Assessment of RLG Trusted Digital Repository Requirements

Reagan W. Moore
San Diego Supercomputer Center
9500 Gilman Drive
La Jolla, CA 92093-0505
01 858 534 5073
moore@sdsc.edu

MacKenzie Smith
MIT Libraries
77 Massachusetts Avenue
Cambridge, MA 02139-4307
01 617 253 8184
kenzie@mit.edu

ABSTRACT
The RLG/NARA trusted digital repository (TDR) certification checklist defines a set of management policies that establish the characteristics of a repository for digital preservation. We explore how these management policies can be mapped onto a set of repository management policies derived from the DSpace/SRB software system. We examine a mapping of the system management policies to rules, the definition of required policy state information, the types of rules that are required to implement the policies, and the desired level of granularity for application of the policies. This approach exposes general design considerations that should be met by repositories audited with the TDR checklist.

Categories and Subject Descriptors
H.3.4 [**Systems and Software**]: Distributed systems, D.2.6 [**Management**]: life cycle, software process models, H.3.6 [Library Automation]: Large text archives

General Terms
Management, Documentation, Verification.

Keywords
Rule-based consistency management

1. INTRODUCTION
The Research Library Group, in collaboration with the National Archives and Records Administration, has published "An Audit Checklist for the Certification of Trusted Digital Repositories" [TDR] [1]. The checklist defines a set of management policies that are organized into criteria for the Organization; Repository Functions, Processes, and Procedures; The Designated Community & the Usability of Information; and Technologies & Technical Infrastructure. The document does not specify the implementation of the management policies. We examine the set of rules and associated state information required to automate the verification of the trusted digital repository. In effect, we attempt to define the set of rules that validate the trustworthiness of a repository.

2. APPROACH
The DSpace digital asset management software [2], in combination with the Storage Resource Broker (SRB) [3] distributed data management software, supports the implementation of a trusted digital repository for long-term preservation. DSpace provides management of standard processing steps for the curation of records. The SRB provides management of the digital entities that may be replicated or distributed across multiple storage systems [4]. Together the two systems provide the capabilities needed to implement a trusted digital repository.

The RLG audit checklist can be applied to the DSpace/SRB system as implemented by a given organization to determine whether all the management policies are being adequately met to insure long-term preservation of the contents. In this paper we go one step further, and seek to have the DSpace/SRB system automatically validate the trustworthiness defined by local policy decisions. Our approach is based upon the characterization of each item in the checklist as a rule that must be processed. For each rule, we identify the state information that must be provided to drive the execution of the rule. We then validate the trustworthiness based upon the state information that is generated by the application of the rule. This provides a mechanism to assert how the trusted digital repository is being managed, and also provides the information that is needed to validate the assertion.

In our design process, we discovered many implications within the TDR criteria that impact the ability to describe trustworthiness. We encapsulate these implications in the following observations:

1. The assessment criteria can be mapped to management policies.
2. Implementation of many of the management policies requires mapping to a set of rules that can be automatically evaluated.
3. Management rules require definition of state information that defines the assertion being implemented, as well as attributes that encapsulate the result of the application of the rule.
4. The types of rules that are needed include:
   - Specification of assertions (setting flags and descriptive metadata)
   - Deferred consistency constraints that may be applied at any time (i.e. a-periodic)
   - Periodic rules that execute defined procedures
   - Atomic rules applied on each operation (access controls, audit trails)
5. The level of management granularity on which the rules are applied encompasses the enterprise level, the archives level, the collection (record-series) level, and the item level. A specification of the multiple levels of management granularity is needed to understand how to apply the TDR assessment criteria.
6. The rule that is applied at each level of granularity may differ, even though the same assessment criteria are being applied. This is one of the most important observations: that each management policy may require the definition of multiple rules that are applied at different levels of granularity.
7. Within the DSpace/SRB environment, additional management policies are needed beyond those specified in the TDR document. These include policies related to business case; security architecture; open source software license; user privacy; retention schedule; disposition; destruction of records; withdrawal of records; risk management; protection – data staging; and audit frequency. These should be evaluated for possible inclusion in a future version of the TDR checklist.
8. The actual implementation of the assessment criteria is dependent upon the persistence of the name spaces on which the management policies are applied [5]. The management policies need to apply to persistent identifiers for users, files, metadata and constraints; the name space for defining management state information; and the name space for physical resources (storage systems, databases).
9. The trusted preservation repository should implement multiple levels of virtualization to enable migration onto new technology without impacting the ability of the system to meet the assessment criteria. In practice this includes both persistent name spaces and standard operations for interacting with storage systems (data virtualization), authentication environments (trust virtualization), and rule expression engines (constraint virtualization) [6].

