MetaArchive at 10: growing, diversifying, innovating

Matt Schultz, Rachel Howard & Deanna Ulvestad
2014 LITA Forum, Nov. 7th, 2014
MetaArchive in a Nutshell

✓ Distributed Digital Preservation (DDP)
✓ Member-Hosted & Member-Driven
✓ Community of Practice & Exchange
✓ Secure Dark Archive
✓ Bit-Level Preservation
✓ Disaster Recovery
✓ Scalable Low-Cost Hardware
✓ Open Source Software
2004

In the beginning...

- 6 regional partners
- Grant-funded
- Preserved Southern digital culture
- Archival units < 1 GB

Bud McCoy with family and dog, Kentucky, ca. 1930. Item number ULPA 1979.33.0026 in the Jean Thomas, The Traipsin’ Woman, Collection, University of Louisville Photographic Archives, Louisville, Ky. [Available online](#).
2014

Growth and evolution

- 50+ international partners
- Self-sustaining and governing
- Preserving varied digital assets
- Evolving standards
Harvesting... and other terms of art

- Data wrangler / Data wrangling
- Archival unit
- Manifest page
- Plugin
- Conspectus

Cooperative community

Mr. Thomas B. McCrocklin. Photo of group, 1944. Royal Photo Co. (photographer). Item number ULPA R 07917 in the Royal Photo Company Collection, University of Louisville Photographic Archives, Louisville, Ky. Available online.
Many hands make light work

Greene County Public Library was housed in the Carnegie building from 1906 – 1978. Xenia, Ohio.
Getting Started

November 2010
Attended 5-day workshop
“Digital Preservation Management”
University of Michigan

August 2011
Started to compare “Dark Archive” repositories/vendors

April 2012
Joined MetaArchive as a Preservation Member

January 2013
Started ingesting collections

September 2013
Serve on the Outreach Committee

March 2014
Serve on the Metadata Working Group

First bookmobile used by the Greene County Public Library from 1948 – 1958. Xenia, Ohio.
Ingest Prep...... Building a Digital Time Capsule

Gathering
Collection & Metadata Files

Naming
Descriptive File & Folder Labels

Organizing
Top Level & Sub Folders / Archival Units

Greene County Courthouse Time Capsule of 1901 opened in 2001. Xenia, Ohio.
Ingest Ready

Bellbrook_Moon_Newspaper_Scrapbook_Collection

BellbrookMoon_scrapbooks_300dpi

- BellbrookMoon_scrapbook_volume1
  - BellbrookMoon_scrapbook_volume1_page001
  - BellbrookMoon_scrapbook_volume1_page002

- BellbrookMoon_scrapbook_volume2
- BellbrookMoon_scrapbook_volume3
- BellbrookMoon_scrapbook_volume4
Ingest Addition

Bellbrook_Moon_Newspaper_Scrapbook_Collection

BellbrookMoon_scrapbooks_300dpi

BellbrookMoon_scrapbook_volume1
  BellbrookMoon_scrapbook_volume1_page001
  BellbrookMoon_scrapbook_volume1_page002

BellbrookMoon_scrapbook_volume2
BellbrookMoon_scrapbook_volume3
BellbrookMoon_scrapbook_volume4

BellbrookMoon_scrapebook_metadata_files
  BellbrookMoon_Everyname_Index_database
  BellbrookMoon_Everyname_Index_database_text_file
  BellbrookMoon_PageLevel_Name_files
  BellbrookMoon_CONTENTdm_XML_file
  BellbrookMoon_Project_Guide_document
Modified Ingest

- GCPL CONTENT dm Server
- MetaArchive Server
  - Distributes GCPL Archive Units
  - Duplicates

- MetaArchive Server
  - Distributes GCPL Archive Units

- MetaArchive Server
  - Distributes GCPL Archive Units
How-It-Works - Technology
How-It-Works - Technology

