Against the Grain

Volume 28 | Issue 1 Article 11

2016

How and Why Data Repositories are Changing Academia

Phill Jones
Digital Science, p.jones@digital-science.com

Mark Hahnel
Figshare, mark@figshare.com

Recommended Citation
Jones, Phill and Hahnel, Mark (2018) "How and Why Data Repositories are Changing Academia," *Against the Grain*: Vol. 28: Iss. 1, Article 11.
DOI: https://doi.org/10.7771/2380-176X.7269

How and Why Data Repositories are Changing Academia

by **Phill Jones** (Head of Publisher Outreach, Digital Science) <p.jones@digital-science.com>

and **Mark Hahnel** (Founder, Figshare) <mark@figshare.com>

Academic and scholarly communication is unquestionably in the process of undergoing a revolution. It seems, however, that the nature of that revolution is still a somewhat open question. Libraries in particular are undergoing not so much a shift in focus but a diversification of roles. Where the library once consisted primarily of a physical building containing curated collections of books, journals and other resources, it is now a diverse set of services ranging from research assessment to technology support to the new frontier of data curation and dissemination.

Why Should Librarians Care about Data Sharing?

The role of the library as manager of collections of information for the use of patrons is still alive and well. Increasingly, however, libraries have been concerned with recording and curating the output of their institutions. This expansion of role has on some level been driven by a shift in the way that scholars are communicating their work and accounting for its value. Arguably, this trend began around 15 years ago with the rise of open access publishing, which itself was made possible by the shift to more scalable electronic journals. Many libraries at the time took an interest in the new publishing model by either setting up central funds for the payment of article processing charges or supporting and educating scholars in how and why to publish open access. Later, institutional repositories provided avenues for green open access and library publishing operations began to develop during the first decade of the 2000s, culminating in the creation of the **Library Publishing Coalition** in 2012. Many library publishing operations, in contrast with traditional university presses, aim to support niche areas of scholarship of interest to their own faculty. However, early suggestions that institutional open access paper repositories may replace the role of traditional publishers have proven to be a bridge too far. One can postulate many reasons for this, but publisher brands and the need to publish in high impact factor journals seem the most likely. This is not the case for the emerging requirements of data dissemination. There are as yet no impact factors or prestige publication outputs. This means that libraries may have another opportunity to play a key role in communicating the academic content that comes out of their institutions.

As the open science movement has grown in momentum over the past decade and a half, scholars have sought new outlets for new types of scientific output. The blogosphere has been used to "publish" work almost in real time, resulting in some noteworthy cases. For instance, **Rosie Redfield** of the **University of British Columbia** documented her attempts to replicate **NASA's** claims of discovering arsenic based life on her blog ahead of publishing them in **AAAS** *Science*, which debunked the claim. However, this sort of blogging/publishing generally acts as a more rapid media for hypothesis driven scientific narratives, similar in concept to traditional articles, rather than a way to make data sets available.

For many people interested in data publishing, what's required is a new infrastructure for communicating data and other research outputs that is separate from hypothesis driven narratives and judged on its own terms. The features of this infrastructure are not entirely clear but we do know that it must be able to cope with large quantities of data. Some data will be in well-codified and well-documented formats, but much of it won't be. Data needs to be discoverable and at least somewhat interpretable, so that it is available for re-use and re-analysis when needed. Finally, there's a need to protect a researcher's ability to fully analyse their own data first through embargos and also to protect commercially or medically sensitive information.

Taking all this together, data publishing seems to be a fairly complicated issue, but one that the library is well-placed to tackle.

Why Researchers Care

There are a number of potential advantages to scholars of sharing their data. Probably the most compelling reason is the apparent citation advantage.[1] Other reasons include requirements from funders, journals and institutions, as well as a personal desire to make science more open.

Many researchers believe that open data is necessary to make scholarship more effective. The academic system does work, but it can be an inefficient machine. The majority of inefficiencies lie in the inability for academics to directly build on the research that has gone before them — to better stand on the shoulders of giants. Increased transparency can also improve academia's ability to self-correct through openness to scrutiny and challenge.

