The MetaArchive Cooperative:

A Collaborative Approach to Distributed Digital Preservation

DR. KATHERINE SKINNER
Emory University

NISO DIGITAL PRESERVATION FORUM:
PLANNING TODAY FOR TOMORROW'S RESOURCES

March 14, 2008

Presentation Overview

- Distributed Digital Preservation
- The MetaArchive Cooperative
  - Lessons Learned:
    - Technical Infrastructure
    - Curating Collections
    - Organizational Infrastructure
  - Future Directions:
    - Standards-based work
    - Building bridges with other digital library systems
    - Building bridges with other preservation solutions

Distributed Digital Preservation

CHALLENGES AND OPPORTUNITIES

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)

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 added security through distribution.

Who is preserving?

Precious few of us…

- The Center for Technology in Government 2006 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
- NEDCC 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 *are* we doing?

- 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
  - LOCKSS
  - MetaArchive
  - CDL
  - FCLA DAITSS
  - Chronopolis SRB
  - ICPSR's LOCKSS-based system
  - Etc.

Among the challenges we face…

- Sheer quantity of information + rapid technical developments
  - Digital universe of 161 exabytes in 2006
- We're still in innovation/experimentation phase
  - ALL of us are still establishing our base infrastructures. We know where we're aiming, but no one is at "TDR" capacity yet
- Standards!
  - They both *guide* and *follow* the innovation phase—tricky balance
- Common problem space in "preservation"
  - *"Is this preserved?"* vs. *"for how long can we be confident of preserving this?"* Acknowledge it's an incremental process.
- Major librarian/curator question: do we conduct our own preservation activities or do we outsource?
  - Ramifications either way; must think through carefully

The MetaArchive Cooperative

MetaArchive Cooperative:

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).

MetaArchive's Members

- To date, seven Sustaining Members:
  - Emory University
  - Georgia Tech
  - Virginia Tech
  - Auburn University
  - University of Louisville
  - Florida State University
  - University of Hull (UK)
- More than 5 dozen institutions currently considering membership at various levels.

