Implementing Trusted Digital Repositories

Reagan W. Moore, Arcot Rajasekar, Richard Marciano
San Diego Supercomputer Center
9500 Gilman Drive, La Jolla, CA 92093-0505
{moore, sekar, marciano}@sdsc.edu

Abstract:

Trusted digital repositories manage the integrity and authenticity of records through multiple generations of technology. They provide mechanisms to validate assertions about trustworthiness and provide the preservation processes that implement the required control and management capabilities. Today there are multiple technologies that can be used to build a digital repository that is capable of maintaining the authenticity and integrity of ingested documents. The approaches range from solutions based on data grids (Storage Resource Broker), to solutions based on digital library systems (DSpace and Fedora), to solutions based on rule-oriented environments (integrated Rule-Oriented Data System). These approaches optimize management of different components of a trusted digital repository. No single system currently provides all of the required functionality. This paper examines how rule-based systems can validate assertions of trustworthiness, presents the infrastructure components provided by the integrated Rule-Oriented Data System, and explores how rule-based approaches can be used to develop a theory of preservation.

Introduction:

The concept of a trusted digital repository can be quantified through the identification of assessment criteria that evaluate trustworthiness. A system that is able to validate the assessment criteria can be considered trustworthy, and thus would be a reasonable environment for the preservation of data for the long term. An initial set of assessment criteria have been proposed by the RLG and the National Archives and Records Administration [1]. Analyses of the assessment criteria and mappings to management policies have been published [2,3,4]. The expectation is that one can define management policies that ensure trustworthiness, define rules that apply the policies, define capabilities that implement the required preservation functions, and define preservation metadata that capture information about the application of the preservation functions. One can then query the preservation metadata to assess whether the assessment criteria have been satisfied. An approach based on assessment criteria attempts to define all of the management policies that are needed to prove that a preservation environment will successfully preserve records.

The assessment criteria are based on traditional preservation principles:

- Authenticity, assertions about the provenance of the records
- Respect du fonds, assertions about the arrangement of the records
- Chain of custody, assertions about the ownership of the records
- Integrity, assertions about the management of the records

Each of these preservation principles defines properties that the preservation environment should preserve. At a minimum, the preservation environment needs persistent names for identifying the records, the archivists, and the storage repositories [5]. Assertions about the management of the records can then be based on attributes associated with the persistent name spaces. Examples include the name spaces that are used by the preservation environment to track provenance (descriptive metadata), integrity (rules, preservation

processes, system state information), chain of custody (archivist names, storage resource names), and respect du fonds (record names). These name spaces need to remain invariant over time, ensuring that an operation performed by a prior archivist can be correctly interpreted. The assessment criteria also require that the functions performed by preservation processes remain consistent over time. This means a preservation environment should be quantified in terms of the actual preservation operations that are performed and the management policies that control the execution of the preservation processes. We thus have two separate contexts that require description; the provenance of the records, and the evolution of the preservation environment.

Integrated Rule-Oriented Data Systems

Rule-oriented data systems define the minimal set of preservation processes and management policies on which a preservation environment can be based. Rule-oriented systems also track changes to the preservation environment. We need to know how the preservation processes we are applying today are related to preservation processes that were applied in the past, to make assertions about integrity and authenticity.

The characterization of the minimal set of preservation management policies is difficult. We need to map from assessment criteria, to the management policies that enforce the preservation assertions, to the preservation capabilities that implement preservation processes. This approach is being implemented in the integrated Rule Oriented Data System (iRODS) [6]. As shown in Table 1, we express the capabilities as sets of micro-services that are applied at the remote storage systems where the records are stored. We express the management policies as rules that control the execution of sets of micro-services. We express the assessment criteria as persistent state information that is generated by application of the rules. The rules are stored in a rule engine and the persistent state information is stored in a database.

