Community-Driven Development of Preservation Services

Funded Project Staff listed in Red and Blue

INTEGRATION & BUS DEV

UNC
- SALT
  - Richard Marciano
  - Chien-Yi Hou
- CDR
  - Dave Pcolar ++

POLICY / RULE DEV

West Virginia University
- Donald Adjeroh
- Frances Van Scoy

RENCI
- Leesa Brieger ++

DICE
- Michael Conway ++
- Reagan Moore
- Antoine de Torcy ++

UNC Libraries
- Steve Barr ++
- Greg Jansen ++

UNC Res. Comp. Svcs
- Bill Schulz ++

SILS Grad. Student team
- Heather Bowden ++
- Alex Chassanoff ++
- Christine Cheng ++
- William Miao ++
- Terrell Russell ++
- Jewel Ward ++

UNC CS Grad. Student team
- Tao Yu ++
- Hao Xu ++

STATE ARCHIVES & LIB

Michigan
- Caryn Wojcik
- Mark Harvey

North Carolina
- Kelly Eubank
- Jennifer Ricker ++
- Amy Rudersdorf ++
- Lisa Gregory ++
- Ed Southern --
- Megan Durden --
- IT
- Dean Farrell ++
- Druscie Simpson
- David Minor
- Chris Black --

Kentucky
- Glen McAninch
- Mark Myers ++

Kansas
- Scott Leonard

New York
- Bonnie Weddle
- Michael Martin ++
- Ann Marie Przybyla

California
- Chris Garmire
- Nancy Lenoil-Zimmelman
- Linda Johnson --
- Laren Metzer
- Renee Vincent-Finch --

UNIVERSITY ARCHIVES

Tufts University
- Eliot Wilczek
- Veronica Martzahl ++
- Anne Sauer

UNC Chapel Hill
- Will Owen ++
- Rich Szary ++

CULTURAL INSTITUTIONS

Getty Research Institute
- Joseph Shubitowski
- David Farneth
- Leah Prescott
- Sally Hubbard --
- Mahnaz Ghaznavi --
- Karim Boughida --

Smithsonian Institution Archives
- Riccardo Ferrante ++

SCHOOLS OF LIB & IS

UNC Chapel Hill
- Cal Lee ++

University of Wisconsin-Madison
- Kristin Eschenfelder ++

Legend	Collaborator Roles
Red	Funded
Blue	Cost-sharing
Brown	"Observer"
Black	None of the above
++	Added after project funded
--	At new institution

Abstract

This paper describes the first phase of the *DCAPE* project and the lessons learned in articulating a community-based development approach for preserva–tion services. The *"Distributed Custodial Archival Preservation Environments"* project*, DCAPE,* was funded by the National Historical Publications and Records Commission (NHPRC) in 2007, in a call for proposals for *"cooperative networks and service providers' projects."* The NHPRC's goal was to encourage the creation of e-records storage, preservation, and access services, and to promote sustainable business models. *DCAPE*'s approach proposed to develop a framework to support *institution-specific preservation policies* (including business models) while providing the economy of scale needed for a cost-effective service. The focus of this paper is on the community-driven nature of the preservation services development process.

Index Keyword Terms—Preservation Services, Trusted Digital Repositories, Policy Management, *DCAPE*, iRODS, SALT

1. Introduction

The goal of the *DCAPE* project is to build a distributed production preservation environment that meets the needs of mid-to-large-sized archival reposi–tories, libraries, and cultural institutions for trusted archival preservation services. The preservation environment builds upon the technologies developed at the University of North Carolina–Chapel Hill (UNC) Renaissance Computing Institute (RENCI) and the data storage infrastructure being installed there. The environment includes a trusted digital repository infrastructure that is assembled from a rule-based data management system, commodity storage systems, and sustainable preservation services. The software infrastructure automates many of the administrative tasks associated with management of archival repositories, including validation and trustworthiness.

Our proposal involves the collaboration of multiple "medium-scaled" preservation communities with the explicit goal of defining the common set of services needed by all participating institutions (state archives and libraries, university archives, cultural institutions, etc.), and the unique set of services that must be tuned to specific mandated policies at each site.

The original NHPRC grant called for the develop–ment of cooperative institutions to provide electronic records preservation services to repositories. A single award of up to $400K was to be made but in the end two awards were granted, one to the Emory-based MetaArchive project for $300K and another to the UNC-based *DCAPE* project for $258K.