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

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

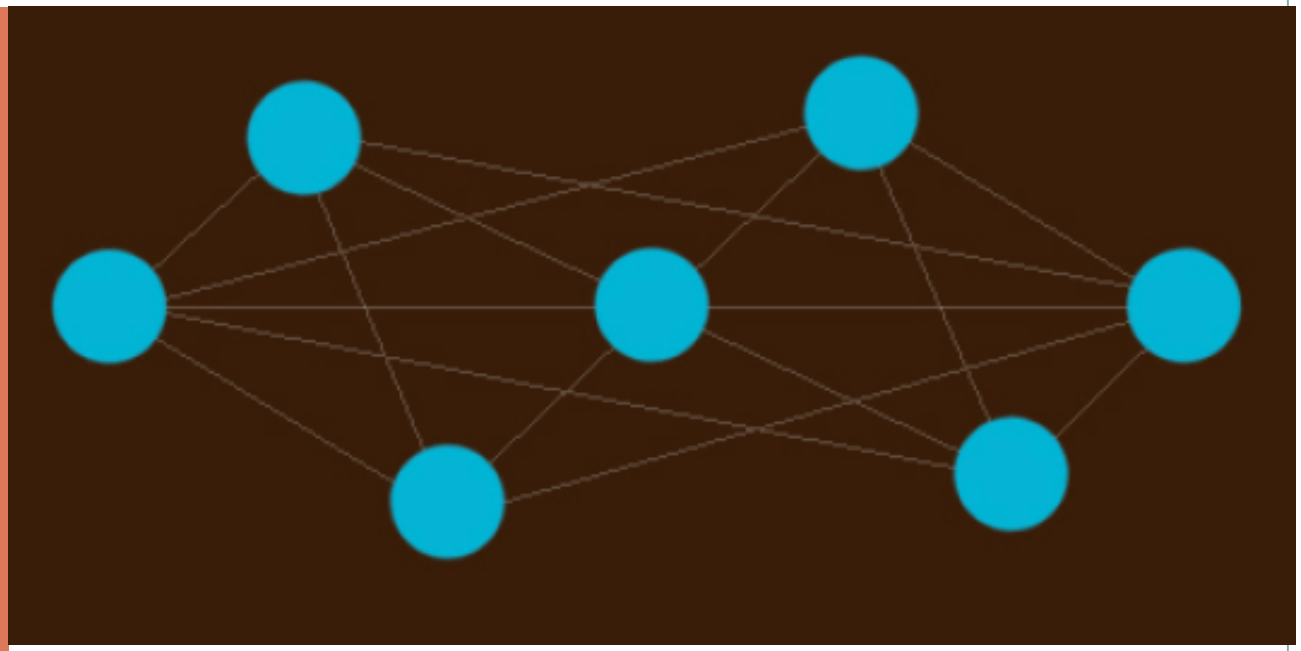
Technical Framework

- LOCKSS-based Distributed Digital Preservation Network
  - Robust, distributed network launched 2004
  - Open Source
  - Built using LOCKSS
    - Digital objects, not just journals
    - Working with larger file sizes
    - Working with more variable collections
  - Fully replicable
    - Others now founding private LOCKSS networks (PLNs), including Alabama, Arizona, Georgia, ICPSR, etc.

Each node of the network is represented here in blue.

All nodes contain copies of a network's ingested 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.

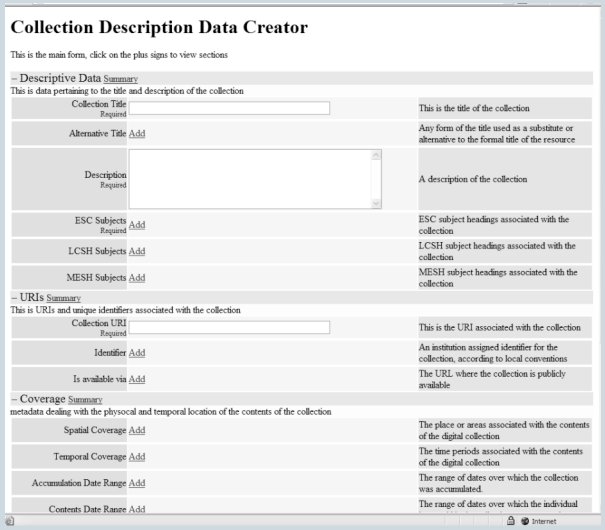


MetaArchive's LOCKSS-based distributed digital preservation framework: Lots of Copies Keep Stuff Safe

Technical Framework

MetaArchive has created software tools to curate its collections

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



Lessons Learned: Technical Framework

- LOCKSS applies well to non-serialized content
- Guaranteeing the authenticity of documents may be best accomplished with a distributed digital preservation archive
- Our technical infrastructure requires systems attention at each Preservation site
- Sustainability requires having multiple people experienced with the management of the system at all times

Curating Collections

- Three "archives" to date:
  - Southern Digital Culture,
  - Electronic Theses and Dissertations, and
  - History of the Slave Trade
- Establishing new archives at member requests
- Curatorial decisions made by the contributing institutions, *not* by MetaArchive
- Can ingest digital objects and their metadata
- Require collection-level metadata for retrieval purposes

Curating Collections

- Ingests from web, OAI, CONTENT dm, DSpace, Fedora
- Preserving more than 200 collections to date

Lessons Learned: Curating Collections

- We need collection-level metadata as a tool and system component—item level not required
- Just because an institution has content doesn't mean that content is ready for ingest.
- Subject- and genre-based archives provide an important structure for distributed archiving
- Preservation begins at creation: the organization of an institution's collections can help or hinder its preservation readiness
- Preservation depends on internal institutional documentation as well

