Columbia's Evolving Research Data Storage Strategy

An experience/position paper for the Workshop on Research Data Management Implementations[*],*
March 13-14, 2013, Arlington

Rajendra Bose, Ph.D., Manager, CUIT Research Computing Services
Amy Nurnberger, Research Data Manager, Center for Digital Research and Scholarship, Columbia Libraries/Information Services
Columbia University, New York NY
March 1, 2013

Introduction

Columbia researchers in all areas generally develop their own storage solutions or use storage infrastructure that is provisioned at the school, department or center level by one of several local IT groups. At least one faculty-led center has a dedicated research computing/IT team that provides data storage and other services to hundreds of users across a number of departments and labs, and another similar group may be formed in 2013.

At the same time, central divisions including Information Technology, the Libraries, and Research Initiatives, together with new faculty committees, have been working to evolve research data storage strategy at the university during the past year. Current planning underscores the complete research data lifecycle and represents a change in philosophy at the university.

In 2011 Columbia's Executive Vice President for Research formed a Research Computing Executive Committee of senior administrative and academic leadership. This group in turn charged a faculty-led "Shared Research Computing Policy Advisory Committee" to design a coordinated governance and financial plan for research computing.

In the past months, a subcommittee of this group focusing on data storage and security has recommended moving forward with a *research storage pilot* project, and this recommendation has been submitted to the parent Research Computing Executive Committee for consideration. If approved, this research storage pilot would initially target a small subset of research faculty, would plan to offer storage at rates roughly comparable to commercial cloud systems such as Amazon S3, and would be tied to a larger central IT division (CUIT) project to acquire an expandable storage system for a variety of institutional needs. With this storage acquisition, CUIT could begin to enact a plan to provide centrally-managed and administered storage to the wider Columbia community. The research storage pilot would also be connected to a *Core Research Computing Facility* construction project to upgrade infrastructure within the university data center[†].

[†] *This project is supported by NIH Research Facility Improvement Grant 1G20RR030893-01, awarded April 15, 2010, and associated New York State funds.*

The Academic Commons repository hosted by the Columbia University Libraries / Information Services' (CUL/IS) Center for Digital Research and Scholarship (CDRS) continues to not only collect and preserve the scholarship of the faculty, but also to make it accessible through search and discovery tools. The scholarly products preserved by Academic Commons are not limited to datasets and raw data, but may additionally include works based on those 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. As the scale and types of data being generated by research are constantly, and respectively, growing and evolving, CUL/IS is developing policies, procedures, communications, and training to accommodate these transformations.

Central IT perspective

In the last decades, CUIT has been responsible primarily for the administrative computing and storage needs for the university. Although CUIT has concentrated activity on managing the campus network as well as financial and student record systems, and does not currently offer storage services suited for research, a five-person research computing group was established four years ago. This group, Research Computing Services, now coordinates centrally-managed high-performance computing resources to some research groups on campus and has worked with others on reviewing their research storage needs.

A legacy service provides all active Columbia associates with a small default amount of home directory file storage on the central Unix system. This service includes regular backups of home directory content with an offsite component. Aside from this home directory storage, researchers are expected to use storage infrastructure provided by their local IT group for day-to-day work.

If these resources are not available or adequate, researchers frequently use their own research funds 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. Many faculty, staff and students now use commercial storage solutions like Dropbox to collaborate with colleagues on research and to backup work documents.

The CUIT Research Computing Services group, in collaboration with several central IT teams, initiated a shared high performance computing (HPC) pilot system in 2009, which now serves roughly 200 researchers. This system includes 75 TB of non-backed-up scratch storage available for users running HPC jobs. Researchers using our HPC system use other file systems over the network to store their data sets and computational results permanently. Faculty committees have recommended future planning to make it easy for HPC users to move data sets from centrally-managed, general research storage to and from HPC scratch storage to accommodate researcher workflows.

Libraries perspective

Historically, academic libraries have focused on the preservation and archiving activities related to the end of the data life cycle. These activities have usually been concerned with

special collections materials, often focusing on the unique or the rare. With the growing attention accorded to research data, libraries have seized the opportunity to employ their expertise in the realms of metadata development, information organization, and curation and begun to explore expansion of their roles in supporting data management in the academy.

Active exploration and discussion of this realm at Columbia was initiated in March 2008 when an e-Science Task Force was convened. The findings of that task force indicated a need for Columbia to invest in a university strategy for research data and the cyberinfrastructure it depends on. Following on these findings, CUL/IS conducted interviews with representative members of the science and engineering faculty to better understand researchers' data management needs. These interviews found that the area of data management is one where there are few best practices and little investment in developing robust services.

In partial response to the e-Science Task Force and the data interviews, Columbia instituted a Research Data Manager position. The Research Data Manager is tasked with: expanding data management planning services; developing practices, outreach, and education for each of the different stages in the data management lifecycle; and initiating policy development around the multiple issues involved in a comprehensive strategy for data management. This is a collaborative role, and the Research Data Manager works with others who share a vested interest in issues around data management and requirements to develop robust services supporting the needs of researchers at Columbia. The establishment of this position emphasizes the increased importance of research data management at Columbia and is indicative of Columbia's dedication to developing research data support and services throughout the digital data lifecycle.

