Research Data Services at the University of Colorado Boulder

by Shelley L. Knuth, Andrew M. Johnson and Thomas Hauser

EDITOR'S SUMMARY

The University of Colorado Boulder's Research Data Services (RDS) is a joint activity of the University Libraries and the Office of Information Technology. Started in 2011 to meet the National Science Foundation's mandate for data management plans with all grant submissions, RDS grew from helping to write the plans to indicating resources to educating researchers on their use and providing individual consultations and reviews. RDS staff also aid researchers with metadata documentation, helping to develop file documentation and later migrating to final formats. Researchers often contact RDS about long-term data storage or for assistance with data dissemination and complying with the requirements of data repositories, funding agencies or scholarly journals. Urging researchers to learn about data management, using a variety of communication channels and venues, is key. Incorporating a data management plan requirement to in-house grants and creating a data management plan competition with financial award have been most successful at promoting the RDS.

KEYWORDS

data curation
documentation
metadata
library technical services
digital repositories

Shelley Knuth is a senior research data specialist at the University of Colorado Boulder in research computing, where she assists campus researchers with their research data needs, including data management, education and storage. She can be reached at shelley.knuth<at>colorado.edu.

Andrew Johnson is research data librarian at the University of Colorado Boulder Libraries where he develops and implements research data services in collaboration with campus partners and co-chairs the campus Research Data Advisory Committee. He can be reached at andrew.m.johnson<at>colorado.edu.

Thomas Hauser is the director of research computing at the University of Colorado Boulder, where he was instrumental in establishing centralized storage for research data and creating the Research Data Services. He can be reached at thomas.hauser<at>colorado.edu.

While research funders and journal publishers now encourage or mandate data management and sharing, researchers are often not formally trained in these practices. As a result, many universities have begun to develop programs to assist faculty, staff and students with these needs. One such effort, Research Data Services (RDS) at the University of Colorado Boulder (CU-Boulder), is a collaborative activity between research computing (RC), a division of the Office of Information Technology, and the University Libraries. Similar to other institutions, the range of data services provided includes assistance with writing data management plans (DMPs), data storage and repository advice, data processing and other needs related to research data management. In addition, RDS has experimented with a variety of novel approaches to outreach and engagement across all disciplines at CU-Boulder and with affiliated institutions in the surrounding area. The history, services, outreach and education efforts of the RDS program at CU-Boulder are described in the sections that follow.

History

RDS was developed in 2011 in response to the National Science Foundation (NSF) requirement for DMPs to be included with all grant proposals [1]. To meet the new requirements, the original RDS was developed initially to help with DMP writing by repurposing existing positions in RC and the libraries. In addition to providing DMP templates via the DMPTool [2], RDS would meet with campus personnel to help them understand the components of the DMP. As a result of a campus-wide Data Management Task Force report [3], a governance structure and additional services were added to RDS to expand offerings, while keeping the primary mission intact.

CONTENTS

< PREVIOUS PAGE

NEXT PAGE >

NEXT ARTICLE >

Two CU-Boulder committees provide oversight and strategic direction for RDS. RDS reports directly to the Research Data Executive Committee (RDEC). The members of this committee consist of RDS staff, the director of research computing, the senior associate dean of libraries and the associate vice chancellor for research (AVCR). This committee develops an implementation plan based on direction provided by members of the Research Data Advisory Committee (RDAC). The RDAC committee consists of members of RDEC plus researchers and support staff from a wide range of disciplines across campus. RDAC also includes data managers and curators from local and regional data centers, including the National Center for Atmospheric Research (NCAR) and the National Snow and Ice Data Center (NSIDC). The purpose of RDAC is to provide strategic direction for RDS and to provide guidance on issues most relevant in specific disciplines.

Services

The CU-Boulder RDS offers several services to its researchers, and most are free of charge. Outreach efforts, described below, have identified new future opportunities to be explored to better reach the research community at CU-Boulder.

