Data Citation: Let's Choose Adoption Over Perfection

Daniella Lowenberg, Rachael Lammey, Matthew B. Jones, John Chodacki, Martin Fenner

DOI: 10.5281/zenodo.4701079

In the last decade, attitudes towards open data publishing have continued to shift, including a rising interest in data citation as well as incorporating open data in research assessment (see Parsons et al. for an overview). This growing emphasis on data citation is driving incentives and evaluation systems for researchers publishing their data. While increased efforts and interest in data citation are a move in the right direction for understanding research data impact and assessment, there are clear difficulties and roadblocks in having universal and accessible data citation across all research disciplines. But these roadblocks can be mitigated and do not need to keep us in a consistent limbo.

The unique properties of data as a citable object have attracted much needed attention, although it has also created an unhelpful perception that data citation is a challenge and requires uniquely burdensome processes to implement. This perception of difficulty begins with defining a 'citation' for data. The reality is that all citations are relationships between scholarly objects. A 'data citation' can be as simple as a journal article or other dataset declaring that a dataset was important to the creation of that work. *This is not a unique challenge*. However, many publishers and funders have elevated the relationship of data that "underlies the research" into a Data Availability Statement (DAS). This has helped address some issues publishers have found with typesetting or production techniques that stripped non-articles from citations. However, because of this segmentation of data from typical citation lists, and the exclusion of data citations in article metadata, many communities have felt they are in a stalemate about how to move forward.

Data citations are targeted as an area to explore in terms of research assessment. However, we do not have a clear understanding of how many data citations exist or how often data are reused. In the last few years, the majority of data citation conversations have been facilitated through groups at Research Data Alliance (via Scholix), Earth Science Information Partners (ESIP), EMBL- European Bioinformatics Institute (EMBL-EBI), American Geophysical Union (AGU), and FORCE11. These conversations have focused primarily on datasets and articles that have DOIs from DataCite and Crossref, respectively, emphasizing the relationship between datasets and published articles. While those relationships are areas that need broad uptake from repositories and publishers alike, they do not illustrate the full picture. Many citations are not being accounted for, namely biomedical datasets with accession numbers and compact identifiers that are not registered through DataCite but readily accessible through resolvers like identifiers.org. There is also a lack of understanding around the citations of datasets in other scholarly and non-scholarly (e.g., government documents, policy papers) outputs.

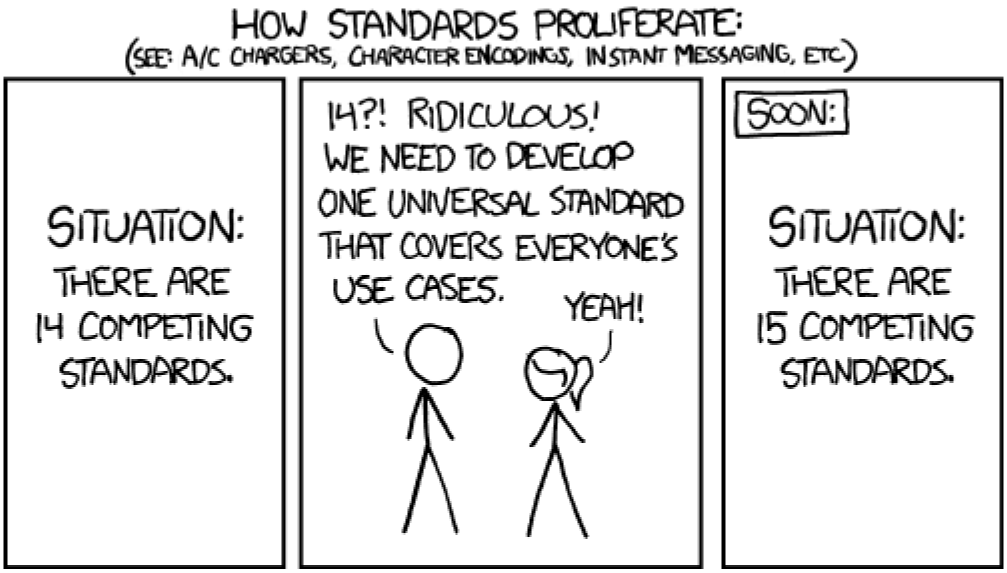
For these reasons, we have tried to ensure that conversations about data citation are not framed solely around the notion of assigning "credit" or around assigning any specific meaning to citations, for that matter. Without a full picture of how many citations exist, how datasets are composed across disciplines, how citation behavior varies across disciplines, and what context the citations are used in, it is impossible and inappropriate to use citations as a shorthand for credit. The community is working towards a better understanding of citation behavior—and we believe we will get there—but we need to be careful and considered in doing so to avoid repeating previous mistakes (e.g., creating another impact factor).

Why data citation is perceived as difficult

Data are complex. As mentioned in our 2019 book, data are nuanced. This means data citations are complex and understanding these nuances are essential for understanding true measures of data reuse. For instance, there is work to be done to understand the role of provenance, dataset to dataset re-usage, data aggregation and derivation, and other ways for measuring usage of datasets without a formal "citation".

Data citations are complex. There is a well-established concept of scholarly citations, using reference lists of citations formatted in a defined citation style. The main challenges with the current approach center on making citations machine-readable using standardized metadata instead of citation styles meant for human readers, as well as on making citations machine-accessible using open APIs and aggregators. These are general challenges to be addressed with citation, but there are additional questions specific to handling data citations: is there a DOI or other persistent identifier and basic citation metadata for the dataset, is there tooling to bring this information into the citation style used by the publisher, should data citations go into the article reference list, what to do when the number of datasets cited in a publication goes into the 1000s or more, should datasets in reference lists be called out as data, etc.

