Cyberinfrastructure Supporting Evolving Data Collections

Maria Esteva
Texas Advanced Computing Center
maria@tacc.utexas.edu

Christopher Jordan
Texas Advanced Computing Center
ctjordan@tacc.utexas.edu

Tomislav Urban
Texas Advanced Computing Cen
turban@tacc.utexas.edu

David Walling
Texas Advanced Computing Center
walling@tacc.utexas.edu

ABSTRACT

The requirements to support large-scale and complex research collections are growing at an accelerated pace. Considering the continuous evolution of the collections, their increasing sizes, the technologies supporting them, and the importance of adequate data management to long-term preservation, a team at the Texas Advanced Computing Center (TACC) developed a cyberinfrastructure to aid researchers in the creation, management, and curation of collections throughout the research lifecycle processes and beyond for access and long term preservation. Collections are maintained on a petabyte-scale data applications facility, and consulting services are available to address data curation needs. In this environment, researchers have the flexibility to build their collections without having to deal with details such as systems administration and hardware migration planning. The cyberinfrastructure facilitates the development of sustainable collections and a seamless transition through data gathering and curation, large-scale analysis, and collections dissemination and preservation.

Categories and Subject Descriptors

H.3.2. **[Information Storage]:** Record Classification; H.3.7 [Digital Libraries]: Collection, Dissemination, Standards.

General Terms

Management, Design, Economics, Reliability

Keywords

Data management, preservation, storage architecture, metadata

1. INTRODUCTION

In the last 10 years there has been a noticeable impulse towards early implementation of data management strategies as a fundamental step towards preserving the digital products of research [1]. Culminating in funding agencies' requirements to include data management plans in the grant proposals [2], this development is driven by the recognized value of digital collections reuse for knowledge dissemination and data-driven discoveries. Currently, many academic libraries are implementing information services to help researchers craft their data management plans as well as providing guidance on how to deposit their collections for long-term preservation, often within their own institutional repositories (IR). While valuable, these services may fall short in addressing complex and large-scale collections' architectures, and in meeting the technical and curatorial needs that emerge during their development stages.

Research data collections are increasingly becoming complex systems, formed by diverse data objects included within layers of software and hardware that provide functionalities needed for analysis and interaction with the data. While research is conducted, raw data from experiments and observations are transformed during analysis, at the same time as new data is being gathered, incorporated into subsequent pipelines, curated, published, and archived. These juxtaposed processes are often difficult to document and some collections growth can be indefinite. Through the research lifecycle, researchers may use different applications to process their data and hardware to store it, resulting in dispersed and often incompatible datasets, which creates research bottlenecks and places data under risk [3]. As collections evolve, so do the standards and the expertise required to support them, which differ across and within science domains. In turn, the technologies on which these collections depend change at fast pace, while the cycles of funding to upgrade and maintain them are uneven. Without data management strategies and an adequate technical environment to facilitate these processes, the possibilities for long-term access and reuse of these collections are challenged. Thus, supporting the continuing and sustainable development of these *evolving collections* requires rethinking the infrastructure and service models.

In 2008 the Texas Advanced Computing Center (TACC) at the University of Texas at Austin (wwww.tacc.utexas.edu), established the Data Management and Collections group (DMC) to work with researchers in the Sciences, Engineering, Social Sciences, and Humanities fields in lifecycle management of research collections. Confronted with the diversity of data and requirements presented by research teams seeking solutions for their collections, we articulated the concept of evolving collections that allows mapping curatorial and technical tasks to the different research stages. Based on this concept, we developed a flexible storage facility and data management services as a cyberinfrastructure to meet the researchers' needs.

Developed in the context of a supercomputing center that enables researchers to conduct large computational tasks, the cyberinfrastructure is designed to seamlessly integrate processes from data gathering, to analysis, dissemination, and preservation. Within such an environment, and throughout the stages during which data is transformed into collections, research teams can conduct curatorial and technical tasks while users access the data. With maintenance and security of the infrastructure provided by TACC, researchers can remain focused on their research without having to deal with day-to-day systems operations. As a

iPRES2011, Nov. 1–4, 2011, Singapore.

contribution to the ongoing discussion in digital curation and preservation, in this paper we present our activities and resources illustrated with vignettes from the collections currently supported at TACC.