- Heterogeneous Hardware
  - Dell
  - IBM
  - iXSystems
  - eRacks
- Low-Cost Options
  - $4,500 or less
- Three-Year Refresh Cycles
- Archive Nodes (LOCKSS)
How-It-Works - Technology

- AWS Central Admin Servers
  - EC2
  - S3
- Network Configs
- Monitoring Services
- Websites & Wikis
- Development
- No Data Storage!
• LOCKSS harvests data via the Web
How-It-Works - Technology

- Interoperates with many/most repository software systems
• Works just great w/simple web servers
How-It-Works - Technology

• Files packaged w/BagIt especially easy! (more on this later)
• LOCKSS requires a simple permission page (html)
• And a collection “plugin” (xml) that tells LOCKSS what to include/exclude
• Collections are retrieved via HTTP and WGET
Collections are retrieved via HTTP and WGET
How-It-Works - Technology

• Collections are retrieved via HTTP and WGET – replicated 7 times
• Then what?
How-It-Works - Technology

- LOCKSS goes to work...
• LOCKSS goes to work...
• Comparing replications...
• Do we have the same bits?
How-It-Works - Technology

• If so great!!!!
• Let’s keep it up!!!
How-It-Works - Technology
• Hey look at that!!! Looking good!!!
• Wait what? You mean there’s a broken file!?!
How-It-Works - Technology

- Now what?
How-It-Works - Technology

• Let’s fix that!
• That ought to do it!
How-It-Works - Technology
How-It-Works - Technology

Archives are groups of collections that belong to a certain topic.

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Title</th>
<th># of Collections</th>
</tr>
</thead>
<tbody>
<tr>
<td>ETD</td>
<td>Electronic Theses and Dissertations</td>
<td>17</td>
</tr>
<tr>
<td>SDC</td>
<td>Southern Digital Culture Collections</td>
<td>137</td>
</tr>
<tr>
<td>EML</td>
<td>Early Modern Literature</td>
<td>3</td>
</tr>
<tr>
<td>GEN</td>
<td>Content that does not fit into any other archive.</td>
<td>94</td>
</tr>
<tr>
<td>NEWS</td>
<td>Newspapers</td>
<td>5</td>
</tr>
<tr>
<td>IH</td>
<td>Institutional History</td>
<td>2</td>
</tr>
<tr>
<td>COMM</td>
<td>Community Generated Content</td>
<td>7</td>
</tr>
</tbody>
</table>

http://conspectus.metaarchive.org
How-It-Works - Technology
# How-It-Works - Technology

