The International Journal of Digital Curation

Issue 1, Volume 2 | 2007

Digital Archive Policies and Trusted Digital Repositories

MacKenzie Smith,

MIT Libraries

Reagan W. Moore,

San Diego Supercomputer Center

June 2007

Abstract

The MIT Libraries, the San Diego Supercomputer Center, and the University of California San Diego Libraries are conducting the PLEDGE Project to determine the set of policies that affect operational digital preservation archives and to develop standardized means of recording and enforcing them using rules engines. This has the potential to allow for automated assessment of "trustworthiness" of digital preservation archives. We are also evaluating the completeness of other efforts to define policies for digital preservation such as the RLG/NARA Trusted Digital Repository checklist and the PREMIS metadata schema. We present our results to date.

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.

SOME RIGHTS RESERVED

Introduction

The MIT Libraries are collaborating with the University of California, San Diego Supercomputer Center on the PoLicy Enforcement in Data Grid Environments (PLEDGE) Project, funded by the US National Archives and Records Administration. The project is investigating the various policies in use by operational digital archives, in-formed by the DSpace repository for digital information lifecycle management and the integrated Rule-Oriented Data System (iRODS) for storage virtualization and digital object persistence. We are identifying and categorizing these policies, and defining associated rules and state information to make them machine encodable and, wherever possible, enforceable. We believe that having archival policies recorded in a standard way will allow digital archives to become both more explicit about the many policies that define what they are and how they are managed, as well as making them more portable and interoperable with other archives.

We have mapped the PLEDGE policies for enterprise, archive, collection, and item levels to the recently published RLG/NARA Audit Checklist for the Certification of Trusted Digital Repositories (TDR) which defines a set of preservation management policies. We evaluate what is missing from the TDR checklist and where those policies translate into multiple operational rules. The TDR document does not specify how the management policies should be applied to a given repository, so we examine the set of preservation capabilities, rules and associated metadata (i.e. state information) required to automate the verification of the trusted digital repository. In effect, we attempt to demonstrate the set of rules that automatically validate the trustworthiness of a repository. Our approach is based on the characterization of each item in the assessment checklist as a rule that must be processed by a preservation system. For each such rule, we identify the state information that must be provided to support the execution of the rule. We can then validate the trustworthiness of the archive based upon the state information that is generated by the application of the rule.

We observe that the mapping of assessment criteria to the management policies planned for the DSpace/Storage Resource Broker (SRB) system is not straightforward, nor one-to-one. Multiple assessment criteria may apply to a particular management policy. We address this issue by examining the mapping from management policies to preservation capabilities (i.e. functional requirements for the preservation system), such as those presented for the NARA Electronic Records Archive. We then map from the preservation capabilities to rules that control application of preservation services. The outcomes from applying the rules are saved as technical preservation metadata that can be examined to determine how well the management policies are being enforced according to the as-sessment criteria.

Finally, we postulate how a preservation environment might be assessed to insure that the system is complete: no required processes are overlooked and no required preservation metadata attributes are missing. This can be accomplished by doing one additional mapping, from the technical preservation metadata to a community standard such as the PREMIS metadata. For each required PREMIS preservation attribute, there should be an associated repository assessment criterion, a related management policy, and a set of rules and technical metadata for the application of associated preservation services. Similarly, for each assessment criteria, there should be corresponding standard preservation metadata that can be examined to determine the trustworthiness of the preservation environment. When the system is closed, that is, an assessment criterion has been defined for each preservation metadata attribute, and a set of preservation metadata exists for each assessment criterion, one can expect the system to provide self-consistent preservation management. We believe this methodology can ultimately lead to implementations of preservation environments that are provably self-consistent.

