Classifying Implemented Policies and Identifying Factors in Machine-Level Policy Sharing within the integrated Rule-Oriented Data System (iRODS)

Jewel H. Ward

SILS & DICE
University of North Carolina at Chapel Hill
Chapel Hill, NC 27599-3360
jewel_ward@unc.edu

Abstract

We examine the development of policy standards for trusted preservation repositories. We briefly discuss the lack of implementation and sharing of computer actionable policies by the communities that manage data. We propose a mixed-methods study in order to empirically examine the motivating and discouraging factors for machine-level policy sharing and to classify implemented policies and identify any patterns using the integrated Rule Oriented Data System (iRODS) as the case study, repository policies and iRODS rules for the content analysis, and data grid administrators as study participants.

Index Keyword Terms**—** Data Management, Machine-level Policy implementation, iRODS, Standards

1. Introduction

Librarians, archivists, ILS researchers and computer scientists have made substantial efforts over the past two decades to define the requirements for implementing a trustworthy digital preservation repository [1,2,3,4,5,6]. Their efforts have culminated in the creation of general guidelines and policies, such as the Reference Model for an Open Archival Information System (OAIS) [7], Trustworthy Repositories Audit & Certification (TRAC) [8], ISO MOIMS-RAC [9,10], and self-audits such as DRAMBORA [11]. These practitioners and researchers have sought to establish and define the characteristics of a trustworthy repository, and then provide the mechanisms to verify that a repository actually does fulfill the established criteria for trustworthiness [9,10].

2. Background and Rationale

One way to audit and verify that the administrators of a digital repository meet and enforce these standards for trustworthiness is to automate the policies within the system itself. For example, the creators of the middleware data grid system, the integrated Rule Oriented Data System (iRODS), designed it so that repository managers and administrators could implement archival policies at the storage-level as "rules" [12]. This system built on previous data grid research [13,14] and experience with the Storage Resource Broker (SRB) [15,16]. One research project, called PLEDGE [17], bridged the gap between written policies that “should” be implemented by identifying which TRAC guidelines “could” be implemented at the machine-level within SRB.

While the primary idea behind establishing trusted repository audit mechanisms is to ensure that a repository actually meets and enforces archival standards [9], the other benefit of using standards is reduced costs [18]. By creating a standard set of machine-level automated repository policies based on archival standards, standard administrative preservation functions can be automated. The amount of human intervention needed in the digital preservation process will be reduced because these policies may then be shared among and between communities. This will streamline the digital archive process and potentially reduce long-term costs, thus aiding the longevity of the archive, as well as the integrity of the metadata and digital content it contains. The fewer resources that archivists require to maintain an archive, the more likely an institution or organization is to maintain it for the indefinite long-term.

In theory, the process of taking written policies and implementing them as machine-actionable rules should be fairly simple. In reality, the process of creating rules out of written guidelines has been slow to catch on within a variety of communities that use preservation repositories. The assumption behind the establishment of standards such as ISO MOIMS-RAC is that these policies will be implemented within repositories. Other practitioners and researchers have assumed that the most streamlined way to implement these standards is at the machine-level [19]. These are, after all, digital repositories. The creators of iRODS assumed that communities would coalesce and create rules and microservices [20]. While this has happened in some communities, it has not happened in others. The possible impediments include the learning curve for new technologies, whether the community has quantified their policies, whether the community has infrastructure that supports policy-based management for digital data, and whether the standard for assessment criteria represents a viable set of policies.

If the theory behind almost two decades of work by librarians, archivists, ILS researchers and computer scientists working on the digital preservation problem is that policies will be implemented at the machine-level, but in reality practitioners aren’t doing this, then some of the basic theories of digital preservation for the past 15 years are at stake.

If practitioners aren’t accepting these theories as “best practices” and implementing them, are the theories wrong, are they untenable, or, is the development of social consensus difficult? The current hypothesis within the iRODS group is that community members who manage preservation repositories have policies that are unique to their community, but they find it challenging to implement written policies at the machine-level, and that only a few communities have implemented more than a core (provided) set of policies. The goal of this study is to empirically examine the motivating and discouraging factors for machine-level policy sharing among iRODS users and partners, and to classify implemented policies and identify any patterns.

