To:
National Science Foundation
2415 Eisenhouwer Avenue
Alexandia, Virginia 22314

Re:
Response to NSF 20-015 Dear Colleague Letter: Request for Information on Data-Focused Cyberinfrastructure Needed to Support Future Data-Intensive Science and Engineering Research

6 December 2019

Dear Dr. Erwin Gianchandani and Dr. Manish Parashar:

On 22 October 2019, you wrote a Dear Colleague Letter to request the community to “provide input to NSF on specific data-intensive S&E research questions and challenges and the essential data-related CI services and capabilities needed to publish, discover, transport, manage and process data in secure, performant and scalable ways to enable that data-intensive research.” This is our response.

Enabling Reliable Content-based Data Citations

Abstract

Scientific data lie at the foundation of discovery. Increasingly, scientific data are stored digitally and are openly available online. Reliable data references are essential to the integrity of our scholarly record and the distributed nature of data requires that we increase management of versioning and identifiers in order to practice reproducible science. But, in spite of evidence to support that URLs and DOIs are unreliable due to link rot and content drift, successful initiatives like DataCite [1] and DataONE [2] continue to rely on them. We propose to adopt commonly used cryptographic hashing techniques to complement existing citation practices to implement reliable data references. The proposed reliable references can be used in concert with existing digital data infrastructures to create data indexing and archiving schemes without relying on a complex web of nondeterministic redirection and indirection associated with common use of URLs and DOIs in citations.

Data-Intensive Research Questions and Challenges

Answering global research questions about life on our planet requires meticulous integration of datasets across geospatial, temporal, institutional and disciplinary domains.

Biology has led the way in open data and open science, but despite the increased availability of open-access information facilities and data publications (e.g., Data Dryad, Zenodo, Global Biodiversity Information Facility), re-use of existing datasets to answer a new research question remains a largely manual, resource intensive and error-prone task that requires significant technical expertise. Especially with data-intensive research, scholars need specialized skills to re-use data and require access to adequate digital network, compute, and storage facilities.

But, even if adequate skills and resources are available, a basic challenge exists: the data of a study need to be found, retrieved, and verified. Current data citation practices rely on including a DOI or a web address and an access date (e.g., https://doi.org/10.123/456, accessed at https://example.org/data.zip on 2019-10-01) in the data citation. Assuming that the DOI or web address still produces data at time of re-use, the authenticity of the retrieved data cannot be easily verified. So, current data citation practices are not immune to two documented phenomena on the internet: link rot (data unavailable) and content drift (data available but changed).

This is why we argue that the most basic challenge in re-use of data-intensive, data-driven research is still how to reliably reference data sources. Reliable data citations not only benefit future generations of researchers, but also facilitate peer-reviews of data-intensive research publications. Implicit in this is the requirement for persistent, unique identifiers that are reliably linked to the referenced dataset.

Data-Oriented Cyber Infrastructure Needed to Address the Research Questions and Challenges

A part of the Unix philosophy: "Expect the output of every program to become the input to another, as yet unknown, program. ", might as well be used to attribute qualities of re-usable research: "Expect the output of every *research project* to become the input to another, as yet unknown, *research project*."

Traditionally, re-use of past research relied on scholars (e.g., read a paper and extract knowledge). Increasing, computer programs are used to complement manual knowledge extraction in order to facilitate data-intensive and data-driven research.

Even though academic publishing and data collections have digitized at staggering rates over the last 20 years, a human reader is still needed to manually retrieve a referenced dataset. For instance, a recommended Zenodo data publication citation:

> Poelen, Jorrit H. (2019). Global Biotic Interactions: Elton Dataset Cache (Version 0.3) [Data set]. Zenodo. http://doi.org/10.5281/zenodo.3523403 .

includes where the publication can be found (at https://doi.org/10....), but does not reliably link to the data itself. In order to retrieve the referenced data, the reader has to trust that the referenced DOI redirects to some web location that contains instructions on how to retrieve and verify the referenced dataset (e.g., the data available through https://zenodo.org/record/3523403/files/elton-datasets.tar.gz). A DOI and associated URL redirects are temporary leases on virtual locations in an ever changing and increasingly fragmented, politicized and commercialized internet. So, even though DOIs are widely adopted, using a DOI in a reference is like directing someone to find a book by pointing to a specific shelf at a specific library. This location-based reference offers no way to find copies or verify the authenticity of the book.