Applying Preservation Assessment Criteria to Management Policies

In 2005 the Research Library Group (RLG) and the US National Archives and Records Administration (NARA) jointly developed *An Audit Checklist for the Certification of Trusted Digital Repositories* (RLG/NARA, 2005) with the intention of providing digital archivists with criteria for assessing the "trustworthiness" of a particular digital repository or archive (often referred to as a trusted digital repository or TDR). The checklist[1] takes a top-down approach to the attributes of trusted digital repositories, drawing on the OAIS reference model (CCSDS, 2002) to define four categories of attribute: related to the organization supporting the digital repository; repository functions, processes and procedures; the designated community and usability of information in the repository; and technologies and technical infrastructure of the repository. The checklist has been applied to a number of operational digital archives and preservation environments, and has become a rendezvous point for discussion about the meaning of, and metrics for, authenticity, integrity, and other aspects of digital preservation.

In contrast, the PLEDGE Project initially developed a set of policies that were drawn from the operations of two preservation environments: the DSpace digital asset management software (n.d.) in use at MIT and other research institutions, and the Storage Resource Broker (SRB) distributed data management software from the San Diego Supercomputer Center[2]. These two systems were made interoperable in a previous project so DSpace can serve as a local data curation system while the contents are stored in the SRB data management system. The identified policies were organized into a structure that reflects the typical data model of any archive: enterprise, archive, collection (or record series), and item. In this model the enterprise is equivalent to the TDR checklist's "organization" category, while the archive, collection and item levels include policies from the "functions, processes and procedures" as well as the "technologies and technical infrastructure" categories of the checklist.

The "designated community and usability" category from the RLG/NARA checklist includes policies that manage the generation of dissemination information. In practice, for general purpose digital archives such as DSpace@MIT, these policies often apply to usage by the broader public (e.g. compliance with US ADA regulations, user privacy policies, user support procedures) rather than by a specific designated community. But specifying one or more designated communities as policies, with or without specific usability expectations, is certainly possible in the framework as "specifications of assertions", similar to local regulatory requirements and other

[1] A new version of the checklist was published in early 2007 (OCLC/CRL, 2007). The authors are in the process of mapping policies to the new checklist and will use the new criteria prospectively.
[2] SRB and iRODS are documented at the San Diego Supercomputer Center website http://www.sdsc.edu/srb/index.php/Main_Page

general contextual information about the archive. A complexity of such general assertions is that their auditability over time is difficult to automate, since it involves assessments by the designated community as to the usability of the digital records at unpredictable points in time with no feedback mechanism.

A goal of the PLEDGE Project is to develop its policies and rules engines in such a way that working archivists can interact with the system to specify the policies and maintain them over time. During a project review by advisors from the archives community it was decided to remodel the PLEDGE policies to conform more closely to the RLG/NARA TDR checklist's organizational structure, since it is already familiar to a number of archivists working with digital content. Policies are generally abstract, so for each high-level policy a "concrete policy" is provided to help explain the abstract policy (and so are descriptive rather than normative). This work has been completed and the policies are now ready to recast as capabilities and associated rules for the two systems that will document and enforce the policies (i.e. DSpace and i-RODS).

Applying Management Policies to Capabilities from the Preservation Environment

It is our assumption that archivists who manage digital archives work primarily at the policy level, with occasional specific constraints on particular collections or records (e.g. a patent hold or an access restriction specified in a donor agreement). However preservation systems necessarily function at the rules level, which are specific to the particular capabilities required of the system, and the rules engine implemented by it. Translations must be made between abstract and concrete policies, and between concrete policies and system requirements, or capabilities, that they imply. These in turn determine the rules that will ultimately enforce the policies in a given system, along with the metadata that the policies define and the rules engines require. This mapping from concrete policies established by archivists at the data curation level (e.g. DSpace) into specific system capabilities that are enforced at the data management level (e.g. SRB) will normally be done by technology experts developing the rules engines, rather than the archivists themselves. We are, however, attempting to standardize the policies, the way that they are expressed (the policy expression language), and the protocol for transferring them between preservation environments. In this way, archivists can set policies at the appropriate level, and have the same policy mapped to multiple rules engines and enforced by multiple preservation environments.

