

The MetaArchive Cooperative

Collaboratively Preserving Our Digital Memory

Katherine Skinner, PhD
Executive Director, Educopia Institute

January 8, 2014



Overview

- History and Current Practices
- Philosophy
- Research
- Practical Application



MetaArchive History

- Distributed digital preservation cooperative
- Founded 2004
- Supported by NDIIPP, NHPRC, NEH, IMLS
- Preservation aims: prevent loss and corruption from human malice/error or from a disaster
- First (known) preservation network to preserve special collections/unique materials

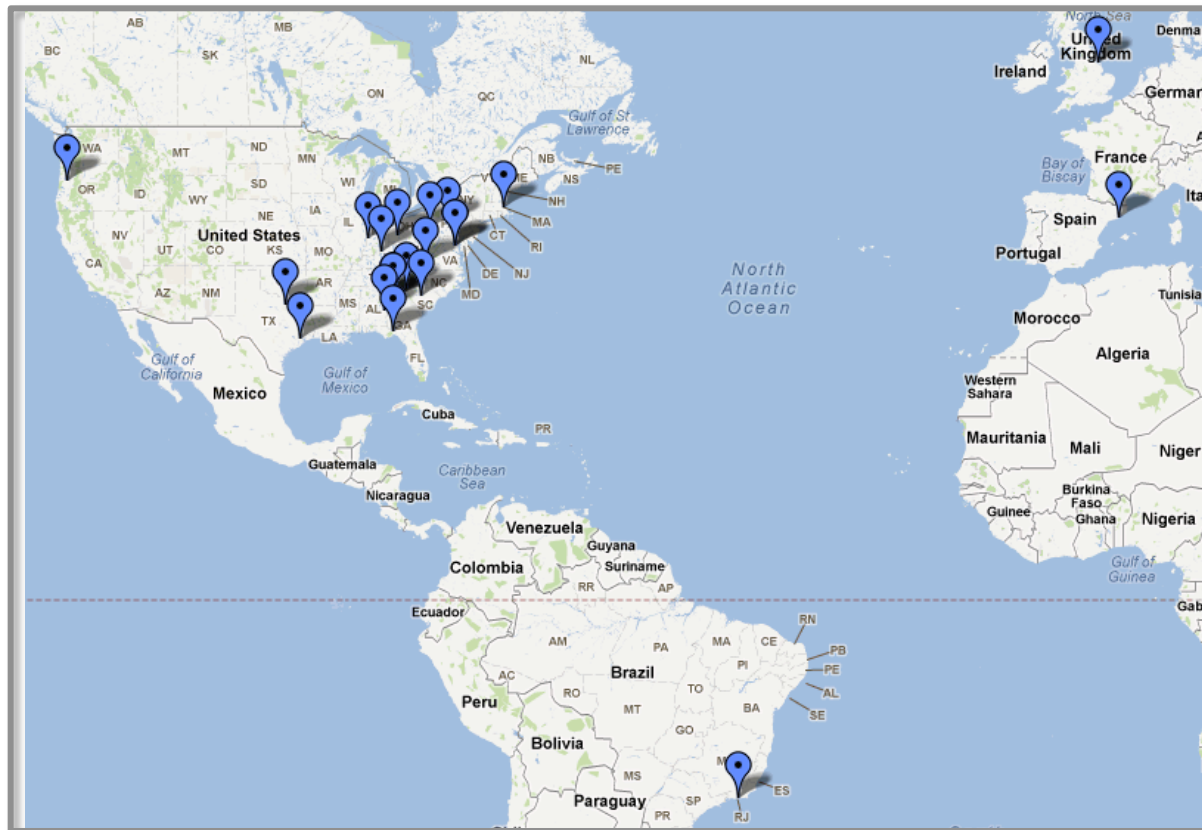
MetaArchive Practices

- Basic processes
 - “Producer” (OAIS) determines curation practices; brings SIPs to MetaArchive (low entry barriers)
 - Multiple copies of AIPs dispersed across geographical, political, and environmental lines
 - Checks and repairs automated across network
 - Supports automated versioning
 - Deaccession cycle versus data deletion

MetaArchive Practices

- More than 50 members in four countries
- Public, private, and academic libraries
- 20 geographically distributed nodes
- 200+ TB network preserving 150+ collections
- Ingest via repository harvest or BagIt
- Regular testing of restoration processes
- TRAC self-audit 2010; will pursue certification under ISO 16363 once there are certified auditors (ISO 16919 not yet approved)

Membership



50 members, 14 states/districts, 4 countries

Skinner 2014

Auburn University
Boston College
Carnegie Mellon University
Clemson University
Consorti de Biblioteques
Universitat de Catalunya
Florida State University
Folger Shakespeare Library
GALILEO Knowledge Repository
Greene County Public Library
HBCU Library Alliance
Indiana State University
Library of Congress
Oregon State University
Penn State University
Pontifícia Universidade Católica do
Rio de Janeiro
Purdue University
Rice University
University of Hull
University of Louisville
University of North Texas
University of South Carolina
University of Tennessee
Virginia Tech

MetaArchive Pricing

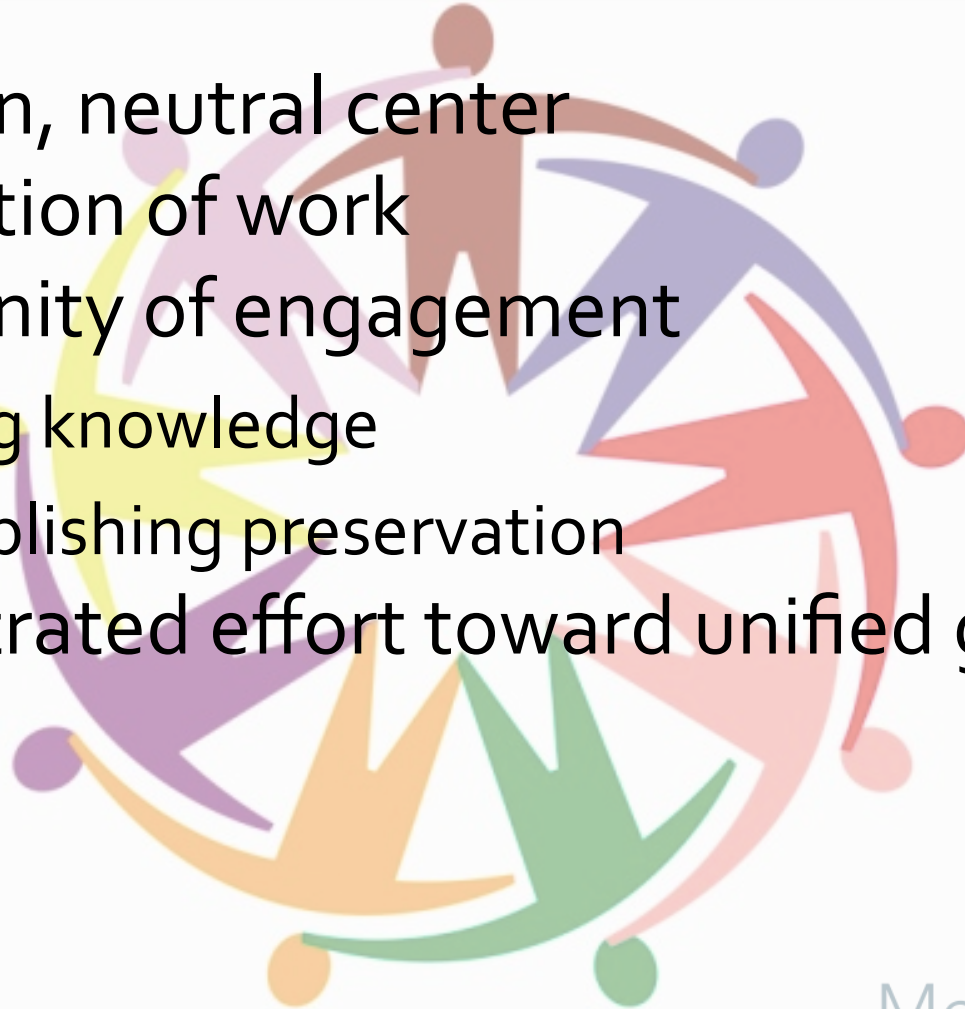
- Three membership levels
 - Collaborative members: \$2.5K/year
 - Preservation members: \$3K/year
 - Sustaining members: \$5.5K/year
- Server cost: <\$5K/term
- Storage cost: \$1K/TB/year

Member Responsibilities

- Undertake a 3-year membership term;
- Take responsibility for content preparation, evaluation, staging, and ingest testing;
- Monitor collections to ensure accurate long-term preservation;
- Host and maintain a MetaArchive cache (server)

MetaArchive Philosophy

- Common, neutral center
- Distribution of work
- Community of engagement
 - Building knowledge
 - Accomplishing preservation
- Concentrated effort toward unified goal



Conspectus home

Conspectus

HelpMy Account (matts)Log Out

Navigation

Home

Content Providers

New Content Provider

Archives

New Archive

Collections

My Collections

New Collection

Plugins

My Plugins

New Plugin

Templates

Users

New User

Reports

Advanced Options

Archives

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

+ New Archive

Acronym	Title	# of Collections
ETD	Electronic Theses and Dissertations	16
SDC	Southern Digital Culture Collections	133
EML	Early Modern Literature	3
GEN	Content that does not fit into any other archive.	58
NEWS	Newspapers	4
IH	Institutional History	2
COMM	Community Generated Content	8

Collection page

[Conspectus](#)[Help](#)[My Account \(matts\)](#)[Log Out](#)

Navigation

[Home](#)[Content Providers](#)

[New Content Provider](#)

[Archives](#)

[New Archive](#)

[Collections](#)

[My Collections](#)[New Collection](#)

[Plugins](#)

[My Plugins](#)[New Plugin](#)

[Templates](#)[Users](#)

[New User](#)

[Reports](#)[Advanced Options](#)

Collection Details

[✎ Edit Collection](#)

Title	Virginia Tech VA-news
Archive	GEN
Plugin	org.metaarchive.AllFromStart2 by Metaarchive
Base URL	http://scholar.lib.vt.edu/Chronicles_ingest
Identifier	238
Description	The Virginia news collection at Virginia Tech. The collection includes the Roanoke Times from January 1992 to February 1997, the Virginian Pilot from 1994 to 1997, and WDBJ7 news transcripts from October 1995 to August 2008
Repository Type	Not specified

Related Items

[Metadata](#)[Plugin XML](#)[Archival Units 1](#)

This collection has no metadata.

Adding Collections

Navigation

- Home
- Content Providers
 - New Content Provider
- Archives
 - New Archive
- Collections
 - My Collections
 - New Collection
- Plugins
 - My Plugins
 - New Plugin
- Templates
- Users
 - New User
- Reports
- Advanced Options

New Collection

Title

The title of the collection, e.g. 'John Hilton Papers'

Description

(Optional) Some text describing the collection's contents

Base url

The base URL to be passed to the LOCKSS plugin for crawling, e.g. <http://your-institution.edu/library/collection/>

Plugin

(Optional) The LOCKSS plugin that will be used to crawl this plugin

Archive

The archive this collection should be categorized under

Repository type

(Optional) What kind of repository this collection is hosted on

Versioning type

(Optional) How this collection's content should be revisited by LOCKSS

Plug-In pathway

[Conspectus](#)[Help](#)[My Account \(matts\)](#)[Log Out](#)

Navigation

[Home](#)[Content Providers](#)

[New Content Provider](#)

[Archives](#)

[New Archive](#)

[Collections](#)

[My Collections](#)[New Collection](#)

[Plugins](#)

[My Plugins](#)[New Plugin](#)

[Templates](#)[Users](#)

[New User](#)

[Reports](#)[Advanced Options](#)

New Plugin

Choose one of the plugin creation options below.