- MetaArchive aims to develop a sustainable digital preservation service for cultural and historical records and a cost-model for providing preservation services based on the Lots of Copies Keep Stuff Safe (LOCKSS) model. In addition, the goal is to integrate LOCKSS with the Storage Resource Broker (SRB) and the Integrated Rule-Oriented Data System (iRODS) data grid technologies, developed by members of the DICE group at UNC Chapel Hill.
- *DCAPE* aims to develop a sustainable digital preservation service for state and university archives and other repositories, and a cost model for providing distributed and customized preservation services based on the iRODS model. The approach allows for the customization of services based on the profile of the archives or collections.

The innovative *DCAPE* approach intends to develop sets of machine-actionable preservation policies, but allow individual communities to customize the behaviors of these policies. Given the limited level of project funding, a collaborative and community-development approach has emerged, as demonstrated by the impressive list of participants and contributors in the project so far. The focus of this paper is on the community-driven nature of the preservation services development process.

2. A Sustainable Development Approach

Beyond the funded project staff, others have participated in conversations and meetings around the project., accounting for some 60 people! This is a reflection of *DCAPE*'s development philosophy of establishing a systematic and sustainable development partnership. We wish to reflect on several aspects of sustainability: (1) NHPRC's sustained investments in building collaborations, as demonstrated by the agency's funding agenda over the last twelve years, (2) the leveraging of community development when funds are limited, and (3) the sustainability of projects beyond the initial funding.

2.1 NHPRC's Sustained Funding in e-Records

NHPRC funded projects have been seminal in initiating and sustaining conversations between technologists and archivists over the last decade. Richard Marciano, principal investigator on *DCAPE*, has been privileged to participate in a series of NHPRC-funded projects starting in 2000 with the Archivists' Workbench.

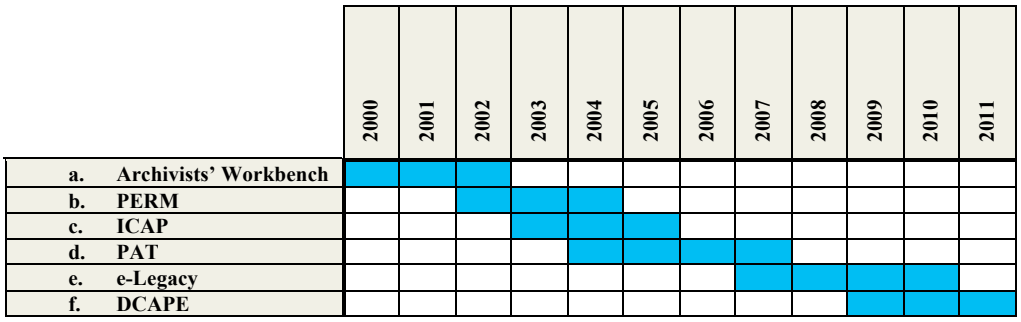

Figure 1: NHPRC-funded projects leading to *DCAPE* at SDSC and UNC (Richard Marciano, PI)

DCAPE participants involved in these earlier projects include: Chien-Yi Hou (ICAP, PAT, e-Legacy), Reagan Moore (PAT, e-Legacy), Caryn Wojcik (Archivists' Workbench, PERM, PAT), Glen McAninch (PAT), Chris Garmire (e-Legacy), Nancy Lenoil-Zimmelman (e-Legacy), Linda Johnson (e-Legacy), Laren Metzer (e-Legacy), Renee Vincent-Finch (e-Legacy), Mahnaz Ghaznavi (PAT), and Karim Boughida (PAT).

The Archivists' Workbench (2000-02) was a three-year project conducted at the San Diego Supercomputer Center at the University of California, San Diego that focused on long-term preservation of and access to software-dependent electronic records. This project featured an archival advisory board consisting of many luminaries in the field: Ken Thibodeau (NARA), Theodore Hull (NARA), Bruce Ambacher (NARA), Phil Bantin (Indiana University), Charles Dollar (UBC), Pat Galloway (UT Austin), Anne Gilliland (UCLA), Peter Hirtle (Cornell), Heather MacNeil (UBC), Tom Ruller (NY State Archives), Lee Stout (Penn State), and Caryn Wojcik (State Archives of Michigan), with technical coordination by Mark Conrad (NARA) and Peter Bloniarz (SUNY at Albany). The input from these experts was significant and had a lasting impact.

Subsequently, the Preserving the Electronic Records Stored in an RMA (PERM) project (2002-04), with the State of Michigan, developed and tested a model for preserving electronic records stored in a records management application that complies with the Department of Defense (DoD) Standard 5015.2. The project evaluated the DoD Standard 5015.2 to determine which features of the RMA standard needed to be retained in any future preservation model.

