Research Data Storage Approaches at Columbia

A position paper for the Workshop on Research Data Lifecycle Management[*]*,
July 18-20, 2011, Princeton University

Rajendra Bose, Ph.D., Manager, Research Computing Services
Danianne Mizzy, MLS, Engineering Librarian, Columbia Libraries/Information Services
Columbia University, New York NY
July, 2011

Introduction

The current approach to research data management at Columbia has been largely shaped by the post-mainframe-era academic tradition of leaving faculty to their own devices—and funding—to carry out their varied and successful research programs within their own labs or disciplinary communities. However, central divisions including Information Technology and the Libraries are becoming more involved with issues of the digital research data lifecycle, which represents a change in philosophy at the university.

Columbia researchers in all areas are generally meant to use storage infrastructure that is provisioned at the school, department or division level by one of several local IT groups. If these resources are not available or adequate, researchers frequently use their own research money to build their own storage solutions, which could include hiring part-time, full-time or shared IT staff (or using research staff as IT staff), in addition to purchasing and configuring hardware and software and designing a backup regimen. Some researchers are also independently investigating commercial storage solutions, such as Dropbox or Amazon S3.

Historically, the central IT division (CUIT) has been responsible primarily for administrative or enterprise storage and computing needs for the university. This includes managing the campus network and large financial and student record systems, but *not* providing research storage. A four-person research computing group, established three years ago, has begun to coordinate and provide central computing resources to some research groups on a pilot basis.

The Columbia Libraries/Information Services' (CUL/IS) Academic Commons repository has been set up to collect, preserve, and make accessible through search and discovery tools the scholarship of the faculty. This research output may include datasets and raw data—the objects required for deposit by funding agencies—as well as materials that help to contextualize that data, such as articles, book chapters, essays, monographs, working papers, technical reports, conference presentations, multimedia creations (e.g., simulations, three-dimensional maps), and other materials in digital formats. CUL/IS is actively developing policies and procedures for data management in response to the scale and type of data deposited.

Central IT perspective

CUIT provides all active Columbia associates with an email account and with what is now viewed as a negligible default amount (less than 100 MB) of home directory file storage (which can be used to create a basic personal homepage) on the central Unix system. Limited, fee-based increases are possible for these home directories. But generally, researchers are meant to use storage infrastructure provided by their local IT group.

The CUIT Research Computing Services (RCS) group, in collaboration with several central IT teams, initiated a high performance computing (HPC) pilot system which has been available to roughly 100

researchers with accounts during the past two years. This pilot introduced the paradigm of faculty from different departments sharing the costs for centrally-managed computing resources. Research faculty have requested to make the pilot system permanent, and in addition to adding hardware to accommodate another 100 accounts, we are planning for further expansion of this HPC service housed in our *Core Research Computing Facility* within the university data center.

As is typical for a shared HPC system, we provide a limited amount (here, 50 TB) of available, non-backed up storage, which is divided among user groups and designated as temporary working or "scratch" storage for running HPC jobs. The assumption is that researchers using our HPC system have other places available over the network to store their data sets and computational results permanently, with an adequate backup facility.

We are learning that many Columbia researchers want large amounts of storage space for their work, and CUIT is starting a project to explore how to establish an affordable, centrally managed research storage service. Various parties will need to discuss and agree on the appropriate financial and technical models to create a sustainable service.

Libraries perspective

Until recently, most academic research libraries have concerned themselves almost exclusively with the end of the data life cycle. They have focused on preservation and archiving activities and have concentrated those efforts on unique or rare textual and visual materials drawn from special collections. With the rise of the open access movement, institutional repositories and e-science, libraries have begun to explore what role they can and should be playing in data management within the academy.

Active discussion on this topic has been underway at Columbia since March 2008, when an e-Science Task Force was convened. The task force's February 2009 final report stated that:

> The highly decentralized, individually administered compute cluster model of the past is no longer able to meet the needs of many researchers, who now require a richer and more powerful tool set. To maintain its place as a leading research institution, Columbia must develop a University strategy for coordinating, supporting, and developing e-science and the CI [cyberinfrastructure] it depends on in order to ensure and optimize the required investments in technological and human resources.

To better understand researchers' needs, in February 2010 CUL/IS conducted data interviews with five representative science and engineering faculty members. Key findings included these observations:

- Little to no use of best practices for data storage and management
- Lack of mandates for what needs to be stored (No consistent guidance from granting agencies)
- At best, academic departments provide local server capacity for data storage
- At worst, data is stored on a single CD or hard disk on a shelf
- Data management is currently dependent on soft money and as datasets grow in size and complexity, sustainability is a serious issue.

To support the Academic Commons institutional repository and Libraries' digital content, in 2008 CUL/IS developed a Fedora-based digital preservation system. CUL/IS stores preservation digital assets on a total of four copies, two on disk and two on tape, with one disk copy in Columbia's data center and one disk copy offsite in the NYSERNet Data Center located in Syracuse, New York. To manage multiple copies, automate migration and replication and provide a policy-based model to

manage the long-term retention and access to digital assets, CUL/IS uses the Sun StorageTek Storage Archive Manager (SAM) software which provides a self-protecting, automated data migration and recovery model. A total of 280 terabytes (TB) of disk and tape storage has been purchased to support the Digital Preservation Storage System, configured with an effective storage capacity of approximately 70TB.

CUL/IS is also working on a memorandum of understanding (MOU) with the new Integrated Earth Data Applications facility (IEDA) at the Lamont Doherty Earth Observatory (http://www.iedadata.org/) whereby CUL/IS will begin to provide a dark archiving service for IEDA data sets. CUL/IS has also agreed to act as a failover agency for long-term archiving of the Socioeconomic Data and Applications Center (SEDAC) Long Term Archive (http://sedac.ciesin.columbia.edu/lta/). In response to the new NSF data management plan requirements, University Librarian James Neal, and then Executive Vice President for Research David Hirsh issued an MOU expanding the size of datasets (up to 50GB) that could be deposited in Academic Commons and subsequently a template was made available to assist researchers in developing data management plans (DMPs). The Center for Digital Research and Scholarship (CDRS) is currently exploring how to support faculty earlier in the data life cycle by piloting a 1TB production server running Alfresco.

CUL/IS is monitoring the needs of the CU research community through the analysis of all active grants and NSF DMPs. As part of its strategic plan, it is committed to building the capacity to provide storage for data for which no disciplinary repository exists. Along with other campus stakeholders, it will continue to make the case to the university administration and to outside funders for increased support for data management. It will also continue to seek opportunities to partner and collaborate, such as its current gold membership in the Duraspace Consortium.

Conclusion

To date, data storage involvement for the central Information Technology and Library divisions at Columbia has split more or less cleanly along the lines of active or working storage versus curation and preservation storage. CUIT provides temporary storage for its new HPC service, and has started to discuss internally how a central data storage service could support the active, day-to-day work of Columbia researchers. CUL/IS has put a robust institutional repository in place to help preserve a large variety of digital research materials and is involved in support for researcher DMPs.

The new and rapidly evolving IEDA data facility at the Lamont campus portends a more complex future at Columbia, where discipline-specific, library and IT expertise will be interwoven to assemble, organize, provide and preserve a large collection of research data within the institutional setting of the university. We anticipate the need for and value of the combined understanding and efforts of the researchers, campus IT leaders, and library/archive specialists at this NSF-funded Workshop on Research Data Lifecycle Management, and related future events, to help direct Columbia's approach to research data storage.