## Archival Unit Details

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collection</td>
<td>HBCU Library Alliance: A Digital Collection Celebrating the Founding of the Historically Black Colleges and University</td>
</tr>
<tr>
<td>Plugin</td>
<td>edu.hbcu.AllFromStart2 by Historically Black Colleges &amp; Universities</td>
</tr>
<tr>
<td>LOCKSS AUD</td>
<td>cek(hbcuAllStart&amp;base_url=<a href="https://3A%2F%2F%2F144.3.125.36%2F%3E111%2Fsplit&amp;start=AUC_0">https://3A%2F%2F%2F144.3.125.36%2F%3E111%2Fsplit&amp;start=AUC_0</a></td>
</tr>
<tr>
<td>LOCKSS Title DB</td>
<td>hbcu: HBCU Library Alliance: A Digital Collection Celebrating the Founding of the Historically Black College and University: start=AUC_0</td>
</tr>
<tr>
<td>Status</td>
<td>Preserved</td>
</tr>
<tr>
<td>Availability</td>
<td>Site Up</td>
</tr>
<tr>
<td>Parameter Values</td>
<td>start</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cache Size</th>
<th>Disk Usage</th>
<th>Last Crawl</th>
<th>Crawl Result</th>
<th>Last Poll</th>
<th>Agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>lmu-erack</td>
<td>0</td>
<td>None</td>
<td>Jan. 18, 2014, 9:02 p.m.</td>
<td>Can’t fetch permission page</td>
<td>Feb. 15, 2014, 1:30 p.m.</td>
</tr>
<tr>
<td>cmu</td>
<td>0</td>
<td>None</td>
<td>Jan. 14, 2014, 5:27 p.m.</td>
<td>Can’t fetch permission page</td>
<td>None</td>
</tr>
<tr>
<td>bo-lbm</td>
<td>0</td>
<td>None</td>
<td>May 30, 2014, 1:36 p.m.</td>
<td>Successful</td>
<td>May 16, 2014, 5:06 p.m.</td>
</tr>
<tr>
<td>cduc</td>
<td>0</td>
<td>None</td>
<td>March 27, 2014, 6:01 p.m.</td>
<td>Successful</td>
<td>May 5, 2014, 9:09 a.m.</td>
</tr>
<tr>
<td>esu-erack</td>
<td>0</td>
<td>None</td>
<td>Oct. 10, 2013, 9:17 p.m.</td>
<td>Can’t fetch permission page</td>
<td>Jan. 18, 2014, 6:08 a.m.</td>
</tr>
<tr>
<td>Hbcu-erack</td>
<td>0</td>
<td>None</td>
<td>March 27, 2014, 8:13 p.m.</td>
<td>Successful</td>
<td>March 24, 2014, 12:03 p.m.</td>
</tr>
<tr>
<td>unt</td>
<td>0</td>
<td>None</td>
<td>May 30, 2014, 11:23 a.m.</td>
<td>Successful</td>
<td>None</td>
</tr>
</tbody>
</table>
# How-It-Works - Technology

## Index of /chronicles/GTWhistle

<table>
<thead>
<tr>
<th>Name</th>
<th>Last modified</th>
<th>Size</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>bag-info.txt</td>
<td>27-Feb-2013 13:15</td>
<td>593</td>
<td></td>
</tr>
<tr>
<td>bagit.txt</td>
<td>27-Feb-2013 13:11</td>
<td>55</td>
<td></td>
</tr>
<tr>
<td>data/</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>manifest-md5.txt</td>
<td>27-Feb-2013 13:11</td>
<td>148K</td>
<td></td>
</tr>
</tbody>
</table>

Bag-Software-Agent: bagit.py <http://github.com/edu/bagit>
Bagging-Date: 2013-02-27
Payload-Oxum: 523630395.2743
Contact-Email: jody.thompson@library.gatech.edu
Contact-Name: Jody Thompson
Contact-Phone: 404-894-9626
External-Identifier: SMARTech:/1053/3790
Source-Organization: Georgia Tech Library
Organization-Address: 704 Cherry St, Atlanta, GA 30332-0900
External-Description: Faculty newspaper issues submitted by the Georgia Tech Library. Files were delivered as born digitally and scanned from print. License information and extracted text information is included.
Bag-Size: 526M

BagIt-Version: 0.96
Tag-File-Character-Encoding: UTF-8

https://github.com/metaarchive
How-It-Works - Technology

Use the options below to request an audit on this archival unit. When the process completes, a new Report will be created.

Bag Checksum Audit

Descriptive text here

Cache

-----

The LOCKSS cache on which to generate hashes.

BagIt Manifest

[Browse...] No file selected.

Upload a BagIt manifest for this AU. Must be named either manifest-md5.txt or manifest-sha1.txt. The hashing algorithm is detected based on the filename.

Request Audit
How-It-Works - Technology

Audit Request Successful

Your request for an audit has been added to the queue.

When the audit completes, a report containing the results will appear on the Reports page and you should receive an email notification at the address listed on your account [matt.schulte@metaarchive.org].

Return to previous page
How-It-Works - Technology

Reports

This page lists reports generated by various processes within the application.