The Incorporating Change Management into Archival Processes (ICAP) project with UCLA (2003-05), examined the issues involved in access to and long-term preservation of active electronic records that are being changed over time by their creators. Prototypes to study the versioning of records were developed.

The Persistent Archives Testbed (PAT) project (2004-07), was a precursor of *DCAPE*. PAT brought together four State Archives: the Michigan Historical Center, Minnesota Historical Society, Kentucky Department for Libraries and Archives, and Ohio Historical Society. The project explored data grid systems to handle large archival data sets and persistent archives technologies. The project made a case for distributed custody – where records remain in the system which created them while simultaneously being in archival custody.

Finally, the e-Legacy project (2007-10), which is still active, is developing hardware and software infrastructure to preserve the state's geospatial records created by the California Spatial Information Library and managed by the California State Archives.

In addition to these undertakings, Caryn Wojcik proposed development of commercial preservation service models (Preservation-as-a-Service). This idea and the earlier NHPRC-funded projects led to the collaborative network of technologists and archivists in *DCAPE*. These projects as well as many other previous NHPRC-funded projects of *DCAPE* participants, have helped bridge archival concepts and new technological advances. The *DCAPE* project builds on and contributes to this legacy of NHPRC supported conversations between archivists and technologists.

2.2 DCAPE's Community Development Approach

The goals of the *DCAPE* project are ambitious: (1) develop a set of policy and service definitions, driven by the requirements of the underlying partners; (2) implement these services; (3) test them with partner collections; and (4) develop business models for sustaining this effort. Also important – the *DCAPE* com–munity development of rule sets using iRODS is a first and sets the standard for other communities. Moreover, *DCAPE* must meet these challenges with limited resources. The NHPRC funding covers only 15% of one programmer. A subcontract with West Virginia University also allows summer time for a graduate student. Given these lofty goals and limited resources, a community-supported development model is key.

This community-driven development model ac–counts for the nearly 60 participants since the start of the project. Some of the leveraging measures taken include (1) creating a new group called Sustainable Archives & Leveraging Technologies (SALT); (2) partnering with Dave Pcolar at UNC Libraries where the Carolina

Digital Repository, UNC's institutional repository, is being developed; (3) establishing a policy/rule development discussion team that includes programmers from the Renaissance Computing Institute (RENCI), the Data Intensive Cyber Environments (DICE) group, UNC Libraries, UNC Research Computing Services, and graduate students from the School of Information and Library Science (SILS) and Computer Science (CS); and (4) assembling additional archivists, librarians, and IT staff from all six state archives and libraries; (5) new university archives – UNC Chapel Hill Libraries; (6) new cultural institutions – Smithsonian Institution's Archives; and (7) experts from two schools of information and library science.

This approach is fraught with challenges. Beyond the limitations of funding described above, there are management challenges associated with a virtual organization where input from individuals and groups is necessary, even as they are not accountable to the grant project. For example, collaborators have come and gone over the course of the project, as indicated by the "Collaborator Roles legend" on the first page. Collaboration with students and staff funded by other grants, but producing open-source software or other services for the *DCAPE* project, raises questions about grant time accounting and ownership of cooperatively created services. The cooperative model also complicates development of *DCAPE* service models and planning of actual management of the services.

2.3 Developing Sustainable Services

A number of business models are possible under the *DCAPE* approach, from hosting services, subscription mechanisms, membership fees, packaging rule sets as business intelligence, etc. We have partnered with UNC's Business School to explore a range of approaches. "Preservation-as-a-Service (PaaS) is a potential business model that may prove viable for *DCAPE*, as the technologies involved become commodities and the costs for significant amounts of storage fall." [1].

3. Development Methodology

Two core teams have been assembled: (1) a User Community Team, made up of the archivist and librarian partners; and (2) a Policy and Rule Development Team, made up of the NHPRC-funded staff developers, and also observers and teams of students from SILS and CS.

In the first six months of the project a Wiki was established. A working group from the User Community Team conducted an assessment of capabilities from the Reference Model for an Open Archival Information (OAIS) that are relevant to the project, based on requirements from their own institutions. This led to a specification with close to 100 policies. A working subset of 26 rules was extracted for pilot work. The team developed a research testbed SLA (service-level agreement) to facilitate loading of records from the partner institutions into a testbed. An assessment of the 26 pilot rules was conducted and related to the CCSDS MOIMS-RAC Working Group's "Audit and Certification of Trustworthy Digital Repositories" draft standard that is currently being developed for submission to the International Organization for Standardization (ISO) [3]. The 26 rules were then expanded to 52, and these rules were mapped back to RAC rules (see Appendix 1).

