The International Journal of Digital Curation
Issue 2, Volume 2 | 2007

Digital Preservation Theory and Application: Transcontinental Persistent Archives Testbed Activity

Paul Watry,
University of Liverpool

November 2007

Abstract

The National Archives and Records Administration (NARA) and EU SHAMAN projects are working with multiple research institutions on tools and technologies that will supply a comprehensive, systematic, and dynamic means for preserving virtually any type of electronic record, free from dependence on any specific hardware or software. This paper describes the joint development work between the University of Liverpool and the San Diego Supercomputer Center (SDSC) at the University of California, San Diego on the NARA and SHAMAN prototypes. The aim is to provide technologies in support of the required generic data management infrastructure. We describe a Theory of Preservation that quantifies how communication can be accomplished when future technologies are different from those available at present. This includes not only different hardware and software, but also different standards for encoding information. We describe the concept of a "digital ontology" to characterize preservation processes; this is an advance on the current OAIS Reference Model of providing representation information about records. To realize a comprehensive Theory of Preservation, we describe the ongoing integration of distributed shared collection management technologies, digital library browsing, and presentation technologies for the NARA and SHAMAN Persistent Archive Testbeds.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. ISSN: 1746-8256 The IJDC is published by UKOLN at the University of Bath and is a publication of the Digital Curation Centre.

Introduction

A Theory of Preservation

The requirement for a theory of preservation is driven by the need to develop data and information management technology that can be used to build persistent archives. For many years, the data grid, digital library, and persistent archives community have each focused on individual aspects of the problem, mainly on the metadata-driven approach set out in the Open Archival Information System (OAIS) Reference Model[1]. The OAIS standard focuses on the ability to access and interpret records through the creation of information. However, it does not provide representation information about the preservation environment. Work has only recently begun on the development of a rule-driven approach that provides a more complete characterization of preservation processes (Moore & Smith, 2007). This approach, it can be argued, will provide a new way to migrate all preservation processes (not just metadata) onto new technologies. As a result, archivists will be able to interact with future unknown technologies and systems so that potentially any information can be interpreted and displayed, guaranteeing authenticity and integrity, over time.

A theory of preservation extends the concept of digital preservation from one that is focused on sending the records (metadata) into the future to one that can also send into the future a description of the environment that is being used to manage and read the records. The true test of a preservation environment is whether it describes the entire preservation information context sufficiently well that the records can be migrated into an independent preservation environment without loss of authenticity or integrity. This requires migrating not only the records, but also the characterizations of the preservation environment context. The new preservation environment would have to apply the same management policies, the same preservation processes, use the same logical name spaces, and manage the same persistent state information. If all of these context components can be expressed and migrated to a new preservation environment, then the preservation context is correctly described (Moore, Arcot, & Marciano, 2007).

The RLG/NARA Trustworthy Repositories Audit & Certification (TRAC): Criteria and Checklist is one of the most advanced statements of a theory of preservation. The criteria separate the preservation metadata into attributes on the storage resources, users, collections and the data, and require the ability to maintain the information context, arrangement, and descriptions of the comprehensive management of records[2]. In effect, it extends the OAIS metadata-based approach to one that can also support representation information for the preservation environment. This requires abstraction mechanisms to maintain preservation properties despite changes introduced by the evolution of technology. The required abstraction mechanisms are now being developed by the grid community and include characterizations of digital structure and semantics; characterizations of standard operations on storage repositories; characterization of management policies; and characterization of standard access mechanisms.

[1] http://nost.gsfc.nasa.gov/isoas/
[2] http://www.crl.edu/PDF/trac.pdf

In their expression of preservation requirements, the TRAC Criteria are highly influential in setting the agenda for related preservation systems and projects. Mappings include the TRAC/NESTOR catalogue criteria for trusted digital repositories crosswalk[3], the Digital Repository Audit Method Based on Risk Assessment (DRAMBORA)[4], and the iRODS rule-based data management system[5]. The TRAC criteria also informed the Digital Preservation Europe (DPE) policies in the area of preservation environment context (Hedstrom, 1991).

This paper reviews the concept of automating archival processes, focusing on the new generation of rule-based collection management systems in order to characterize the structure of records. We discuss how these systems may be used to support a theory of preservation through the application of a "digital ontology", which can be used to represent the structural, semantic, spatial, and temporal relationships inherent within a record (e.g. the context relative to its production). We discuss how the work on digital ontologies is being taken forward through the development of the digital object technology as a description language ("DFDL") and presentation tool ("Multivalent"), which applies the digital ontology in order to interpret the record. We relate this development to the research efforts now undertaken to support knowledge-based archives.

The paper is divided into three sections. The first section will discuss the fundamental research agenda required to manage the evolution of technology. The second section will conduct an assessment of currently available information management technologies that can be used to realize a theory of preservation. The third section will discuss the integration of technologies to achieve this goal in the NARA and SHAMAN preservation prototypes[6].

Fundamental Research Activities

The current research agenda is focused on defining a set of preservation processes and preservation attributes that should be managed by a preservation environment. The activities can be divided into three broad areas:

- The first ("data") focuses on the use of the data grid technologies in order to map OAIS-representation information onto the logical name space. This will provide the data and trust virtualization needed for infrastructure independence. We can currently do this by the use of existing technologies, for example the Storage Resource Broker (SRB) data grid.
- The second ("information") focuses on the characterization of preservation processes as "digital ontologies" (or representation information) that organize the relationships needed to interpret the structure and meaning of digital entities. This provides the ability to apply semantic labels to structures and identify a knowledge community that understands the labels.
- The third ("knowledge") focuses on the characterization of management

[3] http://www.digitalpreservationeurope.eu/resources/?search%5B%5D=42
[4] http://www.repositoryaudit.eu
[5] Reagan Moore has developed the characterization of the TRAC assessment criteria as rules that can be applied by the iRODS data grid.
[6] SHAMAN: "Sustaining Heritage Access Through Multivalent Archiving". Related EU projects include CASPAR: "Cultural, Artistic and Scientific Knowledge For Preservation, Access, and Retrieval" http://www.casparpreserves.eu ; PLANETS "Preservation and Long-Term Access through Networked Services" http://www.planets-project.eu

policies in terms of rules and preservation processes as standard micro-services. This will supply the ability to describe evolution of the preservation environment, and both the procedural relationships that control the application of the micro-services, and the functional relationships that comprise each micro-service. This anticipates the ability to characterize both information and knowledge content for presentation by new applications. The approach is designed to support the levels of abstractions for data, information, and knowledge management in the persistent archive.

Characterizing Data

Current approaches to digital preservation focus on the ability to access and interpret records through the creation of "representation information" (that is, the information required to render, interpret, and understand digital data) as it is defined in the Open Archival Information System (OAIS) Reference Model (ISO 14721:2003)[7]. The OAIS representation information defines the structure and semantic labels of the structures present within the record and the OAIS community maintains the ability to interpret the structural and semantic labels[8]. The OAIS Archival Information Package (AIP) defines the representation information, the Submission Information Package (SIP) defines quality assurance, and the Dissemination Information Package (DIP) defines information discovery.

The data grid community has engineered software to support the OAIS approach. Logical name spaces provide the required abstractions for managing the metadata in a federated grid environment. These include characterizations of digital structure and semantics; characterizations of standard operations on storage repositories; and characterizations of standard access mechanisms. The Storage Resource Broker (SRB) data grid is an example of software that supplies the abstraction for data sets, collections, users, resources, and proxy methods as required for OAIS metadata. The SRB supports data virtualization, or the ability to manage the properties of the shared collection independently of the storage systems where the data are located. Its grid-based approach provides a number of essential concepts for distributed data management, including data replication (or uniform access to a variety of heterogeneous, distributed storage resources including data base management systems, archival storage systems, and filestores). Multiple types of storage resources can be combined into a preservation environment. These are clear requirements for accessing, maintaining, and sharing metadata in a preservation environment.