DATA MANAGEMENT CONSULTING: At the core of RDS are general data management services, particularly those related to assisting campus personnel with writing DMPs for grant proposals. The DMP can be the first exposure researchers have to the concept of data management, and it is important to properly lay the foundation for better data management strategies for the future. Because researchers are often the first, and for a long time only, stewards of their data, instilling good data management practices early benefits everyone. This process involves coaching on using non-proprietary data formats, understanding how the size of data can limit use, proper data collection processes and proper formatting. Often, researchers do not involve RDS until shortly before proposal submission, but this point is still early enough to produce guidance on good data management practices. CU-Boulder RDS uses these opportunities as teaching moments to encourage good practices for the future. The DMP service is primarily centered on one-on-one consulting, but requests to review draft DMPs via the DMPTool also occur.

RDS also assists researchers with metadata documentation. Generally, CU-Boulder RDS wants to encourage researchers to document their data early and often. RDS has discovered that asking researchers to learn a new language, such as XML, to document data usually results in no documentation being written. As such, RDS encourages researchers to utilize whatever means they are most comfortable with to document their data, which oftentimes is a simple text file describing their dataset. By making this process simple for the researchers, the important information about the dataset is more likely to be captured, and the text files can be converted to other formats without the need for further input from the researchers.

DATA STORAGE ARCHIVING AND DISSEMINATION: Many researchers first find RDS when looking for a place to house data for the long term. RDS staff will provide advice on matters related to data storage, archiving or curation, and will assist CU-Boulder researchers with finding a long-term solution for their data depending upon the level of storage they require. Generally, RDS encourages researchers to store data in discipline-specific locations – in essence, where they themselves would go to find data. Some researchers wish to house their data on their own managed systems; while RDS generally does not have the opportunity to connect with these researchers, we still encourage the use of proper data management techniques, no matter the researcher's preferences or needs.

In some cases, discipline-specific storage facilities may not be beneficial to the researcher, may not exist or may not be cost effective. In these instances, we may also assist with in-house solutions. One local resource the RDS utilizes is the PetaLibrary storage infrastructure, which is an NSF subsidized service for storing research data that is managed by RC. The PetaLibrary offers a minimum of 2 terabytes (TB) of storage to any U.S.-based researcher affiliated with CU-Boulder. Each researcher pays a nominal fee, depending on the level of service, to store data on the system. The PetaLibrary has two main classes of storage available, called *active* and *archive*. Active storage keeps data on spinning disk, intended for data that is frequently written or read. The data can be snapshotted as part of this service. A second level of service, archive, is stored on a combination of disc and tape. Under this service data not accessed after a period of time is migrated to tape. Users can also do various combinations of active and archive data storage, depending on their needs. These options include replication or off-site duplicate storage. Additional information on the PetaLibrary is available at https://rc.colorado.edu/resources/storage/petalibrary.

Another point of contact with researchers involves data dissemination. While some researchers willingly share their data, others only share data because of funding agency or journal requirements. In either case, RDS will assist researchers with finding the proper venue. For some, the process of storing data in a well-managed archive or storage facility can also be the venue for data sharing, particularly in proper data repositories. RDS encourages the use of digital object identifiers (DOIs) to ensure long-term access, and we promote data storage, archiving and preservation locations that offer DOIs. RDS also provides advice on how to prepare data for long-term storage, including proper formats and metadata as described above. Issues of data security can also be discussed with RDS, who will refer researchers to the appropriate offices on campus to handle these types of data. CU-Boulder's local resources, such as the PetaLibrary, are currently not equipped to house secure data and do not attach DOIs to data.

Education and Outreach

CU-Boulder RDS spends a large portion of time on outreach efforts to encourage researchers to learn techniques to better manage their data. Despite the increased use of digital data and data management requirements by funding agencies, researchers may not utilize university RDS effectively due to various reasons, such as lack of knowledge about campus data management services or little incentive to improve their data management practices. CU-Boulder RDS has approached these concerns with ideas to promote awareness and to entice the utilization of data management services with great success. To promote awareness, in 2014/2015 RDS set up accounts on social media, began utilizing campus and department newsletters, hosting workshops and contacting department chairs directly via email. We have also made an effort to attend departmental and institute faculty meetings on campus to introduce researchers to RDS services. This outreach has resulted in several consultation requests from faculty.

