Distributed Digital Preservation:

The MetaArchive Approach

Rachel Howard
Digital Initiatives Librarian
University of Louisville Libraries
rachel.howard@louisville.edu

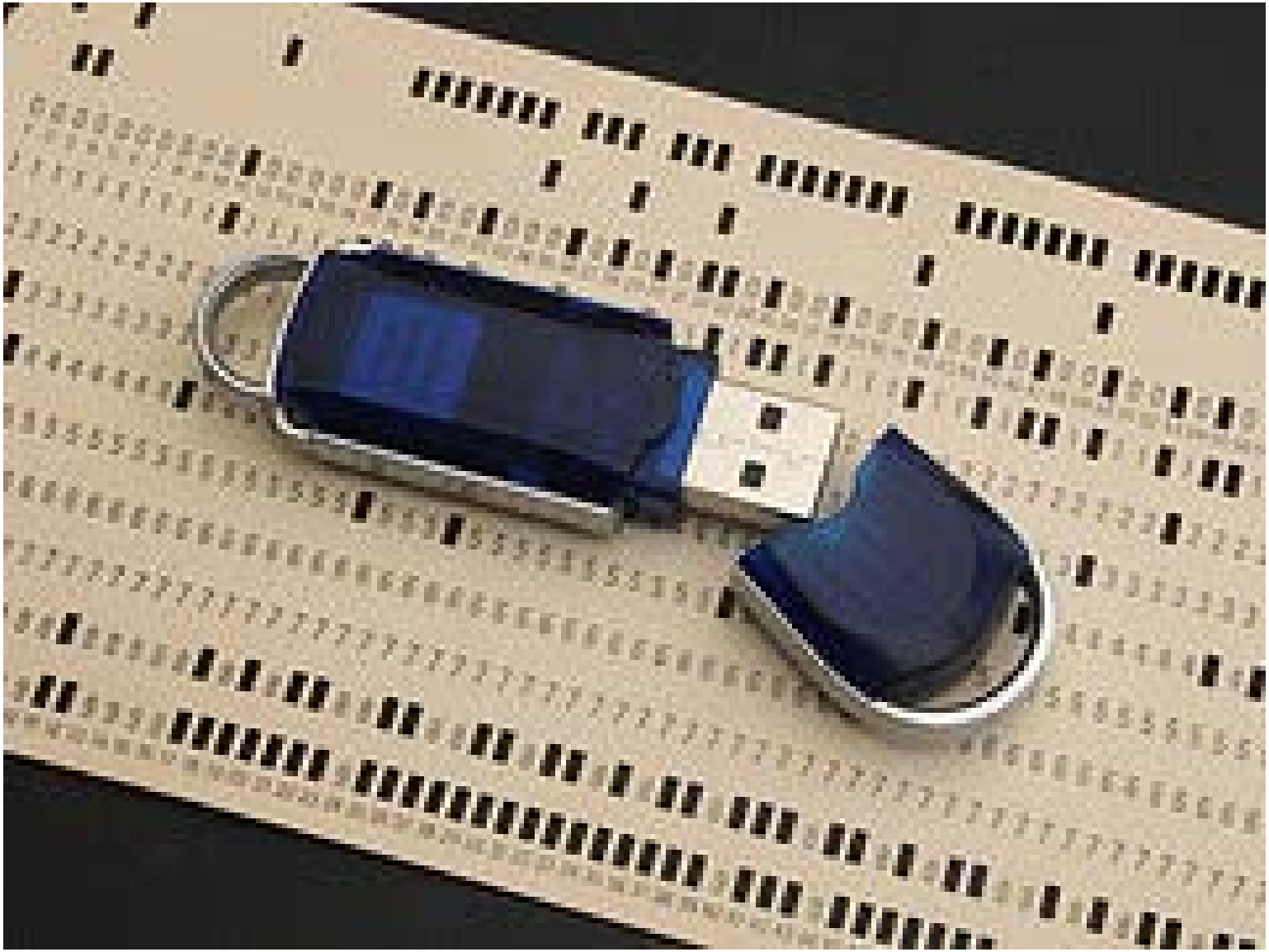
THE GREATEST THREAT

to digital assets is not fire, flood or theft. It's the hazy assumption that cultural heritage institutions have taken the steps needed to preserve them.

