Research Data Lifecycle Management: Tools and guidelines

E. Deumens[1], L. N. F. Taylor[2], R. A. Schipper[3], C. Botero[4], R. Garcia-Milian[4], H. F. Norton[4], M. R. Tennant[4], S. K. Acord[5], C. P. Barnes[6]
University of Florida

Lessons learned

The past two decades have shown that research data can outgrow budgets for storage systems. The national labs and several large research universities have built large file systems with hierarchical storage systems and backup robots to collect petabytes of data. It is now clear that this approach does not work: too much data is collected and it is not known how much of it has value. The amount of data is now too large to weed out the worthless from the precious.

Too often, people responsible for providing support for data management have been preoccupied with the technology of data storage systems and backup robots. Although these issues are far from trivial, the focus in the interaction with the researchers has to be on the process of data management and the services provided by support organizations, such as university IT departments, and by national cyber infrastructure, such as the TeraGrid/XD.

A quick analysis shows that research data comes in a small number of categories with widely different qualitative requirements. These are:

- **Raw data** from experimental observation such as photographic images from microscopes and output from gene sequencing machines. Some data in this category must be kept safe forever, such as when the observation cannot be repeated. Other data in this category may be discarded after the project is complete, such as when the experiment can easily and cheaply be repeated at will.
- **Intermediate data** for analyzing and processing as part of numerical simulations and explorations. This data usually can be discarded at the end of the computation, as soon as the computation has been certified as having completed correctly. For projects that develop and utilize new software and/or computational methods for processing the data, those resources also need to be retained as part of the data for purposes of final data validation and for reproducibility.[i]
- **Final data** including results to be uploaded to national repositories, or to be made available indefinitely to the community through a university or publisher website.

The data in each category can vary in size from small by current standards to enormous. For purely technical reasons the data in these different categories are often

[1] High Performance Computing Center
[2] Digital Library Center
[3] University of Florida Libraries
[4] Health Science Center Libraries
[5] Center for Humanities and Public Sphere
[6] Clinical Translational Research Informatics Program

conveniently lumped together in a single storage system. Because there is no built-in provision for classifying the data correctly again at a later time, when the requirements associated with each category lead to different actions, delete versus keep, the default action of keeping everything leads to insurmountable problems. The data formats and source code of data processing tools need to be preserved with the data. The available infrastructure for keeping data safe does not have sufficient levels of quality to ensure that data is kept safe in a cost effective manner. Research data, unlike enterprise data, comes in different classes of quality and size. The full backup with replication at another site is appropriate for enterprise data, but may not be affordable for some research data of extremely large sizes. Thus, the need arises for more suitable infrastructure arises that is cost effective for such data sets that nevertheless must be kept safe.

Need for a new approach

A new approach is needed that tags and categorizes data from the moment it is created so that automated tools can ensure that the data is processed correctly at all times.[ii] For this approach to be successful, it is necessary that convenient and flexible tools be designed and implemented that help researchers perform the task of creating metadata continuously as an active part of the continued research process. Training and consulting services to assist the researchers will also be required.

Many institutions, research centers and libraries, have started to develop methods, such as data management plans, to address these issues. However, a clear standard of tools and best practices that will scale to the largest data sets has not yet emerged.

Tools for specific needs for specific research fields are available in some cases. For instance, the Data Documentation Initiative 3 Specification (DDI3)[iii] is an international standard for social, business, and economic data that documents the full lifecycle of research data from the beginning of a research project through data dissemination, with controls for disclosure risk and protection of human subjects. The Inter-University Consortium for Political and Social Research (ICPSR)[iv] already provides tools and supports for DDI3 with additional resources also in development. DDI3 was able to develop based on well-established, pre-defined community standards for data collection as with census data and is thus an exception to the norm. As an exception, examples like DDI3, while excellent for specific purposes, are useful in some regards but do not provide clear direction for overall data curation needs.

The issues that we have encountered at the University of Florida are the following:

- Moving data is slow and error prone, especially for large data sets. Making data available through shared file systems and databases across campus-wide and even state-wide high-performance networks is crucial to make the infrastructure efficient, effective, and convenient.
- A general data annotation structure must be developed and established as a national and international standard. Metadata must be created by the researchers, but with tools, training, and help provided by librarians and possibly other data management specialists. It is crucial that metadata management is easy and convenient, or it will drop to the bottom of researchers' priorities.
- A general approach to making metadata searchable across disciplines is needed that is accepted and workable across all disciplines. This is essential to achieve the innovation we are trying to foster.

- Data must be protected and kept private until it is ready for publication. A general federated authentication mechanism is critical to controlling access.[v]
- Preservation technologies must be developed for data designated for long term storage that is more cost effective and more sustainable than backup to tape and restoration from tape.

Funding mechanisms must be developed that have the flexibility to serve many different disciplines, each with their own standards and best practices of paying for resources and services and with appropriate contribution from the university as infrastructure:

1. Health-related research projects have a tradition of paying market rates.
2. Physics, chemistry, astronomy, geography, engineering, and biology research projects can obtain grant funding, but often need subsidies from the university.
3. Research projects in the humanities, sociology, anthropology, linguistics, and history have few opportunities to get funding that is commensurate with the amount of data they process for the length of time required for their research projects. These disciplines may need special grant programs to be set up from the college or university.

A data lifecycle test bed

The University of Florida Research Computing initiative is a multi-disciplinary group working to address the aforementioned issues. It will enable the creation of a data lifecycle infrastructure flexible enough to implement and analyze several approaches using real life research projects. Pilot projects will be selected to be representative of the range of projects on the campus. The researchers will need to be willing to commit some extra effort in working with the staff at the libraries and in the research computing support infrastructure, as the tools are still being developed and procedures are being worked out.

The UF Libraries offer experience in organizing information, interfacing directly with researchers, and developing cyberinfrastructure efforts on campus. Given this experience, the libraries will implement an institution-wide survey to assess researchers' data workflows and data management needs. Librarians will work with researchers to ensure that the tools developed have a user-centered approach, and cover the unique data needs of the diverse UF research community. Library staff will be involved in creating an institutional data management plan template that walks researchers through required metadata elements, including information about timeframe for data retention. The library is currently exploring additional potential roles as partners in the institution's data endeavors.

The High Performance Computing Center will build and manage the storage systems and the high-speed networks to hold and move the data. This effort will be carried out in conjunction with the Sunshine State Education and Research Computing Alliance[vi] which currently supports the data needs of three research projects that are collaborative in nature between three universities in Florida (FSU, UF, and USF).

---

[i] For more on the need and problems relate to software for reproducible research, see: http://www.nature.com/news/2010/101013/full/467775a.html

[ii] In software engineering discipline, version control systems have provided this capability for some time.

[iii] http://www.ddialliance.org/

[iv] http://www.icpsr.umich.edu/icpsrweb/ICPSR/partners/projects.jsp

[v] The InCommon environment seems to be the emerging standard environment to achieve this goal.

[vi] http://www.sserca.org