Start from a Sample

Plugin Identifier:

The plugin identifier as seen by LOCKSS. The prefix shown will be prepended automatically, if applicable.

Plugin Template

☐ All from Start

Crawls everything from a given start path, blah blah blah.

Content Provider

The plugin will belong to the selected content provider.

Continue

- or -

Upload a Custom Plugin

Plugin XML Upload

Browse...

No file selected.

Upload your LOCKSS plugin XML file here. The plugin identifier will be extracted from the uploaded file.

Content Provider

The plugin will belong to the selected content provider.

Continue

Plug in Wizard

[Conspectus](#)[Help](#)[My Account \(matts\)](#)[Log Out](#)

Navigation

[Home](#)[Content Providers](#)

[New Content Provider](#)

[Archives](#)

[New Archive](#)

[Collections](#)

[My Collections](#)[New Collection](#)

[Plugins](#)

[My Plugins](#)[New Plugin](#)

[Templates](#)[Users](#)

[New User](#)

[Reports](#)[Advanced Options](#)

Plugin Template Details

Edit Template

Name	All from Start
Description	Crawls everything from a given start path, blah blah blah blah.

```
<map>
  <entry>
    <string>plugin_config_props</string>
    <list>
      <org.lockss.daemon.ConfigParamDescr>
        <key>start</key>
        <displayName>Start</displayName>
        <type>1</type>
        <size>20</size>
        <definitional>true</definitional>
        <defaultOnly>false</defaultOnly>
      </org.lockss.daemon.ConfigParamDescr>
      <org.lockss.daemon.ConfigParamDescr>
        <key>base_url</key>
        <displayName>Base URL</displayName>
        <description>Usually of the form http://&lt;journal-name&gt;.com</description>
        <type>3</type>
        <size>40</size>
        <definitional>true</definitional>
        <defaultOnly>false</defaultOnly>
      </org.lockss.daemon.ConfigParamDescr>
      <org.lockss.daemon.ConfigParamDescr>
        <key>path</key>
        <displayName>Path</displayName>
        <type>1</type>
        <size>60</size>
        <definitional>false</definitional>
        <defaultOnly>false</defaultOnly>
      </org.lockss.daemon.ConfigParamDescr>
    </list>
  </entry>
  <entry>
    <string>plugin_version</string>
```

Auditing Control

Navigation

[Home](#)
[Content Providers](#)
 [New Content Provider](#)
[Archives](#)
 [New Archive](#)
[Collections](#)
 [My Collections](#)
 [New Collection](#)
[Plugins](#)
 [My Plugins](#)
 [New Plugin](#)
[Templates](#)
[Users](#)
 [New User](#)
[Reports](#)
[Advanced Options](#)

Archival Unit Details

Collection	Virginia Tech VA-news				
Plugin	org.metaarchive.AllFromStart2 by Metaarchive				
Status	Test				
Up/Down	Site Up				
Parameter Values	<table><tr><td>base_url</td><td>http://scholar.lib.vt.edu/Chronicles_Ingest</td></tr><tr><td>start</td><td>VA-news_bag</td></tr></table>	base_url	http://scholar.lib.vt.edu/Chronicles_Ingest	start	VA-news_bag
base_url	http://scholar.lib.vt.edu/Chronicles_Ingest				
start	VA-news_bag				

Preservation Status

Coming soon.

Auditing Tools

Use the form below to request an audit on this archival unit.

When the process completes, a new [Report](#) will be created.

Audit Type

The type of audit.

Cache

The cache to use.

Begin Audit

Research: Practical Application

- NEH & IMLS
 - BagIt Ingest Implementations
 - Bag Splitting & Rebuilding Support
 - Bag Auditing Tools
 - Change Management Support for Bags
 - Curation Micro-Services Ingest Implementations
 - Format Recognition
 - Virus Checking
 - PREMIS Metadata Creation & Management

What is BagIt?

- BagIt is a packaging specification for validating digital content exchanges between 2 or more parties (some people also use it to store content)
 - Not a ZIP or a TAR package - easier to peek inside
 - Creates a manifest (listed inventory) of folders/files and corresponding checksums (md5 or sha1)
 - Contains a data directory of all included folders/files
 - Has an optional metadata file
- There are several open source BagIt tools that will package & validate sent packages

BagIt Structure

Index of /chronicles/GTWhistle

<u>Name</u>	<u>Last modified</u>	<u>Size</u>	<u>Description</u>
 Parent Directory		-	
 bag-info.txt	27-Feb-2013 13:15	593	
 bagit.txt	27-Feb-2013 13:11	55	
 data/	27-Feb-2013 13:11	-	
 manifest-md5.txt	27-Feb-2013 13:11	148K	

BagIt Metadata

```
Bag-Software-Agent: bagit.py <http://github.com/edsu/bagit>
Bagging-Date: 2013-02-27
Payload-Oxum: 523930395.2743
Contact-Email: jody.thompson@library.gatech.edu
Contact-Name: Jody Thompson
Contact-Phone: 404-894-9626
External-Identifier: SMARTech:/1853/3790
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
```

bag-info.txt

BagIt Version & Encoding

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

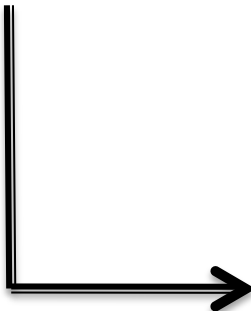
bagit.txt

Baglt Data Directory

 [data/](#)

27-Feb-2013 13:11 -

Index of /chronicles/GTWhistle/data



<u>Name</u>	<u>Last modified</u>	<u>Size</u>	<u>Description</u>
 Parent Directory		-	
 1/	13-Feb-2013 15:36	-	
 10/	13-Feb-2013 15:36	-	
 100/	13-Feb-2013 15:36	-	
 101/	13-Feb-2013 15:36	-	
 102/	13-Feb-2013 15:36	-	
 103/	13-Feb-2013 15:36	-	
 104/	13-Feb-2013 15:36	-	
 105/	13-Feb-2013 15:36	-	
 106/	13-Feb-2013 15:36	-	
 107/	13-Feb-2013 15:36	-	
 108/	13-Feb-2013 15:36	-	
 109/	13-Feb-2013 15:36	-	
 11/	13-Feb-2013 15:36	-	

BagIt Manifest

```
f4fec6201ffdf9be206a55266cc9ae3a data/113
3ad1c735e407411d093f143fc095bed7 data/11
7d9461e55090d989651a4285d064518e data/116
b4f2fa21b3f0bb9e340f8c895be23413 data/114
cd8a45cf835cf367b48f737fc349e2e7 data/112
019c88e86ed59b31561b545594d6a2ef data/115
79fc9baa157f37275795c74f9a7b011a data/630/handle
0fa7b3722cb0e9476be05a4fc136216b data/630/mar10.pdf.txt
423d24b8e01b67cdd09e484b3d9de795 data/630/mar10.pdf
d48e0e1533582023ef65f917089f1343 data/630/license.txt
7f7f9e43306d40e994c6a8af5366918 data/630/contents
289f74ca90c3d2aa2456cdc6d4b6919a data/630/dublin_core.xml
2b24206e8ce8a876419c58ddfc6fa6d5 data/626/handle
1f5f5c56e0501b240e0f081a943707eb data/626/apr7.pdf.txt
423db70fdb75ff11ea4255a3764574a9 data/626/license.txt
c2660aac89ca8728f5fae9593e53c511 data/626/contents
4c0939c2c7cee5ed930899488f34d82d data/626/apr7.pdf
da1292b5e6360d7f494f8583cd3f7433 data/626/dublin_core.xml
29598c1b8973a5889ae055c857d8087a data/568/handle
014ecf88ee2596e49e580f4e8008bb10 data/568/license.txt
da3c6107cb5b6193fa0a48edd929430b data/568/contents
6ca17667e5801d6e8151c775aa9fe320 data/568/dec9.pdf
01dd5ce76312514675c913613a148d0f data/568/dec9.pdf.txt
28d6b2aa45067ad072acf6d491c91ff6 data/568/dublin_core.xml
53555ddc9387e3887963ad194ea02df3 data/665/handle
03d832f6d46c0a094b896e39fbc64f8 data/665/license.txt
7285f923b266d723f3466bd87e86bb11 data/665/contents
d2e6975759a4d4e0701f85b7ale93fb2 data/665/may20.pdf
2a2e024943a572fed38f190b8f0582ec data/665/dublin_core.xml
41f5191beb5ae92b18c75807d4e65324 data/665/may20.pdf.txt
e48f754347ea5c8e80e9d2434364d4a3 data/740/handle
53c8c5995calba906a367b3ec7038894 data/740/license.txt
95e8c438a3759ebd758960bf9f325a8c data/740/contents
95e66705b34c8dd7ea8d4dc88fbfe64f data/740/dublin_core.xml
c211ecdbd6aeb6ddfa318a5501d12604 data/740/whistle_apr21.pdf.txt
ae50309d4e7ed70566a69cecc25a4c18 data/740/whistle_apr21.pdf
5dafbecb0413e5e16fec2500813e8002 data/29/handle
722ea69d344ea5e5416e2f8064e9afc5 data/29/license.txt
76c0df9ef82e9a69ce59f7c40e30c12f data/29/contents
```

manifest-md5.txt

How Do You Build Bags?

- Several Bagit utilities that will “bag” content
– below are the most popular



University of California
CDL
California Digital Library

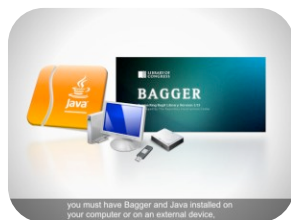
-Java Bag Library (Library of Congress/
CDL)

<http://sourceforge.net/projects/loc-xferutils/files/loc-bil-java-library/3.6/>



-Bagit.py (Ed Summers Python Tools)

<https://github.com/edsu/bagit>



-Bagger-GUI (Library of Congress/CDL)

<http://sourceforge.net/projects/loc-xferutils/files/loc-bagger/>

MetaArchive BagIt Research

- NHPRC Project (2007-2011)
 - Exported 256 AUs (1.4 TB) into “bags” and successfully transferred to Chronopolis
 - Created workflows & scripts for packaging
- NEH Chronicles in Preservation (2011-2013)
 - Ingesting digital newspaper collections as “bags”
 - Bags have no size limitations (can = multiple TBs)
 - Creating scripts to generate smaller “bags” from larger “bags” for MetaArchive Archival Units (AUs)

Benefits of Using BagIt

- Pre-Ingest File Fixity & Inventory
- Improved Post-Ingest Auditing
- Simplified Ingests
 - Out-of-the-box plugin
 - Simple manifest page
 - No need to worry about Archival Units upfront
- Maintains Collection Integrity

MetaArchive Bag Creation

DURASPACE™

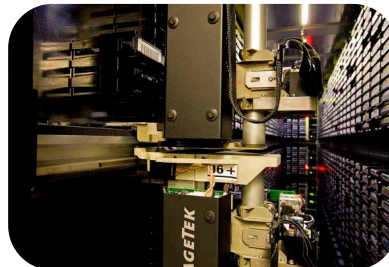
CONTENTdm®

The first thing that a member would need to do is export copies of their digital collections/files

