On Building a Reference Implementation of a Preservation Environment

Reagan W. Moore
Richard Marciano
Arcot Rajasekar
Mike Wan

{moore,marciano,rajasekar,mwan}@diceresearch.org
http://irods.diceresearch.org
http://srb.diceresearch.org
Preservation Environment Reference Implementation

- Starter kit for assembling a preservation environment
- Provides initial
  - Assessment criteria
  - Preservation policies
  - Preservation procedures
  - Preservation clients
  - Preservation framework
<table>
<thead>
<tr>
<th>Data Grid</th>
<th>NSF / NVO</th>
<th>17,800</th>
<th>5,139</th>
<th>51,380</th>
<th>8,690</th>
<th>80</th>
<th>88,216</th>
<th>14,550</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>NSF / NPACI</td>
<td></td>
<td>1,972</td>
<td>1,083</td>
<td>17,578</td>
<td>4,694</td>
<td>380</td>
<td>59,567</td>
<td>7,888</td>
<td>380</td>
</tr>
<tr>
<td>Hayden</td>
<td></td>
<td>6,800</td>
<td>41</td>
<td>7,201</td>
<td>113</td>
<td>178</td>
<td>8,013</td>
<td>161</td>
<td>227</td>
</tr>
<tr>
<td>Pzone</td>
<td></td>
<td>438</td>
<td>31</td>
<td>812</td>
<td>47</td>
<td>49</td>
<td>37,918</td>
<td>25,117</td>
<td>68</td>
</tr>
<tr>
<td>NSF / LDAS-</td>
<td></td>
<td>239</td>
<td>1</td>
<td>4,562</td>
<td>16</td>
<td>66</td>
<td>253,203</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NSF / SLAC-</td>
<td></td>
<td>514</td>
<td>77</td>
<td>4,317</td>
<td>563</td>
<td>47</td>
<td>34,423</td>
<td>3,304</td>
<td>55</td>
</tr>
<tr>
<td>NSF / TeraGrid</td>
<td></td>
<td>80,354</td>
<td>685</td>
<td>2,962</td>
<td>293,907</td>
<td>7,392</td>
<td>3,267</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NIH / BIRN</td>
<td></td>
<td>5,416</td>
<td>3,366</td>
<td>148</td>
<td>22,900</td>
<td>45,728</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NCAR</td>
<td></td>
<td>76,255</td>
<td>436</td>
<td>2</td>
<td>708</td>
<td>76</td>
<td>2</td>
<td>529</td>
<td></td>
</tr>
<tr>
<td>LCA</td>
<td></td>
<td>5,756</td>
<td>80</td>
<td>2</td>
<td>529</td>
<td>708</td>
<td>2</td>
<td>529</td>
<td></td>
</tr>
<tr>
<td>Digital Library</td>
<td>NSF / LTER</td>
<td>158</td>
<td>3</td>
<td>233</td>
<td>6</td>
<td>35</td>
<td>260</td>
<td>42</td>
<td></td>
</tr>
<tr>
<td>NSF / Portal</td>
<td></td>
<td>33</td>
<td>5</td>
<td>1,745</td>
<td>48</td>
<td>384</td>
<td>2,620</td>
<td>53</td>
<td>460</td>
</tr>
<tr>
<td>NIH / AfCS</td>
<td></td>
<td>27</td>
<td>4</td>
<td>462</td>
<td>49</td>
<td>21</td>
<td>733</td>
<td>94</td>
<td></td>
</tr>
<tr>
<td>NSF / SIO</td>
<td></td>
<td>19</td>
<td>1</td>
<td>1,734</td>
<td>601</td>
<td>27</td>
<td>3,053</td>
<td>1,220</td>
<td>27</td>
</tr>
<tr>
<td>NSF / SCEC</td>
<td></td>
<td>15,246</td>
<td>1,737</td>
<td>52</td>
<td>168,931</td>
<td>3,545</td>
<td>73</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LLNL</td>
<td></td>
<td>18,934</td>
<td>2,338</td>
<td>5</td>
<td>182</td>
<td>106</td>
<td>3,267</td>
<td>53</td>
<td></td>
</tr>
<tr>
<td>CHRON</td>
<td></td>
<td>13,497</td>
<td>6,678</td>
<td>5</td>
<td>13,497</td>
<td>6,678</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Persistent</td>
<td>NARA</td>
<td>7</td>
<td>2</td>
<td>63</td>
<td>81</td>
<td>58</td>
<td>5,036</td>
<td>6,410</td>
<td>58</td>
</tr>
<tr>
<td>NSF / NSDL</td>
<td></td>
<td>2,785</td>
<td>20,054</td>
<td>119</td>
<td>8,618</td>
<td>85,004</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UCSD Libraries</td>
<td></td>
<td>127</td>
<td>202</td>
<td>29</td>
<td>5,880</td>
<td>2,094</td>
<td>29</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NHPRC / PAT</td>
<td></td>
<td>2,575</td>
<td>1,051</td>
<td>28</td>
<td>2,575</td>
<td>1,051</td>
<td>28</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RoadNet</td>
<td></td>
<td>4,902</td>
<td>2,674</td>
<td>30</td>
<td>4,902</td>
<td>2,674</td>
<td>30</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UCTV</td>
<td></td>
<td>7,140</td>
<td>2</td>
<td>5</td>
<td>7,140</td>
<td>2</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LOC</td>
<td></td>
<td>6,644</td>
<td>192</td>
<td>8</td>
<td>6,644</td>
<td>192</td>
<td>8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Earth Sci</td>
<td></td>
<td>7,867</td>
<td>707</td>
<td>5</td>
<td>7,867</td>
<td>707</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TOTAL</td>
<td></td>
<td>28 TB</td>
<td>6 mil</td>
<td>194 TB</td>
<td>40 mil</td>
<td>4,635</td>
<td>1,136 TB</td>
<td>216 mil</td>
<td>5,539</td>
</tr>
</tbody>
</table>
Concepts Driving Architecture

