Policy-Driven Data Preservation

Reagan W. Moore
University of North Carolina at Chapel Hill
rwmoore@renci.org
http://irods.diceresearch.org

NSF OCI-0848296 “NARA Transcontinental Persistent Archives Prototype” (2008-2012)
NSF SDCI 0721400 “Data Grids for Community Driven Applications” (2007-2010)
Preservation is an Integral Part of the Data Life Cycle

• Organize project data into a shared collection
• Publish data in a digital library for use by other researchers
• Preserve reference collection for use by future research initiatives
• Compare new reference collection against prior state-of-the-art data
• Enable data-driven analyses that dynamically optimize research
Policy-based Data Management

iRODS - integrated Rule Oriented Data System

• Turn management policies into computer actionable rules
  – Support dynamic rule base updates

• Turn management processes into remotely executable computer procedures
  – Apply procedural workflow at the storage system to filter, subset, manipulate data
  – Minimize the amount of data pulled over the network
  – Automate administrative tasks

• Validate assessment criteria
  – Automate validation of collection properties
  – ISO MOIMS-rac
## Generic Data Management Systems

**iRODS - integrated Rule-Oriented Data System**

<table>
<thead>
<tr>
<th>Data Management Environment</th>
<th>Conserved Properties</th>
<th>Control Mechanisms</th>
<th>Remote Operations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Management Functions</td>
<td>Assessment Criteria</td>
<td>Management Policies</td>
<td>Management Procedures</td>
</tr>
<tr>
<td>Data Management Infrastructure</td>
<td>State Information</td>
<td>Rules</td>
<td>Micro-services</td>
</tr>
<tr>
<td>Physical Infrastructure</td>
<td>Database</td>
<td>Rule Engine</td>
<td>Storage System</td>
</tr>
</tbody>
</table>
Overview of iRODS Data System

User

Can Search, Access, Add and Manage Data & Metadata

iRODS Data System

iRODS Data Server
 disk, tape, etc.

iRODS Rule Engine
 track policies

iRODS Metadata Catalog
 track data

*Access data with Web-based Browser or iRODS GUI or Command Line clients.
Policies

• Administrative
  – Retention, disposition, distribution, replication, deletion, registration, synchronization, integrity checks, IRB approval flags, addition of users, addition of resources

• Ingestion / Access
  – Metadata extraction, logical organization, derived data product generation, redaction, time-dependent access controls

• Validation
  – Authenticity verification, chain of custody, repository trustworthiness, audit trails