

Distributed Digital Preservation Networks Across a Region, Across a State: Stretching LOCKSS

Gail McMillan, Virginia Tech

Martin Halbert, Emory

Aaron Trehub, Auburn

SCHEV LAC

Christopher Newport University

March 14, 2008

Distributed Digital Preservation Networks: MetaArchive Stretches LOCKSS Across a Region

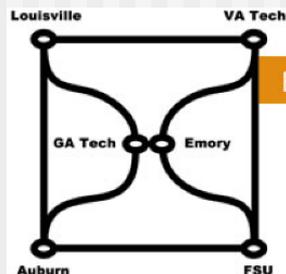
Gail McMillan, Virginia Tech

SCHEV LAC

Christopher Newport University

March 14, 2008

Stretching LOCKSS



Emory Univ. | Ga. Tech. | Va. Tech | Florida State Univ. | Auburn Univ. | Univ. of Louisville

M E T A A R C H I V E

Supported by the Library of Congress

<http://www.metaarchive.org> ■ <http://www.digitalpreservation.gov>



THE ALABAMA DIGITAL PRESERVATION NETWORK
PRESERVING ALABAMA'S DIGITAL RESOURCES

LOCKSS: Cooperative Digital Preservation

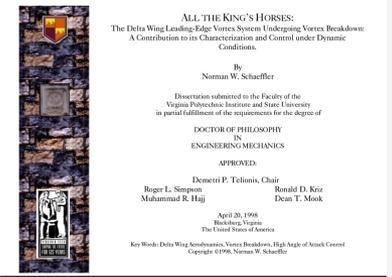
Gail McMillan

Digital Library and Archives, University Libraries
Virginia Polytechnic Institute and State University



SCHEV LAC

Virginia State University
June 10, 2005



Libraries should own, as well as manage, their digital collections

LOCKSS, fundamentally

- Programmatically collects content from a publisher
- Preserves content among LOCKSS and partners' servers
 - Low cost to administer and run
 - Inexpensive computer, free software
 - Audits content and repairs as needed from publisher or partners
- Disseminates content to only the appropriate users
 - Host library's clientele see the content from publisher's site
 - Unless it isn't available from there
 - Provide copies to partners only to audit and repair

Library of Congress Funding: NDIIPP

- National Digital Information Infrastructure and Preservation Program
- Support preservation of significant “born-digital” content at risk: Southern Heritage and Culture
- Three areas of focus
 - Network of preservation partners
 - Architectural framework for preservation
 - Digital preservation research



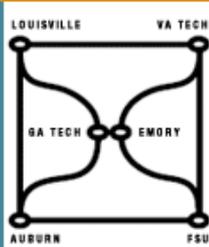
SOUTHERN SPACES

An internet journal and scholarly forum

METAARCHIVE

Supported by the Library of Congress

[about](#)
[documentation](#)
[meetings](#)
[links](#)
[contacts](#)



Welcome to the MetaArchive Project:

The MetaArchive Project is a collaborative venture of Emory University, Georgia Tech, Virginia Tech, Florida State University, Auburn University, University of Louisville, and the Library of Congress. The project is part of the [National Digital Information Infrastructure and Preservation Program](#) (NDIIPP) supported by the Library of Congress.

The partner institutions of this project are engaged in a three-year process to develop a cooperative for the preservation of at-risk digital content with a particular content focus: the culture and history of the American South. The project group members are jointly developing: 1) a prioritized conspectus of at-risk digital content in this subject domain held at the partner sites, 2) a harvested body of the most critical content at the partner sites to be preserved, 3) a cooperative agreement for ongoing collaboration, and 4) a distributed preservation network infrastructure based on the LOCKSS software. The work plan for this project builds on relationships and workflows developed during previous projects of the MetaScholar Initiative and other collaborating consortia.

This site will serve as a central space for disseminating information, reports, and other documents associated with MetaArchive.

METAARCHIVE

A PARTNERSHIP BETWEEN:
 EMORY UNIVERSITY
 THE LIBRARY OF CONGRESS
 GEORGIA TECH
 VIRGINIA TECH
 FLORIDA STATE UNIVERSITY
 AUBURN UNIVERSITY
 UNIVERSITY OF LOUISVILLE

©2004 METASCHOLAR INITIATIVE

MetaArchive Goals

- Create a conspectus of digital content within the subject domain held by the partners
- Distributed preservation network infrastructure based on LOCKSS software
- Harvested body of the most critical content to be preserved (3 TB per institution)
- Develop a model cooperative agreement for ongoing collaboration and sustainability of preservation partners



Key Features of the MetaArchive of Southern Digital Culture

- Distributed preservation strategy
- Flexible organizational model
- Formal content selection process
- Capability for migrating archives
- Dark archiving strategy
- Low cost to deployment
- Self-sustaining incentives
- Simple exchange mechanisms



MetaArchive Conspectus DB

<http://www.metaarchive.org/conspectus/>

- Scope
- Standards
 - Schema
 - Controlled vocabulary
- Database and Conspectus
 - Inventory of Collections
 - Formats
- Prioritizing
 - At risk
 - Data wrangling
- Adapting LOCKSS
- Rights Issues



