Meeting Institutional Goals by Working Beyond Institutional Walls

John Chodacki

Daniella Lowenberg

Follow this and additional works at: <https://docs.lib.purdue.edu/atg>

Part of the [Library and Information Science Commons](#)

This document has been made available through Purdue e-Pubs, a service of the Purdue University Libraries.
Please contact epubs@purdue.edu for additional information.

The Opportunities and Challenges ... from page 11

The existing relationships liaison librarians have with researchers is critical for making researchers aware of library data services and providing support for discipline-specific repositories, journal requirements, and meta-data creation. The CaltechDATA repository is based on Invenio 3, which is an open-source repository system first developed at CERN. TIND, which provides commercial support for Invenio-based repositories, runs the hosted Invenio instance for **Caltech**. For libraries that do not want to run their own repository, they can aid researchers submitting data or software to discipline-specific or available general repositories such as Zenodo, Harvard Dataverse, or Dryad. Many of these repositories provide APIs that can be used to automate submissions and access data for reuse.

Libraries have a unique opportunity to provide solutions for the data and software preservation challenges that plague the scientific community. Maintaining the record of scientific knowledge, which now includes data and software, requires institutional backing to succeed. By developing simple repository services that are compliant with the FAIR principles, partnering with disciplinary repositories to act as storage agents, and working to meet the needs of researchers, libraries can ensure that research data and software remains open and available for years to come.

Bibliography

Andrew, R. L., Albert, A. Y. K., Renaut, S., Rennison, D. J., Bock, D. G., and Vines, T. (2015). Assessing the reproducibility of discriminant function analyses. *PeerJ*. <https://doi.org/10.7717/peerj.1137>

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Sayers, E. W. (2013). GenBank. *Nucleic Acids Research*, 41(Database issue), D36–42. <https://doi.org/10.1093/nar/gks1195>

Berman, H. M., Kleywegt, G. J., Nakamura, H., and Markley, J. L. (2014). The Protein Data Bank archive as an open data resource. *J. Comput. Aided. Mol. Des.*, 1009–1014. <https://doi.org/10.1007/s10822-014-9770-y>

COPDESS. (2019). Commitment Statement in the Earth, Space, and Environmental Sciences. Retrieved July 2, 2019, from <http://www.copdess.org/enabling-fair-data-project/commitment-to-enabling-fair-data-in-the-earth-space-and-environmental-sciences/>

Cousijn, H., Kenall, A., Ganley, E., Harrison, M., Kernohan, D., Lemberger, T., ... Clark, T. (2018). A data citation roadmap for scientific publishers. *Scientific Data*, 5, 180259–180259. <https://doi.org/10.1038/sdata.2018.259>

Hudson-Vitale, C., Imker, H., Johnston, L. R., Carlson, J., Kozlowski, W., Olen-dorf, R., and Stewart, C. (2017). SPEC Kit 354 Data Curation. Association of Research Libraries.

Lee, R. Y. N., Howe, K. L., Harris, T. W., Arnaboldi, V., Cain, S., Chan, J., ... Sternberg, P. W. (2018). WormBase 2017: Molting into a new stage. *Nucleic Acids Research*, 46(D1), D869–D874. <https://doi.org/10.1093/nar/gkx998>

McQuilton, P., Gonzalez-Beltran, A., Rocca-Serra, P., Thurston, M., Lister, A., Maguire, E., and Sansone, S.-A. (2016). BioSharing: Curated and crowd-sourced meta-data standards, databases and data policies in the life sciences. *Database*, 2016. <https://doi.org/10.1093/database/baw075>

Morrell, T. E. (2019). caltechlibrary/caltechdata_usage: Jupyter notebook with visualization of submissions to CaltechDATA (Version v0.0.2). <https://doi.org/10.22002/d1.1250>

Pampel, H., Vierkant, P., Scholze, F., Bertelmann, R., Kindling, M., Klump, J., ... Dierolf, U. (2013). Making research data repositories visible: The re3data.org registry. *PLoS One*, 8(11). <https://doi.org/10.1371/journal.pone.0078080>

Romaine, S. B., Barbara Albee, and Sion. (2019, April 4). Deal or No Deal | Periodicals Price Survey 2019. *Library Journal*. Retrieved from <https://www.libraryjournal.com/?detailStory=Deal-or-No-Deal-Periodicals-Price-Survey-2019>