An Example of Translating Policies to Rules

An archivist determines that a particular item (e.g. a dataset of very high value to the community of scientists who work in that field) should have at least five identical copies made and distributed across geopolitical regions to help insure its long-term availability, while maintaining its authenticity and integrity across all copies. The archivist specifies the policy in DSpace, which records it in a standard way, with a standard policy expression language (PEL), and starts its routine storage procedure. The policy is then translated in the rule set that DSpace requires to enforce the policy, which dictates that one copy be stored locally and a second copy, with its policies, be sent to iRODS for further data storage and management. iRODS receives the item and its policies, translates them into its own rules set, makes the requested five distributed copies, and records the state to trigger a periodic rule to check the authenticity and integrity of each copy once a month.

If a given preservation environment lacks a particular capability that a policy implies, the mapping from management policies to preservation capabilities will fail and the policy will devolve to an assertion that cannot be verified. This defines one essential component of a trustworthy preservation environment, that it support all capabili-ties required to implement assessment criteria.

Preservation Operation Control

Rule Types for Preservation Systems

An early insight of the project was that policies drawn from operational digital archives could be recast as rules that the archival preservation system could enforce with a rules engine. Once such a rules engine exists, the possibility presents itself of automatic enforcement to provide an auditable means of verifying a given system's "trustworthiness" as defined by the RLG/NARA TDR checklist (or the management policies that the checklist includes).

Because of this, the PLEDGE policies were mapped into rule types and the state information needed to record the outcomes of applying the rules. We draw from four categories of rules:

- Specification of assertions about the enterprise, the specific archive, its content, and its legal and regulatory environment. These assertions are often not machine-enforceable, but rather define state information required by lower-level rules. They are most often described at the enterprise level (i.e. across all the archives managed by a particular organization in its particular legal and regulatory environment).
- Aperiodic, or deferred consistency rules are those that can be applied at any time to enforce assertions made about an archive, collection or item. They are normally applied at the archive or collection level, for example, occasionally verifying that every item in the archive has a unique identifier from a particular identification system assigned to it.
- Periodic rules are normally applied at the collection or item level, and are driven by mandates for periodic validation of integrity. An example is a nightly audit of each item in the archive to verify that its checksum has not changed from the night before (a standard integrity check).
- Atomic rules are those which occur on execution of a specific event (in the event-condition-action model) and are most often evaluated at the item level on each execution of a related operation. An example is assigning an approved submission agreement to an item at the point of ingest into the repository (i.e. the item should not be accepted without an approved agreement attached to it, so this rule cannot be deferred).

Mapping Preservation Capabilities to Rules

We can illustrate the rules mapping process by examining the mapping of the Electronic Records Archives capability requirements[3] to rules that can control the execution of preservation processes. For each rule, we can define technical preservation metadata (i.e. state information) that records the outcome of the rule application. Thus for each ERA capability, there is a set of rules, preservation services, and technical metadata that are required for implementation. We can then compare the technical metadata with community standards for preservation metadata, such as that defined by PREMIS. A complete system will have an assessment criterion for each preservation metadata attribute. And for each assessment criterion, technical metadata will exist which can be examined to verify trustworthiness. When all assessment criteria can be expressed as required preservation metadata, and when all preservation metadata have an assessment criteria, the preservation system can be considered self-consistent and complete.

The ERA capabilities list defines 854 key capabilities (or functional requirements) needed for preservation. The capabilities can be loosely organized into categories related to:

- Management of disposition agreements describing record retention and disposition actions
- Accession: the formal acceptance of records into the data management system
- Arrangement: the organization of the records to preserve a required structure (implemented as a collection/sub-collection hierarchy)
- Description: the management of descriptive metadata as well as text indexing
- Preservation: the generation of Archival Information Packages
- Access: the generation of Dissemination Information Packages
- Subscription: the specification of services that a user picks for execution
- Notification: the delivery of notices on service execution results
- Queuing of large scale tasks through interaction with workflow systems
- System performance and failure reports. Of particular interest is the identification of all failures within the data management system and the recovery procedures that were invoked.
- Transformative migration, the ability to convert specified data formats to new standards. In this case, each new encoding format is managed as a version of the original record.
- Display transformation: the ability to reformat a file for presentation.
- Automated client specification: the ability to pick the appropriate client for each user.

