

Collaborative Adventures in Distributed Digital Preservation:



The MetaArchive Cooperative and the Educopia Institute

DR. KATHERINE SKINNER

**Emory University
February 27, 2008**

Presentation Overview



- **Challenges/Opportunities of Distributed Digital Preservation**
- **What is the MetaArchive Cooperative**
- **Strategies we've employed to support, sustain, and grow the Collaborative Network to date**
- **Lessons Learned: Strengths we've found of different organizational structures for accomplishing collaborative goals**

Challenges/Opportunities of Distributed Digital Preservation



What is Digital Preservation:



What is Digital Preservation:



- ***Digital Preservation:*** Managed activities necessary for ensuring both the long-term maintenance of a bytestream and continued accessibility of its contents. (TDR, p.3)
- ***Digital Archive:*** an organization responsible for digital preservation. (OAIS Reference Model)

Goal: the accurate rendering of authenticated content.

What is Distributed Digital Preservation:



- ***Distributed Digital Preservation:*** The distribution, management, and maintenance of digital information over a wide geographical area and over a long period of time—maintaining its viability, authenticity, and accessibility across changing technologies, formats, and user expectations. (*Guide to Distributed Digital Preservation*)

Goal: provide additional security through distribution.

Why do we preserve?



Access in the digital realm over time has a prerequisite: preservation. How can we preserve essential objects?

- Digital data sets?
- Data generated via geographical information systems?
- Digitally recorded image, audio, and video files?
- Official documentation (government, business, etc)
- The digitized/born digital content of digital archives ?
- Electronic Theses and Dissertations?
- Web sites, blogs (e.g., Sept. 11, 2001, Hurricane Katrina, the Va Tech shootings)?
- Email (e.g., Executive correspondence)?

Who is preserving?



Precious few of us...

- **The Center for Technology in Government's Survey and Report**
 - current capacity for digital preservation is very low, approaches are inconsistent, and there is no standard way to prioritize at-risk materials for preservation
- **Northeast Document Conservation Center 2005 online survey**
 - 88% "collecting, acquiring, or creating digital assets," 30% have been backed up *one time or not at all*
 - Devoted 5% or less of their budget to any type of preservation activity, and 9% devoted *none at all*; 66% report no one is responsible for digital pres. activities
- **Stewardship of Digital Assets 2007-2008 surveys**
 - 94.7% report engaging back up strategies, only 21% report even employing off-site storage of backups. 16.7% report that they are creating no metadata for their digital collections
 - 13.6% have a digital preservation plan, and 12% report operating a digital preservation solution

So what have we been doing?



As science and social science data move primarily to digital representations and as much of our communications infrastructure moves from print to digital, there are concomitant needs for preservation of these materials.

- The Consultative Committee for Space Data Systems (OAIS)
- Digital Preservation Management Workshop (Cornell University)
- National Digital Information Infrastructure and Preservation Program (US)
- Digital Curation Centre (UK)
- Digital Preservation Coalition (UK)
- nestor (Germany)
- Australian Partnership for Sustainable Repositories (APSR)

Examples of Preservation Activities



- **Establishing standards/Standards**
 - OAIS Reference Model
 - Preservation metadata (PREMIS)
- **Developing technical infrastructures**
 - LOCKSS
 - PRONOM/GDFR (registries)
 - JHOVE, DROID, New Zealand metadata extractor
 - etc.
- **Developing organizational infrastructures**
 - CLOCKSS
 - MetaArchive
 - CDL
 - FCLA DAITSS
 - Chronopolis SRB
 - ICPSR's LOCKSS-based system

It's hard to preserve!



Challenges include:

- Inception of a new field
- Rapid pace of technological change (hardware, software)
- Instability of digital medium
- Sheer quantity of information
 - Digital universe of 161 exabytes = more than 3M times the info contained in all books ever produced

Buzz about Standards/Standard Practices



- We're still in innovation/experimentation phase
- The trouble with “TDRs” and the like—have criteria before we can satisfy them. Wonderful in terms of setting high bar. Less wonderful in the expectations that this places on current projects and programs.
- Common problem space at the heart of both “sustainability” and “preservation” as we currently use the terms. The question is not so much: “*is this collection preserved?*” but rather “*for how long can we be confident of preserving this collection?*”

Preservation: Collaborative vs. Centralized



Cooperative Model

- Institutional dependence upon each other
- Institutional members driving development decisions
- Geographic distribution with “live” preservation

Central Service Provider

- Institutional dependence upon one central group
- Company driving development decisions
- Often one location with “back ups” stored elsewhere

What is the MetaArchive Cooperative?