In the second six months of the project, the Policy and Rule Development Team met weekly to interpret and map the 52 policies and map them into machine-actionable iRODS rules [2]. This team has implemented the rules using two instances of iRODS, a development testbed and a community testbed. Records from community members can be loaded according to the established SLA in the community testbed.

In the current phase of the project, both teams have come together face-to-face, and an integration team has been put together. The integration team will move to the next step of all the iRODS rules into the *DCAPE* implementations.

4. Summary

In this paper, we introduced the community-driven development methodology we are using to establish *DCAPE* preservation services. While community-development helps to overcome the deficiency of funding available for preservation projects, it also introduces complications in project management. As of early March 2010, we are at the half-way point. While much has been accomplished, much work remains. One significant accomplishment is the development of a set of community preservation rules – rules that are being created for the first time in the context of the project. The business models we aim to provide are predicated on the creation and implementation of these rules. Our experience so far points to the potential for overcoming technical barriers through persistence, flexibility, and cultivation of mutually beneficial collaborations.

5. Acknowledgements

This project is funded by NHPRC Records Projects grant NAR08-RE-10010-08, "Distributed Custodial Archival Preservation Environments", 2008-2011.

6. References

[1] J. Ward, T. Russell, A. Chassanoff, "Building a Trusted Distributed Archival Preservation with iRODS,", poster submission to the iRODS User meeting in Chapel Hill, March 24-26, 2010.

[2] iRODS: Data Grids, Digital Libraries, Persistent Archives, and Real-time Data Systems. http://www.irods.org

[3] Draft Recommendation for Space Data System Practices, CCSDS 652.0-R-1, “Audit and Certification of Trustworthy Digital Repositories”., October 2009.

Appendix I

Initial ISO MOIMS-rac Capabilities and Mapping to *DCAPE* rules
RAC No.'s are from the "Combined Annotated document" Wiki page
http://wiki.digitalrepositoryauditandcertification.org/bin/view/Main/CombinedMetricsDocumentsFollowingFaceToFace
Accessed on Sep. 2009

ISO Item	RAC No.	DCAPE Item	ISO Criteria	DCAPE Machine-Actionable Rule
1	A3.2.2 A5.1.3 A.5.1.4 A5.2		Address liability and challenges to ownership/rights.	Map from submission template to access and distribution controls
2	B1.1	DCAPE 4	Identify the content information and the information properties that the repository will preserve.	Define templates that specify required metadata and parameters for rules that are required to enforce properties
3	B1.1.2		Maintain a record of the Content Information and the Information Properties that it will preserve.	Link submission and policy templates to the preserved collection
4	B1.3	DCAPE 3	Specify Submission Information Package format (SIP)	Define templates that specify structure of a SIP and required content of a SIP.
5	B1.4	DCAPE 1	Verify the depositor of all materials.	Ingest data through a staging area that has a separate account for each depositor.
6	B1.5	DCAPE 6	Verify each SIP for completeness and correctness	Compare content of each SIP against template.
7	B1.6	DCAPE 8	Maintain the chain of custody during preservation.	Manage audit trails that document the identity of the archivist initiating the task
8	B1.7	DCAPE 22	Document the ingestion process and report to the producer	Send e-mail message to producer when process flags are set.
9	B1.8	DCAPE 10	Document administration processes that are relevant to content acquisition.	Maintain list of rules that govern management of the archives
10	B2.1 B2.1.1	DCAPE 13	Specify Archival Information Package format (AIP)	Define templates that specify structure of an AIP and required content of an AIP.
11	B2.1.2		Label the types of AIPs.	Store AIP type with each collection.
12	B2.2	DCAPE 13	Specify how AIPs are constructed from SIPs.	Define transformation rule based on parsing of SIP template and AIP template
13	B2.3 B2.3.1	DCAPE 14	Document the final disposition of all SIPs	Maintain an audit trail for all SIPs
14	B2.4 B2.4.1 B2.4.1.1 B2.4.1.2 B2.4.1.3		Generate persistent, unique identifiers for all AIPs.	Define unique persistent logical name for each AIP