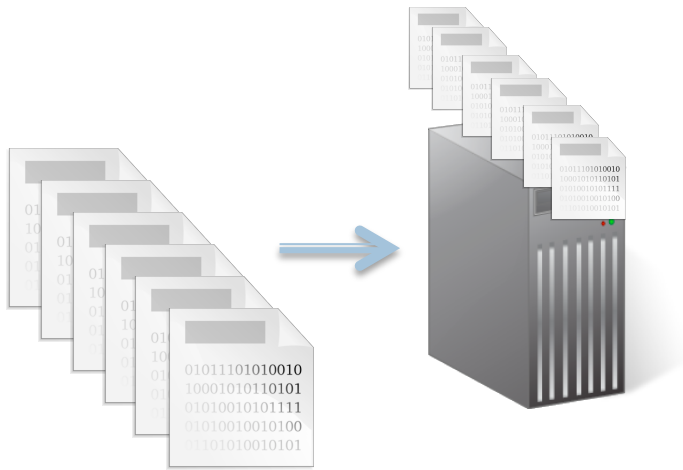
DigiTool



ArchivalWare

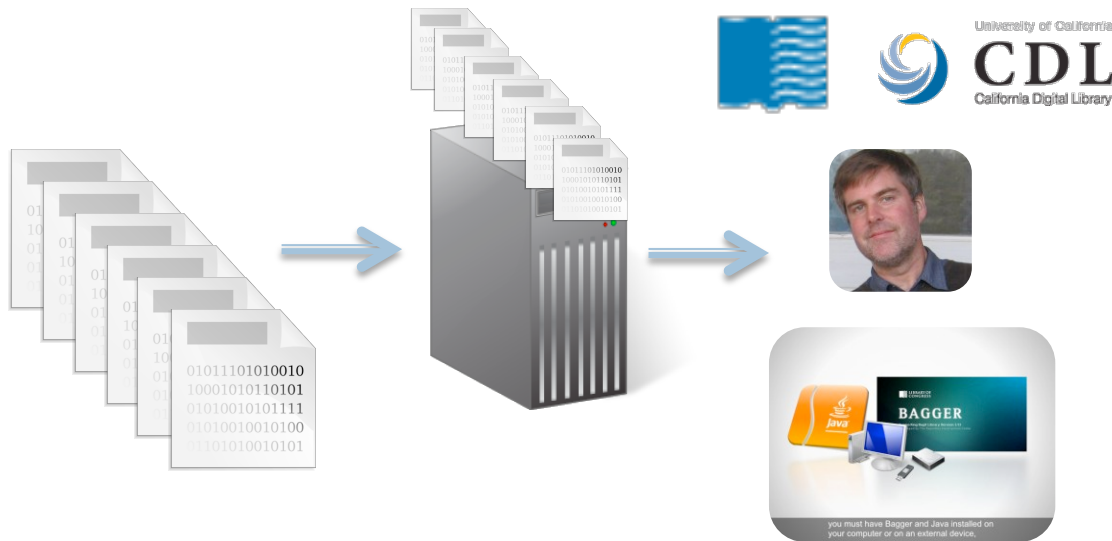


MetaArchive Bag Creation



The member then needs to get these placed on a local staging server for performing Bag creation operations

MetaArchive Bag Creation



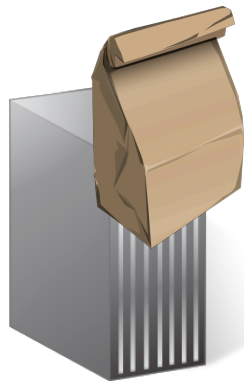
Once staged the member can run any of a number of user-friendly BagIt utilities (Java Bag Library, bagit.py, Bagger-GUI) to “bag” up collections/files.

MetaArchive Bag Creation



This will give the member a “bag” sitting on their staging server. What now?

MetaArchive Bag Creation



Boston College Libraries - Chronicles Project Test Exchanges

- Manifest Page to allow preservation by LOCKSS daemon in the MetaArchive Network
- LOCKSS system has permission to collect, preserve, and serve this Archival Unit.

Collection Info:

- Conspicuous Collection(s): Boston College bcheights-bcshr-bc885
- Source Organization: Boston College
- Contact Name: William J. Donovan
- Contact Phone: 617-552-1570
- Contact Email: bill.donovan@bc.edu
- Organization Address: 140 Commonwealth Avenue, Chestnut Hill, MA 02467 USA

Newspaper issues from "The Heights", "The Sacred Heart Review", and "The 885". Content was digitized to article-level METS and ALTO XML, with JPEG2000 page images and issue-level PDF files. Digital files were produced by Digital Divide Data in 2011 and 2012, to standards suitable for ingest with Veridian software.

```
1 <map>
2 <entry>
3   <string>plugin_config_props</string>
4   <list>
5     <org.lockss.daemon.ConfigParamDescr>
6       <key>start</key>
7       <displayName>Start</displayName>
8       <type>1</type>
9       <size>20</size>
10      <definitional>true</definitional>
11      <defaultOnly>false</defaultOnly>
12    </org.lockss.daemon.ConfigParamDescr>
13    <org.lockss.daemon.ConfigParamDescr>
14      <key>base_url</key>
15      <displayName>Base URL</displayName>
16      <description>Usually of the form http://&lt;journal-name&gt;.com
17    </description>
18    <type>3</type>
19    <size>40</size>
20    <definitional>true</definitional>
21    <defaultOnly>false</defaultOnly>
22  </org.lockss.daemon.ConfigParamDescr>
23  <org.lockss.daemon.ConfigParamDescr>
24    <key>path</key>
25    <displayName>Path</displayName>
```

If the bag can fit into our current AU sizes (1GB-30GB), then we can put a manifest page & plugin in place, define the AU...

Collection

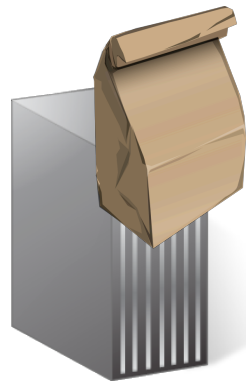
Boston College bcheights-bcshr-bc885

Archive: GEN
Plugin: org.metaarchive.AliFromStart by Metaarchive
Base Uri: <http://ec2-75-101-229-12.compute-1.amazonaws.com:53790>
Identifier: 231
Description: Newspaper issues from "The Heights", "The Sacred Heart Review", and "The 885". Content was digitized to article-level METS and ALTO XML, with JPEG2000 page images and issue-level PDF files. Digital files were produced by Digital Divide Data in 2011 and 2012, to standards suitable for ingest with Veridian software.
Metadata: Undefined Edit (/collections/231/metadata.xml)
Premis: http://admin.metaarchive.org/co_premis/231/
Status: http://data.metaarchive.org/co_status/content_providers/content_provider=10/collection=231/

Archival units

Archival Unit Parameters	Status	#Copies	
path = boston_data.bag_split start = boston_data.bag_0	test	0	site up ⬇
path = boston_data.bag_split start = boston_data.bag_1	test	0	site up ⬇
path = boston_data.bag_split start = boston_data.bag_2	test	0	site up ⬇
path = boston_data.bag_split start = boston_data.bag_3	test	0	site up ⬇
path = boston_data.bag_split start = boston_data.bag_4	test	0	site up ⬇
path = boston_data.bag_split start = boston_data.bag_5	test	0	site up ⬇
path = boston_data.bag_split start = boston_data.bag_6	test	0	site up ⬇
path = boston_data.bag_split start = boston_data.bag_metadata	test	0	site up ⬇

MetaArchive Bag Creation



Boston College Libraries - Chronicles Project Test Exchanges

- Manifest Page to allow preservation by LOCKSS daemon in the MetaArchive Network
- LOCKSS system has permission to collect, preserve, and serve this Archival Unit.

Collection Info:

- Conspicuous Collection(s): Boston College bcheights-bcshr-bc885
- Source Organization: Boston College
- Contact Name: William J. Donovan
- Contact Phone: 617-552-1570
- Contact Email: bill.donovan@bc.edu
- Organization Address: 140 Commonwealth Avenue, Chestnut Hill, MA 02467 USA

Newspaper issues from "The Heights", "The Sacred Heart Review", and "The 885". Content was digitized to article-level METS and ALTO XML files. Digital files were produced by Digital Divide Data in 2011 and 2012, to standards suitable for ingest with Veridian software.

```
1 <map>
2 <entry>
3   <string-plugin_config_props/>string
4   <list>
5     <org.lockss.daemon.ConfigParamDescr>
6       <key-start/>key
7       <displayName-Start/>displayName
8       <type-1/>type
9       <size-20/>size
10      <definitional-true/>definitional
11      <defaultOnly-false/>defaultOnly
12    </org.lockss.daemon.ConfigParamDescr>
13    <org.lockss.daemon.ConfigParamDescr>
14      <key-base_url/>key
15      <displayName-Base URL/>displayName
16      <description-Usually of the form http://&lt;journal-name&gt;.com
17    </description>
18    <type-3/>type
19    <size-40/>size
20    <definitional-true/>definitional
21    <defaultOnly-false/>defaultOnly
22  </org.lockss.daemon.ConfigParamDescr>
23  <org.lockss.daemon.ConfigParamDescr>
24    <key-path/>key
25    <displayName-Path/>displayName
26  </org.lockss.daemon.ConfigParamDescr>
27</list>
28</entry>
29</map>
```

Collection

Boston College bcheights-bcshr-bc885

Archive: GEN
Plugin: org.metaarchive.AliFromStart by Metaarchive
Base Uri: <http://ec2-75-101-229-12.compute-1.amazonaws.com:53790/>
Identifier: 231
Description: Newspaper issues from "The Heights", "The Sacred Heart Review", and "The 885". Content was digitized to article-level METS and ALTO XML files. Digital files were produced by Digital Divide Data in 2011 and 2012, to standards suitable for ingest with Veridian software.

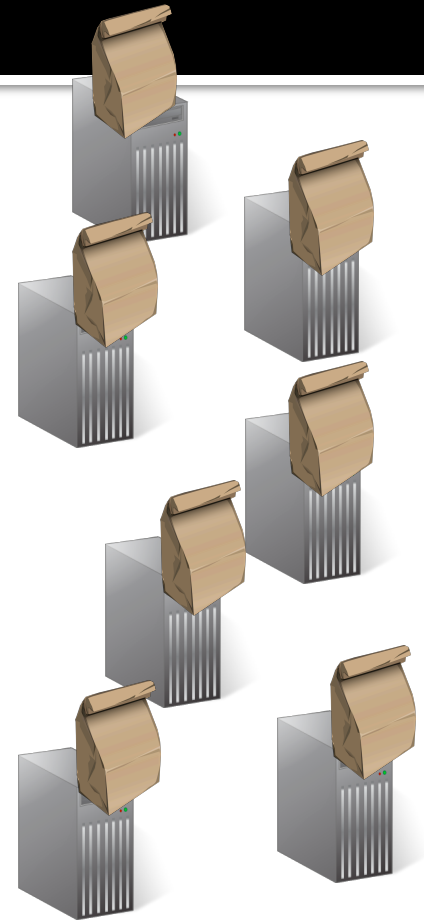
MetaData: Undefined Edit: <http://collections/231/metadata.xml>
Premis: http://admin.metaarchive.org/co_premis/231/
Status: http://data.metaarchive.org/co_status/content_providers/content_provider=10/collection=231/

Archival units

Archival Unit Parameters	Status	# Copies	
path = boston_data.bag_split start = boston_data.bag_0	test	0	site up ⬇
path = boston_data.bag_split start = boston_data.bag_1	test	0	site up ⬇
path = boston_data.bag_split start = boston_data.bag_2	test	0	site up ⬇
path = boston_data.bag_split start = boston_data.bag_3	test	0	site up ⬇
path = boston_data.bag_split start = boston_data.bag_4	test	0	site up ⬇
path = boston_data.bag_split start = boston_data.bag_5	test	0	site up ⬇
path = boston_data.bag_split start = boston_data.bag_6	test	0	site up ⬇
path = boston_data.bag_split start = boston_data.bag_metadata	test	0	site up ⬇

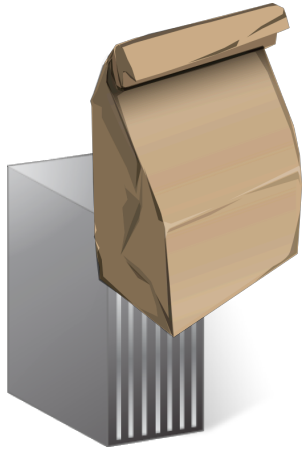
...and off we go!