<table>
<thead>
<tr>
<th>Created On</th>
<th>Requested By</th>
<th>Related To</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sept. 3, 2014, 4:04 a.m.</td>
<td>mats</td>
<td>start = AUC_2</td>
</tr>
<tr>
<td>Sept. 4, 2014, 5:01 a.m.</td>
<td>mats</td>
<td>start = AUC_0</td>
</tr>
</tbody>
</table>
# How-It-Works - Technology

<table>
<thead>
<tr>
<th>PREMIS Event Service</th>
<th>Events</th>
<th>Search</th>
<th>Agents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Event Identifier</td>
<td>5b6b023424784ce4b273c5a532422060</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Identifier Type</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Event Type</td>
<td>auditArchiveUnitHashes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Event Date</td>
<td>May 5, 2014, 8:29 p.m.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Event Details</td>
<td>User mats requested a LOCKSS HashCUS bag manifest comparison audit.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Date Added</td>
<td>May 5, 2014, 8:29 p.m.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linking Agent Identifier Value</td>
<td>conspectus</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linking Agent Identifier Type</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linking Agent Role</td>
<td>success</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Outcome</td>
<td>Task compared 20059 records, encountered 0 errors</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Outcome Detail</td>
<td>Task compared 20059 records, encountered 0 errors</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linked Objects</td>
<td>org/metaarchiveAJFromStart&amp;base_url=http%3A%2F%2Flibmetaarchive%2Elibrary%2Eedu%2Fchronicles%2Fmetaarchive-staging%2Futah&amp;start=ndnp_bag_01_6</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

[ATOMPUB](#)
How-It-Works - Technology

<table>
<thead>
<tr>
<th>Event Identifier</th>
<th>c29433565ee545a79bcd26601ab59fe</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identifier Type</td>
<td>⚫</td>
</tr>
<tr>
<td>Event Type</td>
<td>📁 auditArchivalUnitHashes</td>
</tr>
<tr>
<td>Event Date</td>
<td>📆 May 3, 2014, 12:17 a.m.</td>
</tr>
<tr>
<td>Event Details</td>
<td>🍂 User admin requested a LOCKSS HashCUS bag manifest comparison audit.</td>
</tr>
<tr>
<td>Date Added</td>
<td>🍂 May 3, 2014, 12:17 a.m.</td>
</tr>
<tr>
<td>Linking Agent Identifier Value</td>
<td>⚫ econspectus</td>
</tr>
<tr>
<td>Linking Agent Identifier Type</td>
<td>⚫</td>
</tr>
<tr>
<td>Linking Agent Role</td>
<td>⚫ failure</td>
</tr>
<tr>
<td>Outcome</td>
<td>⚫ failure</td>
</tr>
<tr>
<td>Outcome Detail</td>
<td>⚫ Task compared 6 records, encountered 6 errors</td>
</tr>
</tbody>
</table>
Preserving Diverse Digital Collections
- Electronic Theses & Dissertations (ETDs)
- Digital Newspapers
- Institutional Histories
- Early Modern Literature
- Community-Generated Content
- Research Data
- Digital Audio/Video
MetaArchive Governance

**Steering Committee**
The Steering Committee is the governing body of the MetaArchive Cooperative. It is responsible for overall mgmt, coordination, communication, and reporting efforts.

**Sub-Committees**
- Preservation Committee
- Content Committee
- Outreach Committee
- Technical Committee

**Working Groups**
- Metadata

**Capacity Development Process (September-December, 2014)**
- Cooperative has grown and diversified
- Documenting roles, responsibilities & relationships between Committees, Working Groups & Central Staff
MetaArchive Cooperative

- Started w/ six Southern U.S. institutions focused on digitized special collections
- Now an international network working w/ 50 + Members
  - Research libraries, public libraries, research centers, foundations
- Developing robust digital preservation technologies and methods
- Preserving all manner of digital collections
- Helping libraries maintain control & ownership
Thank You!

Questions?
MetaArchive Cooperative

Matt Schultz
Program Manager
MetaArchive Cooperative
matt.schultz@metaarchive.org
@metaarchive