

PARTNERSHIPS FOR DISTRIBUTED DIGITAL PRESERVATION



Martin Halbert (Emory)

Gail McMillan (Virginia Tech)

Mark Stoffan (FSU)

CNI Fall 2008 Task Force Meeting
Washington, D.C.

Session Overview



1. Review of MetaArchive and recent developments (Martin Halbert)
2. Discussion of the new ETD Archive (Gail McMillan)
3. Overview of process for becoming a new member and repositing content in the network (Mark Stoffan)



Review of MetaArchive and Recent Developments

Martin Halbert (Emory)

Session Questions



- What is the MetaArchive Cooperative?
- What is distributed digital preservation? Why has MetaArchive embraced it as the most critically important strategy for the preservation of digital archives?
- Why did we form it?
- What differentiates it from other efforts?

What led to MetaArchive?



- Planning meetings by librarians and archivists in 2002-2003 on concerns about preserving digital archives
- Sense that we needed to do something practical to help each other preserve our data
- Not based on studies, just the observation of our collective anxieties about keeping our (expensive) digital materials preserved and viable

Gap in Digital Preservation Programs



- 66% of cultural heritage institutions (academic libraries, archives, art museums, public libraries, and other similar kinds of institutions) report that no one is responsible for digital preservation activities.
- 30% of all archives have been backed up one time or not at all.

Source: 2005 NEDCC Survey by Bishoff and Clareson

Need Collaborative Approaches



“The increased number and diversity of those concerned with digital preservation—coupled with the current general scarcity of resources for preservation infrastructure—suggests that *new collaborative relationships that cross institutional and sector boundaries could provide important and promising ways to deal with the data preservation challenge*. These collaborations could potentially help spread the burden of preservation, create economies of scale needed to support it, and mitigate the risks of data loss.”

- The Need for Formalized Trust in Digital Repository Collaborative Infrastructure

Backups vs. Digital Preservation



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

- ***Backups are tactical measures.*** They 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.*** Preserving information over long periods requires systematic attention rather than benign neglect or unthinking actions.

Institutional Repositories vs. Digital Preservation



What differentiates an IR program from a distributed digital preservation (DDP) program?

- ***The IR is not distributed.*** It is a centralized approach aimed at managing information flow within the institution. It typically does not attempt to securely cache prioritized content at multiple geographically dispersed sites.
- ***DDP mobilizes efforts of multiple institutions.*** It 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.

Secure and Distributed Cache Networks



Why are the characteristics of geographical distribution and security so important? This strategy maximizes survivability of content in both individual and collective terms:

- **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.

By creating a collaborative network for secure and distributed preservation, a group can also work together on more complex issues such as format migration.

Both Technical and Organizational Networking are Required



- A single cultural heritage organization is unlikely to have the capability to operate several geographically dispersed and securely maintained servers
- Collaboration between institutions on technological solutions is essential
- Similarly, inter-institutional agreements must be put in place or there will be no commitment to act in concert over time

Shared Archiving Fails without a Pre-coordinated DDP Network in Place



Lessons from the NDIIPP Archive Ingest and Handling Test (AIHT) and other shared archiving experiments:

- Encounter many unexpected incompatibilities because of different systems and data packaging
- Realization that much of the cost in preserving digital material is in coordinating the organizational and institutional imperatives of preservation, and not the technological costs of storage space

MetaArchive Cooperative



A distributed digital preservation cooperative for digital archives

- Established in 2003 under the auspices of and with funding from the National Digital Information and Infrastructure Preservation Program (NDIIPP) of the US Library of Congress
- A functioning DDP network using/building open source software
- Organized as an incorporated nonprofit cooperative of libraries and other cultural memory organizations
- Sustained by organization fee memberships, cooperative agreement with U.S. Library of Congress, and other sponsored funding
- Provides training and models for other groups to establish similar distributed digital preservation networks
- Fosters broader awareness of digital preservation issues
- Designed to address “in-the-trenches” needs of CMOs after environmental scans of other options

MetaArchive Compared with Other Efforts



- MetaArchive is a *cooperative* not a vendor:
 - A *cooperative* is an organization that consists of a group of individuals who have joined together to perform a function more efficiently than each individual could do alone.
 - The purpose of a cooperative is not to make profits, but to improve each member's situation and the situation of the surrounding society.
- MetaArchive is a collaborative association of cultural memory organizations with a nonprofit administration.
- All hardware and software assets are owned by members.
- Membership fees go to a central pool of support for members' co-op activities.