MetaArchive:



The MetaArchive Cooperative (the "Cooperative") is an independent, unincorporated, international membership association.

The Cooperative's purpose is to **support, promote, and extend** the MetaArchive approach to distributed digital preservation practices (<http://www.metaarchive.org>).

Examples of MetaArchive's materials:



- **Born digital and digitized collections**
- **Digital image, sound, and video files**
- **Datasets and Databases**
- **GIS Collections**
- **Websites**
- **Email correspondence**
- **E-journals**
- **Electronic Theses and Dissertations (ETDs)**
- **Encoded texts**

MetaArchive components



- **Technical Infrastructure**
- **Organizational Framework**

Technical Infrastructure



- **Successful deployment of network**
 - Robust, distributed first network launched 2004
 - Fully replicable
 - ✦ Currently founding new networks
 - ✦ Other institutions also founding private LOCKSS networks (PLNs)
 - Open Source
 - Built using LOCKSS
 - ✦ Digital objects, not just journals
 - ✦ Working with larger file sizes
 - ✦ Working with more variable collections



Technical Infrastructure



Created software tools to curate collections

- **Conspectus schema**
 - Based on DC, MODS, CLD, RSLP
 - Mapping to PREMIS
 - Webform
- **Cache manager**
 - Monitors network
 - Generates human-readable reports

Collection Description Data Creator

This is the main form, click on the plus signs to view sections

Descriptive Data Summary
This is data pertaining to the title and description of the collection

Collection Title This is the title of the collection

Alternative Title Add Any form of the title used as a substitute or alternative to the formal title of the resource

Description Required A description of the collection

ESC Subjects Add ESC subject headings associated with the collection

LCSH Subjects Add LCSH subject headings associated with the collection

MESH Subjects Add MESH subject headings associated with the collection

URIs Summary
This is URIs and unique identifiers associated with the collection

Collection URI Required This is the URI associated with the collection

Identifier Add An arbitrary assigned identifier for the collection, according to local conventions

Is available via Add The URL where the collection is publicly available

Coverage Summary
metadata dealing with the physical and temporal location of the contents of the collection

Spatial Coverage Add The place or places associated with the contents of the digital collection

Temporal Coverage Add The time periods associated with the contents of the digital collection

Accumulation Date Range Add The range of dates over which the collection was accumulated

Contents Date Range Add The range of dates over which the individual

Technical Infrastructure

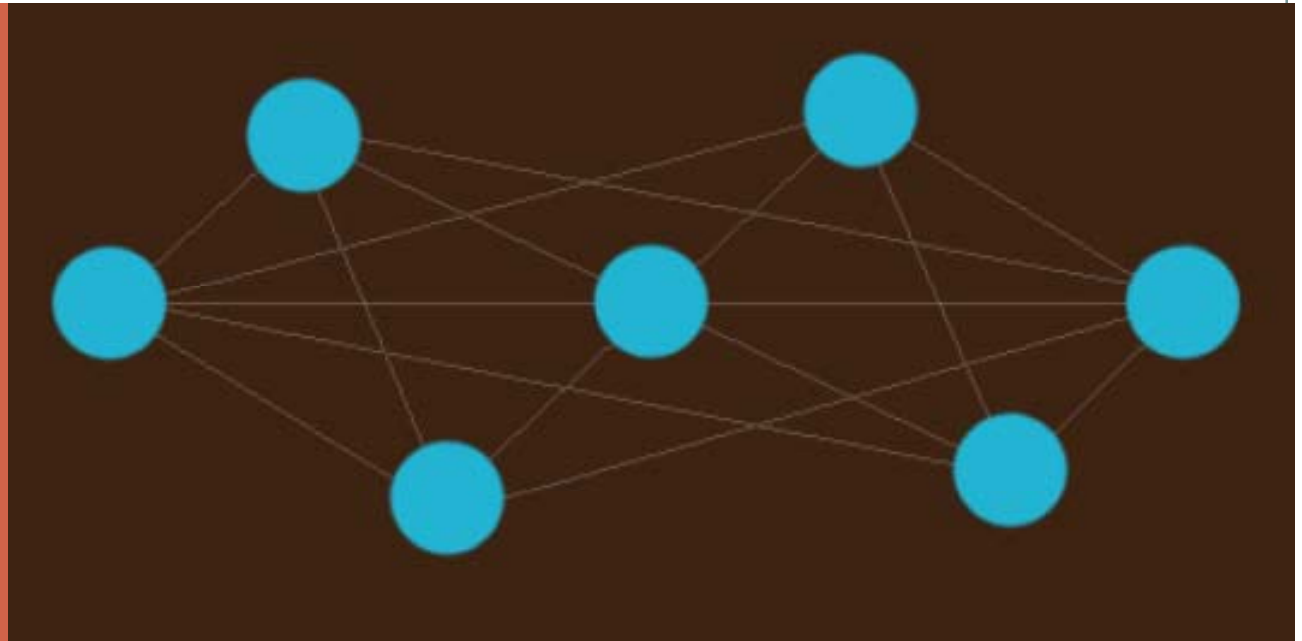
- Preserving more than 200 collections to date
- Harvesting from CONTENT dm, Dspace, Fedora



Each node of the network is represented here in blue.