Most often, we haven't. Which is why the MetaArchive Cooperative is leading a national effort to embrace distributed digital preservation, the future practice of digitally safeguarding the very items that define our culture and identity.

The problem(s):

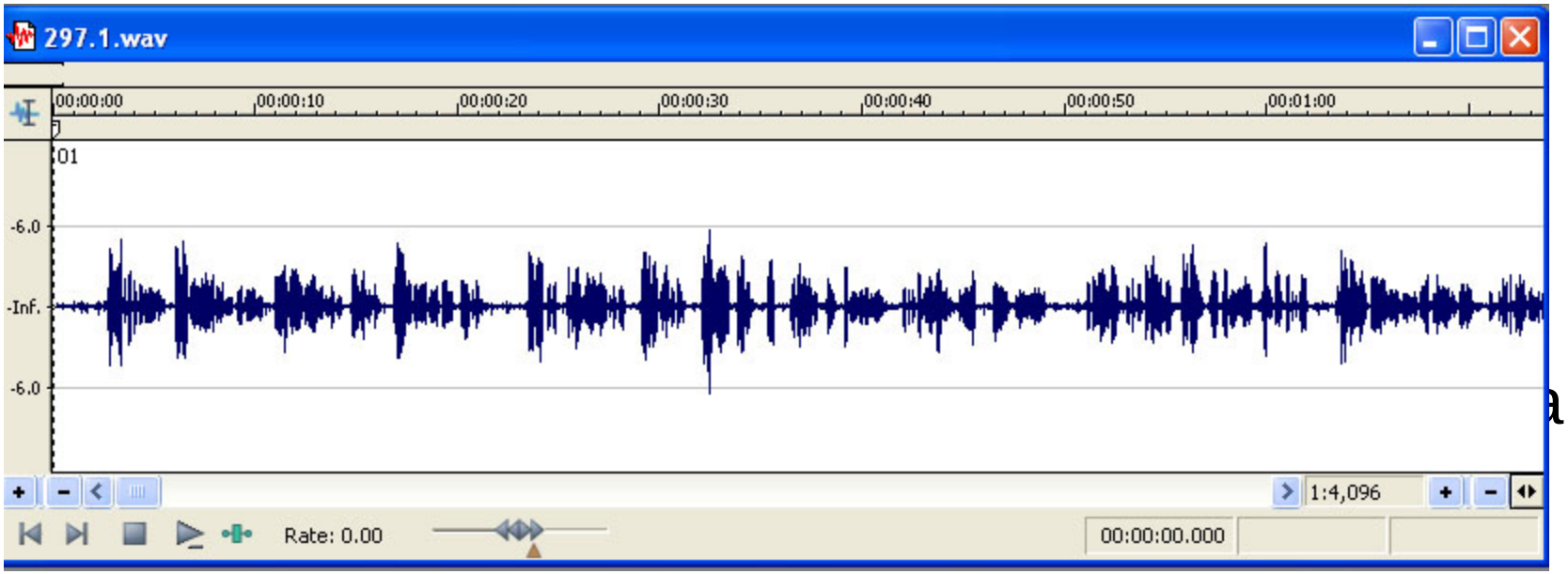
- Digital information is ephemeral.
- Digital information is proliferating.
- Digital preservation requires intention and resources.



“Data storage – old and new.” Made available under Creative Commons 2.0 Attribution License. Available at: http://flickr.com/photos/ian-s/2152798588/in/set-72157602236671297/.

At-risk digital content:

- Web-based projects, exhibitions, and instructional materials with significant content and/or dynamic components.
-
-
-
-
-



Simplest solutions:

- Save files in 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?
- Make multiple 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
- Keep technical and administrative metadata
- Implement a preservation plan

Larger-scale solutions require resources:

- National Digital Information Infrastructure and Preservation Program (NDIIPP):
  - Government funding 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.
  - Overall effort involves more than 100 partners and hundreds of terabytes of data.