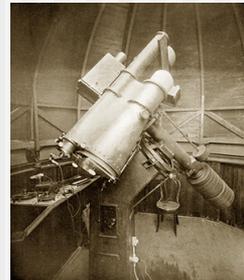
MetaArchive Sample Collections

- Auburn: 4 collections/7.9 GB
 - Extensions pubs, yearbooks (+TIFFs)
- Emory: 10 collections/23 GB
 - Born digital (Southern Spaces), image masters
- FSU: 3 collections/101 MB
 - Juvenile lit, historic photos, 2004 theses
- Georgia Tech: 12 collections/809 MB
 - Digitized special collections, SMARTech, ETDs
- Louisville: 3 collections/17 GB
 - Oral histories, image masters
- VT: 50 collections/1.9 GB
 - Online exhibits, faculty projects, Special Collections



Successful Disaster Recovery Test

- Focused on: Hardware, Content, Network
- Simulated and experienced crashing primary node
- Intentionally damaged content (truncate files)
- Disabled access to plug-ins
- Ran routine tests for “bad disk,” cache manager, conspectus database, yum repository, kickstart script, xml config. file, etc.
- Reconstructed primary node, resurrected network, reconstructed content
- Documented



MetaArchive Delivered

- 2005 Conspectus completed
 Network in operation
 First harvest and caching completed
- 2006 Cooperative model analysis completed
 Cooperative Charter drafted
 Nonprofit host organization formed
- 2007 Workshop for others interested in PLN
 Model replicated in Alabama
 Additional LoC funding
- 2008 Accepting new members



SOUTHERN SPACES

An interdisciplinary journal about the regions, places, and cultures



Southern Spaces is a peer-reviewed, online journal exploring the real and American South and their connections with the wider world. We welcome photographers, and visual artists in such areas as geography, southern or African American, Native, and American Studies, women's studies, LGBT

- contents
- about the forum
- editorial board
- webinars
- search
- FAQs

Dorsett Cox Decade
Arjan Sezing
Atlanta, Georgia
2004



Images Archive

SUPPORTED BY THE
ROBERT W. WELLS
FOUNDATION AND
MORAY UNIVERSITY
© 2014 Moray University

Southern Spaces encourages innovative work in digital scholarship and new media. By providing electronic archiving for our published materials, Southern Spaces makes a commitment to a stable digital presence for internet scholarship. To query about a submission, please contact Senior Editor Allen Tullos or Managing Editor Sarah Tobols for more information or visit our [Frequently Asked Questions](#).



THE METAARCHIVE MODEL: DISTRIBUTED DIGITAL PRESERVATION NETWORKS

Dr. Martin Halbert
Emory University

VIVA/SCHEV LAC Meeting
Christopher Newport University
Tribble Library
Newport News, VA
Friday, March 14, 2008

BASIC QUESTIONS

- What are Distributed Digital Preservation Networks?
- What is MetaArchive?
- What has MetaArchive Phase I accomplished for libraries?
- What does MetaArchive Phase II offer to libraries?

WHAT IS DIGITAL PRESERVATION?

- **Digital Preservation** refers to the systematic management of digital information *over extended (indefinite) periods of time*.
- Unlike the preservation of paper or microfilm, the preservation of digital information demands ongoing attention. This constant input of effort, time, and money to handle rapid technological and organizational advance is considered the main stumbling block for preserving digital information beyond a couple of years.
- Digital preservation can therefore be seen as the set of processes and activities that ensure the continued access to information and many kinds of records, both scientific and cultural heritage, existing in digital formats.

DISTRIBUTED DIGITAL PRESERVATION NETWORKS

- Effective preservation succeeds by replicating copies of content in secure, distributed locations over time
- **Security** reduces the likelihood that any single cache will be compromised.
- **Distribution** reduces the likelihood that the loss of any single cache will lead to a loss of the preserved content.
- A single cultural heritage organization is unlikely to have the capability to operate several geographically dispersed and securely maintained servers
- Inter-institutional agreements must be put in place or there will be no commitment to act in concert over time

BACKUPS/IRS VERSUS DIGITAL PRESERVATION

What differentiates a schedule for data backups from a digital preservation program?

- **Backups are tactical measures.** Backups are typically stored in a single location (often nearby or collocated with the servers backed up) and are performed only periodically. Backups are designed to address short-term data loss via minimal investment of money and staff time resources. Backups are better than nothing, but not a comprehensive solution to the problem of preserving information over time.
- **Digital preservation is strategic.** A digital preservation program entails a geographically dispersed set of secure caches of critical information. A true digital preservation program will require multi-institutional collaboration and at least some ongoing investment to realistically address the issues involved in preserving information over time.