Making data sharable and open has the added benefit of encouraging standards and codification — a vital step to making data machine readable. The power of computers means that data can be interrogated and cross referenced in order to automatically look for correlations between research outputs. Of course, today's artificial intelligence won't enable computers to generate and confirm hypotheses the way a person can, hence the need for academics with subject specific knowledge to build research programs based on machine suggested relationships. Immediately, this provides many more promising avenues to explore across all fields of research in a practice that pharmaceutical companies have been exploiting with computational chemistry for decades.

Barriers to Sharing

The reasons why many researchers choose not to share their data, or share it only upon request through closed systems like email, is less well explored than the benefits mentioned above. Last year, a survey of **Wiley** authors, which was reported on in the *Scholarly Kitchen* by **Alice Meadows**, found that just less than half of researchers choose not to share data.[2] **Wiley** produced a survey infographic, which is linked from the *Scholarly Kitchen* article, which contains a long list of reasons as to why some researchers are reluctant to share. Broadly, there seems to be three overarching themes. The first issue is a fear that sharing data would have negative consequences either because another researcher appropriates data and scoops the original experimenter, or their work gets picked apart and unfairly discredited. The appropriate use of embargoes should mitigate many of those concerns. The second issue is lack of researcher understanding of how to share data. Answers like "My funder/institution does not require data sharing," or "I don't think it was my responsibility" aren't evidence of a positive decision not to share, rather that some researchers are still not yet seriously considering it. It's easy to see how librarians and information professionals can help with that one. Finally, many of the responses speak to a lack of time and resources. This last issue is perhaps the toughest to tackle, so let's look at it in more depth.

Researchers are often juggling many disparate and seemingly unconnected responsibilities, from research to managing their labs and getting grants, to teaching, to university administrative tasks and committees. With such a diverse workload, with so many responsibilities to juggle, it can be challenging to incorporate new workflows. For this reason, simplicity and intuitive workflows are increasingly important. You only have to look at the rising pressure that publishers are under to simplify their submission systems and eliminate author burden, or at the success of simplified search like **Google** to see that researchers often value simplicity and intuitiveness over comprehensive functionality. Against that background, it's not surprising that many researchers are choosing to share data using supplementary materials services offered by publishers despite the fact that in many cases those systems were not designed with data sharing in mind.[3] If data sharing is to become the norm, it will be important to create systems that are not only robust and scalable, but also very simple and time effective to use.

continued on page 24

How and Why Data Repositories ...
from page 22

Data as a First Class Research Object

The idea that datasets should be treated as an equal output to academic articles is a controversial one, but one that funders and advisory committees are beginning to support. Most notably, the **Royal Society's** "Science as an Open Enterprise: Open Data for Open Science" report in 2012[4] suggested that: "Assessment of university research should reward the development of open data on the same scale as journal articles and other publications." This has led to many funders requiring that all data from the research they fund be made openly available.[5] An obvious corollary being that the rewards for open data would need to be comparable with those for traditional articles.

Before we address whether data should have such a status, there's a more fundamental but less obvious question to answer. Just what exactly are data? There are several definitions, but the general theme across disciplines is that data are the digital products of academic research. This can range from digitized field notes in biology to videos of dramatic performances to niche file formats in computational chemistry. The ubiquity of digital scholarship means that any platform for disseminating research should work across the full range of disciplines, with filters applied so that content can be grouped arbitrarily. That is to say, we need persistent file storage, which is discoverable and interpretable by machines and humans alike.

A long-standing problem in academia is that technology has traditionally limited us to one research output type with limited forms of assessment, namely peer review and citation metrics like Impact factor. We are now at a point where all products of research can be released (unless prevented by ethical or commercial reasons). The number of evaluation metrics has exploded to include altmetrics as supplements to citations, as well as open post publication peer review. However, when we look at data, that is, any digital output of research, we have to ask if we can apply the same criteria to a video, as we do to spreadsheet data and how should those criteria differ from the existing criteria for paper publications? Most likely, we will need to define both review and assessment criteria for each type of output. These may be difficult to define and challenging to scale.

