Enterprise Repositories and Federated Archives: The MetaArchive Approach to Distributed Digital Preservation

Tyler O. Walters
Associate Director, Technology & Resource Services
Georgia Tech Library & Information Center

Sun Microsystems PASIG
San Francisco, CA -- May 2008
Briefing on MetaArchive Cooperative

http://www.metaarchive.org

**Project Summary:**
- Eight partner institutions:
  - Emory - Georgia Tech - Florida State – Library of Congress
  - Virginia Tech – Auburn – Louisville – Univ. of Hull (UK)
- Collaborate w/ LC/NDIIPP – $1.2M initial effort to develop cooperative for preservation of digital content, 2004-2009

**Goals:**
1. **Distributed** preservation network infrastructure (LOCKSS)
2. **Conspectus** of digital content held by the partner sites
3. **Harvest** a body of most critical content to be preserved (4 TB)
4. **Cooperative charter** model for collaboration and sustainability

Networks: 1) Southern Culture, 2) ETDs, 3) Trans-Atlantic Slave Trade
Preservation Network Design Precepts

1. Distributed Preservation Infrastructure
2. Peer-to-Peer Network Architecture
   • Each node communicates with all other nodes
   • All nodes are joint custodians
3. Flexible Organizational Model
4. Formal Content Selection Process
5. Capability for Migrating Archives
6. Dark Archiving Strategy (no public access to MA network content)
7. Low Cost to Deployment
8. Self-Sustaining Incentives

*Effective digital preservation succeeds by distributing copies of content in secure, distributed locations over time*
Advantages of Adapting LOCKSS Software

- **Supports “distributed digital replication” approach:**
  - Data integrity checks
  - Rigorous security checks
  - Focused web crawls to gather/ingest content

- **Advantage:** many preservation efforts mix high accessibility online with long-term access (preservation). High accessibility = high costs

- Network content discoverable via metadata search. Conspectus will link to live sites designed for access

- Originally designed for minimal expenditures
  - Low barriers to adoption
  - Inexpensive computers / modest systems administration
Current MetaArchive Technology

Server:
- Dell PowerEdge 1850
- 2x 3.0Ghz/1MB Cache, Xeon 800Mhz / 2Gb Memory

SAN: (could easily be Sun or other hardware)
- PetaBox PowerStore PS4000
- AMD Athlon 64 X2 Dual Core Processor 4600+
- 2GB Memory / 4x1TB SATA HD
- Dell/EMC AX100 Array (single processor)
- 2 TB Storage (12x250 7200 RPM Serial ATA)

OS and LOCKSS:
- Red Hat Advanced Server 4 (release 6) / (Fedora Core 8 – new node)
- Hw/Sw firewalls, access control lists / LOCKSS daemon 1.29.4

Database Driven Conspectus
- MySQL/PHP Interface – Integrated w/LOCKSS Plugin Directory
- Manages Collections within LOCKSS
- Network Administrative Management Tool
Collections Replication

Online Digital Collections

Auburn Yearbooks
Emory Southern Spaces
FSU ETD Collection

LOCKSS for MetaArchive

GaTech Node
Emory Node
FSU Node
Auburn Node
Admin Interface

Georgia Tech Library and Information Center
MetaArchive Conspectus Database

Conspectus Database – Ties in with LOCKSS Central Plugin Repository

Log In Here

Login: 
Password: 

Log In

If you've lost or forgotten your password, click Here.

Login Attempt Failed

Actions

View All RDF Data
View By Institution
View All Valid RDF Data
View Selected Collections
View Conspectus Schema

View Existing Descriptions
Example Conspectus Entry – GaTech ETDs

Collection ID: 32
Alternate Title: Georgia Tech Electronic Thesis and Dissertation Collection
Description: The Electronic Thesis and Dissertation Collection at Georgia Tech covers Fall 2003 through the present. This program is jointly sponsored by the Graduate Studies and Research Office and the Library and Information Center. Georgia Tech is a member of the Networked Digital Library of Theses and Dissertations and participates in the Open Archives Initiative.

ESC Heading: Education
ESC Heading: Transportation
ESC Heading: Art and Architecture
ESC Heading: Law and Politics
ESC Heading: Geography
ESC Heading: Business, Agriculture, and Industry
ESC Heading: Social Class
ESC Heading: Environment
ESC Heading: Recreation
ESC Heading: History, Manners, and Myth

LCSH Heading: Dissertations, Academic -- Georgia -- Atlanta.
LCSH Heading: Electronic dissertations, Georgia, Atlanta.
# LOCKSS ADMIN INTERFACE for METAARCHIVE NETWORK

<table>
<thead>
<tr>
<th>Status</th>
<th>Status Last Updated</th>
<th>Institution Name</th>
<th>IP Address</th>
<th>Reverse DNS Entry</th>
<th>Log Page</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>September 20, 2005, 12:02 pm</td>
<td>Louisville University</td>
<td>meta-vault.library.louisville.edu</td>
<td><a href="http://meta-vault.library.louisville.edu">http://meta-vault.library.louisville.edu</a></td>
<td></td>
</tr>
<tr>
<td></td>
<td>September 20, 2005, 12:02 pm</td>
<td>Auburn University</td>
<td>meg.lib.auburn.edu</td>
<td><a href="http://meg.lib.auburn.edu">http://meg.lib.auburn.edu</a></td>
<td></td>
</tr>
<tr>
<td></td>
<td>September 20, 2005, 12:02 pm</td>
<td>Emory University</td>
<td>ndip.library.emory.edu</td>
<td><a href="http://ndip.library.emory.edu">http://ndip.library.emory.edu</a></td>
<td></td>
</tr>
<tr>
<td></td>
<td>September 20, 2005, 12:02 pm</td>
<td>Florida State University</td>
<td>clockss.lib.fsu.edu</td>
<td><a href="http://clockss.lib.fsu.edu">http://clockss.lib.fsu.edu</a></td>
<td></td>
</tr>
<tr>
<td></td>
<td>September 20, 2005, 12:02 pm</td>
<td>Florida State University</td>
<td>clockss2.lib.fsu.edu</td>
<td><a href="http://clockss2.lib.fsu.edu">http://clockss2.lib.fsu.edu</a></td>
<td></td>
</tr>
<tr>
<td></td>
<td>September 20, 2005, 12:02 pm</td>
<td>Georgia Institute of Technology</td>
<td>ndipplockss.library.gatech.edu</td>
<td><a href="http://ndipplockss.library.gatech.edu">http://ndipplockss.library.gatech.edu</a></td>
<td></td>
</tr>
</tbody>
</table>
Metadata Specification

- Dublin Core Elements & Refinements
- Dublin Core Collection Level Description
- RSLP Collection Level Description
- MODS Physical Description
- MetaArchive Specific Elements
Conclusion / Thank You!

- [www.metaarchive.org](http://www.metaarchive.org)
- [www.digitalpreservation.org](http://www.digitalpreservation.org)
- [www.lockss.org](http://www.lockss.org)

- Tyler Walters
- [Tyler@gatech.edu](mailto:Tyler@gatech.edu)
- 404-385-4489