2. RELATED WORK

Two general models are predominant in data preservation: 1) the centralized model, in which a repository preserves a collection after it is finalized in standard, archival formats; and 2) the decentralized model, in which research teams curate and give access to their own data. Neither is particularly well suited to address the transformations and technical challenges data undergo through the research process. The centralized model cannot address the needs of ongoing research with an evolving collection, while the decentralized one often neglects data management strategies, as these may be too burdensome or technically challenging for the researchers to accomplish [3].

The evolving collections cyberinfrastructure operates in a middle area between these two models. It enables researchers to develop and archive collections in a continuum at any point in their research lifecycle, while they, other users or both perform data analysis and visualization tasks across storage and computing environments. Recently, projects have emerged to allow users to move their data across cloud storage providers and access it as needed, with the added value of including preservation services to assure data integrity and transactions transparency [5]. Like TACC's services, this model proposes maintaining the infrastructure for the users. Differently, it does not yet provide flexibility for researchers to build complex and unique collections architectures and functionalities in the cloud. Thus, as long as the workflows are not fully integrated in the cloud, researchers may have to maintain both local and cloud instances of their data and applications [5]. In contrast, we can provide the same interfaces and/or functionality as DuraCloud on top of our resources. Just as we run database and web servers, we could run DuraSpace and associated applications for users that want those interfaces.

While supercomputing centers are being considered in current data curation discussions [6], their role in data stewardship is not fully developed and their activities and potential are not well known to the community. This paper clarifies the expertise and resources available at TACC, and suggests a scalable model for collections preservation. The cyberinfrastructure has a post-custodial flair [4], as the researcher remains the curator and owner of the collection, while TACC provides the data facility and consulting for as long as it is agreed upon.

3. CYBERINFRASTRUCTURE

3.1 Team Expertise

The DMC group designs, builds, and maintains the data applications facilities, and consults with researchers in aspects of their collections, from creation to long-term preservation and access. Group members are specialized in software application development, Relational Database Management Systems (RDBMS), Geographical Information Systems, scientific data formats, metadata, large storage architecture, system administration, and digital archiving and long-term preservation.

3.2 Infrastructure

Lifecycle collections activities are centralized in a data applications facility called Corral, which provides different technical environments as collections building blocks that users may select according to the functionalities that they need. Corral, consists of 1.2 Petabytes of online disk and a number of servers providing high-performance storage for all types of digital data. It supports databases, web-based access, and other network protocols for storage and retrieval of data. A high-performance parallel file system based on Lustre is directly accessible from all of TACC's High Performance Computing (HPC) resources, enabling mathematical computation and visual analyses of petabyte-scale datasets. Corral's disk subsystem provides 6GB/sec of performance, and each of the web and file servers can move data from between 100MB/sec and 1GB/sec. While data can be moved through multiple servers and services at full speed simultaneously, the system is configured so that no one user or service can consume all the available performance.

For database collections, Corral provides flexibility in terms of the RDBMSs that users may choose from. DB server nodes running MySQL, PostgreSQL, and SQL Server are available. The DB nodes can be easily accessed at high bandwidth by web applications running on Corral's web server nodes. We also maintain open source domain specific databases such as ARK [7] and Specify [8], which support Archaeology and Natural History collections respectively. This infrastructure is useful to researchers from the Pecan Street Project (www.pecanstreetproject.org). The team needed a workflow for ingesting nightly dumps of the usage information of electrical devices from 100+ wired homes in Austin TX, into a database system on which analysis of energy usage can be conducted. A MySQL database is implemented in Corral to serve an ongoing data collection of over 1 billion records, with 5M+ new records added everyday. The size of the datasets and type of general query analysis conducted at this stage, led us to exploring OLAP and column oriented database services for enabling quick analysis, as well as evaluating solid state disks for increasing performance of the SQL queries conducted by the researchers.