The development of the SRB and its coupling to the OAIS reference model was primarily engineered to support the first iteration of the National Archives and Records Administration (NARA) Transcontinental Persistent Archive Prototype in Washington DC, and has informed the development of other preservation infrastructures. Experience in using the SRB software, however, has suggested that most digital preservation initiatives – including the OCLC Preservation Metadata Implementation Strategies (PREMIS)[9] , the NARA Lifecycle Data Requirements Guide (LCDRG)[10] ,

[7] http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=24683
[8] http://www.dcc.ac.uk/events/jorum-2006/JORUM_oais-08022006.ppt
[9] http://www.oclc.org/research/pmwg
[10] http://www.archives.gov/research/arc/data.html

the SHERPA Digital Preservation Project[11], Portico[12], and OAIS Reference Model (among others) – are based on the assumption that the management of preservation metadata is sufficient to maintain a complete preservation environment. In order to guarantee integrity and authenticity, we believe instead that a preservation environment needs to define how the preservation processes that are being applied today and in the future are related to the preservation processes that were applied in the past. Relying solely on the management of metadata – as defined in the OAIS Reference Model – is, in our view, insufficient to make assertions about trustworthiness. Effectively, we need to send into the future not only the information (records), but also a description of the environment (the "context") that is being used to manage and read the records (Moore, Arcot, & Marciano, 2007, p. 5).

In contrast, the Trustworthy Repositories Audit and Certification Criteria (TRAC) have asserted a more complete definition of the preservation environment that includes the preservation processes and management policies of the records. Provision of this information will allow us to quantify the allowed operations, maintain the allowed operations independently of the choice of preservation infrastructure, and track the application of the allowed operations. Mechanisms are now needed to describe both the OAIS metadata as well as the environment that is being used to manage and read the records. This will require defining the abstractions for characterizing the systemic properties about the preservation environment. All of this is beyond the capability of the SRB data grid that focuses on managing the metadata and not the rules that control the environment.

The integrated Rule Oriented Data System (iRODS) under development at the San Diego Supercomputer Center (SDSC) is designed to support the virtualization of current management policies, preservation capabilities, and persistent state information while preserving the ability to execute previous management policies. The automation of management policies will make it possible to schedule and execute processes that support information discovery and knowledge management within a preservation environment. With the appropriate data grid support, we can move from the present OAIS-based situation – that considers preservation from the perspective of standardized data formats and simple metadata mechanisms – to a higher-level conceptual model that supports the characterization of the functions that are implemented by preservation processes and the procedural rules that control the application of the processes . This is referred to as "knowledge-based archives" (Ludascher, Marciano, & Moore, 2001a).

Collectively, the goal is achieved through the integration of the rule-based collection management system (iRODS) – discussed above – with a Virtual Machine technology (Multivalent) that can be used to present and manipulate objects from the original bitstream, without the need to migrate or emulate data. We can then execute discovery or knowledge management services in the preservation environment while, at the same time, assuring authenticity and integrity of the data. The convergence of technologies and standards should lead to better support for both structural and semantic models of archived collections. This includes a "flattened" relational representation ("data"), a structured representation ("information"), and a higher-level

[11] http://www.sherpadp.org.uk/index.html
[12] http://www.portico.org/

semantic representation ("knowledge"). The goal is to maintain the ability to discover, access, and analyze digital objects while the supporting software systems evolve.

The following section discusses in greater detail the concept of information and knowledge content as it applies to the data, such as structure, semantics, context, provenance, and display properties. We argue that it is insufficient merely to copy data at the bit level from obsolete to current media, but the archivist must also create recoverable archival representations that are infrastructure-independent (Ludascher, Marciano, & Moore, 2001b, pp. 9-16). In other words, rather than migrating the digital object through the management of OAIS metadata, we instead manage the evolution of the technology. This guarantees that storage media, storage systems, database backups, and digital object formats will not become obsolete, but instead can be used to generate new knowledge.

Knowledge-Based Persistent Archives

A key challenge for the archivist is to preserve meaning for future generations. The effort to meet this challenge introduces philosophical and epistemological considerations about how to represent anything within the limits of what the language can express. It is already recognized that the goal demanded by some preservation researchers – completeness of collections – is infeasible because completeness is a value judgment that cannot be expressed objectively (Gladney, 2002, note 11). No system can describe itself to this extent. The research effort instead needs to define the minimal set of assumptions that can be used to express a preservation environment, such that it will support information and knowledge representation as an integral part of the archive and the ingestion/migration processes (known as "self-instantiating knowledge-based archives") (Ludascher, Marciano, & Moore, 2001b, pp. 9-16).

The extent to which this is possible has generated considerable debate within the digital preservation community. A commonly held view is that the management of structured and higher level semantic information is the key component of digital preservation research. The archivists have the challenge of preserving the semantic meaning of the terms that they use in the collections to support discovery of individual records (Gladney, 2002, note 9). This will require advances in manipulating structured information and characterizing data management policies that can build on the combined iRODS and Multivalent approach. The key question for research is whether a complete characterization of the information and knowledge within the preservation environment can be achieved, and whether this will meet the needs of the archivist.

Research is now ongoing to define what constitutes a preservation environment and how it relates to the external world. A true preservation environment will require the preservation properties to be maintained independently of the changes occurring in the external world. To what extent this is possible relates to the process of ingest into the preservation environment and the assurance that sufficient information is maintained to ensure complete infrastructure independence for preservation processes. This is a question relating to the philosophy of mathematics and cognitive science. Recent work in this area has focused on the areas of semantic web, artificial intelligence, and rule-based systems for controlling the ingestion of records into the preservation environment.

Characterizing Information and Knowledge Content

The challenge in managing digital entities is not just the management of the data bits but also the management of the infrastructure required to interpret, manipulate, and display these entities or images of reality.

We can use computer science-based specifications to describe what data, information, and knowledge represent. In the simplest possible terms: data corresponds to the bits (zeros and ones) that comprise a digital entity; information corresponds to any tag associated with the bits. The tags are treated as attributes that provide semantic meaning to the bits; knowledge corresponds to any relationship that is defined between information attributes. The types of relationships are closely tied to the data model used to define a digital entity.

At a minimum, these relationships may be logical (the semantic term that can be mapped into an ontology and reasoning done on inferred attributes); temporal (the structure may represent a time stamp that may be used to apply causal relationships); spatial (the structure may represent a coordinate system that can be mapped to a geometry and displayed in a GIS system); procedural (the structure may represent the outcome of a process in a workflow); functional (the structure may represent the result of applying a transformation algorithm); or epistemological (the structure may represent a systemic property of the entire preservation environment) (Moore, Arcot, & Marciano, 2007, p.6). Such relationships define the information context of a preservation environment that can be taken forward to generate knowledge from archives (Boisvert & Tang, 2001).

A theory of preservation can characterize information and knowledge content in terms of a "digital ontology" that can be used to define attributes and assign semantic meaning to the data. The attributes can be tagged as part of the digital object or associated with the digital object. The record and its processing context are both preserved and can be migrated to new technologies. This will make it possible to reapply archival processes, guaranteeing that not only the result of the archival processes can be preserved, but also that a description of the application of the archival processes can be preserved.

The naming conventions used to assign the semantic meaning are currently being defined. The NARA research effort has developed iRODS data grid middleware to characterize the preservation processes applied to records. The SHAMAN research effort will use the Data Format Description Language (DFDL) to define and name the structures present within the record and will use the Multivalent digital object model [13] to parse the DFDL characterization of structures. The intention is to use iRODS software to manage the information repositories and the Multivalent software to provide the logical and physical representation of the digital entity, based on its digital ontology[14].

[13] http://multivalent.sourceforge.net

[14] The concepts of rule-based management and ontology are separate, even though they are related. Thus, an ontology can describe the structures and relationships present within a record (this ontology is an extension of the OAIS representation information). The rules that control the application of micro-services that implement preservation processes can also be thought of as a procedural relationship: the set of procedural relationships can be organized as an ontology, but also viewed as a rule set. The micro-services can be viewed as functional realtionships imposed during preservation processes: again, the functional relationsihps can be organized as an ontology. Thus, we can map between the implementation

Methods for characterizing scientific data are less advanced and there is a need for further development of a data format description language to parse scientific data formats (binary output from application codes). In addition, we also require the addition of tags to the structures that indicate standard physical units (functional relationships), coordinate systems (structural relationships), geometries (spatial relationships), and time stamps (temporal relationships). Work in this area is ongoing in discreet communities for labeling structures of scientific data (e.g. HDF5 support for group and multidimensional array), and some preliminary work is now starting to characterize the relationships between the structures (for example the Open GIS Consortium data model to characterize coordinate transformations).

