“...Their Unknown and Undiscovered Brothers”

Keynote for Private LOCKSS Networks: Community-based Approaches to Distributed Digital Preservation

Martin Halbert, UNT Dean of Libraries
PLN Forum 2010, Boston, MA
Monday, October 25, 2010
Presentation Overview

- Genesis of MetaArchive and other PLNs in early projects with LOCKSS to combat the data loss problem
- Retrospective findings after six years of running the MetaArchive Cooperative
- Growing the distributed digital preservation community
Do Lots of Copies *Really* Keep Stuff Safe?

Practical and philosophical reasons for seeking to combat the data loss problem through distributed digital preservation
“Individual commitment to a group effort -- that is what makes a team work, a company work, a society work, a civilization work.”

- Vince Lombardi
The Data Loss Problem

44% of the sites available on the internet in 1998 had vanished one year later.

From NDIIPP Website on the Importance of Digital preservation (http://www.digitalpreservation.gov/importance/):
A mid-size cultural memory organization engaged in active digitization of its archival collections for a full decade

They scanned at high quality and kept reasonable metadata

They backed up the accumulated scans to tape and stored the tapes...

...in the room next to the server...

...and one afternoon *had a fire and lost it all.*

*Observation: local backups are not a preservation program*
Backups 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.** Preserving information over long periods requires systematic attention rather than benign neglect or unthinking actions.
The 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
Two Historical Strategies for Preservation
Example from the Chirographic (Handwritten) Era: The Nag Hammâdi Library

- Collection of early Coptic texts discovered near the town of Nag Hammâdi in 1945
- Had been buried in the 4th Century CE when censored
- Only extent textual copies from the core of the early Gnostic branch of Christianity

- Survived 15 centuries because they were part of a secure, distributed chirographic network.
Two Broad Strategies for Digital Preservation

- **Centralized repository:** in which all of the key preservation activities take place within one centrally administered repository site
- **Distributed network:** in which the preservation activities occur within a dispersed network environment administered by multiple groups
Distributed Digital Preservation Programs

Engage in three main activities that distinguish them from other preservation approaches:

1. **Replication** of content,
2. **Distribution** of these replicated copies to distinct geographical locations, and
3. **Networking** technologies to connect these replicated copies through routine operations, including checksum comparisons and repair activities
What differentiates an IR program from a distributed digital preservation program?

- **The IR is not distributed.** The IR 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.** 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.
Key Features of DDP Networks

- Do not merely store back-up copies of content in distributed locations, but rather create secure networks in which preservation activities take place.
- Include regular and timely fixity checks of each file across the servers that host its replications and the repair of content in a systemized and well-documented manner when errors occur.
- All copies of a file and all servers in a network must be secure in terms of ownership.
A dedicated preservation network for electronic journals, established with funding from the Mellon Foundation, the NDIIPP, and other groups:

- The pioneering leader in distributed digital preservation
- Very highly distributed geographically across the world, with hundreds of sites
- Available for others to join, both to build onto or to piggyback on
- Fee structure for membership
- No signed agreements between sites
- Motivation to preserve content is based on interest by members in long-term access to online journal content to which they subscribe
- Active development community, with new initiatives with publishers (CLOCKSS) and many other technical advancement directions
Distributed digital preservation efforts are one answer to a 2005 joint National Science Foundation (NSF)/Joint Information Systems Committee (JISC) study that found:

...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.

Examples of Distributed Digital Preservation Programs Building on LOCKSS

- MetaArchive Cooperative
- Persistent Digital Archives and Library System (PeDALS)
- Alabama Digital Preservation Network (ADPNet)
- U.S. Government Documents network
- Council of Prairie and Pacific University Libraries (COPPUL)
- Data Preservation Alliance for the Social Sciences (DataPASS project)
Why an Archive of Archives?

MetaArchive Case Study and retrospective findings after six years
“Coming together is a beginning. Keeping together is progress. Working together is success.”

- Henry Ford
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.
The MetaArchive Cooperative was established in 2003 under the auspices of and with funding from the National Digital Information and Infrastructure Preservation Program (NDIIPP) of the Library of Congress.

It is both a functioning distributed digital preservation network (254 TB) and nonprofit organization for libraries and other cultural memory organizations.

Sustained by cooperative membership fees, NDIIPP contracts, and grants from the National Historical Publications & Records Commission and other groups.