Skinner 2014



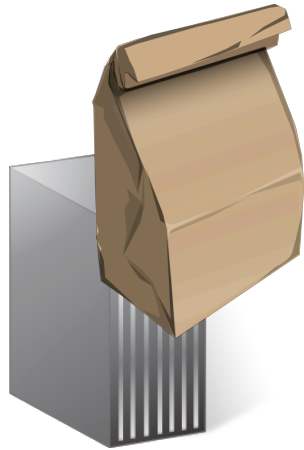
MetaArchive
cooperative

MetaArchive Bag Creation



But what if the bag is
bigger than our current
AU sizes (1GB-30GB)?

MetaArchive Bag Creation



```
$ bag splitbagbysize <BAG> --maxbagsize 30  
$ python bagit-split.py split <BAG>
```

Then we need to apply a custom-built script to split the large bag into several smaller bags.

MetaArchive Bag Creation



MetaArchive Bag Ingest



Boston College Libraries - Chronicles Project Test Exchanges

- Manifest Page to allow preservation by LOCKSS daemon in the MetaArchive Network
- LOCKSS system has permission to collect, preserve, and serve this Archival Unit.

Collection Info:

- Conspicuous Collection(s): Boston College bcheights-bcshr-bc885
- Source Organization: Boston College
- Contact Name: William J. Donovan
- Contact Phone: 617-552-1570
- Contact Email: bill.donovan@bc.edu
- Organization Address: 140 Commonwealth Avenue, Chestnut Hill, MA 02467 USA

Newspaper issues from "The Heights", "The Sacred Heart Review", and "The 885". Content was digitized to article-level METS and ALTO XML, with JPEG2000 page images and issue-level PDF files. Digital files were produced by Digital Divide Data in 2011 and 2012, to standards suitable for ingest with Veridian software.

```
1 <map>
2 <entry>
3   <string-plugin_config_props/>string>
4   <list>
5     <org.lockss.daemon.ConfigParamDescr>
6     <key-start/>key>
7     <displayName>Start</displayName>
8     <type>1</type>
9     <size>20</size>
10    <definitional>true</definitional>
11    <defaultOnly>false</defaultOnly>
12    </org.lockss.daemon.ConfigParamDescr>
13    <org.lockss.daemon.ConfigParamDescr>
14    <key-base_url/>key>
15    <displayName>Base URL</displayName>
16    <description>Usually of the form http://&lt;journal-name&gt;.com
17  /</description>
18  <type>3</type>
19  <size>40</size>
20  <definitional>true</definitional>
21  <defaultOnly>false</defaultOnly>
22  </org.lockss.daemon.ConfigParamDescr>
23  <org.lockss.daemon.ConfigParamDescr>
24  <key-path/>key>
25  <displayName>Path</displayName>
```

We would put
a manifest page
and plugin in place,
define the AUs ...

Collection

Boston College bcheights-bcshr-bc885

Archive: GEN
Plugin: org.metaarchive.AliFromStart by Metaarchive
Base Url: <http://ec2-75-101-229-12.compute-1.amazonaws.com:53790>
Identifier: 231
Description: Newspaper issues from "The Heights", "The Sacred Heart Review", and "The 885". Content was digitized to article-level METS and ALTO XML, with JPEG2000 page images and issue-level PDF files. Digital files were produced by Digital Divide Data in 2011 and 2012, to standards suitable for ingest with Veridian software.
Metadata: Undefined Edit (/collections/231/metadata.xml)
Premis: http://admin.metaarchive.org/co_premis/231/
Status: http://data.metaarchive.org/co_status/content_providers/content_provider=10/collection=231/

Archival units

Archival Unit Parameters	Status	#Copies	
path = boston_data.bag_split start = boston_data.bag_0	test	0	site up ⬇
path = boston_data.bag_split start = boston_data.bag_1	test	0	site up ⬇
path = boston_data.bag_split start = boston_data.bag_2	test	0	site up ⬇
path = boston_data.bag_split start = boston_data.bag_3	test	0	site up ⬇
path = boston_data.bag_split start = boston_data.bag_4	test	0	site up ⬇
path = boston_data.bag_split start = boston_data.bag_5	test	0	site up ⬇
path = boston_data.bag_split start = boston_data.bag_6	test	0	site up ⬇
path = boston_data.bag_split start = boston_data.bag_metadata	test	0	site up ⬇

MetaArchive Bag Ingest

Boston College Libraries - Chronicles Project Test Exchanges

- Manifest Page to allow preservation by LOCKSS daemon in the MetaArchive Network
- LOCKSS system has permission to collect, preserve, and serve this Archival Unit.

Collection Info:

- Conspicuous Collection(s): Boston College bcheights-bcshr-bc885
- Source Organization: Boston College
- Contact Name: William J. Donovan
- Contact Phone: 617-552-1370
- Contact Email: bill.donovan@bc.edu
- Organization Address: 140 Commonwealth Avenue, Chestnut Hill, MA 02467 USA

Newspaper issues from "The Heights", "The Sacred Heart Review", and "The 885". Content was digitized in METS and ALTO XML, with PDF files and issue-level PDF files. Digital files were produced by Digital Divide Data in 2011 and 2012, to standards suitable for ingest with Veridian software.

```
<entry>
  <string-plugin_configuration>
    <list>
      <org.lockss.daemon.ConfigParamDescr>
        <key>start</key>
        <displayName>Start</displayName>
        <type>1</type>
        <size>20</size>
        <definitional>true</definitional>
        <defaultOnly>false</defaultOnly>
      </org.lockss.daemon.ConfigParamDescr>
      <org.lockss.daemon.ConfigParamDescr>
        <key>base_url</key>
        <displayName>Base URL</displayName>
        <description>Usually of the form http://&lt;journal-name&gt;.com
      </description>
        <type>3</type>
        <size>40</size>
        <definitional>true</definitional>
        <defaultOnly>false</defaultOnly>
      </org.lockss.daemon.ConfigParamDescr>
      <org.lockss.daemon.ConfigParamDescr>
        <key>path</key>
        <displayName>Path</displayName>
      </org.lockss.daemon.ConfigParamDescr>
    </list>
  </string-plugin_configuration>
</entry>
```

Collection

Boston College bcheights-bcshr-bc885

Archive: GEN
 Plugin: org.metaarchive.AliFromStart by Metaarchive
 Base Uri: <http://ec2-75-101-229-12.compute-1.amazonaws.com:53790/>
 Identifier: 231
 Description: Newspaper issues from "The Heights", "The Sacred Heart Review", and "The 885". Content was digitized in METS and ALTO XML, with JPEG2000 page images and issue-level PDF files. Digital files were produced by Digital Divide Data in 2011 and 2012, to standards suitable for ingest with Veridian software.

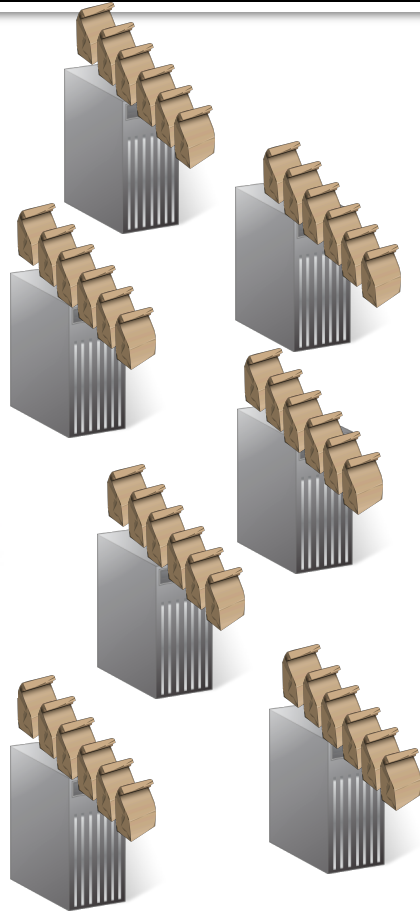
MetaData: Undefined Edit (/collections/231/metadata.xml)
 Premis: http://admin.metaarchive.org/co_premis/231/
 Status: http://data.metaarchive.org/co_status/content_providers/content_provider=10/collection=231/

Archival units

Archival Unit Parameters	Status	#Copies	
path = boston_data.bag_split start = boston_data.bag_0	test	0	site up ⬇
path = boston_data.bag_split start = boston_data.bag_1	test	0	site up ⬇
path = boston_data.bag_split start = boston_data.bag_2	test	0	site up ⬇
path = boston_data.bag_split start = boston_data.bag_3	test	0	site up ⬇
path = boston_data.bag_split start = boston_data.bag_4	test	0	site up ⬇
path = boston_data.bag_split start = boston_data.bag_5	test	0	site up ⬇
path = boston_data.bag_split start = boston_data.bag_6	test	0	site up ⬇
path = boston_data.bag_split start = boston_data.bag_metadata	test	0	36 ⬇