For each capability, we examined the rules that were required to execute data manipu-lation processes, and defined the technical metadata needed to record the outcome of applying the rule. This defined 174 generic rules related to the manipulation of records, the tracking of assertions about submission and disposition agreements, validation of archival information packets, the generation of accounting reports, the management of access controls, etc. The generic rules in turn required the

[3] The Electronic Records Archive capabilities list defines a comprehensive set of capabilities needed to implement a preservation environment (ERA, 2003).

specification of 212 preservation technical metadata attributes. These metadata attributes defined properties that were needed to describe the storage resources, the users of the preservation system, the records, the collections that organized the records, and even the rules themselves. The projection from ERA capabilities to rules on services with technical outcome metadata is being reviewed, but will form the basis for a rule-oriented digital preservation system.

Preservation Operation Outcomes Description

Each policy and rule set that we have defined, as well as those specified by the RLG/NARA TDR checklist, may have associated metadata and state information necessary for its correct interpretation and enforcement. Part of our work has been to identify the policy metadata that should accompany the policy and incorporate it into our policy expression language. Rules engines that enforce archival policies must have some mechanism to record and manage detailed policy management outcomes over time and associate this state information with the relevant policies and preservation operations.

An example of an “organization”-level policy that would be mapped to an assertion rule is the accounting standard to be used by an archival organization/enterprise, as part of a means to determine the economic viability, and so long-term sustainability, of a given digital repository. The definition of this repository policy is that “repository business planning and practices are transparent, compliant with relevant accounting standards and practices, and auditable”. The associated metadata and state information for this policy includes:

- Location X of business planning and practice documentation
- Location Y of accounting standards documentation
- Period P of accounting audit
- Date D of last accounting audit
- Name A of agent responsible for audit

And the concrete policy that derives from this abstract policy might be

- Repository exposes its business planning and practice documentation at [X].
- Repository complies with accounting standards at [Y].
- [A] will audit the repository’s business planning and practice documentation at [X] following accounting standard [Y] on the date determined by [D + P] and update [D].

Another example from a more typical preservation operation of content access is illustrative. The policy is defined so that “the repository ensures that agreements applicable to access conditions are adhered to. Repository access management systems fully implement access policy. Repository logs all access management failures, and staff review inappropriate “access denial” incidents.” This policy might have the following associated policy metadata:

- Location X of access agreements
- Agent A responsible for access management failure review
- Date D of last access management review
- Period P of access management review

And the corresponding concrete policy might be

- Repository enforces access agreements at X
- [A] will review access management failures on [D + P]

This policy lends itself to machine enforceability via a rules engine. To do that, much more specificity is required. The access control rules would be atomic (applied at every attempt to access an item), and would require the following state information for enforcement:

- Roles for types of access permitted
- Names of people with permission for each role
- Flags for access restriction to relevant collection and/or items
- Location of audit trail for attempted accesses (successful or failed)

And the relevant rule would implement access control over all collections and/or items such that only people with roles that permit access to the requested collection or item would be allowed that access, and every access attempt would be logged in an audit trail for future review.

Note that the translation from policy metadata to rule state information is non-trivial. It assumes that the documentation typically referenced by the policy metadata includes significant detail and preferably in a standard way so that rules builders do not need to speculate about the desired system behavior and outcomes. We hope to help this process by further defining the policy metadata through as much system-neutral state information as is possible.