Provides training and models to foster broader awareness of distributed digital preservation and to enable other groups to establish similar networks.
MetaArchive Membership

Current Members
Auburn University
Boston College
Clemson University
Florida State University
Folger Shakespeare Library
Georgia Tech
Indiana State University
Library of Congress
Penn State University
PUC Rio de Janeiro
Rice University
University of Hull
University of Louisville
University of North Texas
University of South Carolina
Virginia Tech

Current Affiliates
NDLTD
SDSC Chronopolis

Slide 22
Collection Variety

- Collections include:
  - Images
  - Text files
  - Multimedia files
  - Datasets
  - Program executables
Building on Top of LOCKSS

- Conspectus Database (Original)
  - Curators enter collection level entries for collections
  - Meant to be used for cooperative prioritization in DDP selection and decision-making activities
  - Not interactive with some key MetaArchive systems (Cache Manager, Ingest Plugins)

- Second Generation Conspectus Database
  - Now in deployment
  - Integrates operation of all network functions
  - Being designed in concert with guidance from other PLNs, hopefully in ways that enable re-use
MetaArchive aims to provide a cost-effective core value proposition through a cooperative structure:
- secure distributed digital preservation of archives
- based on re-use of LOCKSS for archives
- Participation as a MetaArchive node is relatively simple and much cheaper than doing it alone
- Benefits of membership are understandable, bounded, affordable, and address an important institutional gap
Developed a new nonprofit cooperative with guidance from both legal team, librarians, and intellectual property specialists

Created core organizational documents in 2006: charter, membership agreement, papers of incorporation, business plans, etc.

Allows members to understand their commitment and liability clearly
Active Collaborations with Other Efforts

- **LOCKSS** (collaborative development of LOCKSS Cache Manager)
- **SDSC Chronopolis** (PLN/ SRB interoperation testing and bridges)
- **Data-PASS Alliance** (conversations about standards for Private LOCKSS network interoperation and tools)
Hosted first workshop in distributed digital preservation strategies in 2007
- Instructed new MetaArchive members in processes
- Advised other groups considering DDP approaches

Provided advice during the creation of two additional DDPNs
- Alabama ADPN
- Arizona PEDALS
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.
MetaArchive compared with Other Efforts

- MetaArchive is a **cooperative** not a vendor:
  - A *cooperative* (also *co-op*) 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
Effective DDP Implementations

Effective implementations of distributed digital preservation that operate in the collaborative manner require two key elements:

1. A robust technical infrastructure
2. Strong inter-organizational arrangements
Multi-institutional collaboration to build DDP networks

- Most institutions charged with curatorial responsibility for cultural and scientific documentation do not have the capability to operate several geographically dispersed and securely maintained servers.
- Collaboration between institutions has proven essential in distributed efforts to date, and this collaboration requires both organizational and technical investments.
Growing the DDP Community

- Surveys of what Cultural Memory Organizations want as far as Digital Preservation is concerned
- Gaps and Opportunities for the PLN Community
- The Future of Community-based Approaches to Distributed Digital Preservation
“Community cannot for long feed on itself; it can only flourish with the coming of others from beyond, their unknown and undiscovered brothers.”
- Howard Thurman
The MetaArchive Cooperative conducted a survey of CMO’s in 2009 Q1 of *Digital Preservation Practices and Priorities*

- The purpose of the survey was to assess the digital preservation desires, preferences, and needs of CMO’s
- 60 responses
Institutions are now actively acquiring and building significant digital collections, even at small museums and archives.

- Fully 94.8% of respondents indicate that their archives has a growing digital component.
- The average size is 2 TB, with 24 respondents reporting 1 TB or less and 22 reporting between 1 and 20 TB.
- Collections are quickly growing at an average of 540 GB/year.
- Twenty-six respondents report that they expect to add 500 GB or less and 19 anticipate adding more than 500 GB in the next year.
These institutions need to preserve a diverse array of format and genres.