MetaArchive Phase I (2004-2007)



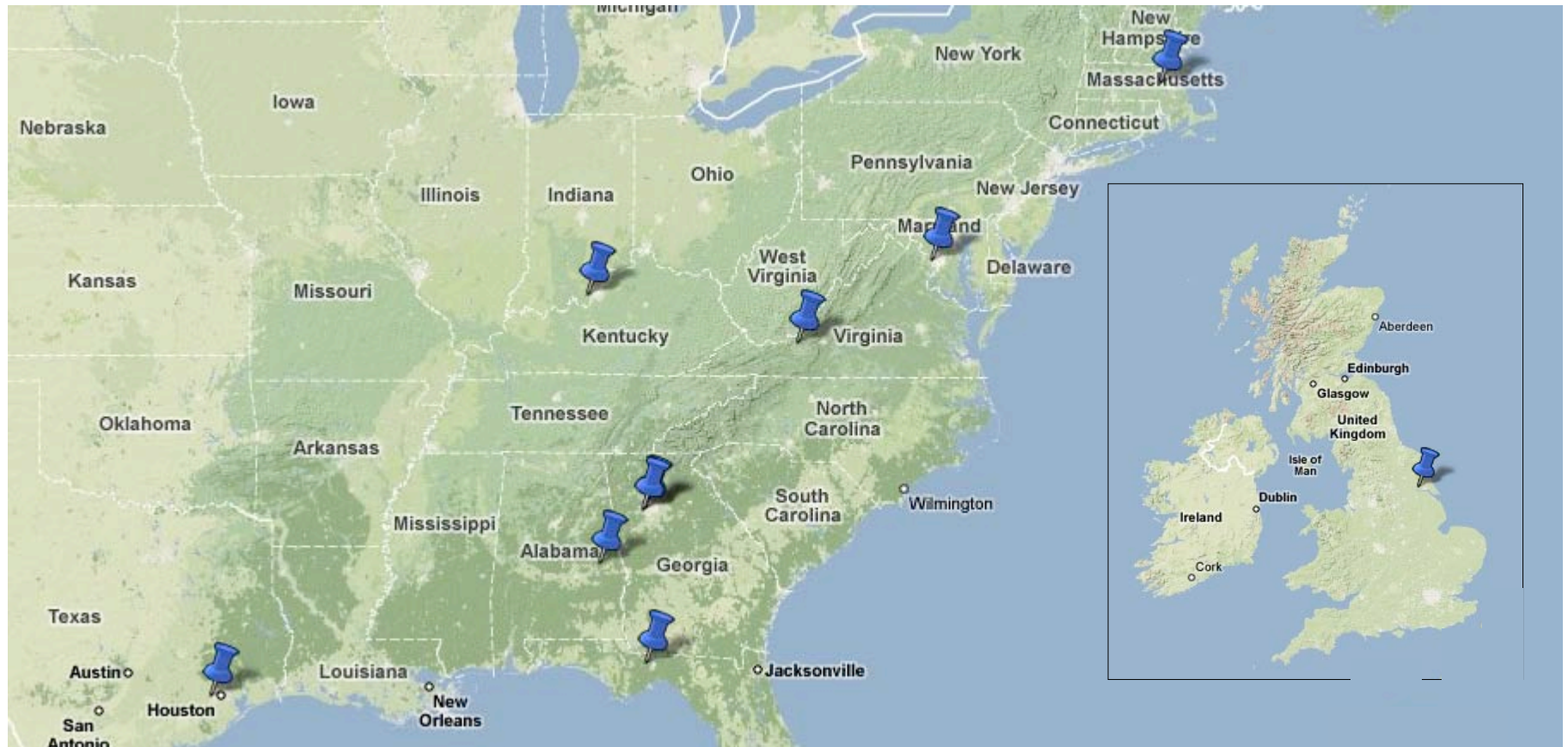
- Developed a functioning network for distributed digital preservation (DDP) used by institutions with shared subject domain focus for mutual benefit
- Developed this technical solution for DDP based on a reuse of LOCKSS technology, in the form of a separate network with higher capacity nodes
- Created a conspectus database to capture collection-level preservation metadata pre-ingest
- Created an administrative nonprofit corporation as an independent legal entity for membership agreements
- Now preserving via DDP more than 650 collections from many different organizations