Data Management Environment	**Conserved Properties**	**Control Mechanisms**	**Remote Operations**
Management Functions	Assessment Criteria	Management policies	Capabilities
Data Management Infrastructure	Persistent State Information	Rules	Micro-services
Physical Infrastructure	Database	Rule Engine	Storage System

Table 1. Characterization of preservation management policies

Based on experience with the Storage Resource Broker data grid [7], we recognize that the operations supported by the remote storage systems do not correspond directly to preservation capabilities. Thus each preservation capability is an aggregation of multiple micro-services, and each micro-service is an aggregation of multiple operations at the remote storage system. One challenge is that even if the same micro-services are executed over time, we still need a standard set of operations that are applied at the remote storage systems. This is ensured through the concept of infrastructure independence. We observe that the set of operations performed by a particular type of storage system differ in either semantics (effect of the operation) or manipulation (different result generated) across vendors. The SRB technology implements standard operations that are storage system independent by creating storage system specific drivers to map from the standard operations to the vendor storage system

protocol. The Storage Resource Broker data grid is an example of a system that provides infrastructure independence for preservation environments.

The iRODS micro-services can define the minimal set of preservation functions that need to be carried forward in time. Similarly, the iRODS rules can define the minimal set of management policies that are needed to enforce trustworthiness. Given a minimal, complete description of the components of a preservation environment, one can then create a rule-based system that is provably correct. The required operations generate the persistent state information needed to validate assertions about trustworthiness.

The iRODS data management system prototype is designed to implement a provably complete preservation environment. The current research effort is exploring:

- Implementation of the ERA capabilities as iRODS rules, micro-services, and state information [8]. A major challenge is the design of the correct level of aggregation of remote storage operations into micro-services. The micro-services need to be simple enough that all preservation capabilities can be implemented from standard micro-services. But if they are too-low level (byte oriented operations) they become difficult to apply. The current assessment has defined 174 rules and a smaller number of micro-services that need to be implemented. The current iRODS environment provides 73 micro-services. The supposition is that if the correct level of aggregation of the standard operations performed at remote storage systems is defined for each micro-service, a minimal number of micro-services will be created that are capable of expressing every required preservation capability. If too fine a granularity of operation is used, the creation of a micro-service becomes onerous. If too coarse a granularity is used, the number of required micro-services increases to characterize every possible combination of standard operations. The tension is between the ability of the rule-oriented environment to minimize the effort required to apply micro-services, and the minimization of the amount of effort needed to create micro-services. The initial set of micro-services is based upon the operations supported by the Storage Resource Broker data grid.
- Implementation of the RLG/NARA assessment criteria for trusted digital repositories. A similar analysis resulted in the identification of 174 management policies needed to express the assessment criteria. A goal of current research is to show that the management policies can be expressed as iRODS rules, micro-services, and state information.
- Comparison of preservation metadata-driven approaches with rule-driven approaches. Examples of the former include PREMIS, the NARA Life Cycle Data Requirements Guide, and the Open Archive Information Standard. Examples of the latter include iRODS and service-based environments such as DSpace and Fedora. IRODS differs from traditional client-driven workflows because iRODS implements server-driven workflows, with the processing done directly at the remote storage system where the data reside.

Besides the ability to characterize the preservation environment, a second implementation requirement is automation of the execution of management policies. We observe that as collections increase in size, the amount of labor required to recover from problems that occur in distributed environments becomes onerous [9]. The iRODS system automates the application of the rules as either atomic operations, executed on each preservation process invocation, or as deferred operations, executed when systems become available, or as periodic operations. The support for deferred operations is needed to handle situations where the remote storage system is temporarily inaccessible. Periodic operations are needed to manage validation of integrity and authenticity. A rule can be written that validates checksums, synchronizes

replicas, and corrects copies that have been corrupted. Similarly a rule can be written that compares provenance metadata to the required metadata for a record series, and then either identifies missing authenticity information, or extracts the required information from submission agreements.

