Policy-based Data Management: Planning Policies & Procedures

Reagan W. Moore
Arcot Rajasekar
University of North Carolina at Chapel Hill

http://irods.diceresearch.org
Policy-based Data Environments - iRODS

- **Purpose** - reason a preservation environment is assembled
- **Properties** - attributes needed to ensure the purpose
- **Policies** - controls for enforcing desired properties,
  - mapped to computer actionable rules
- **Procedures** - functions that implement the policies
  - mapped to computer actionable workflows
- **Persistent state information** - results of applying the procedures
  - mapped to system metadata
- **Assessment criteria** - validation that state information conforms to the desired purpose
  - mapped to periodically executed policies
Policy-based Data Management

The objectives behind a data management application can be enforced through policies that control operational procedures. Standard policy sets can be used to simplify planning. Examples of policy sets are provided for:

- University of North Carolina LifeTime Library
- DataNet Federation Consortium
Policies

• Retention, disposition, distribution, arrangement
• Authenticity, provenance, description
• Integrity, replication, synchronization
• Deletion, trash cans, versioning
• Archiving, staging, caching, data placement
• Authentication, authorization, redaction
• Access, approval, IRB, audit trails, report generation
• Assessment criteria, validation
• Derived data product generation, format parsing
• Federation of independent data grids
Policy-based Data Management

**Client**

**Provider**
- iRODS controlled workflows
  - Storage

**Provider**
- iRODS controlled workflows
  - Storage

**Logical Collection** (data grid)

Consensus on Policies and Procedures Controls the Data Collection
Data Virtualization

Map from the actions requested by the client to multiple policy enforcement points.

Map from policy to standard micro-services.

Map from micro-services to standard Posix I/O operations.

Map standard I/O operations to the protocol supported by the storage system.
Life-Time Library

• Student digital libraries
  – Enable students to build collections of
    • Photographs
    • MP3 audio files
    • Video
    • Class documents
    • Web site archive

• Resources provided by School of Information and Library Science
  – Collections range from 2 GBytes to 150 Gbytes
  – Number of files from 2000 to 12,000
Policy-based Data Management

- Client
- Provider iRODS controlled workflows
- SILS Storage
- Provider iRODS controlled workflows
- RENCI Storage

Life Time Library

Policies selected by SILS
Input from students
Policies

• Integrity
  – Replication
  – Checksums
  – Versioning
  – Strict access controls
  – Quotas
  – Metadata catalog replication
  – Installation environment archiving

• Ingestion
  – Automated synchronization of student directory with LifeTime Library
Enforcement

• Policies enforced by iRODS data grid
  – Some policies should be implemented before digital library creation
    • Versioning
    • Strict access controls
    • Quotas
  – Other policies
    • Replication, when turned on, replicated all initial files
    • Synchronization, student controlled
New Policies

• Students want to control:
  – Staging area into which laptop files are synchronized
  – Automated extraction of metadata – iTunes
  – Tagging of files
  – Integrity validation periodicity
  – Arrangement of files, sorting by type or class
  – Descriptive metadata – Dublin core or reserved words
DataNet Federation Consortium
Data Driven Science

• Implement national data grid
  – Federate existing discipline-specific data management systems to enable national research collaborations

• Enable collaborative research on shared data collections
  – Manage collection life cycle as the user community broadens

• Integrate “live” research data into education initiatives
  – Enable student research participation through control policies

Cyber-infrastructure Partners:
Univ. of North Carolina, Chapel Hill
Univ. of California, San Diego
Arizona State University
Drexel University
Duke University
University of Arizona
University of South Carolina

Science and Engineering Initiatives:
Ocean Observatories Initiative
the iPlant Collaborative
CUAHSI
CIBER-U
Odum Social Science Institute
Temporal Dynamics of Learning Center

National Science Foundation Cooperative Agreement: OCI-0940841
Policy-based Data Management

Each domain implements unique policies in their data grid.

Policies control interactions between data grids.
Types of Federation

• Within a data grid
  – Assemble distributed data into a policy-controlled collection

• Between data grids
  – Establish policies for sharing between communities

• Soft links
  – Register data from another system into a collaboration environment
  – Access data using the remote system protocol
Applications

• International projects
  – International Neuroinformatics Coordinating Facility, Cyber Square Kilometer Array (radio astronomy), Cinegrid (movies)

• National data grids
  – Australia-ARCS, New Zealand, Portugal, UK, France-IN2P3

• Federal agency archives
  – NASA Center for Climate Simulation, National Optical Astronomy Observatories, NSF XSEDE

• Institutional repositories
  – Carolina Digital Repository, UNC-SILS LifeTime Library, Texas Digital Library, French National Library, Broad Institute genomics data grid, Sanger Institute data grid
iRODS - Open Source Software

Reagan W. Moore
rwmoore@renci.org
http://irods.diceresearch.org

NSF OCI-0940841 “DataNet Federation Consortium”
NSF OCI-1032732 “Improvement of iRODS for Multi-Disciplinary Applications”
NSF OCI-0848296 “NARA Transcontinental Persistent Archives Prototype”
NSF SDCI-0721400 “Data Grids for Community Driven Applications”