Final Narrative Report
An NHPRC-funded Geospatial Records Project

"California’s Geospatial Records: Archival Appraisal, Accessioning, & Preservation"

October 1, 2010
Project Overview

The eLegacy project has two interrelated goals necessary to move forward both the Electronic Records program in the California State Archives specifically, and research in electronic records in general. We experimented with hardware and software infrastructure at the California State Archives (CSA) to preserve California's geospatial records stored by the California Spatial Information Library (CaSIL) as part of the California Resources Agency's California Environmental Resources Evaluation Systems (CERES) office. We tested the emerging preservation infrastructure initially developed at the San Diego Supercomputer Center (SDSC) through the NHPRC's funded Persistent Archives Testbed (PAT), which was based on the Storage Resource Broker or SRB.

The NHPRC Persistent Archive Testbed (PAT) (see: https://www.irods.org/pubs/DICE_library-trends.pdf, 2005), linked archives across multiple state institutions. A single data grid was used to link the sites with the metadata catalog maintained at SDSC. Each site used a 1–2 terabyte commodity disk system to house a local copy of its preservation collection. A replica was also kept on tape at SDSC. Each site preserved a different type of material and investigated how to automate the archival procedures that are used for its collection. The sites and types of collections were: Kentucky Department for Libraries and Archives (Web collection), Minnesota Historical Society (spatial data records on land use), Ohio Historical Society (email collection), Stanford Linear Accelerator Center Archives and History Office (high-energy physics project), and the Michigan Department of History, Arts, and Libraries (records stored in a records management application). The focus of the PAT project was on automation of archival processes of appraisal, accessioning, arrangement, description, preservation, and access. While all of the collaborating sites examined how description could be automated, each partner selected a different set of preservation processes to automate. The approach followed in the PAT collaboration was to first put the material to be archived under archivist control by loading it onto the data grid, then developing processing scripts that allowed the organization and description to be characterized. Scripts were developed to extract the preservation metadata and to organize the digital entities. The scripts were then applied to create a new collection within the data grid that provided the appropriate structure and metadata. The digital entities were then replicated onto a tape archive at SDSC using container technology.

The eLegacy project demonstrates how these distributed techniques can be put to use across state government agencies. The scripts we developed in PAT are now expressed as workflows or rules within the enhanced iRODS system, which supports an integrated rule-based engine. This infrastructure was subsequently refined at the University of North Carolina at Chapel Hill and upgraded to iRODS (integrated Rule-Oriented Data System), and tested within the broader technological structure of the state of California. This preservation infrastructure is based on “data grid” and “persistent archives” technology, a type of “data cloud” technology. We initially deployed technology at three sites (CaSIL, CSA, and SDSC) and later on between CaSIL, CSA, and the Renaissance Computing Institute (RENCI) at UNC Chapel Hill when the project was transferred there. This project allowed us to test the distributed and federated management of electronic records transferred from CERES to CSA and replicated at RENCI.
As part of the project, the State Archives also appraised and accessioned the archival geospatial records held by CaSIL into CSA. This accessioning constituted the first structured receipt of electronic records by CSA. We briefly experimented with data conversion and migration of geospatial records, however, this proved to be beyond the scope of our project and an intractable problem for the time being, as many popular formats tend to be proprietary. We experimented with geospatial RSS feeds as a mechanism for transmission of geospatial records and developed Web2.0 archival tagging approaches that automatically trigger iRODS data grid workflows without the need for deep technical knowledge from the appraising archivist or the use of specialized interfaces.

While we were completing eLegacy, the NHPRC-funded DCAPE project (Distributed Custodial Archival Preservation Environments) was under way (see http://dcape.org). DCAPE is building integrated trusted digital repository preservation services on top of iRODS rules. Many of the DCAPE workflows are applicable to the types of automation needed in eLegacy. As such, eLegacy leverages many projects and performance objective #6 (“monitor copies, track changes, and preserve authenticity during the accessioning process”) and the iRODS framework provides most of the hooks needed to implement these tasks.