A related issue when implementing a scalable preservation environment is the management of a large number of records. This is addressed by creating operations that act on sets of records. In practice, the set of SRB operations that are performed at each storage system or database depends upon the level of aggregation that is imposed on each of the logical name spaces used to identify records, archivists, and storage systems:

- Operations on the user logical name space {single user, user group, data grid federation}
- Operations on the storage resource logical name space {single storage system, compound system with a data cache, cluster}
- Operations on the file logical name space {single file, container aggregating multiple files, hierarchical directory of files)
- Operations on user defined metadata {single attribute, hierarchical table, collection}

These multiple levels of aggregation are required for scalability of the preservation environment. A similar set of aggregation levels is being implemented in the iRODS system.

A final design criterion for a preservation environment is support for the evolution of the preservation environment itself. This implies that the preservation management policies and processes can evolve to handle new types of records, new types of provenance metadata, and new legal requirements. To allow the rules and micro-services to evolve, we added support for three more logical names spaces in the iRODS environment. We also added similar levels of aggregation within each name space:

- Operations on the micro-service logical name space {atomic, deferred, periodic}
- Operations on the rule logical name space {single micro-service, set of micro-services, recursive rule hierarchy}
- Operations on the persistent state information logical name space {single attribute, deferred consistency flags for attributes, validation date for appraisal of attributes}

These three name spaces enable management virtualization. One can add new management policies, new preservation capabilities, and new persistent state information without destroying the ability to execute previous management policies. With these three additional logical name spaces, the preservation environment can also evolve. Also, the ability to characterize management policies implies that management policies can be automated, minimizing the labor requirements needed for large collections.

By defining rules that automate the validation of assessment criteria, a preservation environment can be defined that validates its own trustworthiness. It also becomes possible to migrate the records, rules, micro-services, and persistent state information to another independent preservation environment. The management policies that were being applied in the first preservation environment can continue to be applied in the second preservation environment. This means that assertions about authenticity and integrity can continue to be validated as records are moved between different implementations of preservation environments.

Theory of Preservation

Preservation can be thought of as communication with the future. Information that is understood today is transmitted to an unknown system in the future where it will be interpreted and displayed. This paradigm can form the basis for a Theory of Preservation. The theory of

communication is well known. It quantifies the concept of information and describes the mechanisms that optimize the transfer of information. The theory of communication applies when the Sender and Receiver use common technology for data transmission and reception.

A theory of preservation needs to quantify how communication can be accomplished when the Receiver is using different technology than the Sender. This includes not only different hardware and different software, but also different standards for encoding information. Effectively, we need to send into the future not only the information (records), but also a description of the environment that is being used to manage and read the records.

The Receiver is a future preservation system, and hence is linked over time to the original Sender. To maintain the ability to be able to interpret and display the records, the preservation environment must characterize its own evolution, and the impact that preservation environment evolution has on record management. A theory of preservation makes assertions about the ability to maintain the information context, arrangement, and management of records as well as the information context (management policies and preservation procedures) of the preservation environment.

An example of the management of contextual information is defined in the OAIS standard. This focuses on the ability to access and interpret records through the creation of representation information. The representation information defines the structures present within a record and their semantic labels. A designated community is defined that maintains the ability to interpret the semantic labels. However, the OAIS standard does not provide representation information about the preservation environment itself.

The concept of infrastructure independence quantifies assertions about the preservation environment, including not only the name spaces, but also the preservation processes, and the preservation policies. By demonstrating that the preservation environment controls the information context needed to preserve the ability to apply preservation procedures, we can create a theory of preservation, in which the information content of the records and the information context of the preservation environment are communicated into the future.

Godel's theorem proves that no system can be completely self-describing. The theory of preservation needs to define the minimal set of assumptions on which preservation environments are based, and then show how these assumptions are conserved as the preservation environment evolves. We believe a preservation environment is feasible, because the receiver is a controlled environment, whose evolution can be explicitly tracked from the original system used by the sender.