The nature and scale of the challenge is reflected in the Networking and Information Technology Research and Development (NITRD) supplement to the US President's Budget for FY2007[15]. This highlights maintenance of and access to long-lived science and engineering data collections and Federal records as a research priority. The solution to the challenge only partly exists. Research across distributed comunities coordinating their efforts is required to solve the problems.

Technology Assessment

The goal of building shared collections in a preservation environment is the primary focus of the NARA and SHAMAN preservation prototypes. This will require a generic infrastructure that will support preservation and management processes for data in multiple local repositories distributed across multiple institutions. The joint output of the NARA and SHAMAN preservation prototypes will result in international federations of shared collections that support both structured and semantic representations of the data.

The NARA and SHAMAN preservation prototypes aim to solve this challenge through the use of data grids that provide a set of virtualization services to enable the management of data that are distributed across multiple sites and storage systems and a display technology (Multivalent) supplies the ability to present (view) and manipulate structured information independently of infrastructure dependencies. Jointly, these two components supply the basis for managing archival collections at the levels of data, information, and knowledge.

The approach is a significant advance on the OAIS metadata-driven approach, insofar as it can be used to quantify how communication can be accomplished across different hardware environments, different software environments, and different standards for encoding information. We can potentially send into the future both the OAIS information and a description of the environment that is being used to manage and read the records. The ability to interpret and display the records independently of infrastructure constraints forms the basis for preservation environments to support knowledge management capabilities.

(rules, micro-services, and state information) and descriptions of the implementation that organize relationships between the rules, between the micro-services, and between the state information. Information from Reagan Moore (2007).

[15] http://nitrd.gov/

To understand the evolution of the system to this point, and its expected development, will require an assessment of information systems, as they are represented in the data grid, digital library, and persistent archive communities. There is a common goal of generic data management infrastructure that will require the integration of data analysis and knowledge analysis; digital library browsing and presentation; and distributed shared collection management. Much of this integration effort is described in reports issued by the San Diego Supercomputer Center, specifically the writings of Reagan Moore, and SDSC has been working with the various communities to engineer the integration of these technologies[16].

The major initiatives include:

- The Global Grid Forum research groups for manipulating data in distributed storage repositories.
- The Digital Library community for supporting discovery, access, and analysis of materials.
- The Persistent Archive community for preserving the ability to display and manipulate archival objects, while the underlying technologies evolve.

The focus is on the data and information models needed to manage and federate the collections and to migrate them forward into time. This involves the use of information models for describing the data; the ability to distinguish context needed for the data set, for collections, and for access; and support for interoperability across heterogeneous hardware and software systems (Moore, 1999).

We now cover the activities of each community in providing tools to manage the evolution of technology, bringing up to date some of the projections made by Reagan Moore in the 2003-06 SDSC Technical report and relating these to the expected developments of the NARA and SHAMAN preservation prototypes.

Data Grids

The preservation environment is based on the concept of infrastructure independence provided by data grids. This can be interpreted as the ability to manage the properties of the shared records independently of the choice of hardware infrastructure. The properties include the naming conventions, the access controls, the administrative information, and the links to provenance information.

Data grids support federation, the ability to exchange collections between independently managed data grids. This is typically done by projects that manage internationally shared collections. Each institution builds a local data grid and asserts local curation over the contents. The independent data grids can then be federated, with users, files, and resources cross-registered between the independent systems. Thus a remote user may be granted permission to perform selected preservation operations on data in a separate data grid. The choice of federation style is selected jointly by the data grid administrators. Examples include: central archives, in which remote data grids push data to a common preservation system; master data grids in which the files in remote data grids is distributed from the master data grid; chained data grids in which records are replicated from one data grid to the next under administrative control; and peer-to-peer data grids in which nothing is shared and only

[16] This section draws on the work and observations made by Reagan Moore at SDSC and Robert Chadduck at the NARA during the first iteration of the NARA preservation prototype (Moore, 2006).

publicly accessible files may be accessed from a remote site. Thus information in multiple independent metadata catalogues can be synchronized under administrator control.

The required abstraction mechanisms for the evolution of technology are being developed through the committees of the Open Grid Forum (OGF), which meets three times a year to promote interactions between grid researchers and implementers. The most relevant committee is the Preservation Environment Research Group[17], which has identified the components of data grid technology that are essential to the construction of persistent archives (Berlin meeting, March 2004). This includes sets of processes that assert authenticity and integrity of preserved digital entities as the basis for defining what constitutes archival context (e.g. administrative, descriptive, and authenticity attributes associated with each digital entity). The documentation concludes that each of these attributes is the result of the application of a process or set of relationships. Assertions of authenticity then correspond to the identification of relationships that have been satisfied[18].

Related committees active in the Open Grid Forum include the Data Transport Research Group[19], which defines transport standards to ensure interoperability across storage repositories; the OGF Data Access and Integration Working Group[20], which defines the set of operations that should be developed for interacting with database technology; the OGF Format Description Language Working Group, which is defining a Data Format Description Language (DFDL) to describe features of data formats; and the Grid Protocol Architecture Working Group[21], which is characterizing the consistency constraints that are needed to assemble a working grid from differentiated services and distributed state information registries (Moore, 2003). A number of initiatives are investigating the application of data and metadata interchange infrastructure based on OGSA-DAI middleware that supports the exposure of data resources, such as relational or XML databases, on to grids.[22]

The OGF evaluated a number of current data grids to determine the functionality for persistent archives based on the results of user communities for the management of data across heterogeneous, distributed storage resources (Moore & Marciano, 2007). The data grid requirements for persistent archives include data distributed across multiple sites and storage systems; data managed independently of the storage system; consistent management of file properties; persistent identifiers and access controls; and a scalable storage environment. These capabilities support many common usage scenarios – for example, managing the replication of data to mitigate risk of data loss – that are absolute requirements of persistent archives.

Each of these types of infrastructure has been implemented using the Storage Resource Broker (SRB) data grid – developed at the University of California (San Diego Supercomputer Center) – which implements the logical name space that is used to define global, persistent identifiers that are location-independent. Archival services create archival state information that is mapped onto the logical name space. Archival

[17] http://www.gridforum.org/6_DATA/persist.htm
[18] Minutes of the meeting, available from http://forge.ogf.org/
[19] http://www.gridforum.org/6_DATA/transport.htm
[20] http://www.gridforum.org/6_DATA/dais.htm
[21] http://www.gridforum.org/5_ARCH/GPA.htm
[22] http://forge.gridforum.org/projects/dais-wg

processes map provenance, administrative, descriptive, and authenticity attributes onto the logical name space. This supports the automation of archival processes.

The operations supported by the SRB provide a representation of the operations that can be aggregated into a preservation capability. Using the SRB enables the creation of a shared collection that may be distributed across multiple types of storage systems, located in multiple administrative domains. The data may be distributed, but indexed by a centralized metadata catalogue. The curation operations may be applied remotely from multiple sites, but the results are registered into the common metadata catalogue.

Analysis of the approach, however, suggests that the operations supported by the SRB do not correspond directly to TRAC-defined preservation capabilities. The preservation environment requires the ability to manage the representation information associated with the preservation environment itself. This includes a characterization of the management policies, the preservation processes, and the preservation administrative metadata. Curatorship requires managing both sets of properties, those of the records and those of the preservation environment. To do this will require the characterization of metadata, preservation processes, and associated preservation management policies. This approach is being implemented in the integrated Rule Oriented Data System (iRODS) from San Diego.[23]

In the NARA and SHAMAN persistent archive prototypes iRODS is used to demonstrate that the mappings for preservation attributes can be managed consistently, through rules that organize the structure and relationships that are needed to impose consistent update of the information. The ability to manage the relationships makes it possible to reapply archival processes, guaranteeing that not only the result of the archival processes can be preserved, but also a description of the archival processes can be preserved. The application of digital ontologies to digital entities to organize the relationships needed to interpret their structure and meaning makes it possible to guarantee the ability to manipulate digital entities in the future[24].