In order to introduce reliable machine readable data references to promote re-usable research products, the following modifications to existing research practices and infrastructures are needed:

1. Each data citation should reference one, and only one, data file.

2. Each data citation should include a reasonably secure content-based identifier (e.g., Secure Hash Algorithm 2, https://en.wikipedia.org/wiki/SHA-2) in addition to traditional citation elements to allow for unique identification and verification of the referenced data. For example:

> Poelen, Jorrit H. 2019. Global Biotic Interactions: Elton Dataset Cache (Version 0.3, hash://sha256/2d91e00654464c4d0a6df39d0a1b26ffb916659322ae2ba7fcc7264a820040cc) [Dataset] . Zenodo. http://doi.org/10.5281/zenodo.3523403

3. Establish Data Indexes that associate a content-based identifier with one or more locations (e.g., URLs) at which the data was available at some point. These data indexes may be accessed online via web services or offline in data tables. They may be co-located with item 4 (below) or added as a feature to existing data repositories like Zenodo, FigShare or Data Dryad.

4. Establish Data Observatories that find existing datasets, record their provenance, register both with one or more Data Indexes after storing them using a (pre-existing) data repository.

5. Develop and use intuitive on-/offline data management and citation tools, a cross between "git" and reference managers, to facilitate data (citation) management.

6. Provide funding to research solutions for data citation in research

Note that item 1 (above) still allows for referencing a virtually unlimited amount of datasets with a single citation. This can be accomplished using reference nesting: citing a data file that contains one or more (reliable) references.

A frugal, proof-of-concept tool "Preston" (https://preston.guoda.bio) and associated data publications have shown that adopting modifications 1-5 are feasible at scale by re-using existing infrastructures like the Internet Archive and Zenodo in combination with commonly-used open technologies like nginx, rsync and linux.

Other considerations

The distributed nature of data requires that we increase management of versioning and identifiers in order to practice reproducible science. Much of this will require an increase in technical staff at data repositories and education of practicing researchers regarding identifier hygiene [3] and 21st Century data culture. Some of these issues are technical, but most are social.

A technical issue is that managing and referencing data for individual researchers and their labs is still challenging: existing online tools like Google Docs and Dropbox are not built to keep track of scientific datasets and lead to inconsistent collections of data files. Just like citation managers (e.g., BibTex in combination with Zotero/Mendeley) make it easy to reuse references across manuscripts, data management tools can help to find and cite datasets.

A social challenge is to change attitudes about data ownership. As long as scientists using non-open access data sources have an advantage in obtaining funding opportunities, data will continue to be restricted. One suggestion is that in addition to requiring NSF awards to make their data public, there is also a requirement that all data used in the proposal must be public. In this way, individual labs will

have to release data instead of keeping it to themselves for future research. Thereby removing that disincentive to share.

Sustainability will, of course, be an issue. We understand that NSF cannot commit to funding this infrastructure forever, so development of these services and tools should be designed using the Unix philosophy, provide a benefit to the researcher, consider customers outside academia and implement business models (e.g., subscription, freemium product/services, open source consulting) from the very beginning to support long term (>10-50 years) availability of research data.

Sincerely,

Jorrit H. Poelen, Independent, Oakland, CA

Anne Thessen, Oregon State University

Jennifer Hammock, Smithsonian Institution

Katja Seltmann, University of California, Santa Barbara

Melissa Haendel, Oregon State University

Ramona Walls, Bio5 Institute, University of Arizona

Carl Boettiger, University of California, Berkeley

Tobias Kuhn, VU University Amsterdam, Netherlands

Michael J. Elliott, University of Florida

References

[1] Brase, J., 2009. DataCite - A Global Registration Agency for Research Data. 2009 Fourth International Conference on Cooperation and Promotion of Information Resources in Science and Technology. Available at: http://dx.doi.org/10.1109/coinfo.2009.66.

[2] Michener, W.K. et al., 2012. Participatory design of DataONE—Enabling cyberinfrastructure for the biological and environmental sciences. Ecological Informatics, 11, pp.5–15. Available at: http://dx.doi.org/10.1016/j.ecoinf.2011.08.007.

[3] McMurry, J.A. et al., 2017. Identifiers for the 21st century: How to design, provision, and reuse persistent identifiers to maximize utility and impact of life science data. PLOS Biology, 15(6), p.e2001414. Available at: http://dx.doi.org/10.1371/journal.pbio.2001414.