We need a few more concepts for a theory of preservation. In the case of information theory, the fundamental unit of information is a "bit". We need a fundamental unit of "function" on which preservation processes can be based. The candidates for a unit of preservation function are the minimal set of management policies and the minimal set of micro-services required to implement and manage all preservation capabilities. We want to characterize the impact of applying a preservation process as a change of state information, and a transformation function that is applied to the record. For viable preservation processes, we need reversible transformations: the ability to transform back to the original record. This involves characterizing records as follows:

- Every record is a sequence of bits

- Information content is described by defining the structures present in the bit sequence, and then naming the structures. The structure names represent the semantic terms used to define the meaning of the record.
- Knowledge content is defined as relationships between and on the structures. Examples include:
  - Logical relationships. The semantic term can be mapped into an ontology, and reasoning done on inferred attributes (semantic grid).
  - Temporal relationships. The structure may represent a time stamp that may be used to apply causal relationships.
  - Spatial relationships. The structure may represent a coordinate system that can be mapped to a geometry and displayed in a GIS system.
  - Procedural relationships. The structure may represent the outcome of a process in a workflow.
  - Functional relationships. The structure may represent the result of applying a transformation algorithm.

These characterizations of records enable the concept of persistent objects [10]. A persistent object can be created that can be displayed in the future using future technology, even though its internal structures and relationships are based on present-day technology. To enable display of persistent objects, we need one more concept, namely that future manipulations of records can be expressed in terms of the manipulation of the structures and relationships that have been described for each record. We base the ability to display records on two levels of indirection:

- Characterize the standard structures and relationships present within the record
- Characterize the standard operations that can be applied on each type of relationship for each type of record.
- Characterize the manipulations performed by a display application in terms of standard operations on standard relationships. In effect, the display application does not manipulate records. Instead it executes standard operations on standard relationships on standard structures. One can map from the actions of the display application to the standard operations. Given this mapping, any display application can manipulate any record, or at least the structures within the record for which the required relationships are defined.

When we apply a future display mechanism, we map from the operations on structures that the display needs to perform, to standard operations on relationships/structures present within the record. This is a form of infrastructure independence for display applications. Exactly the same indirection mechanisms are used in data grids to support the manipulation of data on heterogeneous storage systems. We map from the operations desired by the applications, to the operations that the remote storage system is able to apply.

The preservation process then consists of the manipulation of structures in records, or the assignment of properties to sets of records, or the establishment of relationships between two records. If we have a defined set of fundamental reversible preservation processes, we can assert that any future preservation environment can transform all records back to their original form. The future preservation environment can correctly interpret the preservation information context from the past and apply the same preservation policies.

A theory of preservation needs to demonstrate epistemological constraints about the internal consistency between the assessment criteria and the rules that control generation of persistent state information. The system is internally consistent if all preservation attributes needed to

quantify the preservation principals (authenticity, integrity, chain of custody, respect du fonds) can be generated or validated through the application of management policies; and if the persistent state information generated by the application of management policies are retained as preservation attributes. We cannot have a situation in which a preservation attribute that is needed for assessing preservation principals cannot be controlled or verified by one of the preservation rules. Nor can we have a situation in which persistent state information generated by application of the rules is not included in the representation of the preservation environment context that is migrated forward into the future.

In current preservation environments, preservation metadata is defined without reference to the associated preservation management policies or preservation capabilities. An approach based on only maintenance of the preservation provenance metadata is only providing half of the required preservation environment. Ideally, a characterization of a preservation environment context should be possible either by characterizing the preservation management policies and the resulting persistent state information, or by characterizing the preservation attributes and the management rules that validate their authenticity and integrity. We should be able to map from a context that is driven by preservation metadata to a context that is driven by preservation management policies, and back.

The acid 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 characterization 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.