MetaArchive Membership



- 11 institutions currently:
 - Emory, GA Tech, Auburn, VA Tech, FSU, Louisville, Hull, Rice, Boston College, Folger, and US Library of Congress
- Doubled in size of membership within past year; plan to double again in next 12 months
- Now undertaking strategic alliances with other membership organizations to provide DDP services (more on alliance with NDLTD in a moment...)

MetaArchive Geographic Coverage



Collection Variety

- Collections include:
 - Images
 - Text files
 - Multimedia files
 - Datasets
 - Program executables



Catalytic Efforts



- Host workshops in distributed digital preservation strategies
 - Instructing new MetaArchive members in network processes
 - Advise other groups considering DDP approaches
- Advised/assisted in creation of two additional DDPNs:
 - Alabama
 - Arizona

MetaArchive Phase II (2007-2010)



- Established additional distributed archives
 - African Diaspora
 - Electronic Theses and Dissertations
 - Early Modern Literature
- New software tools for enhanced conspectus, interoperability with grid-computing, format migration
- Became international with the addition of Hull University in UK
- Upcoming DDP conferences and workshops
- Plan to double in size each year (on average) to reach a robust cooperative size
- With funding from NHPRC will provide consulting and outreach services on the MetaArchive model for DDP services

Membership Levels



- **Contributing Member Sites** are institutions that need to preserve digital content, and therefore decide to contribute content into the preservation network. The preservation network acts for the common good to preserve the at-risk content submitted by the contributing sites. Contributing sites may also be preservation sites.
- **Preservation Member Sites** are responsible for the basic ongoing activity of preserving digital content. At a minimum, every preservation site must include responsible staff and a node server of the relevant preservation network. Preservation sites collectively comprise a preservation network.
- **Sustaining Member Sites** are responsible for steering committee of the Cooperative, technical development of the computer systems that enable the preservation network. Obviously, development sites may also be preservation sites and contributing sites.

Individual Roles



- **Program Managers** are leaders that accept responsibility for coordinating the activities of a digital preservation network.
- **Data Wranglers** are programmers and other technically adept workers that prepare local digital archives for ingestion into a preservation network.
- **System Administrators** are staff members that maintain individual preservation node servers of the relevant preservation network.
- **Selectors** are staff that identify and prioritize content to be preserved. They will most often be knowledgeable concerning the content of an institution's digital archives, and may have been the same individuals that originally created or acquired the archives.



Search this site Go >

My MetaArchive >

- [About MetaArchive](#)
- [Current Networks](#)
- [Resources](#)
- [Tech Tools](#)
- [Events](#)
- [News](#)
- [Contact](#)
- [Join Us](#)
- [All about Distributed Digital Preservation >](#)

THE GREATEST THREAT

to digital assets is not fire, flood or theft. It's the hazy assumption that cultural heritage institutions have taken the steps needed to preserve them.

Most often, we haven't. Which is why the [MetaArchive Cooperative](#) is leading a national effort to embrace [distributed digital preservation](#), the future practice of digitally safeguarding the very items that define our culture and identity. [MORE >](#)

MetaArchive Cooperative receives \$1.2 M in support of additional research and development for its distributed digital preservation services >

[Other news >](#)



Discussion of the Electronic Theses and Dissertations ETD Archive

Gail McMillan (Virginia Tech)

NDLTD/MetaArchive Alliance Pilot Project Goals



- NDLTD = Networked Digital Library of Theses and Dissertations
- Analyze and understand different scenarios for preservation services to NDLTD members
- Test and document procedures and practices
- Model joint NDLTD/MetaArchive
- Planning update with NDLTD Board