There's a lack of consistency in guidance. Despite the growing interest among various stakeholders (researchers, journals, repositories, preprint servers, and others) in supporting data citation, there is no consistency in guidance for and across these groups. These inconsistencies are in respect to how citations should be formatted, how citations should be marked up and indexed, and what the role for each stakeholder should be (especially repositories). Some of this can be attributed to a constant reinvent-the-wheel approach as well as to the wide variety of stakeholder groups and hubs for this information—understandably, people are confused between how Scholix fits with OpenAire, Crossref, and DataCite, nevermind the profusion of other overlapping initiatives and projects in this space that can make it even more difficult to navigate. It is clear that our best way forward is to not consistently reinvent the wheel, spawning new groups and initiatives, but rather to build on the existing work, leveraging the successes of the last decade of investment in data citations, and finding solutions for the more advanced issues at hand. In short: let's focus on developing the most basic, clear guidance, and work upwards from there.


HOW STANDARDS PROLIFERATE:
(SEE: A/C CHARGERS, CHARACTER ENCODINGS, INSTANT MESSAGING, ETC)
SITUATION: THERE ARE 14 COMPETING STANDARDS.
14?! RIDICULOUS! WE NEED TO DEVELOP ONE UNIVERSAL STANDARD THAT COVERS EVERYONE'S USE CASES.
YEAH!
SOON:
SITUATION: THERE ARE 15 COMPETING STANDARDS.

https://xkcd.com/927

There's a tension between data availability statements and data citations. In the last decade, publishers and funders have heavily focused on requiring data availability statements and ensuring that is the way to designate when articles have an associated dataset published. These data availability statements are rarely marked up as a relation in the article metadata or a note of re-use (outside of self citation). If we continue to focus solely on data availability statements as a required first step, which have yet to solve the "machine readability problem", we will lose slim resources that would be better used to think about how each journal publisher can designate data reuse and citations.

Guidance and decision points

Understanding the many intricacies of data, citations, and data citation, we propose the following path forward for our communities to work effectively towards achieving widespread implementation of data citation and data reuse. This path forward begins with making decisions around clear guidance that needs to be provided, shifting focus away from "decision-pending" attitude and moving forward with clear recommendations on the following:

- **Best practices for citing datasets in articles, preprints, and books**. We have multiple sets of best practices. We don't need more guidance documents, we need consolidation and rationalization of the guidance that already exists.
- **Simplifying relationship type complexity.** The complexity of ontologies for relationships is causing unnecessary churn and delays in implementation. Providers should simplify this; however, the community shouldn't wait. We can and should implement viable solutions now. We should be promoting datasets in reference lists as a first viable solution.
- **How non-DOIs are cited.** We have too many conversations happening about DOIs and not enough happening about citation in other identifier communities. These communities need to reach some simple conventions around putting data citations into reference lists with

globally unique PIDs and citation metadata, in order to avoid requiring massive text mining efforts looking for string matches to, for example, “PDB:6VXX”, the identifier for the spike protein for COVID-19.

- **Publisher support for those who are not working with Crossref.** Not all publishers use Crossref services or have the ability to implement Crossref’s approaches to data citations. We need to focus attention on accessible methods for reference extraction (e.g., from PDFs) and larger support for smaller publishers that do not have the resources to retool to fit current guidance.
- **The role for data repositories.** Publishers are key to implementing data citation but data repositories must also focus on declaring relationships to articles and other outputs in their metadata. Data repositories should focus on making their datasets citable through PIDs and declaring robust metadata as well as reporting all known citations and linkages publicly so they can be used for aggregation.
- **Researchers should cite data despite these infrastructure hold-ups.** Regardless of the hurdles to implementing all of the established best practices, the basic fact remains that researchers can currently cite data and they should, using approaches available today.

Choosing adoption over perfection

Perfection is the enemy of good and finding solutions for every complexity of data citation does not need to be a roadblock to get started. We can use a phased approach to begin implementing best practices for data citations right now:

Phase I: basic implementation
Align as much as possible with existing community practices and workflows (e.g., using reference lists)

Phase II: advanced implementation
Address special use cases (e.g., relation types, machine-readable data availability statements, dynamic data, dataset-dataset provenance)

Phase III: beyond data citation
Build infrastructure for other indicators assessing data reuse

While we have dabbled in all three of these phrases already, we are still largely stuck in Phase I, constantly reinventing the same basic wheel that keeps spinning around the same place.

Our focus should be on how to scale these best practices across all publishers and repositories, supporting the diverse research landscape. This includes advancing the conversation beyond the DOI-based focus. Once that happens we can really move forward with building mechanisms for credit and understanding data re-use for research assessment.

Despite the agenda ahead, there are many steps that can be taken right now to continue towards the dreamstate. The community should not wait for infrastructure to be perfect before engaging in data citation support.

This is important, so let's say it again! *The community should not wait for infrastructure to be perfect before engaging in data citation support.*

Data citations are harder when we act like the adoption hurdles are insurmountable, so let's simplify. . Our infrastructure for data citations will continue to improve, use cases will continue to be defined and evolve, and we need as many broad stakeholders as possible to hop on board now and work with us towards comprehensive support for data citation.

This perspective piece was written by members of the Make Data Count team which is funded by the Alfred P. Sloan Foundation. For more information on our initiative, visit https://makedatacount.org.