The emergence of the iRODS system, which integrates ontology management with information management, demonstrates the management of dynamically defined relationships between metadata attributes to support federation of namespaces. The mappings are consistently managed through ontologies that organize the relationships that are needed to impose consistent update of the information. The iRODS supplies the ability to manage the relationships that make it possible to reapply archival processes. This guarantees that not only the result of the archival processes can be preserved, but also a description of the application of the archival processes can be preserved. It makes possible the development of digital ontologies as a combined migration/emulation approach to preservation (Moore, Arcot, & Marciano, 2007).

[23] R. Moore, et al, SDSC Technical Report 2003-02r. http://irods.sdsc.edu/index.php/Introduction_to_iRODS
[24]

Digital Library Technologies

Developments from the digital library and persistent archives communities are used to support data organization, access, and preservation services, for example the ability to support discovery and data analysis of materials organized as collections. Areas of interest include the development of metadata standards for compound objects; the development of metadata interchange standards for retrieving information; the development of technology for the preservation of material; the application of search and collection management technologies; and the application of analysis technologies, as discussed below.

Metadata

One of the major research issues is the development of an understanding of the range of management policies required by diverse communities. The associated remote procedures that are managed by the rules can also be unique and strongly dependent on the data format. Each community will develop a preferred set of management processes and policies.

The current experience is that user communities have expectations for the properties that the shared collection will maintain. They typically expect a measure of consistency across the records (common metadata attributes), a measure of completeness (no missing data sets), a measure of authenticity (provenance information), a measure of integrity (risk mitigation against data loss with a valid copy always available). At the same time, the archivist in control of the preservation environment has expectations about the properties that the preservation processes will maintain. There is usually strong overlap, with a desire for authenticity, integrity, *respect des fonds*, chain of custody, risk mitigation.

Neither has the preservation community reconciled the appropriate preservation model. There are at least four different models, including:

- *Diplomatics (InterPARES)*[25]. This community focuses on authenticity, with retention forever;
- *Life-cycle data management (NARA)*[26]. This community focuses on retention schedules, hierarchical metadata description (record group, record series, folder, item object).
- *Continuum Model (Monash University)*[27]. This community examines preservation of material within the same environment used for active access and manipulation (which is enabled by data grids). The integration of digital library and preservation environments discussed in the Technology Assessment of this paper will lead to an environment similar to the continuum model.
- *Digital Library preservation (DSpace)*[28]. This community is expanding curation services into preservation services, but using digital library standards. The integrations of DSpace and Fedora to the data grids have now been extended to examine policy management issues.

Furthermore, different communities are focusing on different aspects of metadata.

[25] http://www.interpares.org/
[26] http://www.archives.gov/research/arc/lifecycle-data-requirements.doc
[27] http://www.infotech.monash.edu.au/research/groups/rcrg/
[28] http://www.dspace.org/

For scientific data communities, the workflow developers are trying to reach consensus on provenance metadata needed to describe the creation of derived data products. The semantic grid community is instead focused on an approach that is evolving from a description of services based on an agreed vocabulary, to an approach that is based on a characterization of the functions.

Mapping from the archivist expectations to the user community expectations requires strong participation by user communities. The user communities define the desired data formats, the semantics needed for discovery, the desired access mechanisms and the usage policies. Rule-based environments make these characterizations of the management policies explicit. The rules required by each community are typically unique. Rule sets can be defined that are applied to only the data collections created by that community.

As an initial step, efforts are being pursued to map the METS and other metadata standards[29], including the Preservation Metadata Implementation Strategies (PREMIS) data directory[30] and crosswalks with digital library schema, e.g. Dublin Core (ANSI/NISO Z39.85-2007).[31] The result should be the standardization of descriptive metadata to define authenticity.

In order to meet the Trustworthy Repositories Audit and Certification (TRAC) requirements, future work may require METS and other metadata standards to be augmented with the attributes that define the information context of the preservation environment, to allow descriptions of preservation management policies to be migrated into the future. One outcome of the NARA and SHAMAN prototypes will be the specification of criteria for this type of representation information. This will include characterization of the hardware/software infrastructure, or at a higher-level characterization of the operations supported by the environment, or at a higher-level characterization of the micro-services, rules, and persistent state information maintained by the preservation environment.

Further research will be required to integrate all of these aspects into a coherent system. The integration of digital library and preservation technologies must confront integration of digital library curation services (for access and discovery) with preservation archival processes. The first effort is to promote access while the latter is to promote authenticity. Thus, two possibly disjointed sets of metadata may be needed.

[29] The Metadata Encoding and Transmission Standard (METS) provides the ability to characterize the structure of a digital object that can be defined, organized, and maintained independently of the separate components, with support for provenance metadata, administrative metadata, preservation metadata, and structural metadata: http://www.loc.gov/standards/mets/ Archivists are increasingly interested in determining the relationship between the archival state information contained within the METS standard and the requirements of the US Department of Defence 5015.2 preservation standard for records management: http://http://jitc.fhu.disa.mil/recmgt/; and the ISO standard TR 15801 on Records Management and Legal Considerations http://www.iso.org/iso/catalogue_detail?csnumber=29093 .

[30] http://www.oclc.org/research/projects/pmwg/

[31] See the DIFFUSE (Dissemination of InFormal and Formal Useful Specifications and Experiences) European project: http://www.dcc.ac.uk/diffuse/ Dublin Core Metadata Initiative: http://www.dublincore.org/

Preservation Workflows

Digital preservation workflows are used to implement preservation processes with records stored in the information management systems. The NARA preservation prototype implemented a rule-based system (the Producer-Archive Workflow Network – "PAWN") for controlling the ingestion of records into the preservation environment (Smorul, JaJa, Wang, & McCall, 2004). This system was designed for use with the OAIS reference model and the data grid to encapsulate content, structural, descriptive, and preservation metadata.

More recently, San Diego has used the Kepler workflow system[32] to integrate preservation processes into the content production lifecycle, using an existing video production workflow (Arcot, Moore, Berman, & Schottlaender, 2005). This project was designed to abstract the production workflow and the preservation life-cycle management. The Kepler workflow system is also used to collect and maintain provenance information for scientific data and for the basis of a provenance challenge (Altintas, Barney, & Jaeger-Frank, 2006)[33].

Of great interest is a comparison of rules for controlling ingest with the rules needed for maintaining authenticity and integrity. The rules for ingestion have to express interactions between the preservation environment and the external world. The rules for maintaining authenticity and integrity have to show that the preservation properties are being maintained independently of the external world. The question is whether the authenticity and integrity rules operate only on persistent state information being maintained by the preservation environment, and thus are independent of changes occurring in the external world[34]. A research outcome of the SHAMAN preservation project will be to determine the extent to which interactions with the external world can be encapsulated into the ingestion rules, with infrastructure dependencies managed by the drivers that are created to support infrastructure independence.

Search and Collection Management Technologies

Digital library technologies provide standard services for ingestion, access, and display based on the metadata standards and protocols. The integration with the data grid technologies can ensure that digital library system gains the ability to manage collections that exceed the size of the local file system, gain support for replication, and gains the ability to federate with other digital libraries.

The DSpace digital library-SRB data grid integration project demonstrates an example of combining the distributed management of information through the use of digital libraries and the federation of preservation environments through the use of data grids. The project has resulted in DSpace-SRB support for a small subset of the Electronic Record Archives (ERA) capabilities list, devised for the NARA preservation prototype. The DSpace-SRB integration can be used to preserve small collections, but currently lacks the ability to scale to large collections. Further research will be required to determine if management policies can be defined in the DSpace digital library and supported by the iRODS data grid.

[32] http://www.kepler-project.org/
[33] http://twiki.ipaw.info/bin/view/Challenge/FirstProvenanceChallenge
[34] Information from Reagan Moore (2007).