Stall, S., Yarmey, L., Cutcher-Gershenfeld, J., Hanson, B., Lehnert, K., Nosek, B., ... Wyborn, L. (2019). Make scientific data FAIR. *Nature*, 570(7759), 27. <https://doi.org/10.1038/d41586-019-01720-7>

Van Horn, J. D., and Gazzaniga, M. S. (2013). Why share data? Lessons learned from the fMRIDC. *NeuroImage*, 82, 677–682. <https://doi.org/10.1016/j.neuroimage.2012.11.010>

Vines, T. H., Albert, A. Y. K., Andrew, R. L., Débarre, F., Bock, D. G., Franklin, M. T., ... Rennison, D. J. (2014). The availability of research data declines rapidly with article age. *Curr. Biol.*, 24(1), 94–97. <https://doi.org/10.1016/j.cub.2013.11.014>

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data*, 3, 160018. <https://doi.org/10.1038/sdata.2016.18>

Meeting Institutional Goals by Working Beyond Institutional Walls

by **John Chodacki** (California Digital Library, University of California Office of the President) <John.Chaodacki@ucop.edu>

and **Daniella Lowenberg** (California Digital Library, University of California Office of the President) <Daniella.Lowenberg@ucop.edu>

Researchers from across the **University of California (UC)** publish more than 50,000 articles annually. Underlying most of these articles are datasets, many of which have not been published. Even if these datasets were published, the UC system (like any university) does not have the ability to track or index them. While the scale of the UC's research outputs may not be typical of other universities, our story and approach to tackling these issues have been similar to those of other colleges and universities.

In 2014, in an effort to address this problem, **California Digital Library (CDL)** set out to develop an easy submission system on top of our digital preservation repository. That system was called Dash. After receiving a **Sloan Foundation** grant to reimagine Dash as an open source, easy way to publish data, we worked to create an easy and

user-friendly interface for UC researchers to publish and preserve their data at UC. The goal was to get as many datasets (suitable for a general repository) as possible. To attain this goal, our team spent years doing mass outreach to UC researchers, building out new features requested by these researchers, and trying to convince publishers and research workflow systems to integrate with Dash.

The result? Five hundred deposits over three years.

We spent a significant amount of time with researchers to make sure our decisions kept researchers in mind. Despite adopting this researcher-centric approach, we quickly recognized that the project presented several hurdles to executing and building what researchers genuinely value (Narayan & Luca, 2017). So, we realized we had to

continued on page 33

adjust our approach. As we looked inward and evaluated our journey, we arrived at the following three insights, all of which can be applied to research data management (RDM) programs at other universities of varying types and sizes:

1. Researchers are not institutionally focused

Researchers may be influenced by institutional policies and mandates, but the vast majority of them are not aware of or advocates of institutional options. Research is a global enterprise, and the work is focused around disciplines. Ecology researchers are influenced by where other ecologists are publishing and depositing their data. As such, researchers have vocalized their need for community (i.e., discipline-based or cross-disciplinary) solutions (Cragin et al., 2010). To have an institutional data publishing option that collaborators (at other institutions, globally) have not heard of or would want to publish data in is a very real obstacle. Further, to convince researchers to use their institutional offering when they themselves or their collaborators have had success publishing their data in an institution-agnostic, general purpose, or discipline-specific repository is a moot debate. While institutions pour resources into local projects, the value of such resources remains murky to researchers and the adoption rates remain extremely low.

2. Seamless integrations into researcher workflows are not happening

Open Access has gained traction and has become the status quo for many researchers in the last decade, but open data publishing still has a way to go. And the way to drive this adoption is by making it seamless for researchers. While operating and iterating on Dash, we approached many organizations asking them if they could utilize our open APIs to bring in integrations where researchers could publish their data from various workflows. The technology is there and publishers and tools like online lab notebooks understand this need. But they do not invest in this development. Why? Because our project was for a single institution. If this didn't happen for an institution the size of **University of California**, it will continue to not happen for the thousands of others that would love to tackle the same issue.

