New Cooperative Strategies for Distributed Digital Preservation

TYLER WALTERS
UBC Vancouver, March 12, 2010
Presentation Overview

- The MetaArchive Cooperative
  - MetaArchive Components and Lessons Learned
    - Technical Infrastructure
    - Curating Collections
    - Organizational Infrastructure

- Future Directions and Initiatives...
So, what is a “Cooperative” anyway?

- From Dictionary.com:

  “a jointly owned enterprise engaging in the production or distribution of goods or the supplying of services, operated by its members for their mutual benefit…”
MetaArchive: Basic Facts

- Established in 2004, began preserving content for 6 member institutions
- Uses LOCKSS software to provide long-term management for materials in a distributed digital preservation network
- Sustainable organizational framework: Membership organization with a 501c3 host (Educopia Institute)
- 254 TB network capacity (20–50 TB per member, adding more as new members join)
- OAIS Compliant, and as a Trustworthy Digital Repository (2009 TRAC audit available on our site next week)
- Guide to Distributed Digital Preservation – published!
  - 755 downloads since February 23!
Current Members/Contributors
Auburn University
Boston College
Clemson University
Emory University
Florida State University
Folger Shakespeare Library
Georgia Tech
Indiana State University
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
Library of Congress
NDLTD
SDSC Chronopolis

We welcome new members!
The Distributed Digital Preservation Process

1. Assessing DP needs
2. Developing DP plan and policies
3. Selecting content
4. Preparing contracts / MoAs
5. Implementing preservation technology/service
6. Submitting content
7. Monitoring content
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
Technical Framework

- LOCKSS–based Distributed Digital Preservation Network
  - Robust, distributed network launched 2004
  - Open Source
    - For digital objects, not just journals
    - Working with larger file sizes
    - Working with many varied collections
  - Fully replicable technical model

- Others now founding private LOCKSS networks:
  - Alabama/ADPNet, Arizona/PeDALS, ICPSR/DataPASS, western Canada/COPPUL, USGovDocs mirror sites, CLOCKSS, KOPAL (Germany)
Created software tools to curate its collections:

- **Conspectus schema**
  - (Fixity, provenance, content, context)
  - Webform
  - Based on DC, MODS, DC/CLD,
  - UKOLN RSLP
  - PREMIS (future)

- **Cache manager**
  - Monitors network (AUs, collections, inst.-level)
  - Generates human-readable reports
Lessons Learned: Technical Framework

- LOCKSS applies well to non-serialized content
- Guaranteeing authenticity of digital resources may be best accomplished with a distributed digital preservation system approach
- Technical infrastructure requires systems administration 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
  - History of the Slave Trade

- Establishing new archives at member requests
  - e.g. Newspapers, Research Data, Univ. e-Records/Archives

- Curatorial decisions made by the contributing institutions, *not* by MetaArchive

- Can ingest digital objects and their metadata

- Require collection-level metadata for retrieval
Curating Collections

Ingest from web, OAI, CONTENTdm, DSpace, Fedora, and more / Preserving ca. 1,000 Collections to date
Lessons Learned: Curating Collections

- We needed improved collection-level metadata as a basic tool and system component. (Conspectus DB)
- Just because an institution has content doesn’t mean that content is ready for ingest
- 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 Infrastructure

- 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 a preservation solution cannot be dependent on grant funding!

- Sustaining the network demanded longer-term relationship, not dependent on one institution
  - Cooperative Charter and Membership Agreement
Model: Cooperative Association

Cooperative Charter Goals:

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

2. Formalize relationships between member institutions as an effective consortium
Organizational Infrastructure

- **New Host Institution: Educopia Institute, Inc.**

- 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

  - Work with prospective members;
  - Collect, maintain, and distribute funds;
  - Maintain documentation, website, listservs;
  - Organize and host meetings and workshops;
  - Hold members accountable for completing tasks;
  - Foster relationships with other consortia
Organizational Infrastructure

- **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*

*more space can be purchased by GB as needed ($2/GB)
Future: Chronopolis – MetaArchive
Improving Inter-Institutional Preservation

- From Silos to Interoperability...

- Two successful early approaches:
  - Integrated Rule-Oriented Data System (iRODS)
  - Lots of Copies Keep Stuff Safe (LOCKSS)

- Powerful technologies, currently isolated

- Seeking to bridge the gap and foster interoperability
Other Future Initiatives…

- PLN Conference in Boston, Oct. 25–26, 2010
  - Creating a community of distributed digital preservation practitioners

- “Super Node” Initiative under discussion

- Current / Future studies:
  - Understanding the Economics of Digital Preservation
Questions and Comments?

Tyler Walters

404 385 4489

Tyler@gatech.edu

TyWalters1 = Skype / ooVoo/ Gmail / Gtalk

NHPRC

Digital Preservation

National Historical Publications and Records Commission