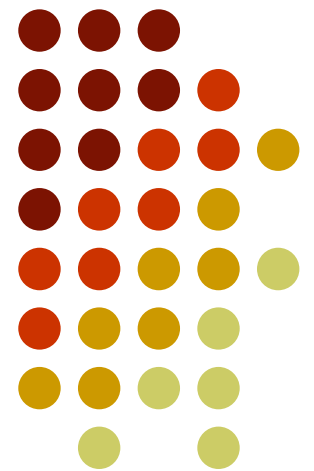


The MetaArchive of Southern Digital Culture

An Approach to Digital
Preservation

Rachel Howard and Delinda Buie
University of Louisville Libraries





The problem:

- Digital information is inherently ephemeral due to rapid change and development of formats and equipment.
- Digital is the way we – scholars, students, businesses, organizations of all types – work now.
- Preserving digital formats requires intention and resources – and, typically, institutional commitment and significant support.





At-risk digital content:

- Web-based projects, exhibitions, and instructional materials with significant content and/or dynamic components.
- Digital media, including video and sound recordings.
- Institutional records or publications created in digital formats.
- Datasets and other primary research materials.
- Personal papers or creative works developed in digital format.



At-risk digital files:

- Materials with uncertain institutional support or unclear lines of responsibility.
- Materials published or developed over time with various sections stored in different digital formats.
- Materials based on older or outmoded technology.

NDIIPP: National Digital Information Infrastructure and Preservation Program



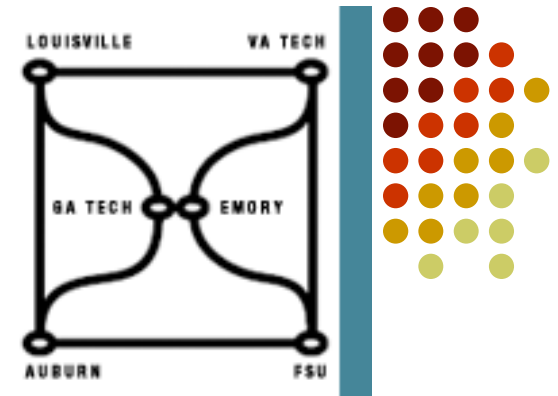
- Federal legislation authorized \$98.2 million to Library of Congress in December 2000 to:
 - Build and support a national network of partners working together to preserve digital content.
 - Identify and preserve at-risk digital content.
 - Support development and use of tools, models, and methods for digital preservation.
 - Develop a national digital collection and preservation strategy.

NDIIPP Preservation Network



- Eight initial awards made in September 2004 in areas such as:
 - Public Television
 - Dot-Com Era Business Records
 - Humanities and Social Sciences Data
 - Geospatial Information
 - MetaArchive of Southern Digital Culture
- Overall effort involves more than 100 partners and 245 terabytes of data.

MetaArchive: The Partners



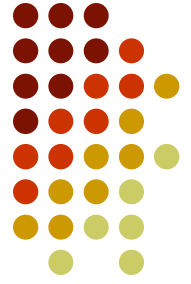
- Emory University (Atlanta, Georgia)
- Georgia Tech (Atlanta, Georgia)
- University of Louisville (Louisville, Kentucky)
- Virginia Tech (Blacksburg, Virginia)
- Florida State University (Tallahassee, Florida)
- Auburn University (Auburn, Alabama)

MetaArchive: The Big Picture



- Establish a distributed digital preservation network for critical and at-risk content relating to the history and culture of the American South.
- Develop a conspectus, or list of targeted collections, to insure preservation of the digital materials most vulnerable to loss and in formats considered most at risk.
- Use open-source LOCKSS (Lots of Copies Keeps Stuff Safe) software, developed at Stanford University, to collect digital content from each other.

