The Opportunities and Challenges of Research Data and Software for Libraries and Institutional Repositories

by **Tom Morrell** (Research Data Specialist, Caltech Library) <tmorrell@caltech.edu>

Long-term availability of data and software generated by researchers is a massive challenge. For libraries, this challenge is also an opportunity to leverage relationships with researchers and utilize expertise in creating metadata and making content available over the long term. At **Caltech**, a strong research data repository was created by keeping the services simple and providing researchers an easy-to-use platform to share data and software (*data.caltech.edu*). In just two years of operation, CaltechDATA has received an impressive number of submissions from over 1% of campus researchers in a wide variety of disciplines. The repository has already powered a discipline-specific data resource, custom visualizations, and allowed for rapid development of many new features.

The publication and data management practices currently used by the research community are clearly insufficient to maximize the value of research, resulting in inaccessible data, non-reproducible data, and worst of all lost data. Even simple measurements that can easily be stored in a text file, such as the length of a bird beak, present significant challenges for data reproducibility and accessibility. A study looking at 20 years of biological organism measurement data found that on average only 20% percent of data files were available when requested, and availability decreased over time (Vines et al., 2014). Even when data are received, they may not be correct or usable by other researchers. In this case, only 13% of papers had data that could be used to reproduce the analysis from the original work (Andrew et al., 2015). This example shows some of the current challenges for accessing and using research data. Larger and more complex types of data and software will prove even more difficult to preserve and reuse.

Depositing data files and software in a repository is the solution to data availability challenges. There are thousands of discipline-specific data repositories that have been developed to store and improve the reusability of research outputs for a specific communities (Pampel et al., 2013). Successful efforts, such as the Protein Data Bank (PDB) for protein structural data, GenBank for genomics data, and WormBase for nematode model organism data, have made large amounts of data available in a standardized way (Benson et al., 2013; Berman, Kleywegt, Nakamura, & Markley, 2014; Lee et al., 2018). However, two major challenges for discipline-specific repositories are scope and funding. The scope of disciplinary-specific repositories is unlikely to be sufficient to meet researcher needs. For example, **Caltech** researchers publish approximately 4,000 peer-reviewed publications annually, and most rely on significant amounts of data and software. Much of this innovative work is interdisciplinary and simply does not fit into existing disciplinary repositories. If a field is still developing, it is nearly impossible to standardize file formats and experiment-specific metadata. Discipline-specific repositories are also often funded through competitive short-term research grants, making long-term funding challenging. For each new grant a justification must be made as to how funding will have a major new impact on the research community. It can be difficult to get funding for the ongoing and unexciting work of maintaining access to data (Van Horn & Gazzaniga, 2013).

The development and promotion of the FAIR (Findable, Accessible, Interoperable, and Reproducible) principles for research data has provided a broader understanding of the requirements for effective data sharing (Wilkinson et al., 2016). Key components of making data FAIR include assigning appropriate metadata, using persistent identifiers, and providing human and machine access to the data. Efforts are underway to enable researchers to make their data FAIR, and publishers are determining how to include data and software citation as part of publications (Cousijn et al., 2018; McQuilton et al., 2016). The COPDESS "Commitment Statement in the Earth, Space, and Environmental Sciences for Depositing and Sharing Data" statement makes recommendations to ensure open data, including having journals stop accepting data files and software as supplemental files and directing researchers to put data and software in appropriate repositories (COPDESS, 2019). This commitment statement is a major step in the right direction and has received signatures from most major publishers. While the current statement is solely for geoscience-related data, the quick adoption by publishers suggests it may easily translate to other fields (Stall et al., 2019). The transition to FAIR data in repositories will result in many questions from researchers, especially since a significant amount of data currently stored in supplemental information or on personal computers will need to find a home.

One approach to the challenge of increasing data deposit demands is to simplify the problem. In the current era of limitless and cost-effective cloud storage, the annual price for storing 100 TB of data in geographically-redundant cloud storage is less than the average cost of a single chemistry journal subscription (Romaine, 2019). Open source repository software like Invenio and Dataverse provides community supported ways of managing data. All institutions have access to the technology to store and make files persist over time, but there are two challenging requirements for a successful repository: collecting files and software from researchers and ensuring that content remains available over the long term. Libraries are uniquely positioned to tackle these challenges since they are experts in storing, archiving and describing materials. They have existing relationships with researchers and deep experience with metadata. Libraries also have a history of preserving content and making thoughtful decisions about retention. Existing institutional funding models for libraries solve the major challenge of long-term sustainability common for disciplinary repositories. Similarly, researchers are likely to be more willing to store their data locally at their own institution. Under the auspices of a university library, all data and software at an institution can be captured by the institution and paid for by the institution.