Most of the related literature examines the creation of the policies themselves. These authors have not examined the actual implementation of the policies within a preservation system; therefore, there is little to no previous direct work in this specific area.

3. What is a Policy?

A policy is “an informal, generally natural language description of desired system behavior. Policies may be defined for particular requirements, such as confidentiality, integrity, availability, safety, etc.” [21].

3.1. Example of a Human Readable Policy

One example of a human-readable policy is *chain of custody*. One part of a chain of custody policy would be, “the repository shall manage the number and location of copies of all digital objects….in order to assert that the repository is providing an authentic copy of a particular digital object.” [9].

3.2. Example of a Machine Readable Policy

One way to implement one part of chain of custody in iRODS is to write a computer actionable rule to audit the objects in the repository. For example:

```
    Get Audit Info By Object
Path||writeLine(stdout,'<?xml version="1.0"
encoding="ISO-8859-
1"?>')##writeLine(stdout,"<audit_trail>")##msiIs
Data(*objPath,*objID,*foobar)##msiGetAuditTrail
InfoByObjectID(*objID,*BUF,*Status)##writeByte
sBuf(stdout,*BUF)##writeLine(stdout,"</audit_trai
l>")|nop
    *objPath=/foo/bar/audit-info.rtf
    ruleExecOut
```

In the example above, the rule creates an XML file, pulls audit trail information by object ID and object Path, and writes it to the XML file. The XML file is machine- and human-readable and lists all operations performed upon the file. This makes it possible to identify the chain of custody of the file through the identification of the persons who manipulated the file.

4. Methods

The proposed study uses mixed methods (qualitative and quantitative). It is a case study in that only users of the iRODS system will be in the sample population studies, but it is primarily a content analysis in that we will be analyzing focus group responses, survey responses, written policies (human language) and machine-level policies (computer code).

4.1. Participants

Subjects must be data grid managers using iRODS. DICE (Data Intensive Cyber Environments) employees at UNC-CH and UCSD will be excluded from the subject pool. Subjects will be recruited via the iRODS-chat mailing list. Participants in the study must be willing to share their core.irb files (rule bases) and written/unwritten policies. They must be willing to be interviewed in person, by phone, or over the Web, and be able to complete an online survey.

4.2. Design

We will begin the study with a series of focus groups. We will analyze the focus group results and develop a questionnaire. We will examine users' written policies for themes, as well as the core.irb files of iRODS users. We will examine what is written vs. what is actually implemented in order to determine any discrepancies, classify the policies, and identify patterns. We will analyze these results, develop hypotheses on usage constraints, relate the hypotheses to existing theory on open source code development, and create a model for implementing machine-level policies. We will test the model by statistical analysis to determine the validity and the strengths of the relationships. Pending the results we will conduct interviews and then more analysis.

5. Results and Discussion

The results of the study will be evaluated in terms of the initial hypotheses, and the design of the study will be somewhat iterative. If necessary, adjustments may be made to the methodology based on the results of the focus group and initial survey. How the results relate to previous research and to the theoretical issues mentioned in the background and rationale will be discussed. The practical implications of the results will also be considered.

Any limitations or delimitations of the initial study will be noted, along with recommendations for how the larger research study may build upon this study. One inherent limitation of this study is that the results will not be generalizable. However, the results should provide some insight into the implementation of policies at the machine level. Finally, future work in this area will be proposed and any results will be summarized.

7. References

[1] C. A. Lynch, "The Integrity of Digital Information: Mechanics and Definitional Issues", *Journal of the American Society for Information Science*, 45(10), pp. 737-744, 1994.

[2] D. Waters and J. Garrett, "Preserving Digital Information", Report of the Task Force on the Archiving of Digital Information, CLIR, Washington, DC, May 1996.