- Still images are cited by respondents as the dominant format (94%). Other popular formats include textual documents (83%), video (76.2%), and audio (74.5%).
- A significant number of respondents also collect email (47.4%), databases (47.5%), and websites (40.6%), and more than a third also collect GIS material (35.5%). Others cited presentation materials, publications, science data, and software code as items that they collect/create.
- Presumably, as these institutions look for preservation solution(s), they will be seeking solutions that can encompass multiple formats and genres of material easily.
Formats Distribution

- Textual documents
- Databases
- Still images
- Video
- Audio
- GIS data
- Web sites
- Email

Slide 38
Platforms and repository structures used by respondents varied widely

- More than half (65%) of the respondents are using an in-house solution to host some or all of their collections.
- The leading repository systems cited were CONTENTdm (9 users, 17%), Fedora (5 users, 9%), DSpace (4 users, 7%), and Access/Excel (3 users, 6%).
- Two each cited SRB and Filemaker as repositories that they currently use. Ten other systems were cited by ten additional respondents.
- This diversity of system types—presumably at least somewhat representative of the overall industry—presents an array of challenges for preservation, especially for distributed replication approaches.
Institutions need assistance in designing suitable policies for managing their digital materials

- Only 21% of respondents state that their institution has a preservation plan for its digital archives.
- Less than half of respondents have any written policies for managing their digital collections (44.8%).
- Of those that have policies at all, most only cover metadata standards (76.9%), back-up strategies (65.4%), conversion of materials from print to digital (61.5%), preservation (57.7%), and acquisition (53.8%).
- Only three institutions (5% of respondents) believed that they had policies that met their institution’s current needs well
Inadequate Preservation Programs

Many of these institutions are not yet even backing-up—let alone preserving—their digital assets

- Only 25 respondents (50% of 50 respondents) stated that they back up all of their digital holdings.
- Of the remaining, less than a quarter (11) back up 75% or more, only 5 more back-up 50% or more, and 8 back up less than 50%.
- Only 10 respondents believed they had in-house expert knowledge regarding digital preservation, 36 felt they had intermediate knowledge, and 8 believed they had only novice-level knowledge of preservation.
Institutions perceive the insufficiencies of their policies, plans, and resources to be the biggest threats to the loss of their digital assets.

- According to survey responses, the greatest threat to collections is insufficient resources for preservation (75%, or 41 institutions), followed by insufficient policies and plans for preservation (51%, or 28 institutions).
- Less troubling to these institutions were such issues as technological obsolescence (31%, or 17 institutions) and the stability of the storage medium (20%, or 11 institutions).
- Institutions need affordable, prescriptive options.
There is widespread interest among cultural memory organizations in participating in cooperative preservation networks

- Three-quarters of respondents (42 institutions) cited interest in participating in a cooperative preservation network, and 89.5% cited interest in “participating in a community-based digital preservation solution.” This stood in contrast to those who indicated interest in preservation services provided by third-party vendors (30.4%, or 17 institutions).
- There is also widespread knowledge of and experience with LOCKSS among the respondents, with almost 50% (28 institutions) citing such knowledge and experience.
- This indicates that the community widely supports the idea of maintaining an active role in collaborative digital preservation activities.
Centralized models are easier to understand and create, are more predominant (often for profit), and yet are more likely to be fragile in the long term.

Collaboration is harder to articulate and operationalize, are less predominant (always nonprofit), and yet may have more strength in the long term.
Networks may build onto other existing preservation networks:

- Can take advantage of previous investments, research, etc. by others
- Requires understanding the rules of each existing network and working out agreements
- Hardest approach to undertake, but might have the most hybrid vigor
May be a need to describe in a DDP reference framework things such as:

- Digital preservation approach (centralized, distributed, formats, genres, etc.)
- Organizational framework and governance (based on existing or new consortia)
- Authentication and authorization models
- Roles and responsibilities
- Technical standards and system architecture
- Funding / economic stability model
Roles in the Stewardship Network

Committed Content Custodians

Communities of Practice and Information Exchange

Services

Capacity Building

Source: “Since we met last year…” Plenary, Martha Anderson, National Digital Information Infrastructure and Preservation Program Annual Partners Meeting 2008
The Future of Community-Based Distributed Digital Preservation

- Contests the view of centralized preservation strategies, especially those that seek to constrain access to shared cultural memory
- Must reach a sustainable balance in funding the shared development of infrastructure
- Offers the only plausible path forward for broadly based long-term access
- Can surmount local failures through broad community awareness
Howard Thurman

from

The Search For Common Ground; An Inquiry Into The Basis Of Man's Experience Of Community.
Thank You!