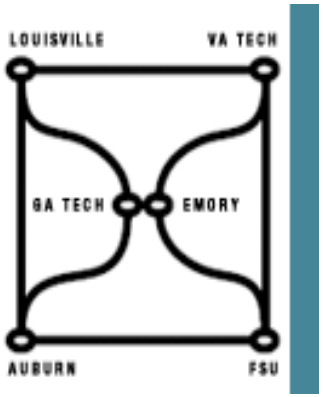
Larger-scale solutions build on working models:

- Lots of Copies Keep Stuff Safe (LOCKSS)
  - Software developed at Stanford University for e-journal preservation
  - 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.
- MetaScholar Initiative
  - Digital library research collaborations led by Emory University

Funding + Open Source Software + Collaboration = MetaArchive

- 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 LOCKSS to collect digital content from each other.
    - Adapting journal concepts (“volumes”) to archival digital materials.

MetaArchive Founding 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)
- [Library of Congress (Washington, DC)]

The Growing Network

- University of Hull (Hull, United Kingdom)
- Boston College (Boston, MA)
- Clemson University (Clemson, SC)
- Rice University (Houston, TX)
- York University (Toronto, Canada)

Private LOCKSS Network (PLN)

- Multiple geographically-dispersed sites host preservation nodes
  - Server is dedicated to collecting materials from every other node, checking to make sure each copy is complete and valid.
  - Participants communicate permission to the LOCKSS system to harvest their materials via a web crawler.
- Disaster recovery
  - A damaged cache can be re-built and re-populated from the identical sets of data at the other nodes.
- Additional modules accommodate non-serialized content
  - Conspectus Database
  - Cache Manager

Documenting collections to harvest: the Conspectus Database

- Database of targeted digital content
  - Cultural heritage of the American South (2004-2007)
  - Format-agnostic
- Includes metadata elements developed specifically for the MetaArchive:
  - Describes the collections
  - Provides information necessary for storage estimates, format migration, location, ownership and rights issues.

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:	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: 1. To whom rights are being issued 2. What rights are specified 3. The resources to which the rights apply 4. Conditions that must be met before rights can be exercised 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 For the MetaArchive project, a controlled list of access categories will be established (Restricted, Unrestricted]
ENCODING SCHEMES:	
OBLIGATION:	Mandatory
DATATYPE:	Character String
MINIMUM OCCUR:	1
MAXIMUM OCCUR:	unbounded
DEVELOPER NOTES:	Two mutually exclusive options (radio buttons): restricted and unrestricted

Preparing items for harvest:

- Define what is to be harvested
  - “Data wrangling”
  - Organize digital files into Archival Units (AUs)
- Provide access to the content and grant permission to harvest
  - Manifest pages (HTML)
- Tell LOCKSS what to harvest and where to find it
  - Plugins (Java)
- Notify partners to harvest new content

Archival Units

- One Volume of an Electronic Journal
- One Year of an ETD collection
- One Year of a scanned yearbook collection
- A folder of archival tif images or sound files

Manifest Pages

The Tin Horn LOCKSS Manifest Page

- 1925
- 1929
- 1930
- 1931

LOCKSS system has permission to collect, preserve, and serve this Archival Unit.

Plugins

- Instruct the LOCKSS software how to crawl and audit content.
- Plugin Repository accessed by LOCKSS box upon startup.

Harvesting collections: the Cache Manager

Journal Configuration

meta-vault.ekstrom.louisville.edu (metaarchive group)

17:25:32 06/05/08, up 4w0d4h

Journal Configuratior
Admin Access Contr
Proxy Access Contro
Proxy Options
Proxy Info
Daemon Status
Contact Us
Help

Add Titles	Add one or more groups of titles
Remove Titles	Remove selected titles
Backup	Backup cache config to a file on your workstation
Restore	Restore cache config from a file on your workstation
Manual Add/Edit	Add, Edit or Delete an individual AU

