**AN EVOLVING MODEL TO MANAGE AND PRESERVE RESEARCH DATA COLLECTIONS AT THE TEXAS ADVANCED COMPUTING CENTER (TACC)**

Maria Esteva, Chris Jordan, Tomislav Urban, David Walling and Weijia Xu.

**Introduction:** In 2008 the Texas Advanced Computing Center (TACC) at the University of Texas at Austin 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 their research collections. The group developed a conceptual model to grasp with the diversity of data, functionalities, and requirements presented by the collections. The “evolving collections” model, allows mapping curatorial and technical tasks to the research stages during which collections are created and analyzed, and beyond when they are published, used, and preserved.

**Evolving Data Collections Model:** While research is conducted, raw data from experiments, observations, and simulations are transformed during analysis and for publication at the same time that new data is being gathered, incorporated into subsequent cycles of analysis, and curated. Without an adequate data management plan, these juxtaposed processes may generate collections in which the relationships between the data and the research stages become blurry. In turn this creates research bottlenecks, and future access and reuse of data may become unfeasible. As collections grow in size and complexity, so does the computing infrastructure needed to support them. Along with the collections, the technologies used to create and study them evolve at a fast pace.

Two general models have emerged in data preservation: 1) the centralized model, in which a repository handles a collection after it is finalized and consigned in standard, archival formats; and 2) the decentralized model, in which research projects curate and give access to their own data. Neither is particularly well suited to address the transformations and technical challenges digital data undergo between data creation/gathering, analysis, and their final publication. The centralized repository 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 for the researchers to accomplish. The challenge resides in managing the collection so that the distinction between raw and curated data is seamless and transparent, facilitating the research process and subsequent data access and reuse.

Differently, the evolving collections model makes available a secure and state of the art environment where researchers can develop collections on an ongoing basis. Leveraging the expertise and technical supercomputing resources available at TACC, we help researchers—with more or less technical expertise—focus on creating sound collections without having to deal with day-to-day systems administration and updates. In the following sections we describe the multidisciplinary team, and the infrastructure and services in place.

**Team:** The DMC group at TACC designs, builds, and maintains data-management and storage resources, and consults with collections’ creators in all aspects of the data lifecycle, from creation to long-term preservation and access. Group members are specialized in Relational Database Management (RDBMS), Geographical Information Systems (GIS), scientific data formats, metadata, large storage architecture, archiving, and long-term digital preservation.

**System’s Architecture:** Corral, the storage and archiving system supporting collection services, 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, a high-performance parallel file system, 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 TACC’s
computational resources, enabling mathematical and visual analysis of petabyte-scale datasets. Corral is configured to provide 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. For users requiring server-based GIS storage, we run ESRI’s ArcGIS Server on our Windows nodes with ArcSDE providing spatial extensions for SQL Server, or PostGIS on top of PostgreSQL.

In Corral, archival collections are managed within iRODS. Off site replication is done in Ranch, a Sun Microsystems StorageTek 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. The close maintenance, parts replacements contracts, and frequent schedule of upgrades of the systems are based on the supercomputing administration model that assures 24/7 service, data security, and good performance.

**Collection Services and Best Practices:** Users may request storage allocation and services to build, manage, and or archive their collections, or disk space to store their data while conducting analysis. Collections that do not involve large computational tasks may also be stored for periods decided by their creators. The allocations model is based on the supercomputing services in which users request allocations to run large data analysis jobs on HPC systems. Allocations are renewed on a yearly bases including or revisiting the collection services needed. Members of the DMC group manage access to the system, consult with researchers, and install software and dependency libraries as needed by a particular collection’s architecture. While varied services are available, the goal is to generalize them across as many collections as possible. For example, we currently maintain two domain specific databases ARK and Specify to support Archaeology and Natural History collections.

Data transfer to the system is achieved from various interfaces and ingest tools, and command line and UI interfaces allow users to query the iRODS metadata catalog for data of interest. The case of the iPlant Collaborative, which brings data from standard plant genetic repositories as well as user submitted data, is illustrative. Services involve developing applications to support large data ingest into iRODS, and a number of web services sitting in front of it to make the data accessible to and from different analysis workflows. In this 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.