All nodes contain copies of a network's harvested content. These nodes then communicate with each other constantly, staying alert for any bit rot or fragmentation of the files they contain.

If one node's copy begins to deteriorate, the other nodes compare their copies to make sure that they agree on the correct content version. Once they reach quorum, they can safely fix the decay.



MetaArchive distributed digital preservation model: Lots of Copies Keep Stuff Safe

Technical Infrastructure



... and that was the *easy(ish)* part!

Strategies we've employed to support, sustain, and grow MetaArchive



Organizational Framework



- **External collaboration**
 - Funding agencies
 - Other Digital Preservation Services (LOCKSS, SRB, etc)
 - Coordinating with Standards bodies and emerging standards
- **Internal collaboration**
 - From project partners to Cooperative members

Collaborative Partnerships: External



- **Collaborations with other entities**
 - Library of Congress National Digital Information Infrastructure and Preservation Program (NDIIPP) – center for expertise
 - National Historical Publications and Records Commission (NHPRC)
 - Tapping into existing arenas (NDLTD)
- **Collaboration with other Digital Preservation groups**
 - Chronopolis SRB
 - LOCKSS
 - Statewide Networks (AL, AZ, GA)
 - ICPSR

Collaborative Partnerships: Internal



- **Began as one six-institution network as part of the NDIIPP MetaArchive project**
 - Library of Congress, Emory University, Ga Tech, Va Tech, Auburn University, University of Louisville, Florida State
- **Sustainability demanded longer-term relationship**
 - Cooperative Charter and Membership Agreement

Collaborative Partnerships



Cooperative Charter and Membership Agreement

- **Two interrelated goals:**
 - To define the mission and operating principles, membership responsibilities, governance structure, and services and operations of the Cooperative, and
 - To formalize the relationships between member institutions as an effective consortium



In 2005, the original project partners of the NDIIPP-funded project decided that their cooperative approach would provide a sustainable framework for distributed digital preservation.

Educupia Institute

2888 Chimney Springs Drive
Marietta, Georgia 30062 Phone 678 461 0664

MetaArchive Cooperative Charter

A charter describing the purposes and aims of the MetaArchive Cooperative, an association dedicated to the preservation of cultural heritage materials that are digital in nature and form

Table of Contents

1. Introduction.....	4
1.1. What is the MetaArchive Cooperative	4
1.2. What is the MetaArchive of Southern Digital Culture	4
1.3. Mission and Operating Principles	4
1.4. Who Should Participate?	5
2. Membership	5
2.1. Eligibility	5
2.2. Types of Membership	5
2.2.1. MetaArchive Sustaining Members	5
2.2.2. MetaArchive Preservation Members	6
2.2.3. MetaArchive Contributing Members	6
2.2.4. MetaArchive Sponsorships	6
2.3. Costs and Fees	6
2.3.1. LOCKSS Alliance Membership	7
2.3.2. Systems Administration and Cache Monitoring	7
2.3.3. Communications	7
2.3.4. Content Provision	7
2.3.5. Administration	8
2.4. Benefits and Responsibilities	8
2.4.1. Benefits	8
2.4.2. Responsibilities	9
2.4.3. Copyright and Intellectual Property	11
2.5. The MetaArchive Membership Agreement	11
2.6. Joining the Cooperative	11
3. Organization and Governance	11
3.1. The MetaArchive Cooperative	11
3.2. The MetaArchive Committees	12
3.2.1. MetaArchive Steering Committee	12
3.2.2. MetaArchive Content Committee	12
3.2.3. MetaArchive Preservation Committee	12
3.2.4. MetaArchive Technical Committee	12
3.2.5. Selection and Terms of Service	12
3.3. Communication	13
3.4. Annual Meeting	13
3.5. Withdrawing from the Cooperative	13
3.6. Procedures for Non-Compliance and Material Breach	13
4. Services and Operations.....	14
4.1. MetaArchive Cooperative Services	14
4.1.1. Digital Preservation Network	14
4.1.2. Digital Collection Disaster Recovery	15
4.1.3. Digital Preservation Network Assistance	15
4.1.4. Security Characteristics of a Preservation Network	16