One of the more popular activities has been the introduction of data workshops across campus. Since May 2014, approximately 15 workshops taught by RDS have provided information on a wide variety of topics, including best practices in data management, how data publishing benefits careers, federal funding agency requirements around data, learning the difference between specific data formats, data transfer and storage, and how to write good data management plans. Promotion of these workshops takes place through campus newsletters, social media, email listservs and emailing the chairs of departments on campus. The most popular workshop to date has been *Best Practices for Good Data Management*, in which we discussed the various components of a NSF DMP and general good data management practices (slides: http://researchcomputing.github.io/meetup_fall_2014/pdfs/fall2014_meetup11_data_management.pdf).

Another activity for increasing visibility across campus is to provide financial incentives. RDS, in conjunction with RDEC, RDAC and the office of the vice chancellor for research (OVCR), developed two funding opportunities for researchers on campus to encourage good data management practices. The first was to add a DMP requirement to an existing internal grant competition – Innovative Seed Grants (ISGs). The ISGs provide seed money for innovative and collaborative research projects, in part to improve faculty competitiveness for federal grants. The addition of a DMP as part of the ISGs forced faculty to give more consideration to how they will manage their data as part of their research. Since faculty from all disciplines apply for these grants, the addition of a DMP requirement expanded outreach to faculty who normally do not apply for grants with data management requirements. In 2014, the first year of this requirement, the DMPs were only reviewed by a committee to provide feedback to the researchers. By 2015, the DMPs were both required and counted as a small part of the applicant's score. The addition of the DMP as part of the ISGs was the most impactful action for promoting awareness of data management issues to-date. RDS received requests to review approximately 20% of

DMPs to be submitted as part of an ISG proposal. OVCR also requested RDS organize educational workshops on writing good DMPs. Approximately 20 faculty members attended these workshops.

The second funding opportunity developed by OVCR and RDS was called the Best Digital Data Management Plans and Practices Competition. This opportunity allows faculty, postdocs, research staff or graduate students to submit a DMP to the challenge, where the winners in five disciplines (arts and humanities, social sciences, life sciences, physical sciences and engineering) receive $2,000 in unrestricted university funds. The purpose of this competition is multifold – first, to promote conversations about good data management practices; second, to collect a bevy of plans that could be distributed as exemplary submissions; and third, to increase awareness of RDS. The 2014 competition drew approximately 15 submissions. A 2015 competition is already underway. Overall, tying data management requirements to financial incentives was the most successful outreach effort organized by RDS.

Summary and Future Work

RDS at CU-Boulder, with support from the OVCR, has made great strides toward promoting data management awareness on campus since its inception in 2011. RDS has developed new funding and educational opportunities and made a strong effort toward promoting services through various outreach venues. RDS hopes to expand future services on campus by providing more tailored services to campus researchers, such as visualization and analytics services, and additional services focused on the humanities and social sciences.

Acknowledgments

The authors acknowledge support for the PetaLibrary storage resources, which were provided by NSF-MRI Grant ACI-1126839. ■

Resources Mentioned in the Article

[1] National Science Foundation. (January 2013). Grant proposal instructions: Chapter II - Proposal preparation instructions. (Document number gpg13001). Retrieved from www.nsf.gov/pubs/policydocs/pappguide/nsf13001/gpg_2.jsp#dmp

[2] Strasser, C., & Cruse, P. (March 2013). The DMPTool and DataUp: Helping researchers manage, archive and share their data. Position paper for the Research Data Management Implementations Workshop, Arlington VA.

[3] Vice Chancellor for Research Data Management Task Force, University of Colorado Boulder. (2012). Research data management at the University of Colorado Boulder: Recommendations in support of fostering 21st century research excellence. Retrieved from http://scholar.colorado.edu/ovcr/1/