Mapping State Information to Preservation Metadata Standards

Once the state information needed by a preservation system's rules engine is defined, it is desirable to compare that state information to relevant standards for preservation metadata. This is both to ensure that all the necessary state information is being kept, as well as to test the metadata standards for completeness and consistency.

We have begun this by comparing state information derived from the NARA ERA preservation system's requirements specification (with 212 elements of state information defined) to the PREMIS metadata specification[4], with its 68 metadata attributes. This resulted in the identification of attributes that were unique to either the PREMIS schema or the NARA ERA requirements, and pointed to areas where the ERA requirements were more specific than PREMIS metadata would support, or vice versa. This will help us to refine the set of state information needed for preservation environments and rules engines, as well as inform the PREMIS work going forward.

We are using this assessment to perform the following end-to-end analysis of a preservation environment based on the DSpace and iRODS technologies. The goal is to map from community standards for assessment criteria and community standards for preservation metadata to preservation capabilities that can be expressed as rules on preservation services. The steps require the definition of:

1) Community standard for assessment criteria. An example is the

[4] PREservation Metadata: Implementation Strategies (PREMIS) defines a metadata schema to support digital preservation activities and digital lifecycle management. The current schema can be found at the PREMIS webpage (PREMIS WG, n.d.).

RLG/NARA TDR.

2) Local data management policies resulting from the assessment criteria in a particular archival setting
3) Generic capabilities required for preservation. An example is the ERA capabilities list
4) Rules for executing the management policies that result from the assessment criteria and that are needed to manage preservation capabilities
5) Micro-services that implement the remote preservation operations
6) Technical information that records the outcomes of application of the rules
7) Community standard for preservation metadata. Examples are the Life Cycle Data Requirements Guide or PREMIS metadata.

We can then show the completeness and self-sufficiency of the whole system, such that:

- For each assessment criteria we can generate appropriate persistent state information. This implies that we can evaluate the assessment criteria by examining the preservation metadata.
- For each preservation metadata attribute, there exists an assessment criterion. This implies there are governing management policies that can be derived from the assessment criteria, and that the application of the management policies results in the set of preservation metadata that are provided by the preservation environment.

Conclusions

With this project we have attempted to present an end-to-end description of the management properties needed in a preservation environment, from assessment criteria through to the rules that express the management policies and the descriptive and technical metadata needed to validate the assessment results. We have secondarily described the iteration between metadata standards and state information needed to track assessment criteria. The purpose of this exercise is to demonstrate how it is possible to develop preservation systems that are subject to rigorous assessment. Automatically generated metadata can be examined through straightforward (and possibly automatic) review to determine trustworthiness. We believe that this will allow preservation environment to scale appropriately in the coming decades.

Acknowledgements

This project was supported by 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

CCSDS. (2002). Reference model for an Open Archival Information System (OAIS). Retrieved on June 14, 2007 from the Consultative Committee for Space Data Systems (CCSDS) website: http://public.ccsds.org/publications/archive/650x0b1.pdf

DSpace. (n.d.). Retrieved on June 14, 2007 from http://dspace.org/

ERA. (2003). Electronic Records Archives (ERA) requirements document. (ERA RFP# NAMA-03-R-0018, attachment 2 to Section J, amendment 0001). Retrieved on June 14, 2007 from http://www.archives.gov/era/pdf/requirements-amend0001.pdf

OCLC/CRL. (2007). Trustworthy repositories audit & certification (TRAC): Criteria and checklist. Retrieved on June 14, 2007 from http://www.crl.edu/PDF/trac.pdf

PREMIS WG. (n.d.). PREservation Metadata: Implementation Strategies Working Group documents. Retrieved on June 14, 2007 from http://www.oclc.org/research/projects/pmwg/

RLG/NARA. (2005). An audit checklist for the certification of trusted digital repositories. Retrieved on June 14, 2007 from http://www.rlg.org/en/pdfs/rlgnara-repositorieschecklist.pdf