.com.

Counting the number of data citations for each item shows that most items are cited only once. There are notable exceptions, such as LY294002 with 9983 citations and A549 with 5883 citations. The corpus treats LY294002 and A549 as accession numbers for molecular sequences, but given that LY294002 is a chemical compound and A549 is cell type it is likely that the most highly-cited data records are not data items, but instead are false matches to biological entities.

Closer examination shows that false matches plague the corpus, for example (with links to annotated sources):

- A museum specimen CR00240699 is mistakenly interpreted as a GenBank accession number, see https://hyp.is/CGTJcM_kEe674TfyvGLC0A/zookeys.pensoft.net/article/21580/download/pdf/287887
- A grant number Y21026 is mistakenly interpreted as a GenBank accession number, see https://hyp.is/HpVXhs9PEe6D2UMxrIdqJw/bmjopen.bmj.com/content/12/9/e054887

To gauge the extent of these errors I analysed two datasets in more detail. The Protein Data Bank (PDB) is the third most cited resource in version 2 of the corpus with 403,412 individual citations. We can check whether these are potentially valid by comparing these citations to a list of known identifiers obtained from the PDB. This enables us to screen out obviously incorrect citations, such as "24hr" being treated as a PDB identifier, see Page, 2024a. Of the original PDB citations 14% were not valid PDB identifiers, leaving 86% which correspond to a possible PDB record (whether they actually are citations would need further investigation).

GenBank has some 3.7 billion sequences (Sayers et al., 2024) so the approach of comparing citations to a list of identifiers does not readily scale to nucleotide sequences. Of a random sample of 1000 cited nucleotide sequences, I was able to find just over half (51%) in the sequence databases. Extrapolating to the 2.2 million sequence citations in the corpus it is likely that 1 million of these nucleotide citations are fictitious.

What does the corpus measure?

Most data items are cited only once. This might imply that most data is of limited interest, but closer examination suggests that these single citations are not citations at all. For example, the vast majority of data in the "Taylor & Francis" repository is published by Informa UK Ltd, which is the parent company of Taylor & Francis. If you visit an article in an Informa journal, such as "Enhanced Selectivity of Ultraviolet-Visible Absorption Spectroscopy with Trilinear Decomposition on Spectral pH Measurements for the Interference-Free Determination of

Rutin and Isorhamnetin in Chinese Herbal Medicine" (Zhou et al., 2021) the supplementary information is stored in Figshare (https://doi.org/10.6084/m9.figshare.14103264.v1). These links between the publication and the supplementary information are treated as "citations" in the corpus. Yet, this is not a citation in the sense of reuse of already published data, rather it records a link between two publication events, those of the publication and its accompanying data. Metrics based on these "citations" do not measure data reuse, but rather data preservation.

Next steps

This preliminary analysis of the Data Citation Corpus suggests that at least a million of the citations are fictitious (citing non-existent accession numbers), and that many of the so-called citations are actually publication events (data being stored in a repository). The title of this article ""The Grand challenge..."" is a nod to the Elsevier Grand Challenge: Knowledge Enhancement in the Life Sciences (https://web.archive.org/web/20081024064958/http://www.elseviergrandchallenge.com/description.html). The work reported in Page (2010) was my entry in that challenge, and the late David Shotton was one of the judges. Given the significant issues with the Data Citation Corpus, perhaps the time is ripe for a similar competition to develop better methods for accurately tracking citations of data.

References

DataCite & Make Data Count (2024a). Data Citation Corpus Data File (Version v1.1) [Dataset]. DataCite. https://doi.org/10.5281/zenodo.11216814

DataCite & Make Data Count. (2024b). Data Citation Corpus Data File (Version v2.0) [Dataset]. DataCite. https://doi.org/10.5281/ZENODO.13376773

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). https://doi.org/10.18653/v1/N19-1423

Karsch-Mizrachi, I., Nakamura, Y., & Cochrane, G. (2012). The International Nucleotide Sequence Database Collaboration. Nucleic Acids Research, 40(Database issue), D33–D37. https://doi.org/10.1093/nar/gkr1006

Page, R. D. M. (2010). Enhanced display of scientific articles using extended metadata. Web Semantics: Science, Services and Agents on the World Wide Web, 8(2–3), 190–195. https://doi.org/10.1016/j.websem.2010.03.004

Page, R. D. M. (2024a, February 20). Problems with the DataCite Data Citation Corpus. iPhylo. https://doi.org/10.59350/t80g1-xys37

Page, R. D. M. (2024b, October 8). The Data Citation Corpus revisited. iPhylo. https://doi.org/10.59350/wvwva-v7125

Peroni, S., & Shotton, D. (2020). OpenCitations, an infrastructure organization for open scholarship. Quantitative Science Studies, 1(1), 428–444. https://doi.org/10.1162/qss_a_00023

Sayers, E. W., Cavanaugh, M., Clark, K., Pruitt, K. D., Sherry, S. T., Yankie, L., & Karsch-Mizrachi, I. (2024). GenBank 2024 Update. Nucleic Acids Research, 52(D1), D134–D137. https://doi.org/10.1093/nar/gkad903

Zhou, P.-R., Tang, Z.-F., Wei, K.-S., Wan, Y., Gao, Y.-M., Liang, Y.-M., Yan, X.-F., Bin, J., & Kang, C. (2021). Enhanced Selectivity of Ultraviolet-Visible Absorption Spectroscopy with Trilinear Decomposition on Spectral pH Measurements for the Interference-Free Determination of Rutin and Isorhamnetin in Chinese Herbal Medicine. Analytical Letters, 54(17), 2750–2768. https://doi.org/10.1080/00032719.2021.1888966