As a part of supporting the entire data life cycle, CUL/IS continues to consider options for collaborative and working data storage spaces. While an earlier pilot of Alfresco has been discontinued, other platforms, such as Google Apps, are being considered. In concert with this, CUL/IS has continued to assist researchers in developing data management plans (DMPs) by maintaining a library of agency and directory specific templates. These DMPs are further supported by the planned increase of fee-free storage for research products, such as data, provided by Columbia's institutional repository Academic Commons. This summer, Academic Commons will be supporting file sizes of up to 10GB free of charge. Funded research data files beyond that will require $10/GB for their storage.

Academic Commons is housed on a Fedora-based digital preservation system. CUL/IS is in the process of moving from a storage system employing disks and tapes—with offsite support by the NYSERNet Data Center located in Syracuse, New York, and managed using the Sun StorageTek Storage Archive Manager (SAM) software—to one based on two Isilon clusters. This will be supported with auxiliary storage through a partnership with Indiana University. Each cluster will contain four nodes providing 432TB of raw disk space, equating to 292TB of usable storage. One cluster will be located in Columbia's data center and the other in NYSERNet Data Center. The data placed in the local cluster is immediately and automatically replicated to the NYSERNet cluster. Additionally, different data classes may be configured to survive some number of local failures beyond the default of two

simultaneous disk failures or a single node failure from the cluster. The partnership with Indiana University permits CUL/IS to place a third copy of the data in their large Hierarchical File System (HFS) setup, which streams data off to tape. This partnership allows CUL/IS not only to have greater geographic distribution for storage, but also to protect against potential problems with the Isilon replication system by having the data stored on a completely separate system.

Building on the Academic Commons' Fedora-based digital preservation system, CDRS has implemented a SOLR-Lucene enterprise search server to index the metadata associated with the repository's contents and provide enhanced discovery and access to its materials. In addition to this, Blacklight, a Ruby on Rails application, was instituted in April 2011 to further increase the visibility of items residing in the repository by better exposing the index metadata to search engines. Further refinements enabling access and enhancing visibility include serializing the metadata in multiple formats (e.g. RSS, ATOM, and OAI-PMH), generating XML site maps, creating structured metaheaders supporting Google Scholar indexing, and implementing schema.org microdata (RDFa).

In addition to the work with Academic Commons, CUL/IS also has memoranda of understanding (MOU) with the Integrated Earth Data Applications facility (IEDA) at Columbia's Lamont Doherty Earth Observatory,[‡] whereby CUL/IS is providing a dark archiving service for IEDA data sets, and the Center for International Earth Science Information Network[§], providing long term storage and access. 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[**]. There continue to be consortia efforts to build a Trustworthy Repositories Audit & Certification (TRAC) certified repository, as well as considerations of participation in the Academic Preservation Trust (APTrust) and the Digital Preservation Network (DPN). CUL/IS will continue to seek opportunities to partner and collaborate, as demonstrated by its gold level sponsorship of the Duraspace Consortium.

In the recent past, CUL/IS has looked to collaborate with its researchers and research centers by assessing needs through monitoring and analyzing active grants and agency required data management plans. Going forward with the knowledge that the needs of researchers are constantly changing, as demonstrated in fields as disparate as neuroscience and digital humanities, CUL/IS is looking beyond its "monitor and respond" pattern of activities. Acknowledging that demands for data storage and guidance will increase, as indicated by the recent memorandum released by the Office of Science and Technology Policy[††] CUL/IS has the goal of anticipating and providing forward-looking support of needs, and being the solution researchers did not know they needed. This goal requires a continued collaboration with campus stakeholders, and a consistent message to university administration and to outside funders that increased support for data management is required. A Research Data Symposium was recently hosted by CUL/IS

[‡] *http://www.iedadata.org/*

[§] *http://www.ciesin.org/*

[**] *http://sedac.ciesin.columbia.edu/lta/*

[††] *http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf*

including national and international research, funding agency and publishing industry representatives. This event confirmed that partnership and collaboration outside of the university is intrinsic to institutional success in data storage and is an imperative part of data storage services.

Faculty-led efforts

Since 2005 the Columbia University Center for Computational Biology and Bioinformatics (C2B2), located on the medical center campus, has included a research computing support team of roughly nine staff. This group provides fee-based services, including data storage for research computing applications as well as central file storage with backup, for about 15 labs, or roughly 200 researchers and staff. C2B2 services are a model for research computing/IT for the Mortimer B. Zuckerman Mind Brain Behavior Institute which will occupy the Jerome L. Greene Science Center building on the newest Columbia Manhattanville campus. The Greene Science Center is scheduled for occupancy in 2016 and is expected to house a total of 800 researchers and administrators.

Concluding Remarks

Efforts by central IT, Library, and Research Initiative divisions to evolve research storage strategy at Columbia having been moving forward over the past year at the same time as faculty-led research centers develop their own research computing teams and resources. These parallel activities will benefit from closer and continued communication and coordination between central university divisions and distributed research IT staff.

In addition, several other major university initiatives—including the creation of an Institute for Data Science and Engineering that will bring in new faculty and a focus on big data issues, the growth of the New York Genome Center that provides computing, data storage and other services for biomedical researchers at member institutions, and the launch of eight global centers outside of New York—promise to keep research storage at the forefront of technology planning at Columbia for many years to come.

The combined understanding and efforts of the researchers, campus IT leaders, and library/archive specialists at the NSF-funded Workshop on Research Data Management Implementions, and related future events, will help shape Columbia's approach to research data storage.