- **A distributed digital preservation cooperative for digital archives**
- Established under the auspices of and with funding from the National Digital Information and Infrastructure Preservation Program (NDIIPP) of the Library of Congress
- A DDP network based on LOCKSS technology, but a separate network with higher capacity nodes
- Sustained by cooperative fee memberships and LC contracts
- Provides training and models for other groups to establish similar distributed digital preservation networks

METAARCHIVE

PHASE I (2004-2007)

- Created distributed archive of southern digital culture between inaugural members: Emory, Virginia Tech, Auburn, Georgia Tech, FSU, and University of Louisville, *enabling the cooperative preservation of more than 120 collections*
- Created an organizational charter, agreements between inaugural members, and founded an administrative nonprofit corporation (Educopia)
- Established a distributed preservation network infrastructure for replication based on the LOCKSS software, together with first version of conspectus database for collection decisions
- Hosted first workshop in distributed digital preservation strategies in 2007
- Assisted in creation of two additional DDPNs in Alabama and Arizona

METAARCHIVE

PHASE II (2007-2010)

- Created second distributed archive (for transatlantic slave trade historical data), and planning an ETD distributed archive
- Became international with the addition of Hull University in UK
- Hosting additional DDP workshops
- Will double in size to 12 members
- With funding from NHPRC will provide consulting and outreach services on the MetaArchive model for distributed digital preservation services



THE ALABAMA DIGITAL PRESERVATION NETWORK

PRESERVING ALABAMA'S DIGITAL RESOURCES

The Alabama Digital Preservation Network (ADPNet)

Aaron Trehub
Director of Library Technology
Auburn University

State Council of Higher Education for Virginia LAC
Christopher Newport University
March 14, 2008

Background

- ADPNet inspired by experience with NDIIPP MetaArchive Project
- IMLS grant: September 2006 through September 2008
- Grant awarded to and administered by Alabama Commission on Higher Education/Network of Alabama Academic Libraries (NAAL) in Montgomery
- Project director at Auburn University Libraries
- Commitments from seven institutions across the state

The objective

To create a low-cost, low-maintenance, sustainable, geographically distributed digital preservation network for libraries, archives, museums, and other cultural heritage organizations in Alabama.

The seven participating institutions

- Alabama Department of Archives and History (Montgomery)
- Auburn University (Auburn)
- Spring Hill College (Mobile)
- Troy University (Troy)
- University of Alabama (Tuscaloosa)
- University of Alabama at Birmingham
- University of North Alabama (Florence)

The network

- ADPNet is a Private LOCKSS Network (PLN)
- Uses off-the-shelf equipment and a standard LOCKSS installation
- LOCKSS servers (nodes) at all seven participating institutions
- Each institution maintains its LOCKSS server
- Each institution contributes content for harvesting and archiving by the network
- Runs on sweat equity, with help from LOCKSS staff

Why Alabama?

- Hurricanes...
- Tornadoes...
- Growing number of rich digital collections (e.g. [AlabamaMosaic](#))...
- Modest financial resources...
- Uneven technical support...
- = Ideal test case for geographically distributed digital preservation network

Why LOCKSS?

- Familiar with it (through MetaArchive Project)
- Simple
- Robust
- Low maintenance
- Cheap (except for membership in the [LOCKSS Alliance](#))
- Good technical support
- Know it works

Costs

- Servers: LOCKSS server and Web server (for making content available to the network)
- Staff time (less than we anticipated)
- Communication (weekly conference calls, project listserv, project Wiki)
- Some travel (mostly in-state)
- The biggie: LOCKSS Alliance membership fee (annual). Supports LOCKSS software development and technical support.

ADPNet content

- ADPNet currently contains 11 collections (“archival units”) from five of seven institutions
- Over 100 gigabytes harvested
- Network capacity: one terabyte
- Plenty of room for more collections
- More collections on the way, including audio and video files

ADPNet administration

- ADPNet is a single-state network
- Folded into existing administrative infrastructure: ACHE/NAAL
- Not a service organization
- No membership fees (but LOCKSS Alliance membership mandatory)
- In-kind contribution: bring up and run a LOCKSS node in the network
- Governance document in the works

ADPNet digital preservation awareness survey

- Sent to academic and public libraries, archives, schools, and state and municipal agencies in Alabama in February 2008
- 79 responses: public libraries largest single group of respondents
- Most important factors in deciding whether to join digital preservation network: reliability, expertise and support, cost, staffing, and preservation of mission-critical collections
- Most people learn about new initiatives from conferences and colleagues, so focus on those

Lessons learned

- Keep it simple
- Keep it cheap
- Don't get fancy
- Low maintenance
- Low administrative overhead
- Take advantage of existing structures and relationships (easier to do with single-state network)

Future plans

- Add more content to the network
- Test disaster recovery procedures
- Recruit more member institutions, including public libraries (e.g. Birmingham Public Library) and museums
- Spread the word

Distributed Digital Preservation Networks and the MetaArchive Model

Contacts

Gail McMillan gailmac@vt.edu
(540) 231-9252

Martin Halbert mhalber@emory.edu
(404) 727-2204

Aaron Trehub trehuaj@auburn.edu
(334) 844-1716 <http://adpn.org/>