Community-Driven Development of Preservation Services

Chien-Yi Hou

School of Information and Library Science (SILS)
Sustainable Archives & Leveraging Technologies Group (SALT)
Data Intensive Cyber Environments (DICE) Center
University of North Carolina at Chapel Hill

ISGC 2010, Taipei, Taiwan
Sustainability: a metaphor for sustainability

Richard Marciano
Sustainability

SALT/”yan”: a metaphor for sustainability

Richard Marciano
Sustainable Archives & Leveraging Technologies Group
Directed by Dr. Richard Marciano

governance

**e-Legacy:** preservation of geo-data and crowd-sourcing content

**T-RACES:** historical GIS

policy

**PoDRI:** Policy-Driven Repository Interoperability
**DCAPE:** Community policies & business models

infrastructure

**TIP:** RENCI+ RTP universities

evolution

---

4
Sustainable Archives & Leveraging Technologies Group

Directed by Dr. Richard Marciano

governance

e-Legacy: preservation of geo-data and crowd-sourcing content

T-RACES: historical GIS

policy

PoDRI: Policy-Driven Repository Interoperability

DCAPE: Community policies & business models

infrastructure

TIP: RENCI+ RTP universities

evolution

evolution
Overview of iRODS Architecture

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 information

*Access data with Web-based Browser or iRODS GUI or Command Line clients.*
• Each rule defines
  • An action for an event
  • Condition
  • Action chains (micro-services and rules)
  • Recovery chains
• Invoked by servers to enforce policies
• Invoked by clients to run workflows on servers
• Rule types
  • Atomic  --  applied immediately
  • Deferred  --  run at a later time in the background
  • Periodic  --  run at a fix time interval
iRODS Rule Example

- **Event**: acPostProcForPut
- **Condition**: $objPath$ like /HOU/home/rods/ISGC/*
- **Action**: msiDataObjRepl($objPath, Resource 2, *result)
- **Recovery**: nop
The PoDRI project investigates the requirements for policy-aware interoperability and demonstrates key features needed for its implementation.
PoDRI Use Cases

New content ingest via Fedora

Update of content or metadata via Fedora

Update of content or metadata via iRODS

Bulk registration of Fedora into iRODS

New content ingest via iRODS

The goal of the **DCAPE** project is to build a distributed production preservation environment that meets the needs of archival repositories for trusted archival preservation services.
INTEGRATION & BUS. DEV.

UNC
@ SALT
  Richard Marciano
  Chien-Yi Hou
@ CDR
  Dave Peolar ++

POLICY / RULE DEV.

@ West Virginia University
  Donald Adjeroh
  Frances Van Seoy

@ RENCI
  Leesa Brieger ++

@ DICE
  Michael Conway ++
  Reagan Moore
  Antoine de Torcy ++

@ UNC Libraries
  Steve Barr ++
  Greg Jansen ++

@ UNC Campus IT Services
  Will Schulz ++

@ SILS Grad. Student team
  Heather Bowden ++
  Alex Chassanoff ++
  Christine Cheng ++
  William Miao ++
  Terrell Russell ++
  Jewel Ward ++

@ UNC CS Grad. Student team
  Tao Yu ++
  Hao Xu ++

STATE ARCHIVES & LIB

Michigan
  Caryn Wojcik
  Mark Harvey

North Carolina
  Kelly Eubank
  Jennifer Ricker ++
  Amy Rudersdorf ++
  Lisa Gregory ++
  Ed Southern --
  Megan Durden --

Kentucky
  Dean Farrell ++
  Druscie Simpson
  David Minor
  Chris Black --

Kansas
  Glen McAninch
  Mark Myers ++

New York
  Bonnie Weddle
  Michael Martin ++
  Maria McCashion
  Ann Marie Przybyla

California
  Chris Garmire
  Nancy Lenoil-Zimmelman
  Linda Johnson
  Laren Metzer
  Renee Vincent-Finch

UNIVERSITY ARCHIVES

Tufts University
  Eliot Wilezek
  Veronica Martzah ++
  Anne Sauer

UNC Chapel Hill
  Will Owen ++
  Rich Szary ++

CULTURAL INSTITUTIONS

Getty Research Institute
  Joseph Shubitowski
  David Farneth
  Sally Hubbard
  Leah Prescott
  Mahnaz Ghaznavi --
  Karim Boughida --

Smithsonian Institution Archives
  Riccardo Ferrante ++

SCHOOLS OF LIB & IS

UNC Chapel Hill
  Cal Lee ++

University of Wisconsin-Madison
  Kristin Eschenfelder ++

Legend: Collaborator Roles:

Red: funded
Purple: cost-sharing
Brown: "observer"
+ ++: added after project funded
-= at new institution
<table>
<thead>
<tr>
<th>ISO Item</th>
<th>DCAPE Item</th>
<th>ISO Criteria</th>
<th>DCAPE Machine-Actionable Rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>ISO 2</td>
<td>DCAPE 3</td>
<td>Identify the content information and the information properties that the repository will preserve.</td>
<td>Define templates that specify required metadata and parameters for rules that are required to enforce properties.</td>
</tr>
<tr>
<td>ISO 3</td>
<td>DCAPE 4</td>
<td>Maintain a record of the Content Information and the Information Properties that it will preserve.</td>
<td>Link submission and policy templates to the preserved collection.</td>
</tr>
<tr>
<td>ISO 4</td>
<td>DCAPE 5</td>
<td>Specify Submission Information Package format (SIP)</td>
<td>Define templates that specify structure of a SIP and required content of a SIP.</td>
</tr>
<tr>
<td>ISO 5</td>
<td>DCAPE 1</td>
<td>Verify the depositor of all materials.</td>
<td>Ingest data through a staging area that has a separate account for each depositor.</td>
</tr>
<tr>
<td>ISO 6</td>
<td>DCAPE 2</td>
<td>Verify each SIP for completeness and correctness</td>
<td>Compare content of each SIP against template.</td>
</tr>
<tr>
<td>ISO 7</td>
<td>DCAPE 3</td>
<td>Maintain the chain of custody during preservation.</td>
<td>Manage audit trails that document the identity of the archivist initiating the task.</td>
</tr>
<tr>
<td>ISO 8</td>
<td>DCAPE 4</td>
<td>Document the ingestion process and report to the producer</td>
<td>Send e-mail message to producer when process flags are set.</td>
</tr>
<tr>
<td>ISO 9</td>
<td>DCAPE 5</td>
<td>Document administration processes that are relevant to content acquisition.</td>
<td>Maintain list of rules that govern management of the archives.</td>
</tr>
<tr>
<td>ISO 10</td>
<td>DCAPE 1</td>
<td>Specify Archival Information Package format (AIP)</td>
<td>Define templates that specify structure of an AIP and required content of an AIP.</td>
</tr>
<tr>
<td>ISO 11</td>
<td>DCAPE 2</td>
<td>Label the types of AIPs.</td>
<td>Store AIP type with each collection.</td>
</tr>
<tr>
<td>ISO 12</td>
<td>DCAPE 3</td>
<td>Specify how AIPs are constructed from SIPs.</td>
<td>Define transformation rule based on parsing of SIP template and AIP template.</td>
</tr>
<tr>
<td>ISO 13</td>
<td>DCAPE 4</td>
<td>Document the final disposition of all AIPs.</td>
<td>Maintain an audit trail for all AIPs.</td>
</tr>
<tr>
<td>ISO 14</td>
<td>DCAPE 5</td>
<td>Generate persistent, unique identifiers for all AIPs.</td>
<td>Define unique persistent logical name for each AIP.</td>
</tr>
<tr>
<td>ISO 15</td>
<td>DCAPE 6</td>
<td>Verify uniqueness of identifiers.</td>
<td>Identifiers uniqueness enforced by algorithm that assigns identifiers.</td>
</tr>
<tr>
<td>ISO 16</td>
<td>DCAPE 7</td>
<td>Manage mapping from unique identifier to physical storage location.</td>
<td>Storage location mapping enforced by iRODS data grid framework.</td>
</tr>
<tr>
<td>ISO 17</td>
<td>DCAPE 8</td>
<td>Provide authoritative representation information for all digital objects.</td>
<td>Define template specifying required representation information.</td>
</tr>
<tr>
<td>ISO 18</td>
<td>DCAPE 9</td>
<td>Identify the file type of all submitted Data Objects.</td>
<td>Apply type identification routine to each object on ingestion.</td>
</tr>
<tr>
<td>ISO 19</td>
<td>DCAPE 10</td>
<td>Document processes for acquiring preservation description information (PDI).</td>
<td>Define rule set that will be applied to extract PDI.</td>
</tr>
<tr>
<td>ISO 20</td>
<td>DCAPE 11</td>
<td>Execute the documented processes for acquiring PDI.</td>
<td>Apply PDI rules specific to a collection.</td>
</tr>
<tr>
<td>ISO 21</td>
<td>DCAPE 12</td>
<td>Ensure link between the PDI and relevant Content Information.</td>
<td>Set PDI extraction flag as part of PDI extraction rules.</td>
</tr>
<tr>
<td>ISO 22</td>
<td>DCAPE 13</td>
<td>Verify completeness and correctness of</td>
<td>Compare AIP against template for required</td>
</tr>
</tbody>
</table>
DCAPE item 2: Virtual Loading Dock
Upon acceptance, content may be ingested into the Virtual Loading Dock before being moved to the preservation area

Event → acSetRescSchemeForCreate | $objPath like /DCAPE/home/rods/LoadingDock/* | Action → msiSetDefaultResc(Resource 1,forced) | Recovery → nop
Virtual Loading Dock

Preservation Area

Replicate

Periodically validate checksums

Producer

Submit

SIP

Virus Check

Consumer

Archived

DIP

AIP

1. Submit

2. Virtual Loading Dock

3. Archive

4. Preservation Area

5. AIP

6. Replicate

7. AIP

8. Virtual Loading Dock

9. Archive

10. Preservation Area

11. AIP

12. Replicate

13. AIP

14. Periodically validate checksums

15. Replicate

16. Archive

17. Preservation Area

18. AIP

19. Replicate

20. AIP

21. Periodically validate checksums

22. Archive

23. Preservation Area

24. Interface

25. Disseminate

26. DIP
Thank you!

chienyi@unc.edu
http://salt.unc.edu