Digital library technologies can be used to associate display functions with each data type, allowing relationships to be imposed on records, and mapping semantic labels on records to a digital ontology. The Fedora digital library and SRB integration, as part of the Dataset Acquisition, Accessibility, and Annotation eResearch Technologies (DART) Project, is designed to implement a preservation environment, with a particular focus on logical relationships.[35] A further research effort is required to determine how Fedora could be used to manage semantics on preservation attributes.

The Open Archives Initiative Protocol for Metadata Harvesting standard (OAI-PMH) is used to support publication of metadata from independent collections into a central repository.[36] OAI provides a mechanism for the access of attributes for manipulations by programmes. The technology has a set of complementary services that are being defined in the Database Access and Integration Services (DAIS) Working Group of the OGF[37]. The group has determined that the integration of digital library technology and grid technology is required to create a standard that can be used within persistent archive. The working group has proposed to implement a DAIS interface that is used to talk to a database and an OAI interface that is used to talk to a registry that lists the databases. Scientific projects investigating this integration include the National Virtual Observatory (NVO)[38].

Digital library technologies can also be used to introduce advanced discovery and data analysis capabilities for the persistent archive. Using the Cheshire digital library framework[39], for example, it is possible to apply knowledge analysis and data analysis tools to a shared collection whose records reside in multiple types of storage systems, at multiple institutions, located in multiple nations (Larson & Sanderson, 2005, 2006; Watry & Larson, 2005). The system provides an interface to many data mining algorithms, including clustering, classification, and association rule mining all of which can be deployed for collections of objects in a preservation environment. Further research is now required to investigate the use of workflow support for search and collection management processes, using both the Kepler workflow system and the Cheshire service-oriented digital library framework which executes processing workflows.

An output of the SHAMAN preservation prototype will evaluate how rules applied by the iRODS can be used to facilitate analysis by the Cheshire system. This will include automated migration of collections by iRODS onto high-performance disk (Teragrid) for Cheshire analyses; application of Cheshire services directly at the remote storage system through an iRODS rule; and a scheduler that defines whether the data should be moved to the compute platform for computation by Cheshire or whether the service should be executed at the remote storage system under an iRODS rule.

For the NARA preservation prototype, the Cheshire system was integrated with the SRB data grid to support distributed searches in a preservation environment. The Cheshire system has now been implanted in terms of processing workflows that can

[35] http://ausweb.scu.edu.au/aw06/papers/refereed/treloar/paper.html
[36] http://www.openarchives.org/OAI/openarchivesprotocol.html
[37] http://forge.ogf.org/sf/projects/dais-wg
[38] http://www.adass.org/adass/proceedings/adass03/P3-8/ (under Section 3.2).
[39] http://www.cheshire3.org/

integrate text and data mining processes in a distributed search environment. For the EU SHAMAN preservation prototype, we will now investigate how compute-intensive applications, such as analysis, can be performed on the client-side with the data migrated to a powerful platform where the computation is done. The outcome of this will be the integration of the client-side workflows (supporting search and collection management technologies) with the server-side workflows of the rule-oriented environment in order to support all scales of computation and data manipulation.

Persistent Archives (Data Preservation)

The development of preservation theory – from one that is focused on sending the metadata into the future to one that can also send into a future a description of the environment that is being used to manage and read the records – is directly traceable to the grid and digital library lineage outlined above. The ability to characterize management policies in terms of rules and preservation processes as standard micro-services is required to fulfill the different preservation strategies as may be required by different preservation communities.

Future preservation environments should be able to define separate preservation strategies for each record series (collection) based on the integration of data grid and digital library technologies. For example, if a collection has a preservation policy of transformative migration to XML, we should be able to define a rule that executes a remote micro-service at the storage system where the records reside that converts from the obsolete format to the new format. A second collection might use an emulation approach: in this case, a rule that controls the display of the records would invoke the correct version of the emulation technologies; for example, Portable Document Format (PDF). A third collection might be based on the use of Multivalent as a technology that supports the presentation of any digital object independently of infrastructure constraints[40]. The goal is to be able to assert which preservation policy is being applied for each collection, and track the application of the preservation policies over time. This means tracking the rule set that was applied, the set of remote micro-services that was executed by the rule set, and the state information that results from applying the rules. An example is an audit trail of all of the transformative migrations that were applied.

To achieve the objective, the persistent archive technologies required for the NARA and SHAMAN preservation prototypes are primarily based on applications of the iRODS rule-based data grid and the Multivalent digital object technology. The former supports the management of relationships between metadata attributes; the latter supports the ability to apply these relationships in order to interpret the digital entity. The two technologies can be used as the basis for supporting the Trustworthy Repositories Audit and Certification (TRAC) criteria and checklist. As a synthesis of the data grid and digital library technologies discussed above, the iRODS and Multivalent approach can be used to support not only the management of metadata, but a description of the environment that is being used to manage and read the records.

[40] The distinctions between the Multivalent approach and metadata-driven approaches (e.g. document emulation, conversion/migration, universal format, and universal computer) are discussed in Phelps and Watry (2005).

For the persistent archive, the goal is to use the iRODS data grid to manage consistently the mappings through ontologies that organize the relationships that are needed to impose consistent update of the information. (This is a step beyond the SRB, which can, in itself, demonstrate the ability to manage the mapping of the appropriate information onto the logical name space, but lacks management virtualization.) The ability of iRODS to manage the relationships makes it possible to reapply archival processes, guaranteeing that not only the result of the archival processes can be preserved, but also their description. The use of Multivalent can then be used to interpret the structure of digital entities and guarantee that they may be manipulated in the future.

Within the digital preservation community, research in the area of presentation tools is partly been based on the concept of a Universal Virtual Computer (UVC)[41], which defines preservation operations at the bit level and can, in theory, characterize the manipulations on records as bit-level operations. This requires a sophisticated programme to correctly interpret the bits. The UVC system can be migrated onto new operating systems, in theory enabling the parsing and manipulation of the record on to new systems. However, the implementation of UVC has a number of recognized deficiencies, particularly in the way that management policies are hard coded into the software. It is currently possible to use UVC to parse relatively simple formats, but not complex ones. To be more effective, the approach needs to implement the ability to map from application actions to bit level operations as generic higher-level operations. This would provide the ability to handle complex document and data formats, which is currently not possible in any meaningful way.

In contrast, the Multivalent digital object technology can already support the generic higher-level operations for manipulating characterization of structures in records. This will enable the interpretation of digital entities for preservation and manipulation while the underlying technologies evolve. The Multivalent approach differentiates between parsing through media adapters and manipulation through the application of behaviours. Once a document or digital objects has been parsed, it can then be manipulated by standard Multivalent behaviours. Multivalent can parse multiple data formats (e.g. PDF, OpenOffice, HTML) and, since Multivalent is written in Java it can be ported onto new operating systems relatively easily.

Multivalent can be thought of as an emulation environment that is written using a higher level language (Java) that separates the problem of parsing from display, and that provides a library of standard operations that can be used to display and manipulate documents and data. The Multivalent architecture is designed to interpret a digital entity based upon a digital ontology that represents the structural, semantic, spatial, and temporal relations inherent within a digital entity (Phelps & Watry, 2005). In this way, it is able to render all records from their original form and guarantee the correct interpretation of the record in future preservation environments. This provides a form of infrastructure independence for display applications.

The characterization of these digital ontologies is being undertaken by the Data Format Description Language (DFDL) research group of the Global Grid Forum, which is developing an XML-based description of the structures present within the structures. Multivalent is used as a presentation tool that applies these relationships, in

[41] http://www.nla.gov.au/padi/topics/492.html

the order defined by the DFDL digital ontology, to interpret the digital entity. In doing so, it will support the presentation and manipulation of digital objects (for example documents) with no dependencies. An example is the use of Multivalent to display Adobe Acrobat PDF documents from the original bitstream (i.e. without requiring the Adobe software to be on the system)[42]. The development of the iRODS technology makes it possible to represent the structural, semantic, spatial, and temporal relationships inherent within a digital entity. The Multivalent object model can then use this information to interpret the digital entity from the original bitstream.