There have been suggestions that peer review is only really of use for data when it is to be reused. There have been examples of serious problems being discovered when researchers have tried to reanalyse data. For instance, in the case of **LaCour** whose fraudulent data was exposed in 2015.[6] However, by the time the fraud came to light, the research had been published in *Science* and covered by the mainstream media so the critical review arguably happened too late.

One interesting development in this space has been the idea of machine readable badges (*http://openresearchbadges.org/*). These are essentially automated or manual markup of content to better describe and accredit research outputs.

Scholarly Publishers and Data

Over the past decade, some traditional publishers have worked with repositories to link raw digitised objects that underlie research to the hypothesis-driven narrative of the article. The goal is to standardize the approach to linking research data to publications, irrespective of the repository, which hosts the data.

Early succesful repositories, such as the **Protein Databank** (*http://www.rcsb.org/pdb/*) and **Genbank** (*http://www.ncbi.nlm.nih.gov/genbank/*) archive molecules and genetic sequences to help reproduce research in the life sciences. Later, generic repositories came to the forefront through projects like **Dryad** (*http://datadryad.org/*), which helped motivate ecologists to make all of their *one-moment-in-time* series data available.

When funders started requiring that data be made available at the point of article publication, academic publishers took steps to help researchers comply with these requirements. Partnerships with repositories such as **Figshare** (*www.figshare.com*) allow journals to preview the digital files embedded within the HTML version of the article. The long-term preservation of the data is contractually maintained and each object is individually citable. Later, some publishers developed *data journals*, like the **Geoscience** *Data Journal* (*http://onlinelibrary.wiley.com/journal/10.1002/(ISSN)2049-6060*) published by the **Royal Meterolical Society**, that allows researchers to publish short descriptive articles, that aren't hypothesis driven, linked to data archived in approved repositories.

In 2014, **Nature Publishing Group** launched *Scientific Data*, which applies traditional peer review to data descriptor articles: "Acceptance for publication is based on the technical rigour of the procedures used to generate the data, the reuse value of the data, and the completeness of the data description."

There are movements to codify standards for data sharing outside of publishers, particularly in the sciences. A good example of this is the **Open Microscopy Environment** project (**OME**, *www.http://www.openmicroscopy.org/*). **OME** develops both standards in microscopy and open source imaging software. Organizations like **Research Data Alliance**, **CODATA**, the **Data FAIRport** initiative and **FORCE11** are working towards standards for data storage, markup and dissemination. The work being carried out by **DataCite** and **ORCID** is of particular interest.[7] This will enable research repositories to automatically update a researcher's **ORCID** profile. This collaboration extends to **CrossRef** so that all academics should be able to sync their publications as well as their data with no extra effort.

Subject Specific and Structured Repositories

Certain disciplines lend themselves more easily to data sharing, such as astronomy, and the *-omics* disciplines. Structured repositories require data to comply with format standards thereby encouraging their adoption. They play a key role in data science as community or funder-driven focal points for collaborative and industrial scale efforts to assemble super-datasets like **Zooniverse's Galaxy Zoo** (*http://data.galaxyzoo.org/*) and the **NIH's GenBank**.

There are a number of libraries and other groups that maintain lists of these types of databases, perhaps most notable are the **Registry of Research Databases** (*www.re3data.org*), which was started in 2012 and is funded by the **German Research Foundation (DFG)** and **Biosharing** (*www.biosharing.org*), which is hosted by **Oxford University**. Encouraging patron participation in these repositories where appropriate is just one way that librarians can assist the open data movement.