Collections requiring long-term preservation are managed within iRODS [9]. Off site replication is done in Ranch, a Sun Microsystems StorageTeck Mass Storage tape system with a capacity of 10 PB, and geographical replication is accomplished through an agreement with Indiana University's Research Computing Division [10]. Close supervision, parts replacement contracts, and frequent schedule of upgrades are in place for maintaining the infrastructure. This model is based on TACC's experience managing systems to assure 24/7 services and data security,

Coupled with performance levels for reliable data transfer between storage and computing resources, the integrated infrastructure of Corral enables flexibility to implement different collection configurations and functionalities. The UT Center for Space Research (www.csr.utexas.edu) uses Corral to store very large sensor, satellite, aerial and radar datasets that they curate for dissemination purposes. Within three days of the 2010 earthquake in Haiti in collaboration with TACC, the repository/file system used for managing CSR data in Corral, was turned into a web repository for sharing data. This allowed CSR to access, organize, retrieve, and post the data required by the emergency operations in the region [11]. This type of quick repurposing allowed a multi-terabyte collection managed through one application, to instantly become accessible through a password-protected web application on another server.

3.3 Services

3.3.1 *Collections Set-up*

Users can request a storage allocation through TACC's user portal and select services to build, manage, and archive their collections.

Storage allocations are renewed on a yearly bases, including revisiting the services needed.

Significant consulting with group members takes place before the data is transferred to TACC. This allows planning data transfers, and deciding what technologies and configurations are needed for a particular collection. Examples of such work entail documenting the existing collection's architecture, guiding data inventories, analyzing data pipelines and improving aspects of their organization, and implementing metadata standards. To set up a collection, and depending on its architecture and lifecycle stage, DMC members manage access to the systems, install database servers, dependency libraries and webservers, and migrate data as needed. Users have access to their deployed web code, but are not burdened with systems administration tasks. A simple case of collection set up is the Oplontis archaeology project (www.oplontisproject.org). To facilitate remote access to a team of international researchers so they could input data and interpretations to a database, we moved an SQL database located at a restricted server to an SQL server on Corral. More complex projects require migrating data and code from commercial databases to ones that run on Corral, and integrating data and metadata from diverse legacy systems.

An ongoing effort is automating and generalizing services. For large image collections like the University of Alaska Museum's Herbarium (www.uaf.edu/museum/collections/herb), as the curators upload the raw files, processing scripts create image derivatives and OCR of labels. In turn, these scripts can be adapted for parallel processing of very large image collections from the UT Libraries in TACC's HPC resources. Rules based services for iRODS are also refactored to use across different collections. Once collections are fully functional and some services are automated, users require less specialized support. As they become their own collections managers, the tasks for our group are related to general data facility administration. When needed, users can submit tickets with requests for support via the TACC users support system.

3.3.2 Data Ingest and Retrieval

Data transfer to the system is achieved from various ingest tools, the selection of which depend on the needs of the users. For users conducting bulk submissions from their desktops to Corral, we developed TACCingest, to move large batches of files in a simple and reliable fashion. The tool was first tested so that Lawrence McFarland, a photography professor with a terabyte sized collection of ~3 GB images, can move large groups of images from his 100 hard-drives to safe storage. Command line and UI tools such as iDROP are also available to transfer data to iRODS, or to query its metadata catalog for data of interest. Data ingest activities may include automated metadata extraction, integrity checking and evaluation of file naming compliance.

The iPlant Collaborative (www.iplantcollaborative.org), which integrates data from standard plant genetic repositories as well as user submitted data, is illustrative of complex data transfers and retrieval due to the amount of users involved, and the functionalities that they request to use their data. Services involve developing applications to support large data ingest into iRODS, and a number of web interfaces to iRODS to make the data accessible to and from different analysis workflows. Provenance metadata is also collected and stored into iRODS using the same web API. In our configuration, digital objects may be accessed from a different workflow from which they originated, repurposed for analysis and or publication, and re-entered to the system as new objects.

3.3.3 Data Preservation and Integrity

Preservation services for the collections stored in iRODS include: rules to generate file checksums, automatic off-site and geographical replication, massive extraction of metadata using FITS [12] and encoded as Preservation Metadata (PREMIS), and finally registering the metadata in the iRODS catalogue.