MetaArchive organizational model: Cooperative Association

Membership Levels and Responsibilities



- **Sustaining Members:**

- Pioneers. \$5,000/year; 3-year term; host node for research, development, and preservation activities; representation on the Steering Committee; access to 40 GB space*

- **Preservation Members:**

- Central preservation partners. \$1,000/year, 3-year term, host node for preservation activities, access to 20 GB space*

- **Contributing Members**

- Smaller institutions that do not want to host the infrastructure but need to preserve their materials. \$200/year, 3-year term, access to 5 GB space*

*more space can be purchased by GB as needed

Collaborative Partnerships



- **Question arose: with whom were we making the agreements/commitments?**
- **Who's in charge of a Cooperative that is comprised of peer institutions?**

Need for a New Catalytic Organization



- Centralized management entity for our consortium
- External organization to administer the cooperative
- Clear leadership and accountability
- Focus is on the cooperative, not on individual institutional goals
- Can forge mutually beneficial relationships with other consortia
- Continuity of programmatic goals
- Enable Cooperative activities undertaken by peer institutions

Educopia



- **In October 2006, we created a 501(c)3 nonprofit organization to address the needs of cultural memory institutions for shared cyberinfrastructure**
 - Distributed digital preservation (dim archiving)
 - Access mechanisms for lighting up dim archives
 - (prospective) digital publishing at consortial level
- **Serves as a catalytic agency**
 - Consortial grant applications
 - Training
 - Consulting and advising services

Educopia and MetaArchive



The Educopia Institute provides administrative services for the MetaArchive Cooperative, including:

- billing member organizations for annual dues;
- maintaining and distributing funds;
- organizing and hosting annual meetings of MetaArchive members;
- holding members accountable for completing agreed-upon tasks;
- hosting workshop programs on digital preservation topics.

Board Members



Officers

- Martin Halbert (president)
- Tyler Walters (treasurer)
- David Seaman (vice-president)
- Rachael Bower (secretary)
- Greg Crane

Staff

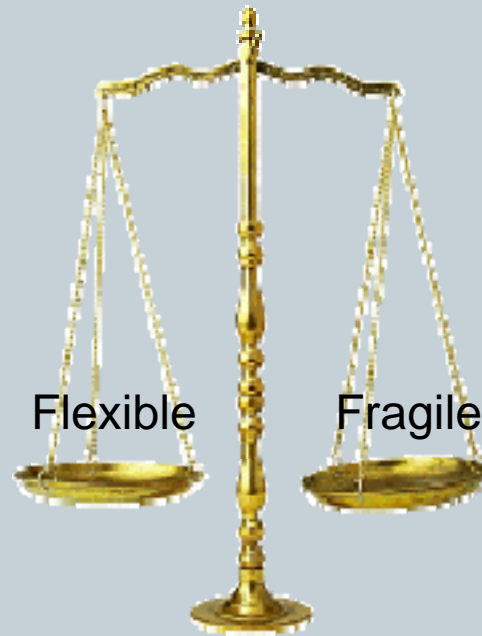
- Katherine Skinner (executive director)

Lessons Learned:



STRENGTHS OF DIFFERENT ORGANIZATIONAL STRUCTURES FOR ACCOMPLISHING DIFFERENT GOALS

The Challenge:



Flexible

Fragile

Collaborative Networking:



- **Management of:**
 - Accountability
 - Legitimacy
 - Conflict
 - Commitment
 - Design

Source: Milward and Provan. *A Manager's Guide to Choosing and Using Collaborative Networks*, 2006

Collaborative Networking:



- **Management of:**
 - **Accountability**

Management of Accountability:



- **who is responsible for what?**
- **who determines who is responsible for what?**
- **who notices when tasks are left undone?**
- **who responds (and how) when tasks are left undone?**
- **what about “free riders?”**

Accountability in MetaArchive



- **Initially, contract with LC—solid guide to work, heavily documented responsibilities**
 - BUT—already encountered challenges: staffing turnover, changes in leadership
 - Also, as with most contracts/grants, one lead institution bore uneven responsibility
- **As we transitioned from project to program, needed to establish a better system**
 - Membership with a central agency enabled clear line of direction; membership agreement spells out both commitment and what happens if that commitment is not met

Collaborative Networking:



- **Management of:**
 - Accountability
 - Legitimacy

Management of Legitimacy:



- how does a cooperative venture earn clout?
- how does it convince new members that it is worth joining?
- how does it convince stakeholders that their work in the larger network continues to be valuable and worthwhile?