Institutional Data Repositories

Institutional data repositories have been historically designed with a view to managing and curating the output of institutions. In that sense, they are intertwined with both research assessment and library publishing efforts; at some institutions, library publishing and data repository services are provided using the same platform.[8] As data dissemination becomes increasingly important, it makes sense to look at some of the work that pioneering library publishing efforts have made in populating and popularizing their repositories.

In her 2001 article *Institutional Repositories: Keys to Success*,[9] **Joan Giesecke**, then Dean of Libraries at the **University of Nebraska-Lincoln**, outlined how they successfully transformed their repository from what she calls a *collection centric viewpoint* which assumes faculty participation and focuses on curation, to one of service provision which focuses on making the repository an attractive place to put content. **Giesecke** notes the danger that institutional repositories can become overly restrictive, focusing too much on the desire to create an orderly collection, thereby unintentionally creating barriers to participation. By adopting the service driven approach of a university press, with a focus on discoverability, dissemination, search engine optimization and improved user experience, **University of Nebraska-Lincoln** were able to grow their traffic from zero to 300,000 uses per month in under five years.

Unstructured or General Repositories

With the growth in popularity of data sharing among academics and the increase in funder mandates, it's clear that all researchers are going to need data sharing solutions. Subject specific and institutional repositories form an overlapping and occasionally incomplete patchwork of coverage for authors looking to place content, particularly data that doesn't fit into the predefined data formats that structured repositories support.

There has been very little research into the volume of data produced by academics. The true scale and nature of research data is unknown as much of it sits on institutional and departmental servers or on the hard drives of computers under researchers' desks. Anecdotally, researchers generally have large personal collections of data in a diverse range of formats.

continued on page 25

How and Why Data Repositories ...
from page 24

As part of **Figshare's** partnership with **Nature Publishing Group** and their journal *Scientific Data*, we've been able to analyze user behaviour and preferences. *Scientific Data* ask researchers to place data in structured data repositories, institutional repositories or both when suitable ones exists. Tellingly, over 30% of data submissions were made to **Figshare**, making it the most used repository. We know from this that the majority of researchers require an unstructured repository for their data. The extent to which this will change over time as codification and structuring efforts proceed is arguable. It is our opinion that there will always be a strong need for unstructured repositories because it is the nature of research that many experiments and techniques are novel and unique.

Where Does this Leave Us?

It has taken longer than expected for the promise of the digital age to begin to make a real difference to the way scholars communicate their work. The persistence of traditional measures of quality are the most likely explanation for academia's apparent conservatism, but with funding bodies increasingly encouraging and mandating the sharing of data, we are finally seeing diversification of what is considered legitimate scholarship.

The publishing industry has made strides over the last decade or so to integrate with institutional, funder and community based repositories. Together with groups interested in the standardization of data formats, a lot of progress has been made to codify formats in many fields. There remains, however a large quantity of data on researchers' hard drives and servers that don't fit into easily standardized formats because the techniques are either new or unique.

There are still many open questions in data publishing, from how to deal with embargoes or sensitive data to how best to assess the quality of the diverse range of digital research outputs. The field of data publishing is still in its formative stages and represents an opportunity for both publishers and libraries to help academics adapt to new requirements.

Endnotes

1. **Piwowar H. A.**, **Day R. S.**, **Fridsma D. B.** (2007) Sharing Detailed Research Data Is Associated with Increased Citation Rate. *PLoS ONE* 2(3): e308. doi:10.1371/journal.pone.0000308

2. **Meadows A.** (2011) *Scholarly Kitchen. http://scholarlykitchen.sspnet.org/2014/11/11/to-share-or-not-to-share-that-is-the-research-data-question/*

3. **Schaffer, T.** and **Jackson, K. M.** 2004. The use of online supplementary material in high-impact scientific journals. *Science & Technology Libraries* 25(1/2):73-85.

4. The **Royal Society Science Policy Centre** report: *Science* as an open enterprise (2012) *https://royalsociety.org/~/media/Royal_Society_Content/policy/projects/sape/2012-06-20-SAOE.pdf.*

5. **Valen D.** and **Blanchat K.** (2015) **Figshare** Blog *https://figshare.com/articles/Overview_of_OSTP_Responses/1367165*.