NDLTD/MetaArchive Alliance: Background



ETD Preservation Survey

- Dec. 2007-April 2008
- 95 institutions responded
- 80% have ETDs
- 73%: No formal preservation plans for ETDs
- Only 27% have preservation plans for ETDs
- 92% interested in DDPN

NDLTD/MetaArchive Alliance: ETD Preservation Plan



- Hardware, software
- Metadata: Conspectus Database
- Organizing ETD collections
- Harvest Frequency
- Institutional Workflow
- Authors' Responsibilities
- Personnel
- Training Opportunities
- Documentation and Reports
- Retrieving from the ETD Archive

NDLTD/MetaArchive Alliance: ETD Preservation Strategy



- Workshops: Aberdeen, June 2008
 - ETD conference
 - Atlanta, October 2008
 - MetaArchive meeting
 - Pittsburgh, June 2009
 - ETD conference

NDLTD/MetaArchive Alliance

ETD Archive Pilot Program



- Oct. 2008: Planning for pilot project, approved additional participants
- Jan. 2009: Begin pilot project
- April 2009: Finish initial tests
- May 2009: Develop proposal for going forward
- June 2009: Present results to NDLTD Board
Make decisions for joint
NDLTD/MetaArchive program

Scenarios to Explore for Preserving ETD Collections



- Hub and Spoke Model
 - Identify small number of NDLTD/MetaArchive service providers who will offer preservation functions for all NDLTD members
- Cloud Model
 - Do not differentiate NDLTD/MetaArchive nodes
 - Build up more joint members

Seeking New NDLTD/MetaArchive Pilot Project Collaborators



- Looking for a few additional participants (already have VA Tech, GA Tech, Boston College, and Rice University)
- Collaborators must join MetaArchive and be willing to participate in leading development of NDLTD/MetaArchive ETD preservation program
- Participants may either run a node or help model contributing member procedures

NDLTD/MetaArchive ETD Archive Pilot Program - How to Participate



- Decide if you want to participate in developing this new collaboration.
- What scenarios and functions are you most interested in analyzing? (Hub preservation nodes? Distributed contributing nodes?)
- If interested, contact Gail McMillan: gailmac@vt.edu

Overview of Process for Joining MetaArchive

(and repositing content in the network)

Mark Stoffan (FSU)

MetaArchive Charter and Membership Agreement



- Charter is a formative agreement that lays out the conceptual roles and responsibilities of participants
- Membership agreement is between new members and MetaArchive Services Group nonprofit corporation
 - Agreement to preserve content for specified period
 - Pledge to not intentionally harm the network

Three Membership Levels

- **Contributing Site Members:** Do not run infrastructure, simply use the network to preserve content
- **Preservation Site Members:** Operate a MetaArchive network node for specified period, using it to preserve content
- **Sustaining Site Members:** Operate a node and participate in leadership of cooperative

Archive Ingest Process



- A “Plugin” is written for collections selected for preservation
- Plugins are programs describing rules and structure for the “archival unit”
- Either local staff or MetaArchive staff write these plugins and install them in the network
- At least 6 dispersed sites are selected for repositing the archival unit
- Caching process begins, with updates following if necessary

Requirements for Operating a Node



- Be able to bring up and maintain a Linux server over time
- Task local staff with both program management and systems administration duties, and preferably data wrangling as well
- Contribute content and monitor system functioning occasionally
- Sign membership agreement and pay membership dues

Factors to Consider



- Recovery of data in the event of loss should be planned for, not put off for another day
- Whose job is this going to be in the organization?
- What are the highest priority items for distributed digital preservation?
- What digital preservation challenges are most important to your institutional setting?
- What expertise can you share with the other members of the cooperative?

Observations

- Be prepared for a front-loaded learning curve.
- Ask lots of questions.
- Make sure you have adequate staff assigned to the project.
- Have good project coordination in place among library IT, archives and digital libraries.
- Be realistic in allocating sufficient staff to get the work done.

Thank You!



- Mark Stoffan (mstoffan@fsu.edu)
- Gail McMillan (gailmac@vt.edu)
- Martin Halbert (mhalber@emory.edu)