Despite many advantages, institutional data storage at a library has traditionally limited the amount of customization available to researchers to support discipline-specific requirements. However, modern repository platforms with persistent identifiers and APIs can balance standardization and customization. Persistent identifiers such as DOIs easily provide federated metadata for discovery and APIs allow access to the underlying data files for customized development. This allows a disciplinary or project repository to easily build custom features on top of data that is stored in the institutional repository. For example, the Total Carbon Column Observing Network (TCCON) has their data service (*tccondata.org*) built on the **Caltech** institutional data repository, CaltechDATA (*data.caltech.edu*). TCCON maintains their own data processing pipeline and website. They have complete control of how their data files are organized, can embed custom metadata within their files, and can provide private access to data to members of the consortium. However, the public access to files is via DOIs that resolve to CaltechDATA landing pages. At the end of the TCCON data processing pipeline, files are transferred to CaltechDATA automatically

continued on page 31

using an API. The data files in CaltechDATA can also be accessed programmatically using an API or included in custom visualizations and other processing pipelines. Even if TCCON ceases operation and the *tccondata.org* site goes offline, all the important data files will still be accessible via CaltechDATA and the **Caltech Library**.

Another challenge for institutional repositories run by libraries is a limited history of receiving large volumes of data submissions. A 2017 survey from **ACRL** found that most library institutional repositories receive one or fewer datasets per month (Hudson-Vitale et al., 2017). Library-managed repositories will need to be more efficient in order to tackle the volume of data anticipated from new researchers being required to provide FAIR data to support publications. CaltechDATA has been in operation since summer 2017 and in two years of operation has received over 1,000 records from more than 1% of campus researchers (including faculty, staff, postdocs, and graduate students) from a broad range of disciplines. A one-page deposit form was developed for CaltechDATA that is straightforward, quick, and makes it easy for all authenticated campus researchers to submit files with metadata based on the DataCite schema. The deposit form shows only the most critical fields to the researcher by default, but a complete set of metadata is made available if they want to build a more complex record. Upon submission, the user immediately gets a DOI that can be included in a publication. All metadata is transmitted to DataCite for aggregation to encourage dataset discovery. Users can also automatically submit software via the CaltechDATA GitHub integration and generate and update records using the CaltechDATA API. Data deposits are encouraged by not having a library approval step, so the submission process can easily scale to thousands of records per year. Although the library does not manually curate record metadata and files, the quality has been remarkably high as the researchers feel responsible for their own CaltechDATA records. Since CaltechDATA records are public, the quality of the submitted data and metadata directly affect the public image of the researcher, encouraging high quality submissions.

Even though the underlying CaltechDATA repository is simple, **Caltech Library** has been able to quickly build new features such as automated metadata updates and visualizations to support researchers and data users. When a dataset in CaltechDATA is cited by a new publication, the publication is automatically linked in the CaltechDATA item's metadata using CrossRef Event Data. The researcher who submitted the CaltechDATA record can also choose to receive an email notification every time their item is cited. Similarly, when a dataset is referenced by a completed thesis in the CaltechTHESIS repository, the thesis is automatically linked in the CaltechDATA item's metadata. Project-specific visualizations, such as our geology thesis map (maps.library.caltech.edu) were developed to show where data were collected over time and to promote Caltech research to the broader community. With the geology thesis map, a visitor views an image of the world and can zoom in to specific sites and retrieve images of the original maps and illustrations generated by Caltech researchers since the 1920s. This feature uses the read API to collect data from the repository. In support of software preservation, Caltech was an early adopter of the CodeMeta standard which allows researchers to provide more complete metadata as part of their code repository. For all software preserved in CaltechDATA the CodeMeta file can be extracted and used to update metadata in the record. CaltechDATA also supports interactive software reuse using Binder, an open-source service that can be used to re-run data analysis in a Jupyter notebook or other programming environment from visitors' web browser (Morrell, 2019). These new features have been developed outside of the repository software stack, and can conceptually be applied to any API-enabled repository. All these new features could be developed quickly as they don't impact the basic functionality of the repository.

By keeping things simple, library data services can be possible for all institutions. At Caltech, 1 FTE is dedicated to the data repository, with support from liaison librarians for outreach and submission support. 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 metadata 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.**, **Olendorf, 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 metadata 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*

continued on page 32

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, Ij. 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*