We also observe that the choice of levels of granularity impacts the types of rules that are needed. The rules used at the enterprise level are typically assertions that define the state information required by rules executed at finer levels of granularity. The deferred consistency constraints are typically applied at the collection level to enforce assertions made on the collection. An example is checking compliance of Submission Information Packages with Service Level Agreement specifications. The periodic rules are applied at the collection (record series) level, and are driven by mandates for periodic validation of integrity. An example would be the validation of integrity every 6 months. The atomic rules are evaluated at the item level on each execution of a related operation. The standard example is the checking of access controls before an operation is performed upon a file.

If additional levels of granularity are defined, such as a record group level, one concern is that additional types of rules may be required. In practice, we expect only these four types of rules. This implies that a rule engine that is capable of executing all four rule types should be able to automate validation of the trustworthiness of a digital repository. As part of an NSF funded information technology research project, SDSC is designing an intelligent rule oriented data systems (iRODS). This system will support the four types of rules that have been identified (specification of assertions, deferred consistency constraints, periodic rules, and atomic rules). Thus the iRODS system should be able to execute the rules that implement the management policies. By examining the results generated by the application of the rules, iRODS will be able to track whether the management policies are being met. This constitutes automation of the validation of trustworthiness of a digital repository.

Finally, we observe that the mapping of the certification criteria to the management policies planned for the DSpace/SRB system is not one-to-one. Multiple assessment criteria may apply to a particular repository management policy. We address this issue by explicitly listing each time when the assessment criteria should be applied, and the additional rules that are applied.

3. RULES

To provide a flavor of the assessment, we list some example rule sets for selected TDR criteria. We select an

example from each level of granularity, including a case where the same TDR criteria must be evaluated at multiple levels of the data management hierarchy.

The left-most column in tables 1-4 is the management numbering scheme used in the DSpace/SRB policy assessment. The numbering scheme uses 1 for enterprise level, 2 for archives level, 3 for collection level, and 4 for item level. The second number identifies the management policy at that level of granularity. The second column lists the corresponding policy. The third column lists the TDR criteria number that most closely corresponds to the management policy.

Table 1. Enterprise Level Rule Example

#	4. Policy layers / types	TDR	Rule or procedure	State info - result of rule application	Description
1.5	Annual review of planning processes		Set / Update descriptive metadata	Timestamp of last planning process review	Annual process to review and adjust business plans
		A4.2	Set / Update descriptive metadata	List of dates of annual review process	Repository has in place at least annual processes to review and adjust business plans as necessary

Table 2. Archives Level

#	4. Policy layers / types	TDR	Rule or procedure	State info result of rule application	Description
2.14	Persistent identifiers		Consistency rule - check that handle was created	List of types of GUID. Lists of locations of handle systems for creating GUIDs	Management of mapping of identifiers to SIPs. Which type are assigned and to what? Are multiple identifiers for an item supported?
		B2.4	Set/update naming specification	Specification of standard naming convention for physical files	Repository has and uses a naming convention that can be shown to generate visible, unique identifiers for all AIPs
		B2.5	Set/update templates	Producer-archive submission pipeline for extracting descriptive metadata on ingest; Template based metadata extraction	If unique identifiers are associated with SIPS before ingest, they are preserved in a way that maintains a persistent association with the resultant AIP.

The 4th column lists the type of rule that is needed. The 5th column lists examples of the state information that are needed for either executing the rule, or for managing the result of the application of the rule. The right-most column provides an explanation of the policy.