MetaArchive: The Big Picture



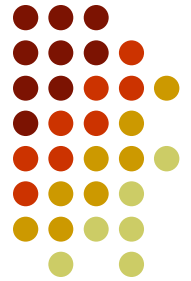
- Ensure sustainability beyond grant funding.
 - Research and draft a cooperative agreement to carry the project beyond the three years funded by NDIIPP, and to encourage new partners to join.
 - Establish standards and guidelines to offer as a model for new networks and collaborations.
- With other NDIIPP-funded projects, help the Library of Congress to raise and begin to answer questions about how to preserve information while protecting the rights of creators.

LOCKSS: Lots of Copies Keep Stuff Safe



- Software developed for e-journals
 - Adapting journal concepts (“volumes”) to archival digital materials.
- Designed to be inexpensive
 - Open source
 - Requires a server but memory keeps getting cheaper.
 - Does require initial support from someone with knowledge of servers and development.

Private LOCKSS Networks



- Multiple geographically dispersed sites host preservation nodes
 - A "node" is a server that is dedicated to collecting materials from every other node, checking to make sure each copy is "right."
 - Participants communicate permission to the LOCKSS system to harvest their materials via a web crawler.
 - Eventually, OAI-PMH as well



MetaArchive: Selecting Content

“...we rely upon curatorial practices of communities of knowledge because those communities are better at collecting and selecting valuable content than communities of policy.”

Abby Smith

MetaArchive: The Conspectus



- Database of targeted digital content relating to the American South
 - A conspectus is harvested with the digital content for each collection.
- Includes metadata elements developed specifically for the MetaArchive:
 - Based on Dublin Core, Research Support Libraries Programme RSLP (UKOLN), Western States (CDP), and IMLS Digital Collections & Content Collection Description Metadata Schema.
 - Describes the collections and provides information necessary for storage estimates, format migration, location, ownership and rights issues.



Preservation metadata

- Preservation metadata = the information a repository uses to support the digital preservation process. (PREMIS)
 - Maintaining viability, renderability, understandability, authenticity, and identity in a preservation context.
- Preservation metadata includes elements of all types of metadata:
 - Administrative (including rights and permissions)
 - Technical
 - Structural
- Digital provenance (the history of an object) and relationships (especially among different objects within the preservation repository) are important.

Conspectus data elements:



[Access Rights](#)
[Accrual Periodicity](#)
[Accrual Policy](#)
[Accumulation Date Range](#)
[Alternative Title](#)
[Associated Collection](#)
[Associated Publication](#)
[Bytes](#)
[Cataloged Status](#)
[Catalogue or description](#)
[Collection Size](#)
[Contents Date Range](#)
[Creator](#)
[Custodial History](#)
[Description](#)
[Format Characteristics](#)
[Institution Collection Identifier](#)
[Is Available Via](#)

[Language](#)
[LOCKSS Manifest Page](#)
[Manifestation](#)
[MetaArchive Collection Identifier](#)
[OAI Provider](#)
[Publisher](#)
[Recommended Harvest Procedure](#)
[Rights](#)
[Risk Factors](#)
[Risk Rank](#)
[Spatial Coverage](#)
[SubCollection](#)
[Subject](#)
[SuperCollection](#)
[Temporal Coverage](#)
[Title](#)
[Type](#)

Preservation metadata in the Conspectus



- Administrative
 - Access Rights
 - Custodial history
 - Associated Collection
 - Manifestation
 - Access
 - Preservation
 - Replacement
- Descriptive
 - Title
 - Subject
- Structural
 - Format characteristics



LABEL:	Access Rights
NAME:	[dcterms:accessRights]
DEFINED BY:	http://purl.org/dc/terms/dcterms
SOURCE DEFINITION:	Information about who can access the resource or an indication of its security status.
PROJECT DEFINITION:	A statement of any access restrictions placed on the collection, including allowed users, charges, etc.
COMMENTS / EXAMPLES:	<p>The World Intellectual Property Organization, the MPEG-21 initiative, and others currently are jointly developing a Rights Data Dictionary and Rights Expression Language to adequately express:</p> <ol style="list-style-type: none">1. To whom rights are being issued2. What rights are specified3. The resources to which the rights apply4. Conditions that must be met before rights can be exercised <p>However, these standards are not yet to the point of being a recommended standard. For more information on current choices and emerging standards for expressing digital access rights, see Karen Coyle's 2004 Rights Expression Languages: A Report for the Library of Congress</p> <p>For the MetaArchive project, a controlled list of access categories will be established (Restricted, Unrestricted]</p>
ENCODING SCHEMES:	
OBLIGATION:	Mandatory
DATATYPE:	Character String

