ARTICLE
http://dx.doi.org/10.18243/eon/2016.9.11.7

Open Data: Repositories and Policies

By Meredith Morovati (ORCID 0000-0002-5383-7643)
Executive Director
Dryad
director@datadryad.org

I am excited to be able to provide some background about data publication and why I think this topic deserves your attention. I have now been the Executive Director of Dryad since the fall of 2014. I have a background in association and board management, but I also have a previous history working in the publishing field, mainly with academic publishers in journal departments. When I was in publishing, Open Access (OA) was not yet a business model. It was not until shortly after leaving the industry for a time, that OA journals started to appear. Fast-forward to 2016 and the organization for OA publishers recently celebrated its 100th member and the spate of OA journals continues to grow rapidly with mega-journals such as PLOS, *Scientific Reports*, and *PeerJ*. What a difference a decade makes.

At some point, OA moved from "new trend" to real business model. I'm not sure if this graduation was due to a specific moment in time or it was the result of a slow realization that what was new and noteworthy had over time become accepted by most. Open data has not *quite* had this moment except that in some cases and in some fields, it absolutely has.

Open data is making data that underlies scientific research openly to anyone who wishes to access it. Data are at the core of science and if we are to be able to reproduce research, then it goes without saying that we need to access that research. Moreover, the open data movement is the desire to fix something we have known for a long time: that informally sharing data between peers doesn't work. If you saw David Crotty's recent post in The Scholarly Kitchen, you may have gotten a kick out of the cartoon that sadly reflects common experience when a researcher wants to access data.

Data Policies

Another way of looking at this is problem is with a more serious lens and provides a stark contrast of challenges surrounding open data versus OA publishing. If a scholarly article is not open, you may still access the information in a library or by paying. But if data is not formally archived in a repository, it could be lost forever. Think of that. This is a very serious issue for science. If data are not archived and linked back to the article, then 17% of all data that science is based upon will be lost every year. It was this very realization and concern that gave the founders of the Joint Data Archiving Policy (JDAP) their passion. This simple clear-cut policy stated:

> *[Journal] requires, as a condition for publication, that data supporting the results in the paper should be archived in an appropriate public archive, such as [list of approved archives here]. Data are important products of the scientific enterprise, and they should be preserved and usable for decades in the future. Authors may elect to have the data publicly available at time of publication, or, if the technology of the archive allows, may opt to embargo access to the data for a period up to a year after publication. Exceptions may be granted at the discretion of the editor, especially for sensitive information such as human subject data or the location of endangered species.*

JDAP was adopted in a joint and coordinated fashion by a group of editors in the evolution and life sciences fields in 2011. The result was that no one was penalized unnecessarily—the journals had a consistent policy and so authors were encouraged to comply. In 2011 there were some specialized and community repositories in place to fill some of the needs. Dryad was created to serve the data that fell outside the scope of those existing repositories and to continue to support the

community. Since then, Dryad has been focused on data that links to and supports scholarly literature, and Dryad is committed to curating the data and providing it openly. Now, just five years after this policy, open data is a point of pride for many in the field of evolution and life sciences and beyond. What a difference a lustrum makes.

Now with requirements from funders for open data, policies on how to handle data are becoming common at various journals and large publishers. Policies are being spread through large publishers like Springer Nature and through Wiley's data sharing service. Elsevier even went as far as to acquire start up repository, Mendeley data, as its preferred repository to safeguard the data associated with Elsevier articles. The result? If you don't have some kind of data policy, you are lagging behind the times.

So what is in a data policy and how do you go about making one? Well, there are many answers to this and some are starting to point a light on the lack of standardization in policies across journals and publications. I suspect that standard policies such as those from Springer Nature will become the norm eventually. But, some fields will always have unique needs. For instance, archeology and museum studies have many samples that might be destroyed during the research. Therefore, methods and notes might be the most important to preserve.

There are some very basic building blocks to a policy. For instance, many instruct researchers to archive at the time of manuscript submission. Essentially this requires an author to indicate where the data are when submitting a paper. What data should be preserved can usually be described as that data which support the arguments in the paper. This is not all data collected. And, generally, open data is easier to provide than closed. Dryad publishes nearly all our data under a CC0 license and there are many reasons for this. But, the main reason is that CC0 does not actually affect the legal status of the data, since facts in and of themselves are not eligible for copyright in most countries. And, publishing data under CC0 does not relieve you of the requirement for citing data.

Citing Data

More and more editors are focusing on practices of how to cite data. This is an essential area that publishers and editors can get on board immediately. The Joint Declaration of Data Citation Principles have been widely endorsed and work is ongoing to provide more instructions on how to adhere to these principles. But, the main tenets of these principles are:

- importance;
- credit and attribution;
- evidence;
- unique identification;
- access;
- persistence; and
- specificity and verifiability.

Under these principles, data are considered important scholarship in their own right that deserve credit and uphold the arguments in a paper. Data need to have persistent and unique identifiers that will outlive the lifespan of any single author or article and should include metadata that are both human understandable and machine readable.

In fact, Crossref recommends that the data are to be cited in the reference section of the article itself. This is commonly referred to as self-citation. And, other citation placement can look like these examples. This is very similar to the citations that editors already handle daily that point to other articles or sources. But, care will have to be taken to make sure that you don't strip out what looks like an orphan citation.

The adoption of open data has from similar desires surrounding OA. And, while some of it may seem technical and complicated, starting with a basic policy and stringent checks on citations is an appropriate place for editors and publishers to start.