3. Lack of name recognition

Institution-based resources do not have brand recognition. While libraries may not be interested in competing in a popularity contest with researchers, the reality on the ground is that general repositories (both commercial and non-commercial) have gained adoption because they are suggested and promoted by colleagues, publishers and funders (not to mention marketing teams at for profits). As a result, many researchers do not feel they have ownership of institution-based resources. Instead they frequently feel more loyalty to community tools, or even commercial products, that have ambassador and other programs. As trite as it sounds, we are adding an additional hurdle to the success of our research data publishing initiatives by promoting institutionally focused tools that lack a clear and recognizable brand name.

Looking Outward for Community Success Stories

By looking inward and evaluating our projects, we were able to fine-tune our success metrics and pinpoint the hurdles. We quickly realized we could not go at it alone and embarked on evaluating options in the community. We wanted to find organizations that were best positioned to help us overcome the hurdles and could also be aligned with our values of openness and responsible stewardship. One project that clearly covers these bases is Dryad.

Dryad (datadryad.org) is a data publishing repository that was launched by researchers in 2009. Since then, Dryad has not only published but also curated (in accordance with FAIR principles) 28,000 datasets in hundreds of disciplines, from over 900 global journals. Dryad has published more data from **University of California** each year than Dash could have ever reached. The reasons? Dryad is researcher owned and recognized, integrated into publisher workflows, and endorsed by funder and publisher policies.

In addition, Dryad shares our own institutional values of proper curation, compliance, access, and preservation of research data. So, in May 2018, we partnered with Dryad to drive adoption of research data publishing and tackle the barriers we had faced with Dash:

• Curation of Published Research Data

Dryad already understands the unique challenges of data curation. They have a team of expert curators who go through every submission, verifying that the research data going to Dryad are in fact usable so that Dryad does not become a dumping ground for various research outputs. Institutions may vary in their capacity for or prioritization of data curation processes (i.e., the ten UC campuses exhibit such variances). But with Dryad we can satisfy the campuses that would like all curation to be handled externally, and we also can engage with the data curation programs at campuses that would like to get involved in the process. By thinking of Dryad as casting a wide net and catching data publications, bringing them into the institutional resources, we can eliminate the resources spent trying to convince researchers to put their data in an institution-based service and rather meet them where they are.

• Connections to the Larger Ecosystem

Dryad took off not only because it was promoted by research communities, but also because funders and publishers have trusted and promoted Dryad to their researchers who are focused on meeting these mandates. Additionally, revisiting conversations with publishers and tools providers about upgrading integrations to seamless API interactions between their platforms and Dryad has been easier because they see the value in integrating with a general, global option for researchers.

• Brand Recognition & Mass Adoption

As mentioned above, Dryad has been adopted by various research communities and is a known entity in many fields. The difference between a publisher or funder saying “use your institutional offering” versus “use www.datadryad.org” is incomparable. Adoption has not been a “crisis” for Dryad as it has been for every institutional repository intending to publish data like we do for articles. We will have much more success advocating for a place that is known and trusted than we will for a new service emerging in this space.

Our Path Forward

For **CDL**, our decision was to retire Dash and fully support Dryad as our institutional data publishing platform. But that may not be the solution for other campuses. Dryad offers ways for campuses to couple the wide net of Dryad with local solutions. That level of flexibility and community-led decision making was another reason why we knew it was a good fit.

Of course, other hurdles will arise as the data publishing and open science space grows and matures, but we believe that shifting our focus externally to support a community our researchers have already endorsed is our best way forward for UC. Instead of thinking about this change as giving up on our institutional solution, we are looking to meet researchers where they are at, effectively leveraging our institutional values and services in a way that doesn't interfere with their workflows. By supporting a larger community effort that meets the needs of our researchers, we can successfully invest in research data publishing.

References

- Cragin, M. H., Palmer, C. L., Carlson, J. R., and Witt, M.** (2010). Data sharing, small science and institutional repositories. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 368(1926), 4023-4038. doi:10.1098/rsta.2010.0165
- Narayan, B., and Luca, E.** (2017). Issues and challenges in researchers' adoption of open access and institutional repositories: a contextual study of a university repository. *Information Research: an international electronic journal*, 2017, 22 (4). Handle: <http://hdl.handle.net/10453/121438> 🌳