In Table 1, there are two items listed for the policy entitled "Annual review of planning processes." The first row is the criteria as proposed within the DSpace/SRB system. The second row lists the corresponding TDR criteria.

Table 3. Collection Level

#	4.Policy layers / types	TDR	Rule or procedure	State info - result of rule application	Description
3.9	Service level agreements for collections		Set / Update flags	Flag for specification of type of service level agreement	Maintain a service level agreement for each collection. Specify required descriptive metadata by SIP type.
		A5.1	Set / Update descriptive metadata	Deposit agreement for storage of data specifying access, replicas, consistency checks	If repository manages, preserves, and/or provides access to digital materials on behalf of another organization, it has and maintains appropriate contracts or deposit agreements.

In Table 2, two assessment criteria from the TDR checklist should be applied to the management policy for persistent identifiers. The types of rules that are needed include both deferred consistency checking as well as setting of state information needed for rule validation. The persistent identifiers require mapping to the identifier used in the Archival Information Package (AIP) from the identifiers specified in the Submission Information Package (SIP).

In Tables 3 and 4, the same TDR criterion (A5.1) is applied at multiple levels of granularity. Rule A5.1 was applied at both the collection and item level. In addition to managing the service level agreement that specifies the required consistency checks, metadata is also needed to allow changes to the service level agreement to occur. For

the item level rule, we also listed the additional TDR criteria that were applied. This indicates that multiple assessment criteria are applicable for a given policy. The validation of the data format requires checking rules related to Service Level Agreements, AIP definitions, allowed transformative migrations, and association of metadata with each file.

Table 4. Item level:

#	4. Policy layers / types	TDR	Rule or procedure	State info - result of rule application	Description
4.2	Format		Periodic rule - check consistency with required formats	List of supported formats and flag for SLA support level for each	Whether file format is accepted, preservation SLA for each accepted format; Also any requirements for quality within format (e.g. compliance with TIFF 6.0 acceptance specs)
		A5.1	Consistency rule - check that deposit agreement exists	Deposit agreement for storage of data specifying access, replicas, consistency checks	If repository manages, preserves, and/or provides access to digital materials on behalf of another organization, it has and maintains appropriate contracts or deposit agreements.
		B2.1	Consistency rule that AIP definition exists	Statement of characteristics of each AIP	Repository has an identifiable, written definition for each AIP or class of information preserved by the repository
		B2.2	Consistency rule - check allowed transformative migration is performed	Criteria for allowed transformative migrations	Repository has a definition of each AIP (or class) that is adequate to fit long-term preservation needs
		B3.9	Set / Update descriptive metadata: Consistency check for changes to allowed transformative migrations	Procedure for updating transformative migration strategy: Audit trail of changes; Consistency check for changes to migration strategy	Repository has mechanisms to change its preservation plans as a result of its monitoring activities
		B4.2	Consistency rule - check required metadata	Validation that minimum descriptive metadata is present	Repository captures or creates minimum descriptive metadata and ensures that it is associate with the AIP

The full assessment of the TDR criteria takes 13 pages to print in 8-point type. The complete mapping is available upon request. Please contact Reagan Moore at moore@sdsc.edu for a copy or visit the project website at http://simile.mit.edu/pledge/.

4. ACKNOWLEDGMENTS

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". 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.

5. REFERENCES

[1] Audit Checklist for Certifying Digital Repositories, http://www.rlg.org/en/page.php?Page_ID=2076

[2] DSpace digital repository, http://www.dspace.org/

[3] Storage Resource Broker data grid, http://www.sdsc.edu/srb

[4] Moore, R., A. Rajasekar, 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, Vol. 93, No.3, pp. 578-588, March 2005.

[5] Moore, R., "Building Preservation Environments with Data Grid Technology", American Archivist, vol. 69, no. 1, pp. 139-158, July 2006.

[6] Moore, R., R. Marciano, "Technologies for Preservation", chapter 6 in "Managing Electronic Records", edited by Julie McLeod and Catherine Hare, Facet Publishing, UK, October 200