L O T S O F C O P I E S K E E P S T U F F S A F E ™

Daemon 1.30.3 built 28-Apr-08 09:59:34 on narses2, Linux RPM 1

Overview of preservation caches

caches · collections · archival units · polls · crawl statuses · disks
system info · status info

Caches Overview: 13 total, 13 up, 0 down, 0 error, 0 not monitored

Name	Peer ID/Detai;s	Hostname/Admin	Status	Version	Last refresh
aub	/caches/basic_info/7	rbdadmin.lib.auburn.edu	0w:6d:16h:56m:15s	1.33.4	2008-10-03 05:46:36 UTC
em	/caches/basic_info/8	ndiip.library.emory.edu	1w:1d:14h:19m:30s	1.33.4	2008-10-03 05:47:09 UTC
fsu	/caches/basic_info/9	clockss.lib.fsu.edu	1w:1d:13h:13m:14s	1.33.4	2008-10-03 05:47:33 UTC
gt	/caches/basic_info/10	ndiiplockss.library.gatech.edu	1w:1d:12h:50m:49s	1.33.4	2008-10-03 05:47:55 UTC
lou	/caches/basic_info/11	meta-vault.library.louisville.edu	1w:1d:13h:15m:31s	1.33.4	2008-10-03 05:48:22 UTC
vt	/caches/basic_info/12	metaarchive.lib.vt.edu	1w:1d:15h:38m:6s	1.32.4	2008-10-03 05:48:47 UTC
em-cap	/caches/basic_info/13	metaarchive.library.emory.edu	1w:1d:15h:44m:33s	1.33.4	2008-10-03 05:49:07 UTC
vt-cap	/caches/basic_info/14	metaarchive2.lib.vt.edu	1w:1d:15h:36m:30s	1.32.4	2008-10-03 05:49:28 UTC
hull	/caches/basic_info/15	metaarchive.hull.ac.uk	1w:1d:15h:44m:40s	1.33.4	2008-10-03 05:49:57 UTC
aub-cap	/caches/basic_info/16	metaarchive.lib.auburn.edu	0w:1d:12h:35m:11s	1.33.4	2008-10-03 05:50:21 UTC
fsu-cap	/caches/basic_info/17	clockss2.lib.fsu.edu	1w:1d:13h:27m:56s	1.33.4	2008-10-03 05:50:56 UTC
gt-cap	/caches/basic_info/18	ndiiplockss2.library.gatech.edu	1w:1d:15h:50m:38s	1.33.3	2008-10-03 05:51:49 UTC
lou-cap	/caches/basic_info/19	a70782.library.louisville.edu	0w:2d:12h:44m:40s	1.33.4	2008-10-03 05:52:12 UTC

Collaboration requires communication

- Committees:
  - Steering
  - Content
  - Preservation
  - Technical
- Communications:
  - Conference calls (1/week)
  - Steering Committee meetings (2/year)
  - Listserv(s)
  - Wiki for document development
  - Participation in NDIIPP meetings

Sustaining and growing the collaboration

- Flexible organizational model
  - Charter broadly defines mission, goals, and activities of the Cooperative
  - Membership Agreement details responsibilities of members of Cooperative
  - Establishment of nonprofit organization, MetaArchive Services Group, to administer Cooperative
    - Minimal overhead.
- Improving and expanding existing collaboration
  - Evolving standards and guidelines to offer as a model for new networks and collaborations
  - Enhancing technology, tools, and services
  - Wide applicability to a range of institutions and digital content
- Spreading the word
  - Outreach to libraries, archives, and museums
- Ongoing exploration of projects to investigate and advance digital preservation.

Membership types and fees

- All membership types presuppose membership in the LOCKSS Alliance (rates based on Carnegie classification) and a 3-year commitment
- Sustaining Members
  - Leadership role
  - Operate a node
  - Contribute more content/year to be harvested
  - Cost: $5K/year or $12K/3 years
- Preservation Members
  - Operate a node
  - Contribute some content/year to be harvested
  - Cost: $1K/year
- Contributing Members
  - Contribute minimal content/year to be harvested (can buy more space)
  - Cost: $300/year

Further reading

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