Beyond basic bit level preservation and technical metadata gathering, we also address the domain scientists' conception of data preservation. In the case of archaeology collections, preservation is strongly associated with integrity, which involves maintaining the relationships between the objects found in a same context in the excavation. To assure that the archived data could render a representation of the site, The Institute of Classical Archaeology (www.utexas.edu/research/ica) selected to have two collection instances within Corral. A presentation instance resides on the ARK database and web site, which provides interactivity features and the possibility for users to study data objects in relation to their geospatial location and to the researchers' interpretations. The archival instance, stored in a hierarchical directory structure in iRODS, and replicated in Ranch, preserves contextual relationships between the raw images—and their versions—of the objects found on the excavation and their correspondent documentation, which is generated on the site and through the research lifecycle. These relationships are preserved through a context code recorded in the digital objects file names and in their metadata records, so that when one object is retrieved, all of the related objects are retrieved as well. In this way, if the ARK database ceases to be supported, the archival instance will serve to reconstruct the site.

3.3.4 Descriptive Metadata

When possible, to facilitate collections organization and avoid manual metadata entry, descriptive metadata is automatically extracted from the collections record-keeping system at ingest. This process requires the existence of an informative and regular file naming and or directory labeling across the collection. It also involves previous work mapping the descriptive data points to standard metadata schemas such as Dublin Core (DC) or Visual Resources Association (VRA) Core. The latter results from consulting with our team and training users on the required standards and practices. Implemented as an iRODS rule, a Jython script parses directory labels and file naming conventions as files are ingested to iRODS. The extracted descriptive metadata is packaged along with the technical metadata, as a METS document and registered with the iRODS metadata catalog [13]. This process is being implemented in McFarland's collection that for a long time has used a systematic naming convention including image title/terms, its geographical location and type of camera codes, and version control number. To access his files, he may search by any of these elements.

3.3.5 Web Access and Services

Corral provides multiple web servers and supports most popular web application languages, including PHP, Java/Tomcat, and Python. Applications are hosted within a shared environment to minimize administrative overhead, although many collections utilize virtual domains within the web server. Data stored within the iRODS environment can be made openly accessible via a well-defined URL, or can be password protected and accessed via WebDav. Because the same data can be made accessible at high performance from several server nodes, both file-centered web services and web applications are configured for automatic failover across multiple nodes, thus ensuring a high level of availability. Examples of data collections websites hosted on Corral include OdonataCentral (www.odonatacentral.org) and Fishes of Texas (www.fishesoftexas.org), both of which utilize image and file services, dynamic web applications, and databases to manage catalogs of specimen data.

3.3.6 Curation

Data curation activities happen at the domain science level and at the general collections level. Researchers as curators gather, analyze, interpret, edit, and preserve the data that through those processes becomes a collection, and DMC members technically enable these activities. General curation services include establishing agreements with researchers and tracking and managing services and collections. Researchers determine when the collection is finalized and decide how to provide access, and we work with them to define the appropriate protocols for open access or to implement access restrictions.

4. ADMINISTRATIVE MODEL

As a service and research organization, TACC offers up to five terabytes of free storage space and basic collection services to researchers on campus, and there is a fee structure for collections requiring more storage space. To support complex collection services, the group faces the same limitations and possibilities as the researchers that create the collections. Thus, the group participates from grant proposals with research partners, and provides services in exchange for funding staff hours. In addition, campus organizations use the data facilities as a dark archive for annual fees, which are used to purchase hardware. TACC's cyberinfrastructure intends to surpass the uncertainties of future research funding by embracing the notion that if a collection is built soundly, it will be used and supported, or it can be easily transferred to other archives or managed within other systems. The data storage facility is now entering its 4th year in production and is planned to grow by at least 5 Petabytes of capacity over the next year. There are currently ~500TB of data stored on Corral, 43TB are under iRODS management, and over 40TB of data are in MySQL databases alone. We also have 52TB of data on Ranch, the majority replicas of the data under iRODS management.

