Editorial

Data sharing and data citation: join the movement!

Rachael Lammey
Head of Community Outreach, Crossref; rlammey@crossref.org; ORCID 0000-0001-5800-1434

If you attended the EASE Conference in Bucharest in July 2018, you might remember a session on data citation, presented by Bart Verberk from Springer Nature, Heather Tierney from ACS & COPE and me, from Crossref. If you weren't there, we talked about data sharing policies and support for these, ethical issues around research data, and the practicalities of citing data and registering these citations with Crossref.

In the months since that event, we've seen the conversation around data publication and data citation continue to grow. Enabling data to be clearly and persistently identified and linked to the journal articles it supports is key for the data to be properly acknowledged, credited and tracked within the research process.

As an editor, it's valuable to be able to look at the data associated with an article you're reading. This can help provide more details behind the ideas the the paper puts forward. As a researcher, you may want to share this information to allow others to check your processes, to help them reproduce, extend or avoid duplicating your research, and also to see who is citing, downloading or viewing the research you are publishing. To bring the sharing and citation of data into the mainstream, authors, editors and publishers need to take steps to include this in their regular workflows. This is easily said, but what does this look like in practice?

The research community has been motivated to share ideas and experiences on how to introduce and implement a policy on data sharing and data citation. A recent article[1] by Jones, Grant and Hrynaszkiewicz (2019) explores the approaches taken by two publishers, Taylor & Francis and Springer Nature, in communicating and integrating their data sharing policies into standard journal workflows.

There are similarities in the approaches taken by these two publishers which may prove helpful to those considering implementing policies of their own. Their publisher policies are tiered, from a basic approach which encourages data sharing, citation and the provision of a data availability statement,[2] through to an open and fully FAIR[3] approach where data must be made freely and openly available under a suitable licence, and data citation and data availability statements are mandatory.

Both publishers have journals adopting different levels of data sharing within these tiers, and they acknowledge that there are differences in willingness to share data across different disciplines. Springer Nature found that "life scientists were the most likely to make their data available in a repository, while those in the physical sciences were least likely to".[1] There are also differences in the perceptions of individual editors and society officers. Taylor & Francis reports that some see the initiative as "important and valuable", but some were worried about the risk of 'scooping' and wary of added impact on editor and reviewer workloads.

Some also raised the question of credit, and how researchers will be acknowledged for sharing their data. The Make Data Count Project[4] is doing important work in this area, and aims to achieve the following:

- to publish a new COUNTER code of practice regarding how data usage and impact should be meaningfully measured and reported;
- to deploy a data-level metrics (DLM) aggregation and publication service based on the open-source Lagotto software and hosted by DataCite;
- to integrate the DLM service with new exemplar data sources and clients; and
- to perform advocacy and training regarding the importance and use of DLM to develop engaged communities of practice and support.[5]

Key to realising the first three of these outputs is for data to be cited consistently and accurately by publications. This is why even the basic Taylor & Francis and Springer Nature policies encourage data citation, and the more open tiers require it. In terms of building out this piece of the picture, a group of industry stakeholders has published a roadmap[6] on how data citation can be implemented into standard journal/publisher workflows, by working through four steps.

The first step is to develop a data policy that includes data citation. Many publishers such as those mentioned publish their policies openly on their websites to serve as a guide for others. The next stage is to explain to authors how they should be citing data, via the Instructions for Authors and/or in the manuscript submission system. Next is to update internal workflows, ie making sure that this information is being carried between internal systems and that communications to copy editors and typesetters are clear on what to do with this type of reference.

Finally, publishers should include these citations in the metadata that they send to Crossref.[7] This basically involves remembering to include links to data in reference lists or adding them in the 'relations' section of the metadata. Once these citations are sent to Crossref (or DataCite[8] for data repositories), they are then disseminated to anyone interested in using this information to link articles and data and to see how data are being used. Event Data[9] by Crossref and DataCite provide an openly available central place to retrieve usage information and

citations for datasets. This information is already being used to display citations to data in repositories like Dryad,[10] and in databases like Scholexplorer[11] (an aggregation and query service fed by the Scholix open information ecosystem) and in Elsevier's Scopus database.

Having links to data displayed via these endpoints has the knock-on effect of further increasing data visibility and usage. It also helps researchers (and their funders) actually see this usage and get credit for it, helps initiatives like Make Data Count attempt to make meaningful assumptions around that usage, and helps editors and publishers see how their collections feed through the research ecosystem. Finally, it helps readers get access to the data that support the research they are reading. The benefits are manifest, so we encourage the community to support this valuable initiative to recognise the importance of data as a key component of the scientific record.

References

1 Jones L, Grant R, Hrynaszkiewicz I. Implementing publisher policies that inform, support and encourage authors to share data: two case studies. *Insights: the UKSG journal*. 2019;32. DOI: https://doi.org/10.1629/uksg.463

2 Research data policies [Internet]. Data Availability Statements | Authors | Springer Nature. Available from: https://www.springernature.com/gp/authors/research-data-policy/data-availability-statements/12330880 (accessed 7 May 2019)

3 FAIR data [Internet]. Wikipedia. Wikimedia Foundation; 2018. Available from: https://en.wikipedia.org/wiki/FAIR_data#cite_note-FAIR_principles_2016-1

4 Lowenberg, Daniella. [Internet]. Data, a first-class research output. Available from: https://makedatacount.org/ (accessed 7 May 2019)

5 About [Internet]. Data, a first-class research output. 2017. Available from: https://makedatacount.org/about/ (accessed 7 May 2019)

6 Cousijn H, Kenall A, Ganley E, *et al*. A data citation roadmap for scientific publishers. *Scientific Data*. 2018;5:180259. DOI: https://doi.org/10.1038/sdata.2018.259

7 Lammey, Rachael. [Internet] Data Citation: what and how for publishers - Crossref. www.crossref.org. 2018. Available from: https://www.crossref.org/blog/data-citation-what-and-how-for-publishers/ (accessed 9 May 2019)

8 DataCite [Internet]. DataCite homepage. Available from: https://datacite.org (accessed 8 May 2019)

9 Event Data - Crossref [Internet]. www.crossref.org. 2017. Available from: https://www.crossref.org/services/event-data/ (accessed 8 May 2019)

10 Dryad Digital Repository [Internet]. Dryad. Available from: https://datadryad.org/ (accessed 8 May 2019)

11 The Data Literature Interlinking Service [Internet]. ScholeXplorer. Available from: http://scholexplorer.openaire.eu/ (accessed 9 May 2019)