Legitimacy in MetaArchive:



- Initially, just seeking to establish a preservation network; later, once that network was successfully running, began to seek new partners.
 - Could do so through new grant applications, but then the structure remains flimsy and time bound
- In order to grow and to encourage others to adopt the MetaArchive methodology, needed to have a central group to take responsibility
- In order to negotiate with other important groups (e.g., repository systems), needed to be more than an loosely affiliated group of university libraries

Collaborative Networking:



- **Management of:**
 - Accountability
 - Legitimacy
 - **Conflict**

Management of Conflict:



- **how do you settle disputes between stakeholders?**
- **who has the authority to settle such disputes?**

Management of Conflict in MetaArchive:



- **Initially, reliant on contract with LC for the original project. As morphed into membership organization, needed to have a clear directing agent that could settle disputes.**
 - Every problem has multiple solutions. How do you know which solution to pursue? Central management agency helps to provide that leadership

Collaborative Networking:



- **Management of:**
 - Accountability
 - Legitimacy
 - Conflict
 - **Commitment**

Management of Commitment:



- how do you entice other organizations to commit to—and follow through with—work?
- salaries/schedules are not in the hands of the lead org, even when a lead org exists
- what motivational tactics available?

Commitment to MetaArchive:



- **Membership fees and tiered membership system**
- **Membership agreement specifies expectations**
- **Prestige factor to help with motivation**

Collaborative Networking:



- **Management of:**
 - Accountability
 - Legitimacy
 - Conflict
 - Commitment
 - **Design**

Management of Design:



- **Distributed, self governing**
- **Centralized, lead organization**
- **Centralized, formed management entity**

Management of Design:



- **Distributed, self governing**
- **Centralized, lead organization**
- **Centralized, formed management entity**

Management of Design:



- **Distributed, self governing**
 - Founded with formal or informal institutional agreements
 - No lead institution—all are equal partners
 - Very flexible; very fragile
 - One bad apple spoils the barrel

Management of Design:



- Distributed, self governing
- Centralized, lead organization
- Centralized, formed management entity

Management of Design:



- **Centralized, lead organization**
 - One institution (also a partner) serves as the lead for the collaborative
 - Provides clear line of command
 - That line of command has limited authority
 - Institutional goals may collide with (or unduly influence) network goals

(where we started...and we were very lucky to have excellent partners!)

Management of Design:



- **Distributed, self governing**
- **Centralized, lead organization**
- **Centralized, formed management entity**

Management of Design:



- **Centralized, formed management entity**
 - External organization to manage the network
 - Clear leadership and accountability
 - Focus is on the network, not on individual inst. goals
 - Can forge mutually beneficial relationships with other consortia

(where we are now...more formal arrangement, but we've left plenty of room for smaller projects and experiments affiliated with MetaArchive at other design levels.)

Benefits from central design in MetaArchive:



- Administrative apparatus separate from members
- Clear commitments and responsibilities
- Clear leadership and accountability
- No blurring of individual members' goals and the cooperative's direction (as can happen with centralized lead organization)
- Leverage for forging external partnerships
- Joint applications for sponsored funding don't get hit by "double overhead"
- Continuity of programmatic goals

Resources



- Pardo, Theresa A., G. Brian Burke, and Hyuckbin Kwon, “Preserving State Government Digital Information: A Baseline Report,” (July 2006). http://www.ctg.albany.edu/publications/reports/digital_preservation_baseline/
- Claeson, Tom. "NEDCC Survey and Colloquium Explore Digitization and Digital Preservation Policies and Practices" *RLG DigiNews*, 10:1 (February 2006). http://www.rlg.org/en/page.php?Page_ID=20894#article1
- Consultative Committee for Space Data Systems, “Reference Model for an Open Archival Information System (OAIS)” (Jan 2002). <http://public.ccsds.org/publications/archive/650x0b1.pdf>
- RLG/OCLC, “Trusted Digital Repositories: Attributes and Responsibilities” (May 2002). <http://www.oclc.org/programs/ourwork/past/trustedrep/repositories.pdf>
- Gantz, John F., David Reinsel, Christopher Chcute, Wolfgang Schlichting, John McArthur, Stephen Minton, Irita Xheneti, Anna Toncheva, and Alex Manfrediz. 2007. “The Expanding Digital Universe: A Forecast of Worldwide Information Growth Through 2010.” IDC and EMC White Paper, available at <http://www.emc.com/collateral/analyst-reports/expanding-digital-idc-white-paper.pdf> (accessed on December 14, 2007).

Questions and Comments?



Katherine Skinner
kskinne@emory.edu
404 783 2534