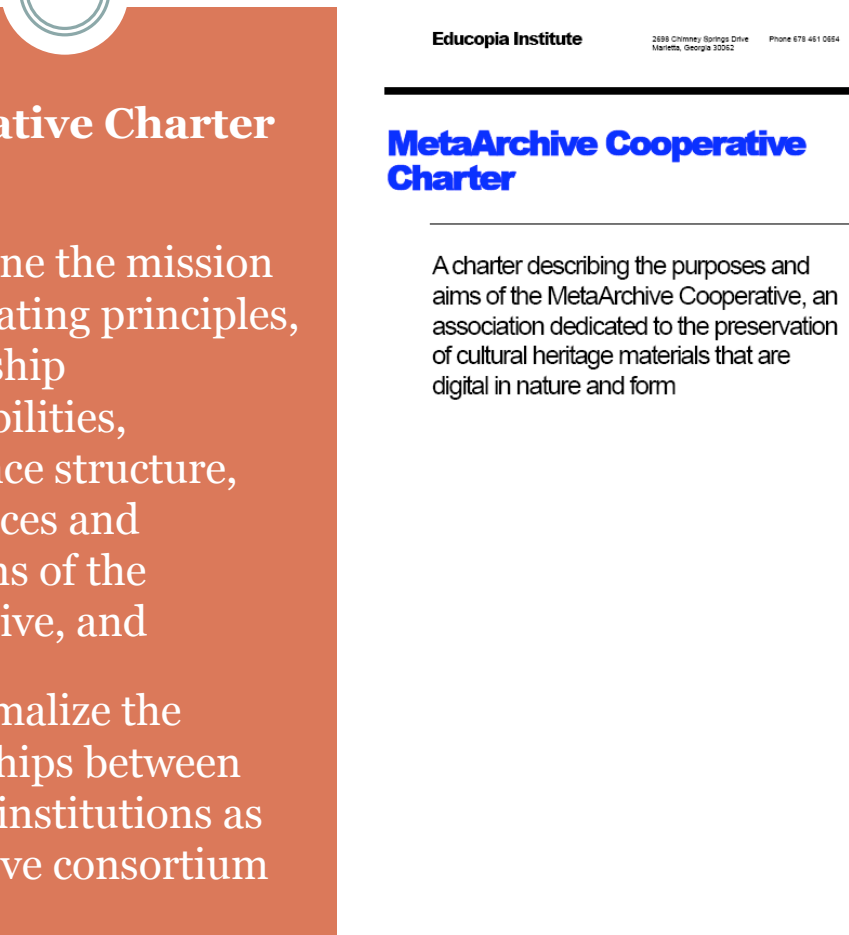
Organizational Framework

- Began as one six-institution network as part of the Library of Congress NDIIPP MetaArchive project
  - Emory University, Georgia Tech, Virginia Tech, Auburn University, University of Louisville, Florida State
- Quickly realized that preservation solution cannot be contingent on contract or grant funding!
- Sustaining our network demanded longer-term relationship
  - Cooperative Charter and Membership Agreement

Cooperative Charter Goals:

1. To define the mission and operating principles, membership responsibilities, governance structure, and services and operations of the Cooperative, and

2. To formalize the relationships between member institutions as an effective consortium



Educopia Institute 2698 Chimney Springs Drive Marietta, Georgia 30062 Phone 678 461 0854

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

MetaArchive organizational model: Cooperative Association

Organizational Framework

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

Organizational Framework

- In October 2006, we created the Educopia Institute, a 501(c)3 nonprofit organization to address the needs of cultural memory institutions for shared cyberinfrastructure
  - Distributed digital preservation (dark and dim archiving)
  - Access mechanisms for lighting up dim archives

Organizational Framework

The Educopia Institute provides services for MetaArchive, including:

- working with prospective members;
- collecting, maintaining, and distributing funds;
- maintaining documentation, website, listservs;
- organizing and hosting meetings and workshops;
- holding members accountable for completing agreed-upon tasks; and
- fostering relationships with other consortia.

Lessons Learned: Organizational Framework

- Benefits to MetaArchive:
  - Administrative apparatus separate from members
  - Clear leadership and accountability
  - No blurring of individual members' goals and the cooperative's direction (as can happen with lead organization or a distributed model)
  - Leverage for forging external partnerships
  - Joint applications for sponsored funding don't get hit by "double overhead"
  - Maintain continuity of programmatic goals

Future Directions

Standards-based work

- Trustworthy Digital Repository—continuing to build toward certification
- PREMIS-based work for our collections-level metadata
- Establishing automated means of extracting and recording technical metadata for each file (using an existing tool, under evaluation)
- Extra work on authenticity and on establishing migration pathways for our archives' contents

Building bridges: other DL systems

- Of particular concern: building technical—and where of interest to both parties, organizational—bridges between MetaArchive's LOCKSS-based system and the most common DL repository systems
  - DSpace, Fedora, CONTENT dm, DigiTool, and others

Building bridges: other preservation solutions

- NHPRC award to build such a bridge between MetaArchive's LOCKSS-based system and SRB
- PREMIS mapping and tests of export and ingest procedures between MetaArchive and other digital preservation systems

Final note: Standards & Digital Preservation

- There are so many! And we are so young!
- Not all Standards apply to all systems
  - Make selections based on the system's goals and the institution's goals
- Flexibility vs. Strength
  - Flexibility allows more participation, but we have to draw the line somewhere in order to strengthen our approaches
- As we build this new field, the best Standards enable:
  - *conversation in a common language* (i.e., OAIS) and
  - *portability of content* (i.e., PREMIS, in the "*s*"tandards sense)

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/
- Clareson, 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

Executive Director, Educopia Institute
Digital Projects Librarian, Emory University
kskinne@emory.edu
404 783 2534