MetaArchive: Harvesting



- Harvest digital resources from each other to rest in “dark archives” on the multi-terabyte servers purchased with project funds and located at each institution.
- Test the system by:
 - Harvesting a variety of file types and sizes;
 - Simulating security breaches; and
 - Simulating a disaster at one institution in order to re-build and re-populate the cache from the identical sets of data at the other five.
- Establish a framework for MetaArchive rights management.
 - According to current copyright law, even making the six digital copies necessary for a dispersed redundant dark archive could be interpreted as infringement.

Copyright / Section 108 and the MetaArchive



- Current copyright law allows archives not holding copyright to make three preservation copies.
- At this point we are archiving material for which we have copyrights.
- “Orphan works” – works for which we cannot ascertain copyright holders, or works where the copyright holder cannot be located.



MetaArchive: Collaborating

- Committees:
 - Steering – for coordination, communication and reporting
 - Content – to organize, develop and select content
 - Preservation – to work on content retention and transfer, acquisitions practices, metadata maintenance, and migratability
 - Technical – to develop and maintain the server architecture and software



MetaArchive: Collaborating

- Communications:
 - Weekly, hour-long conference calls
 - Twice-yearly meetings of the Steering Committee
 - Development of documents via Wiki
 - Participation in NDIIPP Partner and Affinity Group meetings

MetaArchive: Evolving Partnership



- Membership criteria, with various categories to ensure broad applicability:
 - Sustaining
 - Preservation
 - Contributing
- Roles and responsibilities
- Sustainability plan, including financial
- Creation of non-profit entity Educopia

MetaArchive: Operating Principles

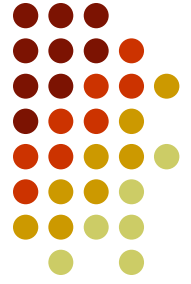


- Commitment to:
 - Long-term preservation of a corpus of critical cultural heritage content.
 - Storage and maintenance in migratable formats and data structures.
 - Standards for metadata and content.
 - A cooperative, peer-to-peer approach to selecting content of shared value, and mutual support of content with a particular, critical, value to individual institutions.

MetaArchive: Operating Principles



- Wide applicability to a range of institutions and digital content.
- Minimal overhead.
- Straightforward mechanisms for collaboration.
- Ongoing exploration of projects to investigate and advance digital preservation.
- Application of LOCKSS software as the principal system for distributing copies of replicated content in secure, distributed locations over time.



Archival formats

- Non-proprietary
- Uncompressed (or at least not lossy)
- In widespread use
- Usable across platforms
- Examples:
 - Images: tiff (jpeg2000)
 - Audio: wav, aiff (mac)
 - Text: plain text (txt); xml; pdf-a
 - Video: motion jpeg, Motion jpeg2000?

Multiple Copies (Lots of Copies!)

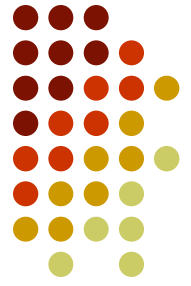


- Preferably, have a copy on a server that is backed up.
- Have another copy on Gold CD
- Keep the CD somewhere distant from the server
- External hard drives (good for video)
- Small collectors – bank safety deposit boxes

Food for thought...



- Digital preservation (and digital libraries) will no longer be separate from preservation (and libraries) – the content we preserve and provide access to is, increasingly, in digital formats.



Further reading

- MetaArchive - <http://www.metaarchive.org/>
- LOCKSS - <http://www.lockss.org/>
- NDIIPP - <http://www.digitalpreservation.gov/>
- PREMIS - <http://www.loc.gov/standards/premis/>
- Digital Preservation Management Tutorial - http://www.library.cornell.edu/iris/tutorial/dpm/eng_index.html