We also provide consulting assistance in data management, metadata implementation, and RDBMS integration. We stress the need to document the collections architectures and help conducting collection surveys and designing record-keeping systems. To facilitate large collection processes we have implemented a variety of automated services. For image collections such the University of Alaska Museum’s Herbarium, processing scripts perform large-scale image reformatting and OCR of labels as Museum curators upload them. Preservation services for collections stored in iRODS include rules to generate file checksums, automatic off site and geographical replication, and extraction of PREMIS and MIX encoded metadata using metadata extraction tools. To capture the collection’s descriptive information we extract DC metadata from the collection’s record-keeping system and register it with the iRODS metadata catalog. All metadata is packed as a METS document and stored in iRODS.

**Curation:** In this context, curation activities happen in parallel at the domain specific and the general collection levels. At the domain specific level, researchers as curators gather, analyze, edit, and interpret the data that forms the collection throughout the research lifecycle. They also determine when their entire collection or its portions are finalized, and the way in which they want to provide access. Depending on the needs, our group may simply provide a robust
environment (hardware and software) in which researchers build and or experiment with their collections. This is the case with the Center for Predictive Engineering and Computational Sciences. PECOS research involves running large number of simulations with different parameters and comparing and contrasting the results. The project uses Corral to store the different simulation packages including the model and the output data for validation purposes. In this case, the domain scientists trace and document the provenance of data, models, experiments, and results within a system that provides a seamless transition between data storage, rapid access to computation resources, results evaluation, and back to storage.

DMC members are also involved in domain specific curatorial tasks. Such is the case with the collection of the Institute of Classical Archaeology, which uses iRODS, RDBMS and GIS services in Corral. In collaboration, we developed a metadata integration system that allows identifying the relationships between multiple data pieces belonging to objects found in a same archaeological (physical) context. Implemented as iRODS rules, the system automatically gathers information from the recordkeeping system in which the files are stored at ingest, from third party metadata extraction tools, and from the ARK database in which specialists input knowledge and interpretation information. The final collection’s metadata structure preserves the site’s documentation according to archaeology best practices allowing users to search for an object and gather all the contextual data.

General collections curation activities include establishing agreements with researchers, and tracking and managing the collections and the services with which the group is involved. To better accomplish these tasks we are developing a collection’s catalogue, based on DDI and DC 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.

**Business Model:** As part of a service and research organization at the University of Texas at Austin, we offer a base of free storage space and basic collection services. To support more complex services, we face some of the same limitations and possibilities as the researchers that create the collections. Like them, we need to obtain grant funding and or provide services in exchange for staff hours and hardware purchases. As collections evolve into stable architectures, our relationship with researchers change from a very involved one to provide basic services that can be attended via TACC’s user consulting system. This has been the case with ICA, which after three years of intense collaboration now manages most of its collection’s functions. The evolving model 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. Our system is now entering its 4th year in production and is planned to grow by at least 5 Petabytes of capacity over the next year

**Conclusion:** The evolving model enables building and archiving collections in a continuum, while performing data analysis and visualization tasks across storage and computing environments. The flexibility of the model contrasts with that of institutional repositories, in which static collections are deposited for good after they are finalized and publishable. Instead, this model has a post-custodial flair. Researchers at any stage of their collections’ lifecycle may require assistance on an array of digital curation tasks and technical implementations, while remaining their owners and curators. The model stresses automating services and the possibility to use state of the art technology to provide analytical access. Across the differences in services and needs of the collections, the system offers a robust infrastructure including systems administration, data authenticity checking and replication, and network security.
References:

1 This section contains parts of previous articles written in collaboration with the researchers of the Institute of Classical Archaeology. See:
   Jessica Trelogan, Adam Rabinowitz, Maria Esteva, Stephen Pipkin, David Walling

   Maria Esteva, Jessica Trelogan, Adam Rabinowitz, David Walling, Stephen Pipkin.
   http://test.imaging.org/ScriptContent/store/epub.cfm?abstrid=43763
3 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.
4 ARK, Archaeological Recording Kit, (Version 0.6). Retrieved from http://ark.lparchaeology.com