[3] D. Bearman and J. Trant, "Authenticity of digital resources: towards a statement of requirements in the research process", *D-Lib Magazine*, The Corporation for National Research Initiatives (CNRI), Reston, VA, June, 1998.

[4] Council on Library and Information Resources, "Authenticity in a Digital Environment", Council on Library and Information Resources, Washington, DC, 2000.

[5] Research Libraries Group, "Trusted Digital Repositories: Attributes and Responsibilities an RLG-OCLC report", Research Libraries Group, Mountain View, CA, 2002.

[6] Research Libraries Group, "An Audit Checklist for the Certification of Trusted Digital Repositories, Draft for Public Comment", Research Libraries Group, Mountain View, CA, 2005.

[7] Consultative Committee for Space Data Systems, "Reference model for an Open Archival Information System (OAIS) (CCSDS 650.0-B-1)", National Aeronautics and Space Administration (NASA), Washington, DC, 2002.

[8] Online Computer Library Center, Inc. (OCLC) and Center for Research Libraries (CRL), "Trustworthy Repositories Audit and Certification: Criteria and Checklist v.1.0.", OCLC and CRL, Dublin, OH & Chicago, IL, 2007.

[9] Consultative Committee for Space Data Systems, "Audit and Certification of Trustworthy Digital Repositories (CCSDS 652.0-R-1)", National Aeronautics and Space Administration (NASA), Washington, DC, 2009.

[10] Consultative Committee for Space Data Systems, "Requirements for Bodies Providing Audit and Certification of Candidate Trustworthy Digital Repositories (CCSDS 000.0-R-0)", National Aeronautics and Space Administration (NASA), Washington, DC, 2009.

[11] Digital Curation Centre and Digital Preservation Europe, "DCC and DPE Digital Repository Audit Method Based on Risk Assessment (DRAMBORA)", Digital Curation Centre (DCC) and DigitalPreservationEurope (DPE), 2007.

[12] A. Rajasekar, M. Wan, R. Moore, and W. Schroeder, "A Prototype Rule-based Distributed Data Management System", paper presented at a workshop on "next generation distributed data management" at the High Performance Distributed Computing Conference, June 19-23, 2006, Paris, France.

[13] R. Moore and A. Merzky, "Persistent archive concepts", paper presented at the 7th Global Grid Forum, Tokyo, Japan, March 4-7, 2003.

[14] R. Moore, "Building Preservation Environments with Data Grid Technology", *American Archivist*, 69(1), Society of American Archivists, Chicago, IL, pp. 139-158, 2006.

[15] R. Moore, "Evolution of Data Grid Concepts", paper presented at the workshop on "data" at the 10th Global Grid Forum, Berlin, Germany, March 9-13, 2004.

[16] Moore, R., "Persistent Collections", in S.H. Kostow & S. Subramaniam (Eds.), *Databasing the Brain: from Data to Knowledge (Neuroinformatics)*, pp. 69-82, John Wiley and Sons, Hoboken, NJ, 2005.

[17] S. MacKenzie and R. Moore, "Digital Archive Policies and Trusted Digital Repositories", paper presented at the 2nd International Digital Curation Conference, November 21 - 22, 2006, Glasgow, Scotland.

[18] The Science and Technology Council, *The Digital Dilemma Strategic Issues in Archiving and Accessing Digital Motion Picture Materials*, The Science and Technology Council of the Academy of Motion Picture Arts and Sciences, Hollywood, CA, 2007.

[19] J. Hunter and S. Choudhury, "Semi-automated Preservation and Archiving of Scientific Data Using Semantic Grid Services", in Proceedings of the IEEE International Symposium on Cluster Computing and the Grid, May 9-12-2005, Cardiff, UK.

[20] R. Moore, A. Rajasekar and M. Wan, "Data Grids, Digital Libraries and Persistent Archives: an Integrated Approach to Publishing, Sharing and Archiving Data", in Proceedings of the IEEE (Special Issue on Grid Computing), 93(3), IEEE Computer Society, Washington, DC, 2005.

[21] NASDAQ, "Credit Card Glossary", NASDAQ, New York, NY, 2011.