The expectation is that we can develop a theory of preservation. Its components are:

- Definition of the persistent name spaces
- Definition of the operations that are performed upon the persistent name spaces
- Characterization of the changes to the persistent state information associated with each persistent name space that occur for each operation
- Characterization of the transformations that are made to the records on each operation
- Demonstration that the set of operations is complete, enabling the decomposition of every preservation process onto the operation set.
- Demonstration that the preservation management policies are complete, enabling the validation of all preservation assessment criteria.
- Demonstration that the preservation environment is complete, enabling the maintenance of authenticity and integrity.
- The assertion is then: if the operations are reversible, then a future preservation environment can recreate a record in its original form, maintain authenticity and integrity, support access, and display the record.
- A corollary is that such a system would allow records to be migrated between independent implementations of preservation environment, while maintaining authenticity and integrity.

The iRODS rule-based environment is a first step towards the creation of a trustworthy digital preservation repository. Finally, we observe that the technology used to implement a preservation environment provides the basic capabilities needed to implement digital libraries

[11]. It is possible to build generic infrastructure that supports both digital libraries and persistent archives.

Acknowledgement

The research was supported by funding from the National Archives and Records Administration under NSF cooperative agreement 0523307 through a supplement to SCI 0438741, "Cyberinfrastructure; From Vision to Reality", and by the National Science Foundation grant ITR 0427196, "Constraint-based Knowledge Systems for Grids, Digital Libraries, and Persistent Archives". The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the National Science Foundation, the National Archives and Records Administration, or the U.S. government.

REFERENCES

1. RLG/NARA Audit Checklist for Certifying Digital Repositories, http://www.rlg.org/en/page.php?Page_ID=2076
2. R. Moore, and M. Smith, "Assessment of RLG Trusted Digital Repository Requirements," Joint Conference on Digital Libraries workshop on "Digital Curation & Trusted Repositories: Seeking Success", Chapel Hill, North Carolina, June 2006.
3. M. Smith, R. Moore, "Digital Archive Policies and Trusted Digital Repositories", proceedings of The 2nd International Digital Curation Conference: Digital Data Curation in Practice, November 2006, Glasgow, Scotland.
4. R. Moore, M. Smith, "Automated Validation of Trusted Digital Repository Assessment Criteria", submitted to Journal of Digital Information, January 2007.
5. R. Moore, "Building Preservation Environments with Data Grid Technology", American Archivist, vol. 69, no. 1, pp. 139-158, July 2006.
6. A. Rajasekar, M. Wan, R. Moore, and W. Schroeder, "A Prototype Rule-based Distributed Data Management System", High Performance Distributed Computing workshop on "Next Generation Distributed Data Management", Paris, France, May 2006.
7. C. Baru, R, Moore, A. Rajasekar, and M. Wan, "The SDSC Storage Resource Broker," Proc. CASCON'98 Conference, Toronto, Canada, Nov.30-Dec.3, 1998, p. 5.
8. The Electronic Records Archive capabilities list defines a comprehensive set of capabilities needed to implement a preservation environment, and can be examined at http://www.archives.gov/era/pdf/requirements-amend0001.pdf
9. R. Moore, M. Wan, and A. Rajasekar, "Storage Resource Broker: Generic Software Infrastructure for Managing Globally Distributed Data", Proceedings of IEEE Conference on Globally Distributed Data, IEEE Computer Society, Piscataway, New Jersey, June 28, 2005, pp. 65-69.
10. R. Moore, "The San Diego Project: Persistent Objects," Archivi & Computer, Automazione E Beni Culturali, l'Archivio Storico Comunale di San Miniato, Pisa, Italy, February, 2003.
11. R. Moore, A. Rajasekar, and M. Wan, "Data Grids, Digital Libraries and Persistent Archives: An Integrated Approach to Publishing, Sharing and Archiving Data", Special Issue of the Proceedings of the IEEE on Grid Computing, IEEE Computer Society, Piscataway, New Jersey, March 2005, Vol. 93, No.3, pp. 578-588.