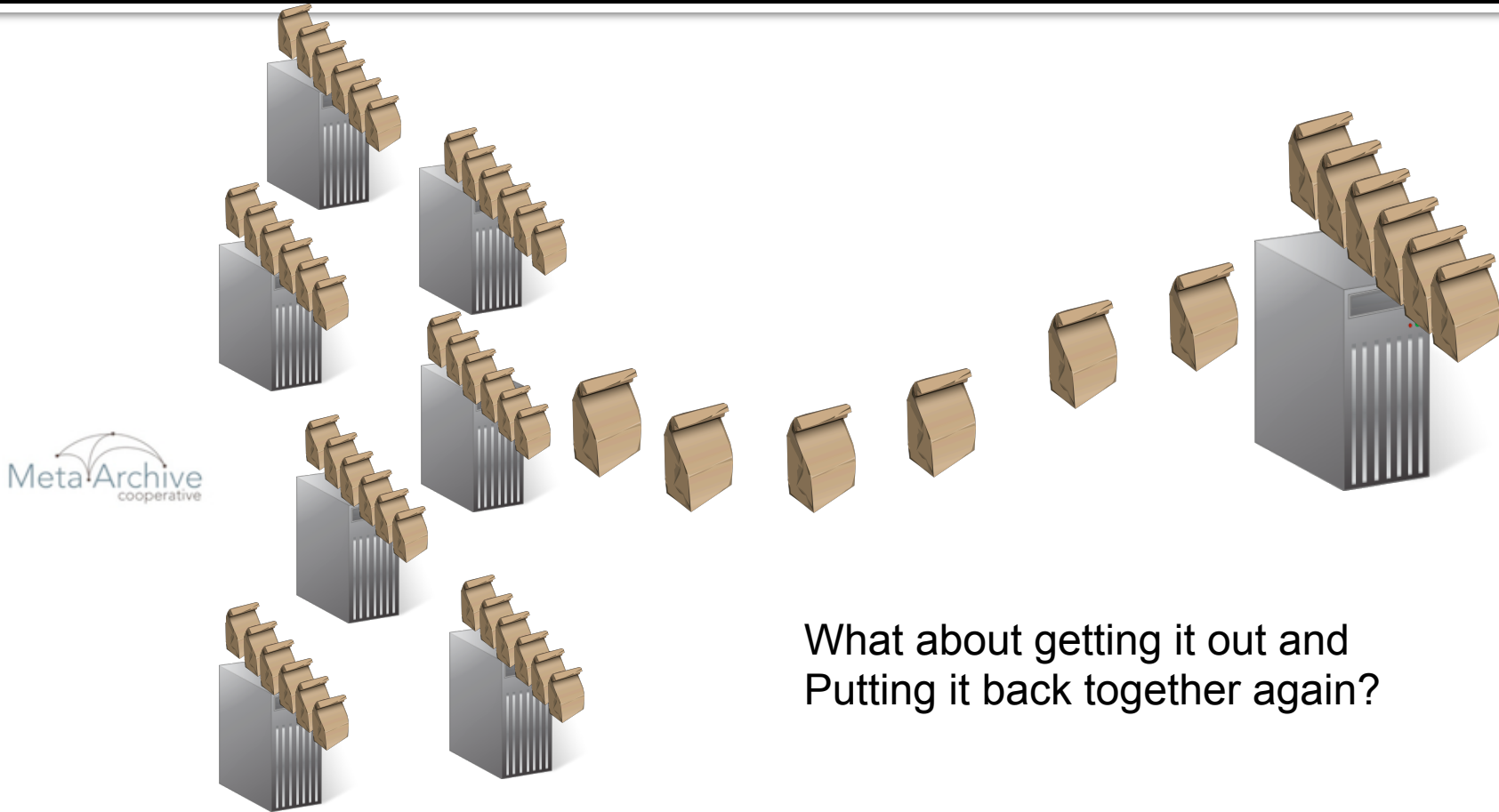
...and off we go!

MetaArchive Bag Export



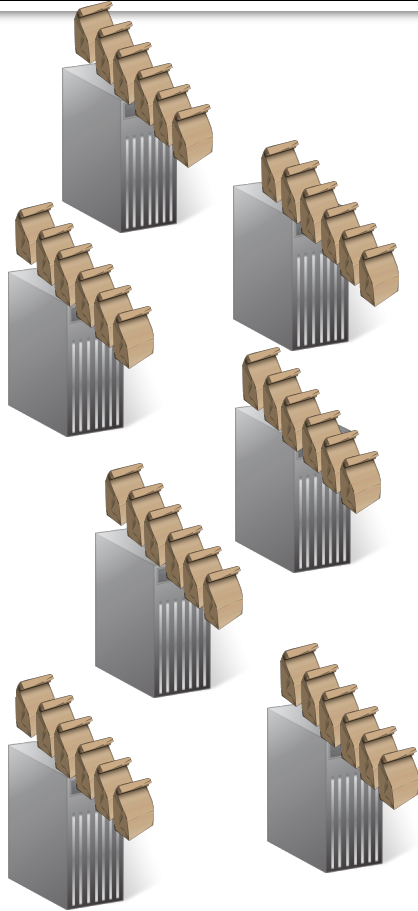
So...we know getting it in works!

MetaArchive Bag Export



What about getting it out and
Putting it back together again?

MetaArchive Bag Rebuilding



\$ python bagit-split.py unsplit <DIRECTORY
CONTAINING BAGS>



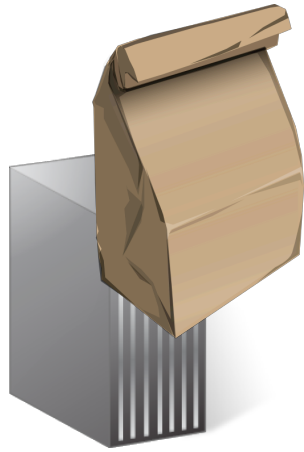
MetaArchive Bag Rebuilding

1. Make sure collection AUs have 100% agreement
2. Use LOCKSS Export Content feature on one cache (uncompressed zip) – download to processing server
3. Run Bag Unsplit Utility
 1. `$ python bagit-split.py unsplit <DIRECTORY CONTAINING BAGS>`

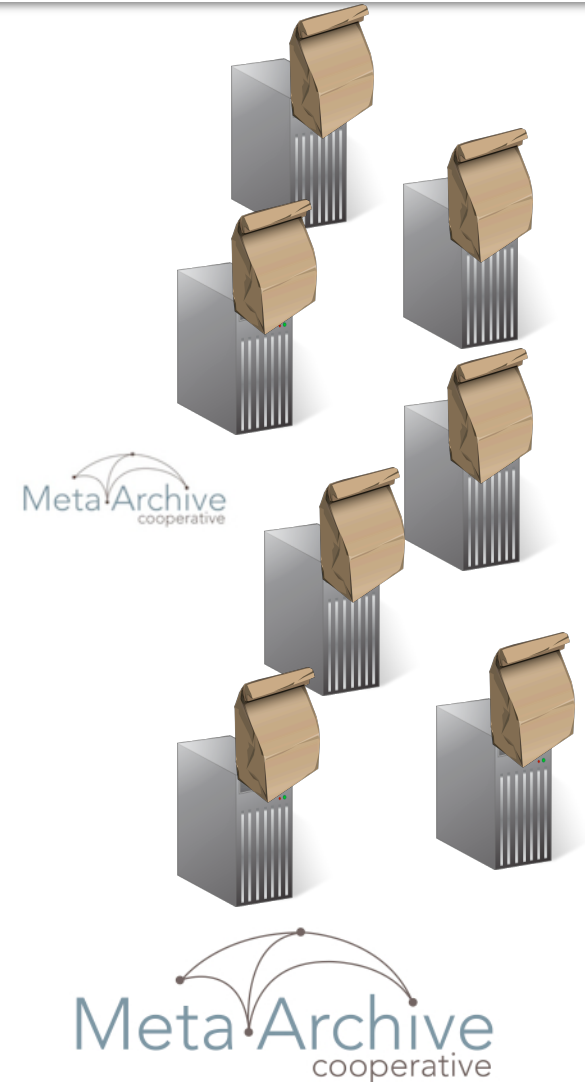
MetaArchive Change Management

- MetaArchive members change collections over time
 - Add files
 - Change existing files
 - Remove files
- LOCKSS Versioning
 - Automated re-crawling for changes
 - Replaced versions are stored but not voted/polled on
- MetaArchive Versioning Policy
 - Semi-automated and scheduled re-crawling for changes
 - Improves uniformity and agreement levels across copies

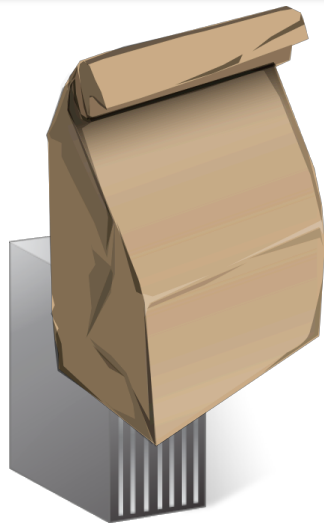
MetaArchive BagIt Change Management



Member Staging
Server
&
Bag

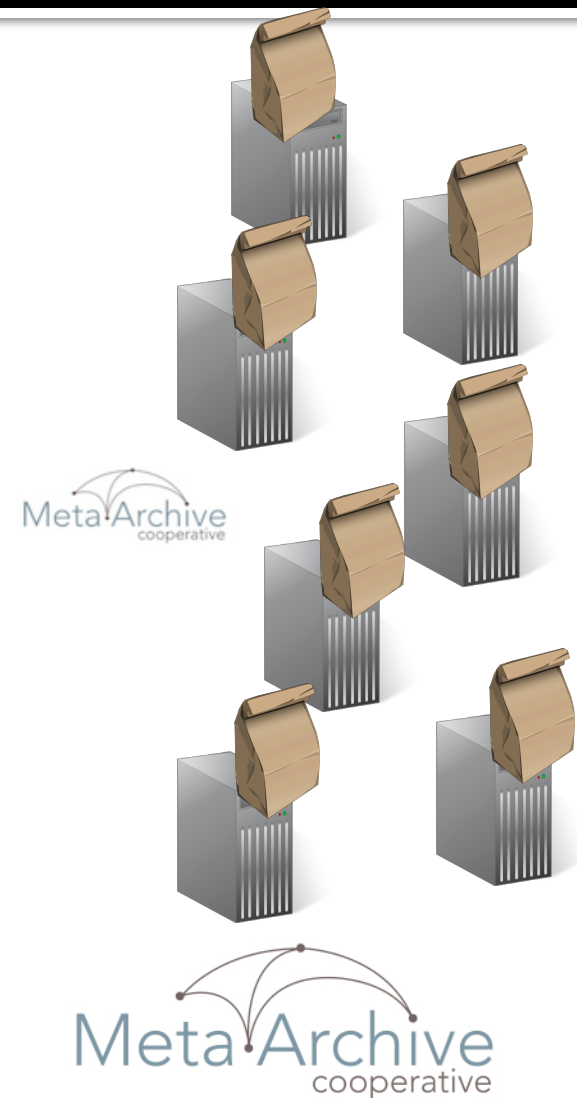


MetaArchive BagIt Change Management



1. New Files!
2. Changed Files!
3. Deleted Files!

Member Staging Server
&
Bag w/changes



BagIt Change Management

- Three basic methods:

1. Re-Bag Approach

- Swap out old bag for a new updated bag
- Easier when managing a single storage environment

2. Bag Merge Approach

- Sync & update existing bags according to diffs
- Easier when there are fewer reps to manage

3. Change Bag Approach

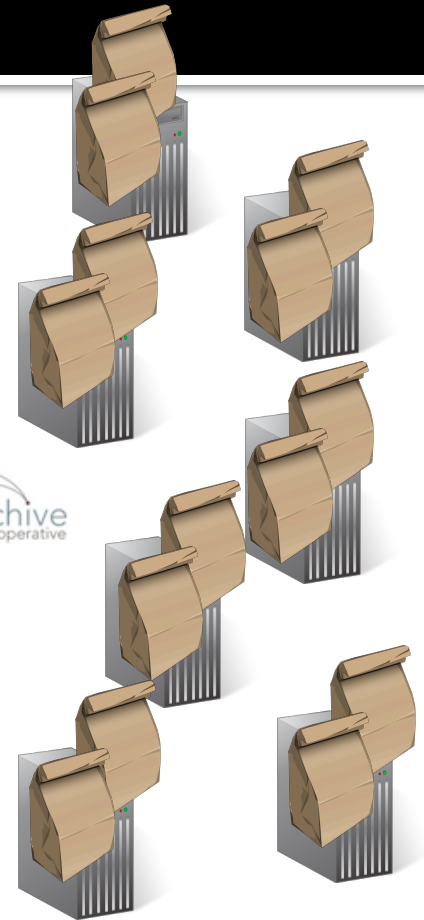
- Bag and ingest only the files that have changed
- Easier when storage is diverse and there are more reps

MetaArchive BagIt Change Management



Member Staging Server
&
Bag w/changes

Re-Bag
Approach?