5. DISCUSSION AND FUTURE WORK

Acknowledging the evolving nature of data collections, of the research process, and of computing technologies, we present cyberinfrastructure that supports the development and management of sustainable collections. Conducting collection activities within a consistent and flexible data-intensive environment, streamlines the research workflow and enables the implementation of unique functionalities that enhance collections use and reuse. Importantly, it protects the data during those processes and beyond, and eliminates issues of scale for conducting these processes.

To handle the growing number of collections and to better accomplish general curation activities we are developing a collection's catalogue. The database schema is based on Data Documentation Initiative (DDI) and other metadata standards, as well as on elements that we created specifically to trace events (such as those governed by rules and others) and services. The catalogue will also provide an interface to fulfill collections agreements.

The limitations to this cyberinfrastructure are not technical but administrative. Not a library or an archive with the mandate to acquire and preserve collections indefinitely, TACC can provide long-term preservation services for as long as there is an agreement with the collection's curator. For example, the Institute of Classical Archaeology indicated that when and if the Institute cannot support the collection, it should be transferred to the custody of UT General Libraries.

And yet, while libraries and archives are acquiring data collections, they don't have yet the cyberinfrastructure for evolving collections nor the capabilities to host and service very large ones. We are currently collaborating with the UT Libraries and UT System to combine our mutual capabilities in service of the long-term preservation of evolving research collections.

6. REFERENCES

[1] National Science Board. 2005. Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century. National Science Foundation. Retrieved 7/15/2011 from, http://www.nsf.gov/pubs/2005/nsb0540/

[2] National Science Foundation. Dissemination and Sharing of Research Results. Retrieved 7/15/2011 from: http://www.nsf.gov/bfa/dias/policy/dmp.jsp

[3] Esteva, M., Trelogan, J. Rabinowitz, A., Walling, D. Pipkin. S. 2010. From the Site to Long-Term Preservation: A Reflexive System to Manage and Archive Digital Archaeological Data. Proceedings of the IS&T's Archiving 2010 Conference, June 1-4, 2010, Den Haag, The Netherlands. Retrieved 7/15/2011 from: http://www.imaging.org/IST/store/epub.cfm?abstrid=43763

[4] McKemmish, S. 1997. "Yesterday, Today and Tomorrow: A Continuum of Responsibility." In Records Continuum Research Group, Retrieved 7/15/2011 from: http://www.sims.monash.edu.au/research/rcrg/publications/recordscontinuum/smckp2.html

[5] DuraCloud. Legacy Documentation. Retrieved 7/15/2011 from:https://wiki.duraspace.org/display/DURACLOUD/DuraCloud+Legacy+Documentation

[6] MacKenzie S. 2010. Managing Research Data at MIT: Growing the Curation Community one Institution at a Time. Keynote IDCC 10, 6-8 December 2010, Chicago, USA. Retrieved 7/15/2011 from: http://www.vimeo.com/17662208

[7] Archaeology Recording Kit. Retrieved 7/15/2011 from: http://ark.lparchaeology.com

[8] Specify Software Project. Retrieved 7/15/2011 from: http://www.specifysoftware.org

[9] iRODS. Data Grids, Digital Libraries, Persistent Archives and Real time Data Systems. Retrieved 7/15/2011 from, https://www.irods.org/index.php/IRODS:Data_Grids,_Digital_Libraries,_Persistent_Archives,_and_Real-time_Data_Systems

[10] Robert McDonald, Chris Jordan, et al..2011. An iRODS Based Data Replication Service for Institutional Data Curation. Poster presentation at the 6th International Conference on Digital Curation, 6 – 8 December in Chicago USA.

[11] Dubrow, A. 2009. Urgent Computing Aids Haiti Relief Effort. Retrieved 07/06/2011, from: http://cms.tacc.utexas.edu/news/feature-stories/2009/urgent-computing-aids-haiti-relief-effort/

[12] FITS, File Information Toolset. Retrieved 7/15/2011 from: http://code.google.com/p/fits/wiki/tools

[13] Walling, D. Esteva, M. 2010. Automating the Extraction of Metadata From Archaeological Data Using iRods Rules, 6th IDCC 10, 6 – 8 December, Chicago USA. To be published in the International Journal of Digital Curation.