6. **Stemwedel J. D.** (2015) *Forbes http://www.forbes.com/sites/janetstemwedel/2015/06/01/reasons-to-keep-discussing-the-lacour-and-green-retraction/.*

7. **Laure Haak L.** (2015) **ORCID** blog *https://orcid.org/blog/2015/10/26/auto-update-has-arrived-orcid-records-move-next-level.*

8. **Jones P. B.**, **Wesolek A.**, **Scherer D.**, **Watkinson A.** (2015). A Game of Spot the Difference: Librarians, Repository Managers and Publishers. Presentation slides. *http://works.bepress.com/andrew_wesolek/30/*

9. **Giesecke J.** (2011), "Institutional Repositories: Keys to Success" Faculty Publications, **UNL** Libraries. Paper 255. *http://digitalcommons.unl.edu/libraryscience/255*

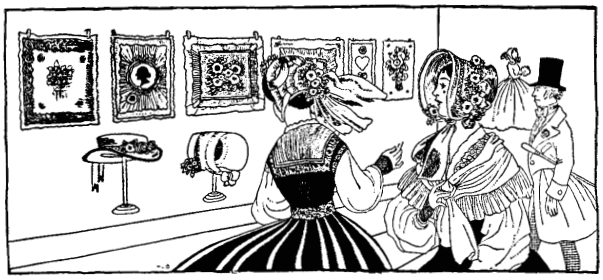
Everything Evolves, Even Publishing

by **Jason Hoyt, PhD.** (CEO and Co-founder, PeerJ) <jason@peerj.com>

and **Peter Binfield, PhD.** (Publisher and co-founder, PeerJ) <pete@peerj.com>

We sometimes hear that for all the promise of the Internet, it is a shame that it has yet to impact scholarly communication in the same way it has other industries. One could argue this point quite effectively: prestige still dominates; the journal name matters just as much as it always has; the same legacy publishers still control most of the literature; Open Access is just a small fraction of all articles, etc., etc. Meanwhile, in other industries it is easy to spot how the old guards have changed and new names have sprung up: **Google**, **Wikipedia**, **Amazon**, **Uber** and **Facebook** to name just a few.

On the other hand, does anyone believe Open Access is going away? Will data not become more widely available? Will tools to make publishing faster never be developed? Why have "megajournals" appeared in the past ten years and not just survived, but become the future revenue model for new and old publishers? Why are scholarly societies struggling after decades/centuries of thriving? Why are governments and funders making Open Access mandates? These events contradict the notion that the Internet hasn't changed things in an "unmovable" 300 year-old industry. Indeed, the evidence actually suggests that we are in the midst of a change so expansive that we don't quite know how to adapt to it.



We take comfort in the way things worked in the past, as they had slowly developed in manageable timetables over the 20th century. There was certainty in how to communicate science, who to trust, or what to do for academic career progression. We now live in an era with an alluring future, but one that raises new concerns:

How will we fund scholarly output? How much should we make open, and how? Is publishing Open Access a bet on the future, or will it negatively affect my students or my career?

What the last ten years or so have done is to open our minds to questions that many of us never anticipated having to find solutions for. It could be argued that just as the Internet has made us more globally aware, so academia has grown more concerned with its impacts outside of the ivory tower. The decentralization that occurred with the World Wide Web makes it clear how we affect those around us, and this has influenced our professional lives in a similar way. It's not that scientists are only just now waking up to the fact that they can be open, they just didn't realize it was possible until recently. Our policies and infrastructures are unprepared for these changes, just as much as our readiness to leave the comfort of the past.

There Would be no Open or Mega-Journals without the Internet

Just as the printed journal was a forgone conclusion of the printing press, so too was Open Access and the megajournal a natural by-product of the Internet. Perhaps someone

continued on page 26