15	B2.4.1.4 B2.4.1.5		Verify uniqueness of identifiers.	Identifier uniqueness enforced by algorithm that assigns identifiers
16	B2.4.2		Manage mapping from unique identifier to physical storage location.	Storage location mapping enforced by iRODS data grid framework
17	B2.5	DCAPE 19	Provide authoritative representation information for all digital objects.	Define template specifying required representation information.
18	B2.5 B2.5.1	DCAPE 7	Identify the file type of all submitted Data Objects.	Apply type identification routine to each object on ingestion.
19	B2.6 B2.6.1		Document processes for acquiring preservation description information (PDI)	Define rule set that will be applied to extract PDI.
20	B2.6.2		Execute the documented processes for acquiring PDI.	Apply PDI rules specific to a collection.
21	B2.6.3 B2.7 B2.7.1 B2.7.2 B2.7.3		Ensure link between the PDI and relevant Content Information.	Set PDI extraction flag as part of PDI extraction rules.
22	B2.8	DCAPE 14	Verify completeness and correctness of each AIP.	Compare AIP against template for required content.
23	B2.9	DCAPE 17	Verify the integrity of the repository collections/content.	Periodically evaluate checksums and compare with original checksum value.
24	B2.10 B3.1 B3.2	DCAPE 21	Record actions and administration processes that are relevant to AIP creation.	Maintain an audit trail of processing steps applied during AIP creation.
25	B4.1		Specify storage of AIPs down to the bit level.	Identify form of container used to implement an AIP.
26	B4.1.1		Preserve the Content Information of AIPs.	Manage replicas of each AIP
27	B4.1.2		Actively monitor the integrity of AIPs.	Periodically evaluate checksums.
28	B4.2 B4.2.1	DCAPE 21	Record actions and administration processes that are relevant to AIP storage.	Maintain an audit trail of processing steps applied during AIP storage.
29	B4.2.2	DCAPE 18	Prove compliance of operations on AIPs to submission agreement.	Parse audit trails to show all operations comply with submission rule template
30	B5.1	DCAPE 24	Specify minimum descriptive information requirements to enable discovery.	Define submission template for required descriptive metadata.
31	B5.2	DCAPE 11	Generate minimum descriptive metadata and associate with the AIP.	Apply rule to extract metadata specified within submission agreement.
32	B5.3 B5.3.1		Maintain link between each AIP and its descriptive information.	Package descriptive metadata within the AIP as an XML file
33	B6.1	DCAPE 9	Enforce access policies.	Authenticate all users, authorize all operations
34	B6.1.1	DCAPE 23	Log and review all access failures	Periodically parse audit trails and

			and anomalies.	summarize access failures
35	B6.2	DCAPE 26	Disseminate authentic copies of records	Define template to specify creation of a Dissemination Information Package (DIP)
36	C1.1.2	DCAPE 15	Maintain replicas of all records, both content and representation information	Periodically snapshot metadata catalog, and maintain at least two replicas
37	C1.1.3	DCAPE 12	Detect bit corruption or loss.	Periodically validate checksums
38	C1.1.3.1	DCAPE 16	Report all incidents of data corruption or loss and repair/replace lost data	Periodically synchronize replicas, and generate and store report
39	C1.1.5	DCAPE 19	Manage migration to new hardware and media	Replicate AIPs to new storage system
40	C1.1.6		Document processes that enforce management policies	Maintain copy of the rule base and micro-services used for each collection
41	C1.1.6.1		Document changes to policies and processes	Version policies and micro-services
42	C1.1.6.1.1		Test and evaluate the effect of changes to the repository's critical processes.	Version state information attributes.
43	C1.2.1		Synchronize replicas	Periodically synchronize replicas
44	C2.3		Delineate roles, responsibilities, and authorization for archivist initiated changes	Define archivist roles and limit execution of preservation procedures to the archivist role
45	C2.4 B2.5.2		Maintain an off-site backup of all preserved information	Federate two independent iRODS data grids and replicate digital holdings
46	B2.5.3		Maintain access to the requisite Representation Information.	Manage Representation Information as metadata attributes on each record
47	B6.2.1 C1.1.1 C1.1.1.1 C1.1.1.2 C1.1.1.3 C1.1.1.4 C1.1.1.5 C1.1.1.6		Maintain and correct problem reports about errors in data or responses from users.	Parse audit trails for unsuccessful operations and design appropriate micro-service recovery mechanisms
48		DCAPE 24	Provide a search interface.	
49		DCAPE 5	Perform a virus check.	
50		DCAPE 2	Implement a loading dock.	
51		DCAPE 20	Migrate records to new formats.	
52		DCAPE 25	Create and certify Dissemination Information Packages.	