Building a Reference Implementation of a Preservation Environment

Reagan W. Moore
Richard Marciano
Arcot Rajasekar
University of North Carolina, Chapel Hill

Mike Wan
Wayne Schroeder
University of California, San Diego

rwmoore@email.unc.edu
http://irods.sdsc.edu
Reference Implementation

• Turn-key system
  • Provide fundamental capabilities needed for preservation
  • Starter kit for building a production system

• Scalable system
  • Support both small and large collections
  • Support extensions for new data formats, new information syntax, new preservation policies

• Infrastructure independent approach
  • Enable use of your preferred vendor products
Collaborative Research

• NARA Transcontinental Persistent Archive Prototype (TPAP)
  • University of North Carolina, Chapel Hill
  • University of Maryland
  • University of California, San Diego

• NSF Software Development for CyberInfrastructure (SDCI)
  • University of North Carolina, Chapel Hill
  • University of California, San Diego

• EU Sustaining Heritage Access through Multivalent ArchiviNg (SHAMAN)
Reference Implementation

• Assessment criteria
  • Needed to validate preservation properties
  • Without validation, difficult to prove that the archives are being managed correctly

• Policies
  • Needed to manage the preservation environment
  • Consist of controls on the preservation procedures

• Procedures
  • Operations performed upon records - appraisal, accession, description, arrangement, preservation, access
  • Data management tasks - risk mitigation
Software Components

- Assessment Criteria
- Preservation Policies
- Preservation Procedures
- Data Grid

Storage  Database
Choosing Assessment Criteria

- Validate that the preservation environment enforces required preservation properties
  - Authenticity
  - Representation information
  - Integrity
  - Chain of custody
  - Original order
  - Retention and disposition
  - Security

- Trustworthiness - TRAC Trustworthy Repositories Audit and Certification
Preservation Policies

• Specify condition for whether and when a preservation procedure will be executed
• Specify how to compose a desired procedure from standard operations (micro-services)
• Specify how to recover from errors when executing the operations in a distributed environment (recovery procedures)
Preservation Procedures

• NARA Electronic Records Archive capabilities list
  • Specifies 852 capabilities that a preservation environment should provide

• Mapping to rules and procedures
  • About 200 separate procedures that can be combined to implement all of the capabilities
  • About 170 metadata attributes (state information) that are tracked as the procedures are executed
Using an iRODS Data Grid - Details

User asks for data
Data request goes to iRODS Server
Server looks up information in metadata catalog
Catalog tells which iRODS server has data
1\textsuperscript{st} server asks 2\textsuperscript{nd} to perform operation
The 2\textsuperscript{nd} iRODS server applies rules
State information is stored in the metadata catalog
Example Reference Implementation

- DSpace digital library interface
  - Provides standard access mechanisms
- Fedora digital library middleware
  - Enforces relationships on records
- iRODS data grid
  - Manages preservation policies and validates assessment criteria
- PostgreSQL database
  - Supports metadata catalog
- Sun Thumper
  - Modular storage system
For More Information

Reagan W. Moore
University of North Carolina, Chapel Hill
rwmoore@email.unc.edu

http://irods.sdsc.edu/