- **Infrastructure Independence**
  - Virtualization mechanisms needed to manage distributed data
  - Virtualization mechanisms needed to establish trust
  - Virtualization mechanisms needed to enforce management policies
  - Virtualization mechanisms needed to validate assessment criteria

- **Scalability**
  - Manage collections with 100 files or 100 million files
  - Manage collections with 10 Gigabytes or 10 Petabytes

- **Federation**
  - No single preservation environment is sufficient
  - Must be able to migrate records between environments
# 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>Policies</td>
<td>Procedures</td>
</tr>
<tr>
<td>Data Management Infrastructure</td>
<td>Persistent State</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>

**Data grid – Management virtualization**

**Data grid – Data and trust virtualization**
Preservation Assertions

- **Authenticity**
  - Maintain links between representation information, provenance information, descriptive information and record

- **Integrity**
  - Ensure original bits have been preserved through replicas, checksums, synchronization

- **Chain of Custody**
  - Ensure preservation procedures controlled by an identified archivist

- **Original Arrangement**
  - Maintain the relationships between records

- **Trustworthiness**
  - Verify assessment criteria
  - Trustworthy Repositories Audit & Certification (TRAC): Criteria and Checklist
Distributed Preservation Framework

- Archivist initiates preservation procedure
- Request goes to iRODS Server
- Server looks up information in catalog
- Catalog tells which iRODS server has data
- 1st server asks 2nd to perform procedure
- The 2nd iRODS server applies policies
Distributed Management Framework

- Rule Engine
- Metadata Catalog
- Execution Engine
- Execution Control
- Messaging System
- Data Transport

Virtualization

Persistent State information

Policy Management

Server Side Workflow
Preservation Procedures

• Appraisal
• Accession
• Arrangement
• Description
• Preservation
• Access

• Examples of 200 executable procedures definable from the Electronic Records Archive capabilities list
Multiple Policy Control Levels

- Access level
  - DSpace - accession procedures
- Digital library middleware level
  - Fedora - relationships between record
- Data grid
  - iRODS - administrative policies
- Storage system
  - Physical control
Preservation Environment Reference Implementation

- Assertions
- Policies
- Services
- Objects
- Administration
Example Starter Kit

- Assertions (TRAC)
- Policies (iRODS rules)
- Services (DSpace)
- Objects (Fedora)
- Administration (iRODS)
# Federation of Preservation Environments

<table>
<thead>
<tr>
<th>Federation Policies (iRODS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assertions (TRAC)</td>
</tr>
<tr>
<td>Policies (iRODS)</td>
</tr>
<tr>
<td>Services (DSpace)</td>
</tr>
<tr>
<td>Objects (Fedora)</td>
</tr>
<tr>
<td>Administration (iRODS)</td>
</tr>
</tbody>
</table>

| Assertions (ISO)          |
| Policies (Drambora)       |
| Services (ERA)            |
| Objects (METS)            |
| Administration (SRB)      |
Assertion

- Data management applications apply many of the same procedures
  - Data format parsing
  - Metadata manipulation
  - Data administration tasks
- Each application applies different management policies
  - Migrate records between data life cycle stages by changing the management policies
  - Can create generic infrastructure that can be used to implement collections, digital libraries, persistent archives
Reference Implementations

- Data grids
  - **Share data** - organize distributed data as a collection
- Digital libraries
  - **Publish data** - support browsing and discovery
- Persistent archives
  - **Preserve data** - manage technology evolution
- Real-time sensor systems
  - **Federate sensor data** - integrate across sensor streams
- Workflow systems
  - **Analyze data** - integrate client- & server-side workflows
For More Information

Reagan W. Moore
University of North Carolina, Chapel Hill
rwmoore@renci.org

http://irods.dicereresearch.org