Generating Metadata for Digital Preservation

Ardys Kozbial
UCSD Libraries

Arwen Hutt
Brad Westbrook
UCSD Libraries

David Minor
Don Sutton
San Diego Supercomputer Center

Archiving 2009
May 5, 2009
Chronopolis
(funded by NDIIPP)

• Creating a 3-node federated data grid at SDSC, NCAR and UMIACS

• Funded to hold up to 50 TB of data (150 TB of disk space)

• Using BagIt and Storage Resource Broker (SRB) protocols to transfer data

• Using the following monitoring tools: Auditing Control Environment (ACE), SRB Replication Monitor, SRB System Monitor

• Analyzing metadata that is created by the various parts of the system

• Writing best practices documents for clients and partners
Chronopolis

- Chronopolis is being developed by a national consortium led by SDSC and the UCSD Libraries.

- Initial Chronopolis nodes include:
  - **SDSC and the UCSD Libraries at UC San Diego**
  - **University of Maryland Institute for Advanced Computer Studies (UMIACS)**
  - **National Center for Atmospheric Research (NCAR) in Boulder, CO**
Institutions and Roles at the Nodes

**SDSC, UMIACS, NCAR**
- Storage and network support
- Transmission packaging modules
- Complete copy of all data
- Network testing
- SRB support (SDSC, UMIACS)
- Advanced data services (UMIACS)
  - ACE: **Auditing Control Environment** to ensure the long-term integrity of digital archives

**UCSD Libraries**
- Metadata expertise
- DIPs (Dissemination Information Packages)
Institutions and Roles: Data Providers

California Digital Library

- 5 TB of data
- Web-at-Risk Project
- Crawls of political and government Web sites
- ARC files, uniform size
- BagIt protocol for data transfer

ICPSR

- 8 TB of data
- 40 years of social science research
- Millions of files
- Currently using SRB
Institutions and Roles: Data Providers

NCSU

• 6 TB of data
• State and local geospatial data
• BagIt protocol for data transfer

SIO

• 1 TB of data
• 50 years of data from SIO research cruises
• Currently using SRB
Storage Resource Broker (SRB)

- Supports shared collections of data that can be distributed across multiple organizations and heterogeneous storage systems. The SRB can be used as a Data Grid Management System that provides a hierarchical logical namespace to manage the organization of data (usually files).
Monitors
(developed at UMIACS)

• Auditing Control Environment (ACE)
  ACE is a system that incorporates a new methodology to address the integrity of long term archives using rigorous cryptographic techniques. ACE continuously audits the contents of the various objects according to the policy set by the archive, and provides mechanisms for an independent third-party auditor to certify the integrity of any object.