Finally, a consulting study proposed a methodology for developing organizational support and participation in establishing federated preservation networks. The study was conducted through interviews of key stakeholders and addressed the often neglected, yet crucial aspect of political and people adoption, a topic we had not built into our initial proposal but felt compelled to address in the end, thus covering both technological and cultural aspects of distributed and federated records networks. We hope this study offers valuable insights on distributed and automated processing infrastructure and its potential adoption and impact within state government.

We wish to thank NHPRC for sponsoring this study and for its support of eLegacy’s complex lifecycle as the project migrated between institutions and principal investigators (Lucy Barber in 2005 – who left in 2006 to join NHPRC, Nancy Lenoil-Zimmerman in 2006, CA State Archives Director, Richard Marciano in 2007 at SDSC, then at UNC Chapel Hill in 2008), as positions, availabilities, and academic landscapes shifted. If anything, this represents the human corollary of the lifecycle of e-records © and our partnering with the consultants on studying the organizational and participatory adoption of distributed infrastructure within State Government represents another important facet of the significance of integrating human and technological processes.
Performance Objectives:

**Performance Objective #1.** Submit complete reports and grant products.

Grant products include:
- Project website: [http://salt.unc.edu](http://salt.unc.edu) (click on Governance and then on e-Legacy)
- Final Report: this report
- A model for the development of new Web 2.0 archival appraisal tools (“archival crowdsourcing”) that seamlessly trigger preservation workflows. See movie at: [http://salt.unc.edu/eLegacy/docs/mov/mov.htm](http://salt.unc.edu/eLegacy/docs/mov/mov.htm) (also linked to the Final Report tab)
- Outcome of performance objective #2: Establishment of hardware and software infrastructure between the California State Archives, CERES, and UNC Chapel Hill
- Outcome of performance objective #4: “Appraising Geospatial: Strategies and Guidelines”, see “Final Report” website tab
- Partial outcome of performance objective #7: Consulting Report on devising a methodology for developing organizational support for establishing federated preservation networks, see “Final Report” website tab

**Performance Objective #2.** Establish the necessary hardware and software infrastructure to preserve geospatial records stored by the California Spatial Information Library.

Data grid hardware and software were installed in several branches of the CA State Government: at the California State Archives (CSA), and at the CA Natural Resources Agency (CERES), with a third remote node initially at the San Diego Supercomputer Center at UC San Diego, and eventually at the Renaissance Computing Institute (RENCI) at UNC Chapel Hill. These three nodes form a distributed storage network or “data cloud” where geospatial records can be transferred to, managed, and replicated.
UNC staff initially installed the Storage Resource Broker software and loaded records and eventually upgraded the data grid software from SRB to iRODS. All servers were updated at the CA State Archives and at CERES with the assistance of IT staff at the State of CA. An iCAT catalog at UNC/RENCI was also set up.

- Hardware installation at CSA and CERES consisted of two 4TB grid bricks. These were shipped in August 2007 and installation was fine-tuned in the fall of 2007.
- Further exploration for using existing underground fiber channels to connect the two buildings at the State of California (CaSIL, CSA) for high-speed transfers between the two buildings was conducted by David Harris. This plan, while attractive, proved to be too ambitious for the scope of this project.

**Performance Objective #3.** Using results of the Persistent Archives Testbed (PAT) project, identify, install and test hardware and software resources necessary to set up a “geospatial archivist grid” distributed across appropriate state agencies.

In the PAT project, an archivist grid was set up by putting storage boxes at the various state archives (Michigan, Minnesota, Ohio, and Kentucky) and Stanford University. A shared registry or metadata catalog was set up at SDSC with remote storage there as well. Scripts were written that helped with the automation of archival processing (appraisal, accessioning, arrangement, description, preservation, and access). This approach ended up being cumbersome as the scripts were external to the SRB data grid and had to be developed and maintained by technical staff.

In the case of the “geospatial archivist grid” we set up, we had the advantage of upgrading the software environment to iRODS. This allowed the scripts to now be part and parcel of the data grid infrastructure. Workflows or rules in iRODS reside in the data grid and are stored and executed at each remote node (CSA, CaSIL, or RENCI in our case). This also allows for more flexible triggering of the rules. We chose to use Web2.0 external triggers and developed a transmission approach based on the use of geospatial RSS feed transfers.