For any given data type, a "media adaptor" is built which transforms the object into the internal structure of the Multivalent interface. Media adaptors are code components that translate concrete document formats into runtime data structures. The primary data structure is the document tree, which represents the entire content of a document (as a scroll or a page at a time) including everything from the text and images, to scripts, to the semantic structure (hierarchy and attributes), to the physical layout. Active (programmatic) elements or a specific document or a document genre, such as hyperlinks or outline opening and collapsing, are implemented by behaviours, which are program code with complete access to the document contents. The particular behaviours that apply to a document or genre are listed in XML-format hubs.

For the NARA and SHAMAN preservation prototype, the Multivalent technology (Java program) and media adaptors are archived, along with whatever is needed to migrate the Java Virtual machine into the future. Emulation then consists of supporting the original operations for manipulating the digital entity. Migration consists of porting the Java Virtual Machine to new technology as needed. Unlike present migration strategies, the digital entity remains unchanged, while making it possible to apply new operations that become available in new versions of Multivalent.

The ability of this approach to handle structured scientific and engineering data is now beginning to be assessed. As stated above, the current NARA and SHAMAN preservation prototypes are engaged with the challenges in curating scientific and engineering data collections, with the goal of being able to archive the data in the context that can best ensure its future usability. We need to support a future usage scenario that might be something unanticipated by the repository builder or archivist. There are a number of research questions relating to the process: for example, will it be possible to define a set of behaviours that display scientific data using Multivalent?

Implementation Scenario

The purpose of the NARA and SHAMAN testbed is to develop and implement policies and software that will achieve this goal by making it possible to migrate not only the records, but also the characterization of the preservation environment context itself. This is the defining metric for describing the preservation environment that is required by both industry and government to fulfill compliance according to the EU Directive on Privacy and Electronic Communications (2002/58/EC) and the new Federal Rules for Civil Procedure (FRCP) in the United States.

The implementation under NARA consists of seven independent data grids that manage data distributed across storage resources at seven institutions (primary copy stored at NARA I in Washington DC; replicated copy at University of Maryland for

[42] http://multivalent.sourceforge.net ; http://bodoni.lib.liv.ac.uk/fab4/

improved access and disaster recovery; deep archive at SDSC; laboratory at NARA II in College Park, Maryland; laboratory at the Rocket Center in West Virginia; collaborator at the University of North Carolina, Chapel Hill; collaborator at Georgia Tech). This is now being used to demonstrate the management of technology evolution, the preservation of web crawls, the automated extraction of authenticity metadata, and a producer-archive submission pipeline. The future implementation for the SHAMAN prototype will consist of three independent data grids in the United Kingdom, Portugal, and Germany with plans to connect with the NARA prototype to form a transcontinental persistent archive testbed.

Although the NARA prototype represents the current state of the art in digital preservation technologies, it also points to the necessity of a new generation of technologies, derived through research advances, which will fulfill its preservation goals, including:

- Authenticity, the assertion that provenance descriptive metadata and integrity metadata remain inextricably linked to the electronic records across all archival processes, and that the provenance metadata has not been altered;
- Integrity, the assertion that the electronic records have not been corrupted, and that the archival chain of custody has been enforced and is audited;
- Infrastructure independence, the assertion that the preservation environment can be maintained across arbitrary evolution of the infrastructure components

The SHAMAN builds upon the NARA prototype in its development of iRODS, DFDL, and Multivalent technologies: a joint effort will be made to extend support for structured and semantic representations of data. This will require the extensions of the iRODS capabilities to characterize the structure and relationships within records, identify the standard operations that can be performed on those relationships, and map from the actions executed by a display application to the standard operations[43]. A parallel effort will be undertaken to extend both the Data Format Description Language (DFDL) and Multivalent digital object technologies to parse and render the data and its relationships. The research effort will focus on the concept of digital ontologies as a new migration/emulation approach to preservation.

The research and development results for NARA I are demonstrated in the research prototype persistent archive. The demonstrations currently include access to electronic records replicated across the three data grids comprising the persistent archive using the SRB data grid and Cheshire system; accession of sample collections through the PAWN producer-archive pipeline developed at the University of Maryland; accession of web crawls; presentation of the archives using the Multivalent technology, and validation of persistent archive holdings. For future NARA testbeds and the SHAMAN prototype, we expect to extend the data analysis and knowledge analysis capabilities to support discovery across independent data collections through the integration of ontologies.

The NARA and SHAMAN research-based, long-term approach has provided a comprehensive, trustworthy means of addressing what is not only a moving target, but one which is rapidly growing both quantitatively and in complexity, and along paths

[43] Information from Reagan Moore (2007).

that are not wholly predictable (Thibodeau, 2001). An example of this is a recent requirement to preserve content, structure, and context over a variety of media, not just documents. The rise of "podcasting" in recent years is one instance of this which probably was not anticipated at the beginning of the project, but which will need to be accommodated in subsequent iterations of the prototype and final service.

Ultimately, the NARA and SHAMAN systems should be able to preserve the content, structure, and context of any data, including specific analysis tools characteristic of eScience or cyberinfrastructure. This is a major challenge that will require better adoption of well-defined data formats and well-defined semantics; it will also require means of recording provenance so that – for example – the precise conditions of scientific experiments may be repeated with confidence.

The basic NARA and SHAMAN services are expected to comprise 854 capabilities as defined by the TRAC. Over 200 different management policies can be defined to control the capabilities. The current research effort defines each capability in terms of rules that govern actions performed on the data and assertions made about the collections (Moore, et al., 2006). In supporting each of these rules, the NARA prototype makes use of the data grid technologies integrated with digital library and preservation technologies to manage:

- the accession of electronic records,
- standard preservation processes to create the archival forms of the electronic records and extract authenticity and integrity metadata,
- templates to define required preservation structured information,
- workflow management systems to apply templates to record series, and
- constraint systems to implement preservation policies.

These integrated technologies are closely coupled with the knowledge and data analysis capabilities supplied by the University of Liverpool. Originally devised independently of the NARA prototype, the Liverpool-based tools now interpret digital entities for presentation and manipulation in ways that are based on the SRB and iRODS data grid technologies. They also form (along with related digital library technologies such as Fedora, CITRIS, DSpace) an active area of research, particularly regarding the specification of digital library data management policies, which have required modification of the data grid management mechanisms to support manipulation of structured information.

In the context of the NARA and SHAMAN prototypes:

- The rule-oriented environment (iRODS) provides the abstraction mechanisms for managing evolution of storage and information repositories;
- The digital library and knowledge analysis systems preserve the ability to manage, access, and analyze the data.
- The Multivalent preservation technology provides the ability to interpret digital entities for presentation and manipulation while the underlying technologies evolve.

The technologies, taken together, represent a common vision for the preservation of all components of a persistent archive and illustrate the feasibility of long-term access and display of digital entities. The challenge within the perspective of the

NARA and SHAMAN prototypes is to automate all aspects of data discovery, access, management, and manipulation. While we are currently able to demonstrate the automation of archival processes at scale on the access of existing NARA digital holdings (and their registration into the persistent archive), we are only now beginning to understand what is required to develop constraint-based collection management systems and the further challenge to develop the concept of digital ontologies as an approach to preservation.

Implementing Management Policies to Validate the Trustworthy Repositories Audit and Certification (TRAC)

Advanced preservation approaches not only need to address the immediate goal of managing and preserving records – assuring authenticity and integrity – but also integration of this capability into an infrastructure that supports the ability to express management policies, for example services that query persistent state information to validate trustworthiness assessment criteria.

The challenge is four-fold:

- To implement a lifecycle information management infrastructure that will guarantee the ability to maintain the information context, arrangement and management of records;
- To implement the ability to parse, display, and manipulate digital objects independently of any infrastructure constraints;
- To integrate data analysis and discovery tools to enable users and administrators to find desired information within this environment; and
- To enable the future creation of knowledge through the generation and application of inference rules in a scalable inference engine.