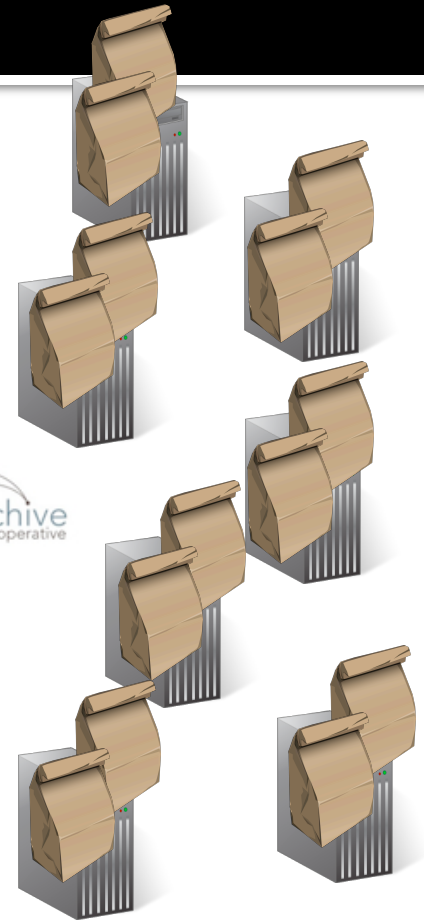


MetaArchive BagIt Change Management

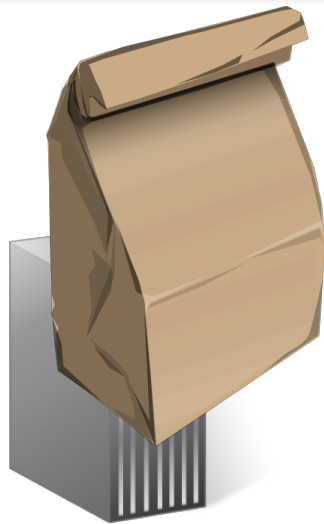


Member Staging Server
&
Bag w/changes

~~Re-Bag~~
~~Approach?~~



MetaArchive BagIt Change Management



Member Staging Server
&
Bag w/changes

~~Re-Bag~~
~~Approach?~~
Bag Merge
Approach?

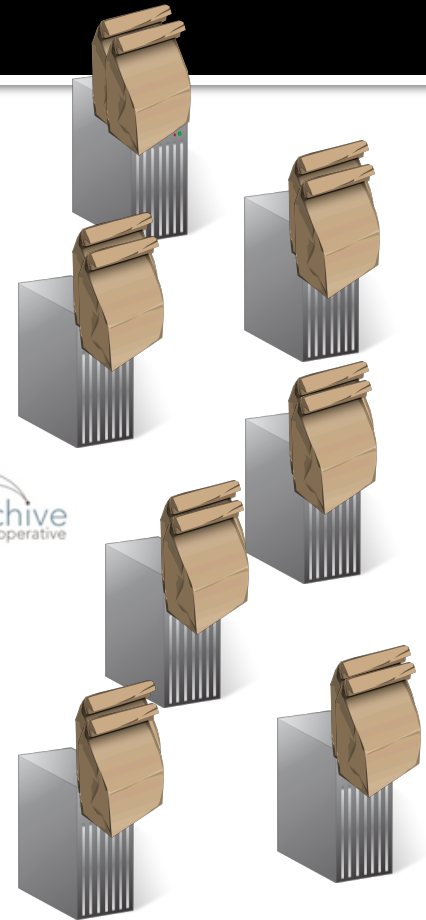


MetaArchive BagIt Change Management



Member Staging Server
&
Bag w/changes

~~Re-Bag~~
~~Approach?~~
~~Bag Merge~~
~~Approach?~~



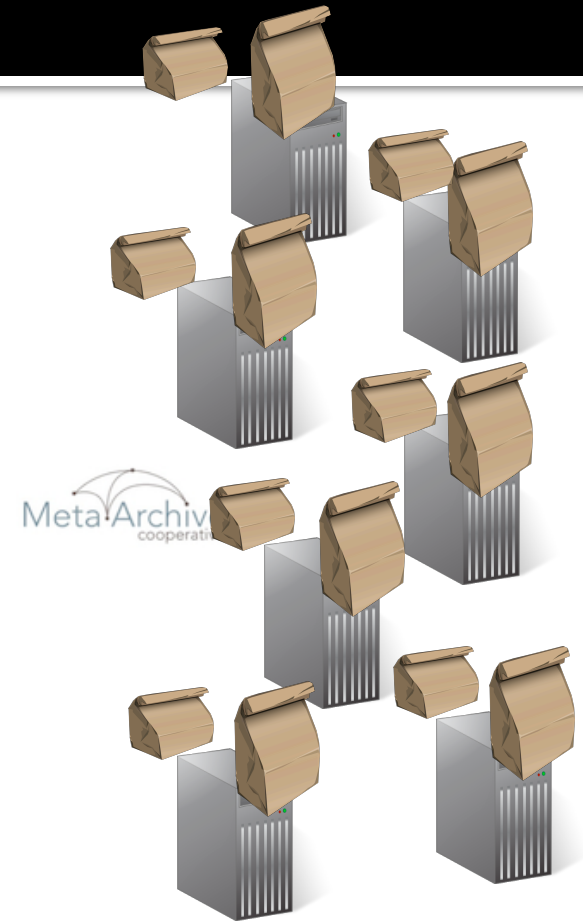
MetaArchive BagIt Change Management



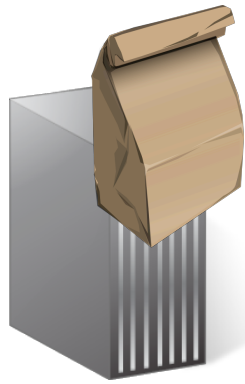
Member Staging Server
&
Bag w/changes

~~Re-Bag~~
~~Approach?~~
~~Bag Merge~~
~~Approach?~~

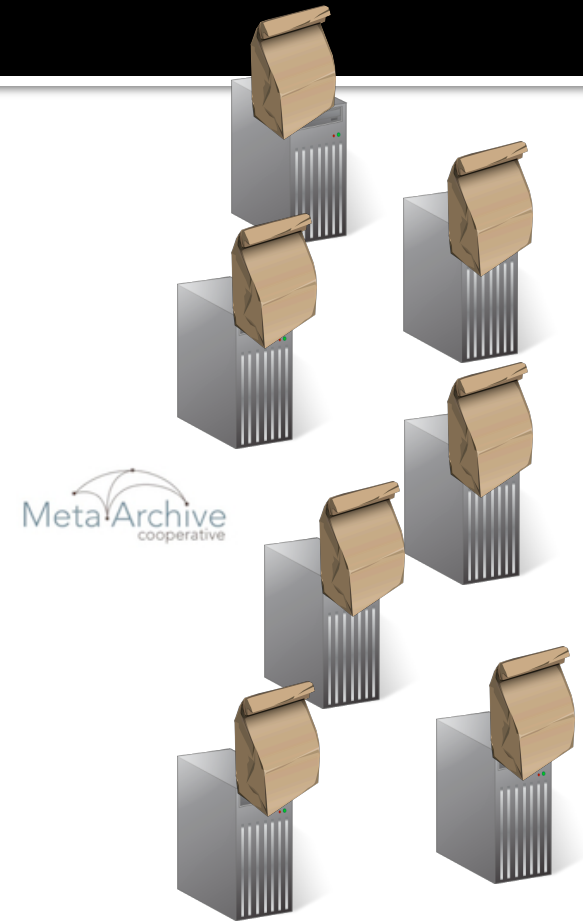
Change Bag
Approach?



MetaArchive BagIt Change Management



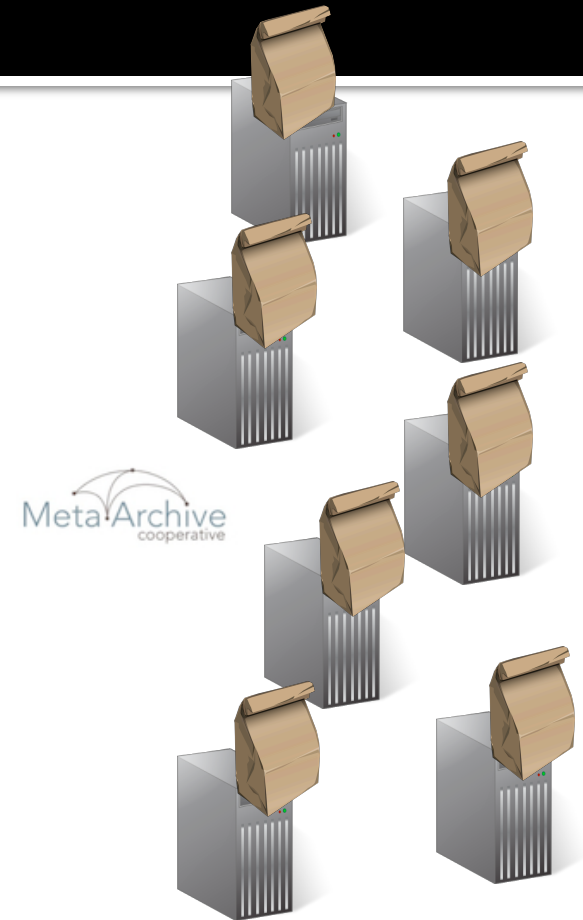
Member Staging Server
&
Staged Bag



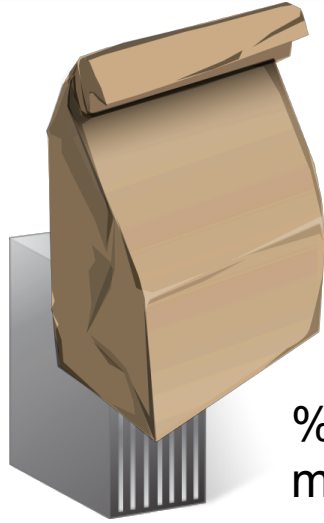
MetaArchive BagIt Change Management



Member Staging Server
&
Bag w/changes



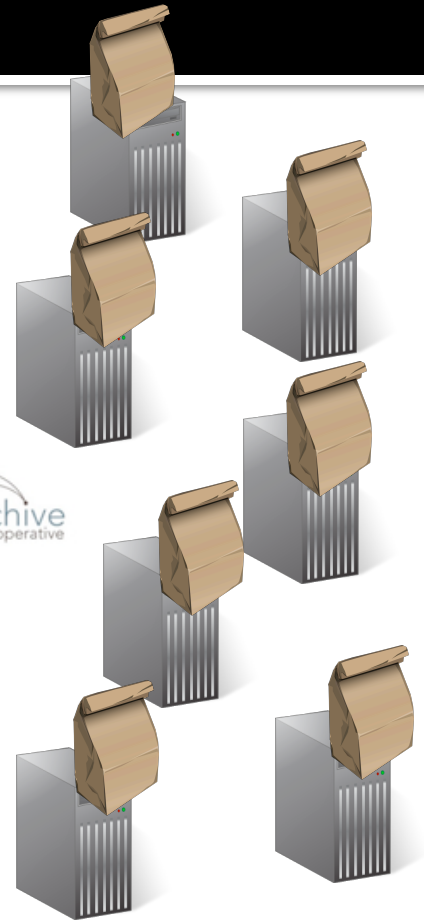
MetaArchive BagIt Change Management



```
% python bag-diff.py /path/to/bag
```

```
% python bag-diff.py -m /path/to/old/  
manifests/dir /path/to/bag
```