RSS is a standardized XML file format for content providers to publish their content. It is also a web feed format. CaSIL experimented with the availability of geoRSS feeds to publish the availability of new geospatial datasets.

An example of a geospatial RSS feed can be found in the next diagram:
Fig. 2: an example of a CaSIL geoRSS feed for Digital Raster Graphics

Fig. 3: an RSS reader client application (in this case Google Reader)
A user of CaSIL’s system could choose to make use of CaSIL/CERES geospatial services, by subscribing to new geospatial RSS feeds and choosing to view / manage them through a feed reader such as Google Reader. In Fig. 3, the user has subscribed to three Digital Raster Graphics (DRG) data feeds of varying resolutions. The RSS reader allows the user to tag individual feeds and share them with other friends in his / her social networking space of friends. In this instance, we demonstrate the use of a tag called “preserve”. The scenario we considered is one where “circle of friends” could be accessioning archivists from different groups for example, working together (in a crowd sourcing fashion) to co-tag geospatial feeds. When a threshold has been reached or voting consensus (e.g. 4 out 5 archivists deem the records to be preservation-worthy), the records are flagged as “Preserved”. In the example below, there are two friends and the first feed has been tagged “Preserved” (in this instance, it requires a unanimous vote).
The flagging of records as “Perserved” would be the trigger that would initiate a data grid workflow or rule. One of the “friends” in this network is an iRODS automated process, which just listens for a voting consensus to be reached. As soon as a decision is reached, an automated iRODS workflow is invoked. This workflow, in our test implementation, automatically parses the XML RSS feed, downloads the records to the CERES node, replicates them to the CSA node, and adds a third copy on the RENCI node, and decorates those records with the parsed metadata from the RSS feed (the iCAT catalog gets those updates through user-defined AVU triplets). The workflow could be arbitrarily complicated though and produce any other number of side effects.

Our approach is essentially a type of mashup, where the combination of RSS + Feed Reader + Data Grid, puts into motion a preservation workflow process.

Fig. 4: Mashup approach
Fig. 5 summarizes the flow of command. Note that the sharing and tagging is asynchronous and distributed and that the iRODS preservation process is completely decoupled from the Web2.0 social tagging design.

**Performance Objective #4. Appraise CaSIL’s geospatial records and identify an item and a series of records.**

CSA and CaSIL staff identified a final list of geospatial records to be transferred into the iRODS data grid. The initial list comprised:

- County Boundaries (1:24000) [http://gis.ca.gov/meta.epl?oid=21384](http://gis.ca.gov/meta.epl?oid=21384)

The State Archives accessioned additional series of records, from modified set of appraised records. The main obstacle was the absence of readable records from both the original and modified list. Additional series were selected for accessioning based on the same appraisal criteria.
assessments, and for our purposes, those records that could be accessed were taken into account. A more complete study is provided from the project website Final Report tab, at: http://salt.unc.edu/eLegacy/docs/FinalReport/appraisal_of_geospatial_records.pdf

**Project staff explored access interfaces for data stored on the distributed architecture. Several candidate web browsers were considered**

This tool replaced the iRODS Rich Web Client used earlier in the project. A sample workflow was examined, including adding metadata and replicating the receiving process at the State Archives. The metadata template selected for the additional datasets resembled that information that is captured when a non-electronic record is first received/accessioned at the State Archives. This process was compared with the generalization of the RSS feed approach studied earlier in the project.

**CSA and UNC staff examined how lessons learned translate to other electronic records scheduled for transfer to CSA**

While the experience of doing research and design on the micro level has been valuable, given CSA’s resources taking on the role of observer in the future seems to be the most effective and beneficial strategy. Participation in the DCAPE project has allowed CSA to submit input and expertise from an archival perspective, without the need to stress their internal and external sources.

**Performance Objective #5. Following experimentation, adopt a reliable format standard for geospatial records to be preserved by the California State Archives, develop a set of transmission requirements for geospatial records, and utilize these requirements to transmit selected archival records series from CaSIL to CSA.**

We had many discussions on appropriate geospatial formats. This is a challenging topic as formats span a wide range from proprietary to open source. We also experimented early on with Safe software (http://www.safe.com), which allows for over 250 spatial formats to be converted.