The long-term goal is an "engineering" theory of data management. We decompose required capabilities into sets of micro-services, decompose management policies into sets of rules; decompose assessment criteria into queries on persistent state information. Given these three spaces, we then can then show that mappings between the three spaces in terms of actual operations are complete, consistent, and closed. It will be essential to prove that the preservation environment does not introduce any dependencies within the preserved material on choice of technology. A way to prove infrastructure independence will be to migrate the preserved material to an independent preservation technology (using alternate technology choices), and then migrate the preserved material back to the original preservation environment without loss of authenticity or integrity.

This requires a collection management infrastructure that can migrate the preserved digital entities to new hardware systems without changing the name spaces used to manage the digital entities; new software systems without affecting the protocols used to access the data; new access protocols without affecting the use of legacy storage systems; new encoding formats without losing information content. The NARA prototype has already demonstrated that data grids provide the required capabilities, including support for replication; federation of multiple independent preservation systems to mitigate risk of data loss; latency management across wide area networks; and bulk operations for scalability. These requirements were published

as the specification for the NARA Electronic Records Archive (ERA) as part of the Vendor Implementation process.

Based on the Trustworthy Repositories Audit and Certification criteria, we are now at the point where we can define the assessment criteria, the management policies, and the standard operations (micro-services) for a verifiable preservation environment and quantify these using technologies that are present today. With the NARA and SHAMAN projects, we are now able to use the approach to demonstrate closure, completeness, and consistency in ways that can be used to migrate all preservation processes (not just metadata) onto new technologies, as follows:

- *Closure*, that for every micro-service that is executed, required state information is created that can be queried to validate an assessment criteria;
- *Closure*, for every assessment criteria there is an associated rule that controls validation;
- *Closure*, for every rule the required micro-services are present;
- *Completeness*, the set of functionality provided by the system provides the required capabilities;
- *Consistency*, for any upgrade to the system, we can verify that the management policies remain consistent, and that the assessment criteria can still be met.

Vendor Implementation

A single data grid still has elements of risk. The preservation environment needs to consider federation of independent data grids or independent preservation environments. We are currently proposing, through the NARA, that we collaborate on a demonstration of infrastructure independence as soon as possible. This would include the migration of records from the SHAMAN into the NARA research prototype persistent archive, and the validation of assertions on authenticity and integrity. We would hope to extend this demonstration to the vendor (Lockheed-Martin) for delivery of the NARA ERA service. An alternative proof could be demonstrated using available technologies from alternative companies. In the case of IBM, for example, this would comprise HPSS archive or similar tape storage environment; DB3 database for managing the preservation metadata (authenticity, integrity, infrastructure state information), OASIS protocols for managing services; and Java clients for access. A demonstration that would prove infrastructure independence would need to provide the infrastructure to tie the components together, build the data grid technology to manage replicas, federate preservation systems, provide bulk operations for data management, provide support for the preservation name spaces (users, files, metadata, access constraints).

The research component for the NARA prototype is kept separate from the production deployment through the commercial vendor (Lockheed-Martin), with the result that the implementation is derived primarily from research advances. The worth of the research program is the impact that it can make in terms of informing NARA Electronic Record Archives (ERA) of new concepts, technologies, and capabilities that have been demonstrated in the prototype NARA persistent archive. The ability to demonstrate new technology at scale is essential to validate the worth of the research activities and their implementation.

The NARA has adopted an incremental approach to vendor implementation. The first part, the base contract period, was awarded to Lockheed-Martin in 2005 in competition on the basis of their system design and working prototype. Following this, the project will adopt an incremental approach with the initial operating capability (IOC) delivered in 2007 and final operating capability delivered in 2011. At this time the vendor is committed to a production version of the NARA prototype technologies that will be revised on an ongoing basis during the life of the project. Continuing research initiatives are:

- Automate archival processes and manage the consistent mapping of provenance, administrative, descriptive, and authenticity attributed onto the logical name space (University of Maryland, SDSC);
- Use data grids to manage the relationships of these mappings to organize relationships which are location-independent (SDSC);
- Use preservation technologies that can guarantee the discovery, access, presentation, and manipulation of any document or data independently of any infrastructure (University of Liverpool).

The NARA prototype has fostered positive partnerships between industry and academia. Some of the distributed storage technologies described in this paper are now being transformed into enterprise versions for use by companies such as EMC, Rolls Royce, etc. which serve as examples of technology transfer to commercial providers.

Relevance of Project

The NARA management policies are relevant to a broad number of related initiatives which solicit the use of data grid technologies to support federation of preservation environments. These include:

- **Library of Congress NDIIPP Project; NSF Chronopolis**. These projects solicit the use of data grid technologies to support federation of preservation environments[44].
- **NSF National Science Digital Library**. SDSC is supporting a persistent archive of educational material retrieved through web crawls of the content registered into the NSDL repository. The NSDL persistent archive is an application of the NARA-funded preservation technology in support of a specific education collection. The University of Liverpool Cheshire and Multivalent technologies are providing the means of analyzing the data[45].
- **NSF Teragrid.** SDSC is applying the NARA-funded preservation technology to support a replication environment for local scientific data sets. The application required development of additional SRB interfaces to interact with the unique archival storage systems used by Teragrid, SDSC, and the National Center for Atmospheric Research (NCAR)[46].
- **NSF DIGARCH (Digital Preservation and Lifecycle Management) Program.** Both SDSC and the University of Maryland participate in demonstrations of the preservation of additional types of multi-media collections. In each project, specific collections are being preserved[47].

[44] http://www.digitalpreservation.gov/
[45] http://www.dli2.nsf.gov/
[46] http://www.teragrid.org/
[47] http://www.sdsc.edu/srb/projects/digarch

More generally, the demand for rule-oriented data management occurs across many scientific disciplines, as well as within multiple data management communities, including:

- The Hierarchical Data Format (HDF) group to support manipulation of HDFv5-encoded scientific data sets[48];
- The Open Source Network for a Data Access Protocol (OpenDAP) community, which uses a separate standard for parsing the syntax and semantics of scientific data[49];
- The DataCutter middleware infrastructure community which applies stream filters for manipulating scientific data[50];
- The Data Format Description Language (DFDL) community whose representations can be used to characterize the data structures of the above communities[51].

As eScience communities gravitate to well-defined data formats and well-defined semantics, we should be able to map application-specific analysis tools to the NARA and SHAMAN technologies. In the future, our approach will be to pick a community and support their library calls for accessing data. There are now several eScience communities sufficiently well organized to do this:

- *Astronomy IVOA*: web services for manipulating Flexible Image Transport System (FITS) files interacting with catalogs[52];
- *Cognitive Science*: Sharing of access-controlled human subject data;
- *Oceanography*: OpenDAP mechanisms to explicitly manipulate registered data formats to extract physical data. This gives support for real time sensor data, including distributed across multiple research facilities[53];
- *NASA Earth Observing System (EOS):* HDF5 library calls for manipulating data structures for NASA[54];
- *UK eScience data grid*: specification of collection-dependent disposition and versioning options[55];
- *DSpace digital library*: Rights management and trusted digital repository assessment criteria[56].

The NARA and SHAMAN technologies have an increasing relevance for a broader range of media, for example the television and telecommunication domains which require the preservation life-cycle to mesh seamlessly with content production through the use of workflows that automate accession, description, organization, and preservation of different media types.

48 http://hdf.ncsa.uiuc.edu/
49 http://www.opendap.org/
50 http://datacutter.osu.edu/
51 http://forge.gridforum.org/projects/dfdl-wg/
52 http://www.ivoa.net/
53 http://www.opendap.org/
54 http://eospso.gsfc.nasa.gov/
55 http://www.e-science.clrc.ac.uk/
56 http://www.dspace.org/

Summary

The technologies required for the implementation of the NARA and SHAMAN prototypes assume the creation of a generic preservation environment. This infrastructure is being generated through the integration of advances in the digital library, data grid, and persistent archives communities. The development and implementation of the technologies in a persistent archive testbed focus on managing the evolution of technology, rather than just characterizing record provenance through representation information. The approach is primarily based on use of the iRODS data grid and Multivalent digital object model with additional support for digital library services to support discovery and analysis of data. True infrastructure independence implies that any of the above components could be swapped for their equivalents with no loss of authenticity. The long-term goal is an "engingeering" theory of data management. We decompose required capabilities into sets of micro-services, decompose management policies into sets of rules, decompose assessment criteria into queries on persistent state information. Given these three spaces, we then show that mappings between the three spaces in terms of actual operations are complete, consistent, and closed. The technology used to provide infrastructure independence should be applicable to other types of data management systems and domains. The work we have done to date will enable us in the future to focus on eScience and cyberinfrastructure communities that have well-defined data formats and well-defined semantics.