Member Staging Server
&
Bag w/changes



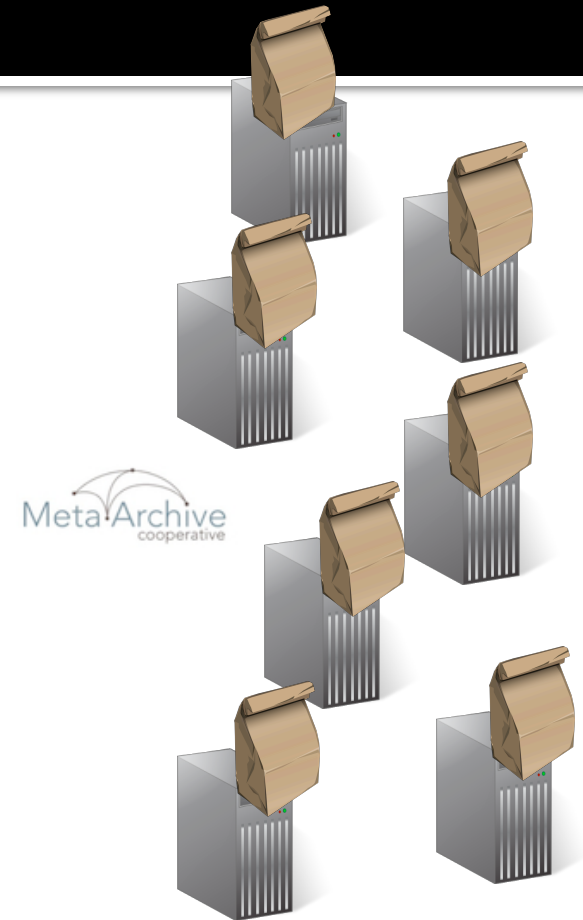
MetaArchive
cooperative

MetaArchive
cooperative

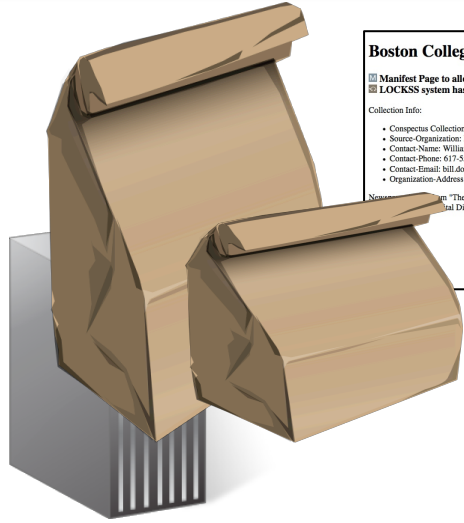
MetaArchive BagIt Change Management



Member Staging Server
w/Bag
plus a new
Updates & Changes Bag



MetaArchive BagIt Change Management



Boston College Libraries - Chronicles Project Test Exchanges

- Manifest Page to allow preservation by LOCKSS daemon in the MetaArchive Network
- LOCKSS system has permission to collect, preserve, and serve this Archival Unit.

Collection Info:

- Conspicuous Collection(s): Boston College bcheights-bcshr-bc885
- Source Organization: Boston College
- Contact Name: William J. Donovan
- Contact Phone: 617-552-1570
- Contact Email: bill.donovan@bc.edu
- Organization Address: 140 Commonwealth Avenue, Chestnut Hill, MA 02467 USA

Newspaper issues from "The Heights", "The Sacred Heart Review", and "The 885". Content was digitized to article-level METS and ALTO XML, with JPEG2000 page images and issue-level PDF files. Digital files were divided into 2011 and 2012, to standards suitable for ingest with Veridian software.

```
1 <map>
2 <entry>
3 <string>plugin_config_props</string>
4 <list>
5 <org.lockss.daemon.ConfigParamDescr>
6 <key>start</key>
7 <displayName>Start</displayName>
8 <type>1</type>
9 <size>20</size>
10 <definitional>true</definitional>
11 <defaultOnly>false</defaultOnly>
12 </org.lockss.daemon.ConfigParamDescr>
13 <org.lockss.daemon.ConfigParamDescr>
14 <key>base_url</key>
15 <displayName>Base URL</displayName>
16 <description>Usually of the form http://&lt;journal-name&gt;.com
17 </description>
18 <type>3</type>
19 <size>40</size>
20 <definitional>true</definitional>
21 <defaultOnly>false</defaultOnly>
22 </org.lockss.daemon.ConfigParamDescr>
23 <org.lockss.daemon.ConfigParamDescr>
24 <key>path</key>
25 <displayName>Path</displayName>
```

Collection

Boston College bcheights-bcshr-bc885

Archive: GEN
 Plugin: org.metaarchive.AliFromStart by Metaarchive
 Base Uri: <http://ec2-75-101-229-12.compute-1.amazonaws.com:53790/>
 Identifier: 231
 Description: Newspaper issues from "The Heights", "The Sacred Heart Review", and "The 885". Content was digitized to article-level METS and ALTO XML, with JPEG2000 page images and issue-level PDF files. Digital files were produced by Digital Divide Data in 2011 and 2012, to standards suitable for ingest with Veridian software.

MetaData: Undefined Edit: <http://collections/231/metadata.xml>
 Prems: http://admin.metaarchive.org/co_premis/231/
 Status: http://data.metaarchive.org/co_status/content_providers/content_provider=10/collection=231/

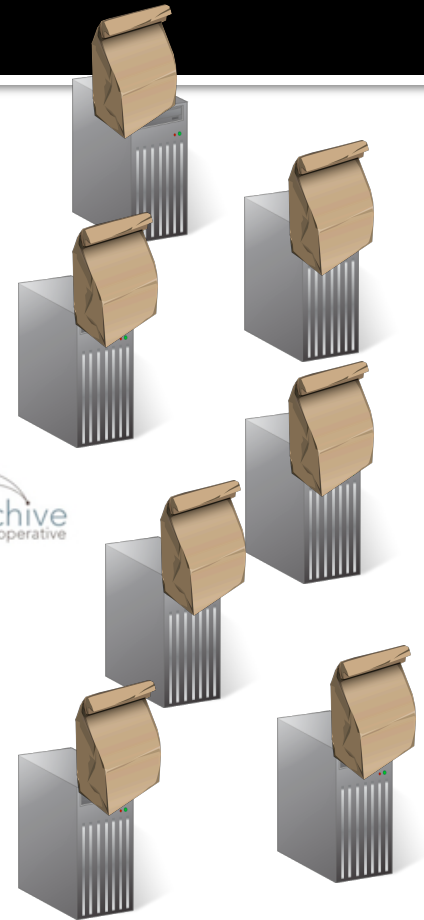
Archival units

Archival Unit Parameters	Status	# Copies	
path = boston_data.bag_split start = boston_data.bag_0	test	0	site up Ⓜ
path = boston_data.bag_split start = boston_data.bag_1	test	0	site up Ⓜ
path = boston_data.bag_split start = boston_data.bag_2	test	0	site up Ⓜ
path = boston_data.bag_split start = boston_data.bag_3	test	0	site up Ⓜ
path = boston_data.bag_split start = boston_data.bag_4	test	0	site up Ⓜ
path = boston_data.bag_split start = boston_data.bag_5	test	0	site up Ⓜ
path = boston_data.bag_split start = boston_data.bag_6	test	0	site up Ⓜ
path = boston_data.bag_split start = boston_data.bag_metadata	test	0	site up Ⓜ

54

Member Staging Server
w/Bag
plus a new
Updates & Changes Bag

Skinner 2014



MetaArchive BagIt Change Management

Boston College Libraries - Chronicles Project Test Exchanges

- Manifest Page to allow preservation by LOCKSS daemon in the MetaArchive Network
- LOCKSS system has permission to collect, preserve, and serve this Archival Unit.

Collection Info:

- Conspectus Collection(s): Boston College bcheights-bcsh-885
- Source Organization: Boston College
- Contact Name: William J. Donovan
- Contact Phone: 617-552-1570
- Contact Email: bill.donovan@bc.edu
- Organization Address: 140 Commonwealth Avenue, Chestnut Hill, MA 02467 USA

Newspaper issues from "The Heights", "The Sacred Heart Review", and "The 885". Content was digitized to article-level METS and ALTO XML, with JPEG2000 page images and issue-level PDF files. Digital files were digitized in 2011 and 2012, to standards suitable for ingest with Veridian software.

```
1 <map>
2 <entry>
3 <string>plugin_config_props</string>
4 <list>
5 <arg.lockss.daemon.ConfigParamDescr>
6 <key>start</key>
7 <displayName>Start</displayName>
8 <type>1</type>
9 </list>
10 </entry>
11 </map>
12 <!-- Additional
13 <!-- defaultOnly>
14 <!-- ConfigParamDescr>
15 <!-- ParamDescr>
16 <!-- displayName>
17 <!-- form http://&lt;journal-name&gt;.c
18 </string>
19 </entry>
20 </list>
21 </arg.lockss.daemon.ConfigParamDescr>
22 </entry>
23 </map>
24 </entry>
```

Collection

Boston College bcheights-bcsh-885

Archive: GEN
Plugin: org.metaarchive.AliFromStart by Metaarchive
Base Uri: <http://ec2-75-101-229-12.compute-1.amazonaws.com:53790/>
Identifier: 231
Description: Newspaper issues from "The Heights", "The Sacred Heart Review", and "The 885". Content was digitized to article-level METS and ALTO XML, with JPEG2000 page images and issue-level PDF files. Digital files were produced by Digital Divide Data in 2011 and 2012, to standards suitable for ingest with Veridian software.

MetaData: Undefined Edit: <http://collections/231/metadata.xml>
Premis: http://admin.metaarchive.org/co_premis/231/
Status: http://data.metaarchive.org/co_status/content_providers/content_provider=10/collection=231/

Archival units

Archival Unit Parameters	Status	# Copies	
path = boston_data.bag_split start = boston_data.bag_0	test	0	site up Ⓜ
path = boston_data.bag_split start = boston_data.bag_1	test	0	site up Ⓜ
path = boston_data.bag_split start = boston_data.bag_2	test	0	site up Ⓜ
path = boston_data.bag_split start = boston_data.bag_3	test	0	site up Ⓜ
path = boston_data.bag_split start = boston_data.bag_4	test	0	site up Ⓜ
path = boston_data.bag_split start = boston_data.bag_5	test	0	site up Ⓜ
path = boston_data.bag_split start = boston_data.bag_6	test	0	site up Ⓜ
path = boston_data.bag_split start = boston_data.bag_metadata	test	0	55 Ⓜ

Member Staging Server
w/Bag
plus a new
Updates & Changes Bag

Skinner 2014

MetaArchive
cooperative

MetaArchive
cooperative

MetaArchive BagIt Change Management



Boston College Libraries - Chronicles Project Test Exchanges

- Manifest Page to allow preservation by LOCKSS daemon in the MetaArchive Network
- LOCKSS system has permission to collect, preserve, and serve this Archival Unit.