• Replication Monitor
  The SRB Replica monitor is a simple webapp that will watch registered directories and ensure that copies exist at designated mirrors. The monitor stores enough information to know if files have been removed from the master site and when the last time a file was seen. In addition any action that the webapp takes on files is logged. The monitor does NOT do any type of integrity checking, this is the responsibility of additional components.
GENERATING METADATA
<table>
<thead>
<tr>
<th>Object_Type</th>
<th>Element Description</th>
<th>Element Name</th>
<th>Required</th>
<th>Data Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>MCAT_Collection_Type</td>
<td>internal collection ID</td>
<td>DATA_GRP_ID</td>
<td>NOT NULL</td>
<td>VARCHAR2(250)</td>
</tr>
<tr>
<td>srb_NdLipchorn.mdas_td_data_grp</td>
<td>collection name</td>
<td>DATA_GRP_NAME</td>
<td>NOT NULL</td>
<td>VARCHAR2(1000)</td>
</tr>
<tr>
<td>parent collection name</td>
<td></td>
<td>PARENT_GRP_NAME</td>
<td>NOT NULL</td>
<td>VARCHAR2(1000)</td>
</tr>
<tr>
<td>collection owner</td>
<td></td>
<td>COLL_OWNER</td>
<td></td>
<td>NUMBER(38)</td>
</tr>
<tr>
<td>collection creation dtstamp</td>
<td></td>
<td>COLL_CRTIMESTAMP</td>
<td></td>
<td>VARCHAR2(32)</td>
</tr>
<tr>
<td>comments on collection</td>
<td></td>
<td>COLL_COMMENTS</td>
<td></td>
<td>VARCHAR2(700)</td>
</tr>
<tr>
<td>collection lock type</td>
<td></td>
<td>COLL_LOCK</td>
<td></td>
<td>NUMBER(38)</td>
</tr>
<tr>
<td>collection lock owner</td>
<td></td>
<td>COLL_LOCKOWNER</td>
<td></td>
<td>NUMBER(38)</td>
</tr>
<tr>
<td>collection lock expiry dtstamp</td>
<td></td>
<td>COLL_LOCKEXPIRY</td>
<td></td>
<td>VARCHAR2(32)</td>
</tr>
<tr>
<td>collection link</td>
<td></td>
<td>COLL_LINK</td>
<td></td>
<td>NUMBER(38)</td>
</tr>
<tr>
<td>collection modification dtstamp</td>
<td></td>
<td>COLL_MODTIMESTAMP</td>
<td></td>
<td>VARCHAR2(32)</td>
</tr>
<tr>
<td>original collection name</td>
<td></td>
<td>ORG_COLL_NAME</td>
<td></td>
<td>VARCHAR2(1000)</td>
</tr>
<tr>
<td>ACE_Sequence_Type</td>
<td>sequence name</td>
<td>SEQ_NAME</td>
<td>NOT NULL</td>
<td>VARCHAR2(50)</td>
</tr>
<tr>
<td>ACE_Collection_Type</td>
<td>sequence count</td>
<td>SEQ_COUNT</td>
<td></td>
<td>NUMBER(38)</td>
</tr>
<tr>
<td>ID</td>
<td></td>
<td>ID</td>
<td>NOT NULL</td>
<td>bigint(20)</td>
</tr>
<tr>
<td>STORAGE</td>
<td></td>
<td></td>
<td></td>
<td>varchar(255)</td>
</tr>
<tr>
<td>CHECKPERIOD</td>
<td></td>
<td></td>
<td></td>
<td>int(11)</td>
</tr>
<tr>
<td>DIRECTORY</td>
<td></td>
<td>DIRECTORY</td>
<td>NOT NULL</td>
<td>varchar(255)</td>
</tr>
</tbody>
</table>
ET-1 Service Level Agreement

Data

DP
ET-1 Service Level Agreement

Node 1

ET-2 Acquisition Transfer
ET-1 Service Level Agreement

ET-2 Acquisition Transfer

ET-3 Acquisition Validation

ET-4 Acquisition Registration to SRB

ET-5 Acquisition Registration into ACE

Node 1

Data

Manifest

ACE

MCAT
ET-1 Service Level Agreement

ET-2 Acquisition Transfer

ET-3 Acquisition Validation

ET-4 Acquisition Registration to SRB

ET-5 Acquisition Registration into ACE

ET-6 Inter-Node Inventory Check

ET-7 Acquisition Replication

Node 1

DP

Node 2

Node 3

ACE

Replication Monitor

MCAT

Manifest

Data
Next Steps

Metadata

• Add elements to SRB where necessary

• Define Chronopolis services
  – Reporting to data providers
  – Frequency of monitoring

• Create schema

• Test data return to data providers
Next Steps
Chronopolis Portal Development

• Joint effort
  • Partners are designing the interface
  • Data providers are providing feedback

• Multi-level views of the system based on usertype

• Shows the status of the system components

• First rollout to the project team in May 2009
Chronopolis Credits

• SDSC
  – Fran Berman
  – Richard Moore
  – David Minor
  – Jim D’Aoust
  – Don Sutton
  – Phong Dinh
  – Emilio Valente

• UCSD Libraries
  – Brian Schottlaender
  – Luc Declerck
  – Ardys Kozbial
  – Brad Westbrook
  – Arwen Hutt

• NCAR
  – Don Middleton
  – Michael Burek

• UMIACS
  – Joseph JaJa
  – Mike Smorul

• Library of Congress
  – Martha Anderson
  – Mike Ashenfelder

• CACI
  – Mike Ivey
http://chronopolis.sdsc.edu
ET1 – Service Level Agreement
ET2 – Acquisition Transfer

ET-1 Service Level Agreement
ET-2 Acquisition Transfer
ET-3 Acquisition Validation
ET-4 Acquisition Registration to SRB
ET-5 Acquisition Registration into ACE
ET-6 Inter-Node Inventory Check
ET-7 Acquisition Replication
ET-8 File Integrity Check

Node 1
DP
Data

Node 2
Replication Manager

Node 3

ACE

Manifest

MCAT