Acknowledgements

The research has been sponsored by the Joint Information Systems Committee (JISC), National Science Foundation (NSF), and the National Archives and Records Administration (NARA) during the period 1999-2007. Grateful acknowledgment is given to Reagan Moore and colleagues at the SDSC Data Intensive Computing unit at University of California, San Diego; Robert Chadduck (NARA); Thomas A. Phelps (developer of Multivalent); Ray R. Larson (UC Berkeley); colleagues at the University of Liverpool (Robert Sanderson, Clare Llewellyn, John Harrison, Fabio Corubolo, Catherine Smith).

References

Arcot, R., Moore, R., Berman, F., & Schottlaender, B. (2005). "Digital Preservation Lifecycle Management for Multi-media Collections", in *Digital Libraries: Implementing Strategies and Sharing Experiences*, pp. 380-84. Lecture Notes in Computer Science, Vol. 3815/2005. DOI 10.1007/11599517.

Altintas, I., Barney, O., & Jaeger-Frank, E. (2006) Provenance collection support in the Kepler Scientific Workflow System. In *Provenance and Annotation of Data, International Provenance and Annotation Workshop, IPAW 2006, Chicago, IL, USA, May 3-5, 2006*. Retrieved December 4, 2007 from http://www.ipaw.info/ipaw06/proceedings/CameraReady_s5_2.pdf

Boisvert, R.F., Tang, P.T.P. (Eds.). (2001). The architecture of scientific software, IFIP TC2/WG2.5. *Working Conference on the Architecture of Scientific Software, October 2-4, 2000, Ottawa, Canada, IFIP Conference Proceedings, vol. 188*, Kluwer.

Gladney, H.M. (2002). Perspectives on trustworthy information. *Digital Document Quarterly*, *vol. 1, No. 3*. note 9. Retrieved December 4, 2007 from http://home.pacbell.net/hgladney/ddq_1_3.htm#_edn9

ibid. note 11. Retrieved December 4, 2007 from http://home.pacbell.net/hgladney/ddq_1_3.htm#_edn11

Hedstrom, M. (1991) Understanding electronic incunabula: A framework for research on electronic records (UEI), cited in Digital Preservation Europe (DPE) Research Roadmap, Project no. 034762, Section B1. *The American Archivist 54, 3 (Summer 1991):* 334-54 (Cox).

Jeffrey, S., & Hunter, J. (2006). A semantic search engine for the Storage Resource Broker. *3rd Semantic Grid WorkshopGGF16, Athens*. Retrieved December 4, 2007 from http://www.itee.uq.edu.au/~eresearch/projects/dart/workpackages/si3.php http://www.itee.uq.edu.au/~eresearch/projects/dart/outcomes/semanticsrb.php

Larson, R., & Sanderson, R. (2006). Cheshire3: Retrieving from tera-scale grid-based digital libraries. *SIGIR 2006*: 730). Retrieved December 4, 2007 from http://portal.acm.org/ft_gateway.cfm?id=1148343&type=pdf

Larson, R., & Sanderson, R. (2005). Grid-based digital libraries. *JCDL 2005*, 112-113. Retrieved December 4, 2007 from http://ieexplore.ieee.org/xpls/abs_all.jsp?arnumber=4118524

Ludascher, B., Marciano, R., & Moore, R. (2001a). Preservation of digital data with self-validating, self-instantiating knowledge-based archives. *ACM Sigmod Record*, *Vol 30, Issue 3*, pp. 54-63. Retrieved December 4, 2007 from http://portal.acm.org/citation.cfm?id=603876

Ludascher, B., Marciano; R., & Moore, R. (2001b). Towards self-validating knowledge-based archives. Research Issues in Data Engineering, 2001, pp. 9-16. In *Research Issues in Data Engineering, 2001. Proceedings. Eleventh International Workshop on*. , pp. 9-16. ISBN 0-7695-0957-6. DOI: 10.1109/RIDE.2001.916486. Retrieved December 4, 2007 from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=916486

Moore, R. (2003). Common consistency requirements for data grid, digital libraries, and persistent archives. Submitted to the 12th High Performance Distributed Computing conference, Seattle, WA (June 2003). Retrieved December 4, 2007 from http://grid.lbl.gov/GPA/GGF7_Data_Consistency.Word95.pdf

Moore, R. (1999). Persistent archives for data collections. National Partnership for Advanced Computational Infrastructure (SDSC). Submitted in 1999 as part of the Archival Workshop on Ingest, Identification, and Certification Standards (AWIICS). Retrieved December 4, 2007 from http://nost.gsfc.nasa.gov/isoas/awiics/

Moore, R. (2006). *Research on persistent archives.* SDSC Technical Report 2003-06. Retrieved December 4, 2007 from http://www.sdsc.edu/NARA/Publications/data-preservation.doc

Moore, R, et al. (2006). *Constraint-based knowledge systems for grids, digital libraries, and persistent archives*. SDSC Technical Report 2005-9.

Moore, R. & Marciano, R. (2007). Prototype preservation environments. Paper submitted to *Library Trends*.

Moore, R., Arcot, R., & Marciano, R. (2007). Implementing Trusted Digital Repositories. Retrieved December 4, 2007 from http://ww.ils.unc.edu/digccurr2007/web-abstracts/moore_abstract_6-4.pdf

Moore, R. & Smith, M. (2007). Automated validation of trusted digital repository assessment criteria. *Journal of Digital Information*, Vol. 8, No 2 (2007). Retrieved December 4, 2007 from http://dspace.mit.edu/html/1721.1/39091/Moore-Smith.htm

Moore, R., Baru, C., Arcot, R., Ludaescher, B., Marciano, R., Wan, M., et al. (2000, March). Collection-based persistent digital archives. *D-Lib Magazine, 6 (11).* Retrieved December 4, 2007 from http://www.dlib.org./dlib/march00/moore/03moore-pt1.html

Phelps, T.A., & Watry, P. (2005). A no-compromises architecture for digital document preservation. *ECDL*, 2005: 266-277. Retrieved December 4, 2007 from http://multivalent.sourceforge.net/Research/Live.pdf

San Diego Supercomputer Center (SDSC).(1999) DOCT Technical Report, Section 2.4. Retrieved December 4, 2007 from http://npaci.edu/DICE/Pubs/replication.doc

Smorul, M., JaJa, J., Wang, Y., & McCall, F. (2004). *PAWN: Producer-Archive Workflow Network in support of digital preservation*. CS-TR-4607, UMIACS-TR-2004-49. Retrieved December 4, 2007 from http://umiacs.umd.edu/research/adapt/papers/UMIACS-TR-2004-49.pdf

Smorul, et al. (2002). Producer-archive interface methodology: Abstract standard, Consultative Committee for Space Data Systems. CCSDS-651.0-R-1, Red Book, December 2002. Retrieved December 4, 2007 from http://public.ccsds.org/publications/archive/651x0b1.pdf

Thibodeau, K. (2001, February). Building the archives of the future. *D-Lib Magazine*, *7, (2)*. Retrieved December 4, 2007 from http://www.dlib.org./dlib/february01/thibodeau/02thibodeau.html

Watry, P., & Larson, R. (2005). Cheshire framework white paper: Implementing support for digital repositories in a Data Grid environment. In *Proceedings of the IEEE Conference on Globally Distributed Data (2005)*, pp. 60-64. Retrieved December 4, 2007 from http://cheshire.berkeley.edu/Cheshire_Sardinia.pdf