Collection Info:

- Conspectus Collection(s): Boston College bcheights-bcsh-885
- Source Organization: Boston College
- Contact Name: William J. Donovan
- Contact Phone: 617-552-1570
- Contact Email: bill.donovan@bc.edu
- Organization Address: 140 Commonwealth Avenue, Chestnut Hill, MA 02467 USA

Newspaper issues from "The Heights", "The Sacred Heart Review", and "The 885". Content was digitized to article-level METS and ALTO XML, with JPEG2000 page images and issue-level PDF files. Digital files were produced by Digital Divide Data in 2011 and 2012, to standards suitable for ingest with Veridian software.

```

1 <map>
2 <entry>
3 <string>plugin_config_props</string>
4 <list>
5 <org.lockss.daemon.ConfigParamDescr>
6 <key>start</key>
7 <displayName>Start</displayName>
8 <type>i</type>
9 <size>20</size>
10 <definitional>true</definitional>
11 <defaultOnly>false</defaultOnly>
12 </org.lockss.daemon.ConfigParamDescr>
13 <org.lockss.daemon.ConfigParamDescr>
14 <key>base_url</key>
15 <displayName>Base URL</displayName>
16 <description>Usually of the form http://&lt;journal-name&gt;.com
17 </description>
18 <type>3</type>
19 <size>40</size>
20 <definitional>true</definitional>
21 <defaultOnly>false</defaultOnly>
22 </org.lockss.daemon.ConfigParamDescr>
23 <org.lockss.daemon.ConfigParamDescr>
24 <key>path</key>
  <displayName>Path</displayName>

```

Collection

Boston College bcheights-bcsh-885

Archive: GEN
 Plugin: org.metaarchive.AliFromStart by Metaarchive
 Base Uri: <http://ec2-75-101-229-12.compute-1.amazonaws.com:53790>
 Identifier: 231
 Description: Newspaper issues from "The Heights", "The Sacred Heart Review", and "The 885". Content was digitized to article-level METS and ALTO XML, with JPEG2000 page images and issue-level PDF files. Digital files were produced by Digital Divide Data in 2011 and 2012, to standards suitable for ingest with Veridian software.

Metadata: Undefined Edit (/collections/231/metadata.xml)
 Premis: http://admin.metaarchive.org/co_premis/231
 Status: http://data.metaarchive.org/co_status/content_providers/content_provider=10/collection=231

Archival units

Archival Unit Parameters	Status	# Copies	
path = boston_data.bag_split start = boston_data.bag_0	test	0	site up
path = boston_data.bag_split start = boston_data.bag_1	test	0	site up
path = boston_data.bag_split start = boston_data.bag_2	test	0	site up
path = boston_data.bag_split start = boston_data.bag_3	test	0	site up
path = boston_data.bag_split start = boston_data.bag_4	test	0	site up
path = boston_data.bag_split start = boston_data.bag_5	test	0	site up
path = boston_data.bag_split start = boston_data.bag_6	test	0	site up
path = boston_data.bag_split start = boston_data.bag_metadata	test	0	56

MetaArchive
cooperative

MetaArchive
cooperative

Boston Collection

 Manifesto Page to all
 LOCKSS system has

Collection Info:

- Conspectus Collection
- Source Organization
- Contact Name: Willis
- Contact Phone: 617.6
- Contact Email: bll@
- Organization Address

Newspaper issues from "The Boston Herald" were produced by Digital Library of America.

 Manifest Page to allow preservation by LOCKSS daemons in the MetaArchive Network
 LOCKSS system has permission to collect, preserve, and serve this Archival Unit.

Collection Info:

- Conspectus Collection(s): Boston College bcheights-bchsr-bc885
- Source-Organization: Boston College
- Contact-Name: William J. Donovan
- Contact-Phone: 617-552-1370
- Contact-Email: bill.donovan@bc.edu
- Organization-Address: 140 Commonwealth Avenue, Chestnut Hill, MA 02467 USA

Newspaper issues from "The Heights", "The Sacred Heart Review", and "The 885". Content was digitized to article-level METS and ALTO XML, with JPEG2000 page images and issue-level PDF files. Digital files were produced by Digital Divide Data in 2011 and 2012, to standards suitable for ingest with Veridian software.

```
1 <prop>
2 <entry>
3 <string>plugin_config_props</string>
4 <list>
5 <org.lockss.daemon.ConfigParamDescr>
6 <key>start</key>
7 <displayName>Start</displayName>
8 <type>1</type>
9
```

...functional>
...defaultOnly>
...ConfigParamDescr>
...ConfigParamDescr>
...displayName>
...form http://<journal-name>...</p></div>

M Boston College [bcheights-bcshr-bc885](#)

Archive:	GEN
Plugin:	org.metaarchive.AllFromStart by Metaarchive
Base URL:	http://ec2-75-101-229-12.compute-1.amazonaws.com:53790 
Identifier:	231
Description:	Newspaper issues from "The Heigher", "The Sacred Heart Review", and "The 889". Content was digitized to include METS and ALTO XML, with JPEG2000 page images and issue-level PDF files. Digital files were produced by Digital Divide Data in 2011, and 2012, to standards suitable for ingest with Veridian software.
MetaData	Undefined  (collections/231/metadata.xml)
Premis	http://admin.metaarchive.org/co_premis/231 
Status	http://data.metaarchive.org/co_status/content_providers/content_provider=10 ?collection=231">?collection=231 

Archival units

Archival Unit Parameters		Status	✖ Copies	
path = boston_data_bag_split	test	0	site up	df
path = boston_data_bag_0				
path = boston_data_bag_split	test	0	site up	df
path = boston_data_bag_1				
path = boston_data_bag_split	test	0	site up	df
path = boston_data_bag_2				
path = boston_data_bag_split	test	0	site up	df
path = boston_data_bag_3				
path = boston_data_bag_split	test	0	site up	df
path = boston_data_bag_4				
path = boston_data_bag_split	test	0	site up	df
path = boston_data_bag_5				
path = boston_data_bag_split	test	0	site up	df
path = boston_data_bag_6				
path = boston_data_bag_split	test	0	site up	df
path = boston_data_bag_metadata				



MetaArchive
cooperative

MetaArchive BagIt Change Management



Boston College Libraries - Chronicles Project Test Exchanges

- Manifest Page to allow preservation by LOCKSS daemon in the MetaArchive Network
- LOCKSS system has permission to collect, preserve, and serve this Archival Unit.

Collection Info:

- Conspectus Collection(s): Boston College bcheights-bcsh-885
- Source Organization: Boston College
- Contact Name: William J. Donovan
- Contact Phone: 617-552-1570
- Contact Email: bill.donovan@bc.edu
- Organization Address: 140 Commonwealth Avenue, Chestnut Hill, MA 02467 USA

Newspaper issues from "The Heights", "The Sacred Heart Review", and "The 885". Content was digitized to article-level METS and ALTO XML, with JPEG2000 page images and issue-level PDF files. Digital files were produced by Digital Divide Data in 2011 and 2012, to standards suitable for ingest with Veridian software.

```

1  <map>
2  <entry>
3    <string>plugin_config_props</string>
4    <list>
5      <org.lockss.daemon.ConfigParamDescr>
6        <key>start</key>
7        <displayName>Start</displayName>
8        <type>i</type>
9        <size>20</size>
10       <definitional>true</definitional>
11       <defaultOnly>false</defaultOnly>
12     </org.lockss.daemon.ConfigParamDescr>
13     <org.lockss.daemon.ConfigParamDescr>
14       <key>base_url</key>
15       <displayName>Base URL</displayName>
16       <description>Usually of the form http://&lt;journal-name&gt;.com
17     </description>
18     <type>3</type>
19     <size>40</size>
20     <definitional>true</definitional>
21     <defaultOnly>false</defaultOnly>
22   </org.lockss.daemon.ConfigParamDescr>
23   <org.lockss.daemon.ConfigParamDescr>
24     <key>path</key>
25     <displayName>Path</displayName>

```

Collection

Boston College bcheights-bcsh-885

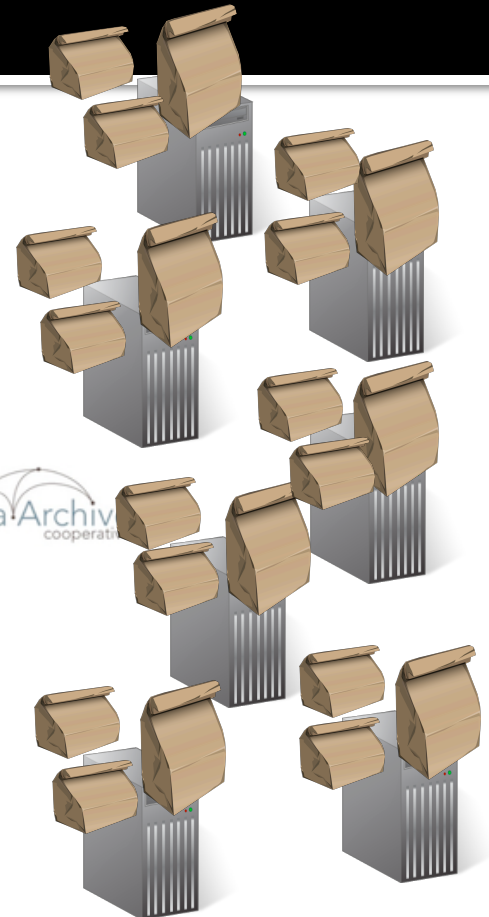
Archive: GEN
 Plugin: org.metaarchive.AliFromStart by Metaarchive
 Base Uri: http://ec2-75-101-229-12.compute-1.amazonaws.com:53790
 Identifier: 231
 Description: Newspaper issues from "The Heights", "The Sacred Heart Review", and "The 885". Content was digitized to article-level METS and ALTO XML, with JPEG2000 page images and issue-level PDF files. Digital files were produced by Digital Divide Data in 2011 and 2012, to standards suitable for ingest with Veridian software.

Metadata: Undefined Edit (/collections/231/metadata.xml)
 Premis: http://admin.metaarchive.org/co_premis/231
 Status: http://data.metaarchive.org/co_status/content_providers/content_provider=10/collection=231

Archival units

Archival Unit Parameters	Status	#Copies	
path = boston_data.bag_split start = boston_data.bag_0	test	0	site up
path = boston_data.bag_split start = boston_data.bag_1	test	0	site up
path = boston_data.bag_split start = boston_data.bag_2	test	0	site up
path = boston_data.bag_split start = boston_data.bag_3	test	0	site up
path = boston_data.bag_split start = boston_data.bag_4	test	0	site up
path = boston_data.bag_split start = boston_data.bag_5	test	0	site up
path = boston_data.bag_split start = boston_data.bag_6	test	0	site up
path = boston_data.bag_split start = boston_data.bag_metadata	test	0	site up

58



MetaArchive
cooperative

MetaArchive
cooperative

MetaArchive BagIt QA Tools

- `find-bad-files.py`
 - Recursively scans a directory for filenames that violate a set of naming standards meant to prevent problems when ingesting collections into LOCKSS over HTTP.
- `lockss-manifest-validate`
 - a tool to verify checksum hashes produced by LOCKSS against hashes provided by a BagIt manifest document.

Get the Scripts

<https://github.com/MetaArchive>

- find-bad-files.py
- bag-split.py
- lockss-manifest-validate
- bag-diff.py

Questions

Katherine Skinner

404 783 2534

katherine@educopia.org

www.metaarchive.org