**CSA staff accessioned records into the iRODS collection management system, including the additional metadata determined to be required for geospatial data.**

The California State Archives accessioned a series of geospatial records held at CaSIL/Department of Resources. The series was selected based on one of four appraisal approaches the State Archives have used over the history of the project. This is further documented in the Archives’ document entitled “Appraising Geospatial Records”.

The records were accessioned from the Department of Resources via a hard drive and directly downloaded to the node located at the California State Archives. Those records are then part of the grid and can be replicated in a seamless fashion.
The California State Archives accessioned/downloaded an additional geospatial dataset from Cal-Atlas/CaSIL, a geospatial clearinghouse of geospatial records administered by the Department of Resources and transferred a copy to the node located at RENCI via the iRODS Rich Web client browser. The records were transferred using a web browser provided by RENCI. California Archives staff added sample metadata to the ingested record. CSA staff is still examining the issue of additional metadata required for geospatial records.

CSA and UNC staff document the data sets stored at CSA and replicate with iRODS at an additional location within the state’s infrastructure and at UNC. Replication explored automated parsing of XML RSS feeds through iRODS policies based on additional requirements developed with federation partners. This represents a generalization of the RSS feed approach studied earlier in the project.

We are looking at adding an iRODS replication rule that can be associated to a watched folder (whenever a file is accessioned into that folder replication can occur automatically). RSS feeds were used earlier for accessioning.

One example which may be useful is the site that is acting as the prototype for the coming state data site similar to the data.gov federal site. The testbed is at data.sgc.ca.gov. Like a lot of government sites the Strategic Growth Council web site (www.sgc.ca.gov) has a data link, in their case near the top right. In the future most California state sites may have this type of link to their raw data and document catalog.

When people do searches of the data.sgc.ca.gov site the results of their search contain an RSS feed link near their upper right and corner. While some of the details are still being worked out, the idea is that the link behind the orange RSS feed icon is the search which was just executed. If a person adds that link to their feed reader and new records are added which meet their search criteria those records show up in their RSS feed. An archivist could watch that feed and evaluate the new items.

**Performance Objective #6.** Use the Storage Resource Broker (SRB) to monitor copies, track changes, and preserve authenticity during the accessioning process.

We are no longer using SRB and as mentioned earlier this performance objective is being handled through the NHPRC-funded DCAPE project (see [http://dcape.org](http://dcape.org))

**Performance Objective #7.** Establish the necessary requirements for the description of and authentic access to these geospatial electronic records, and identify how the technological infrastructure, programmatic and Archives’ employee skills developed for this project can provide models for the appraisal, accessioning, and description of other electronic records.
The contractor conducted interviews, analysis and documentation of data architecture and iRODS rule-driven record producer use requirements to establish a model federated preservation network. Starting with the Inter-agency Ecological Program and spreading out to their data exchange partners, the consultants mapped the needed data exchanges between the partners, including any intersections in their digital data preservation requirements and interests. This is preliminary to developing the model data exchange agreements supporting possible development of a state-wide federated digital preservation network.

This final study focused on developing a methodology for developing organizational support for establishing federated preservation networks.

When we completed the design of the technology to support federated digital preservation of geospatial data, the remaining crucial question was how we would deal with the people problem. Solid technology was not enough. What we needed was a repeatable process for bringing organizations together to participate in collaborative digital preservation networks. In order to accomplish this we initiated a research project seeking to develop a methodology for documenting the information exchange and management needs of a group of organizations causing them to recognize the tremendous value of federated data management and the need for federated data preservation as part of that effort. Instead of attempting to have digital preservation fight for a piece of these organizations’ shrinking resource pie, the intent was to expand the scope of the data management analysis enough to expand the pie making digital preservation a natural benefit of a better understood and more valuable total solution.

David Harris from CERES concluded: “We are extremely happy with the result of this effort. Not only did we end up with a model for federated data exchange network development but we were able to watch the application of this approach shift the perception of the partners in the subject group causing them to have a new understanding the possibilities for their work. The subject initiative was the California Inter-agency Ecological Program and we are continuing to work with them to move them toward a shared service model including digital preservation which meets their particular needs.”