Chronicles in Preservation
Preserving Digital News & Newspapers
ALA 2013

Matt Schultz

Nick Krabbenhoeft

Chronicles in Preservation

- About: NEH grant-funded study (2011-2014)
- Objective: To study, document, and model data preparation and distributed digital preservation for digital newspaper collections
- www.metaarchive.org/neh

- Content Partners
  - Boston College
  - Clemson University
  - Georgia Tech
  - Penn State
  - University of North Texas
  - University of Utah
  - Virginia Tech
- DDP Partners
  - Chronopolis
  - University of North Texas
  - MetaArchive

EDUCOPIA INSTITUTE

Why Digital Newspapers?

- At-risk and valuable scholarly content genre
- Success of the USNP & NDNP programs – cataloging, digitizing, archiving & providing access to public domain newspapers
- Success of research carried out by CRL
- Digitized and born-digital newspaper collections have been created with a variety of
  - standards
  - metadata
  - data models
  - technologies

Research Questions

- What is the spectrum of preservation readiness from essential to optimal?
- How do curators exchange digital newspapers in distributed ways for preservation?
- What are the strengths and challenges of performing distributed digital preservation for digital newspapers?

Deliverables

- **Guidelines for Digital Newspaper Preservation Readiness** – Recommendations for essential and optimal action for curating collections
- **Comparative Analysis of DDP Frameworks** – Analysis based on ingests from the Content Partners into the 3 DDP systems.
- **Interoperability Tools** - Documentation of tools to improve curation of existing collections.

EDUCOPIA INSTITUTE

Guiding Principles

- Don’t Reinvent the Wheel
- Use What Is Already Working
- Improve It

EDUCOPIA INSTITUTE

Tools & Resources

BagIt

Description Service

identify, validate and extract

DAITSS Description Service

PREMIS Event Service

Mark Phillips
University of North Texas
Denton
TX 76205, USA
+1 (940) 565-2415
Mark.Phillips@unt.edu

Matt Schultz
Educopia Institute
Atlanta
GA 30309, USA
+1 (616) 566-3204
Matt.Schultz@metaarchive.org

Kurt Nordstrom
University of North Texas
Denton
TX 76205, USA
+1 (940) 369-7809
Kurt.Nordstrom@unt.edu

UNT PREMIS Event Service

	Level One (Protect Your Data)	Level Two (Know Your data)	Level Three (Monitor Your Data)	Level Four (Repair Your Data)
Storage and Geographic Location	• Two complete copies that are not collocated • For data on heterogeneous media (optical disks, hard drives, etc.) get the content off the medium and into your storage system	• At least three complete copies • At least one copy in a different geographic location • Document your storage system(s) and storage media and what you need to use them	• At least one copy in a geographic location with a different disaster threat • Obsolescence monitoring process for your storage system(s) and media	• At least 3 copies in geographic locations with different disaster threats. • Have a comprehensive plan in place that will keep files and metadata on currently accessible media or systems

NDSA Levels of Preservation

EDUCOPIA INSTITUTE

BagIt

- Digital newspapers have a range of legacy collection structures & conventions
- BagIt is file packaging format for storing and transferring data.
- Provides a simple data model
  - A data directory
  - A manifest inventory of the bag with checksums for all objects within
  - Metadata about the bag

- bagit.py
  - Python-based BagIt tool
  - Released in 2010
  - https://github.com/edsu/bagit
- Bagger
  - Java-based BagIt tool w/ GUI
  - Released 2012
  - http://sourceforge.net/projects/loc-xferutils/files/loc-bagger/

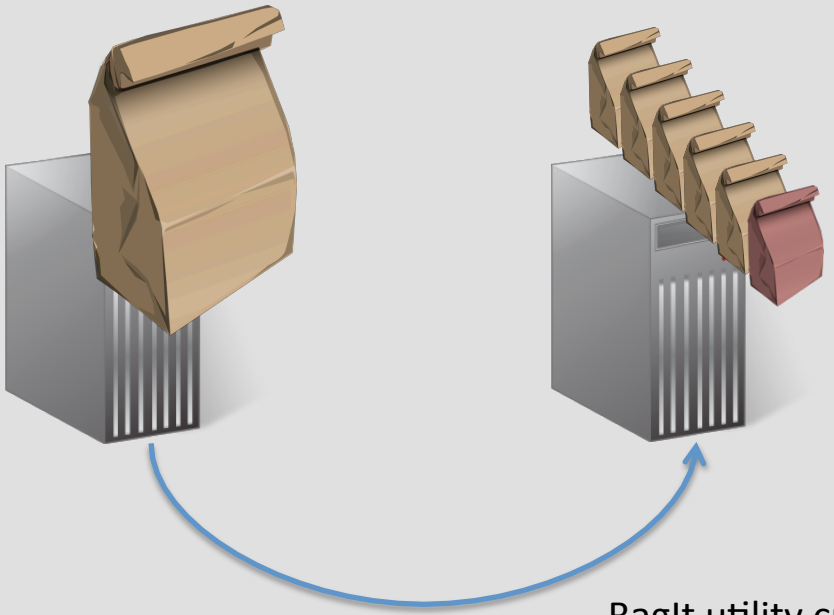
Exchanging Collections

- BagIt made it easy to group diverse collection data and package it with preservation value
- Each project partner bagged and sent 30-300GB of data according to BagIt usage instructions (made available in the project).
  - GUI was key
    - Partners preferred Bagger over bagit.py
  - Large bags require dedicated resources
    - Partners staging data on staff workstations ran the utility overnight in order to avoid interruptions
  - Bags require curation
    - BagIt utilities grab system files like .DS_store thumbs.db

Splitting Bags

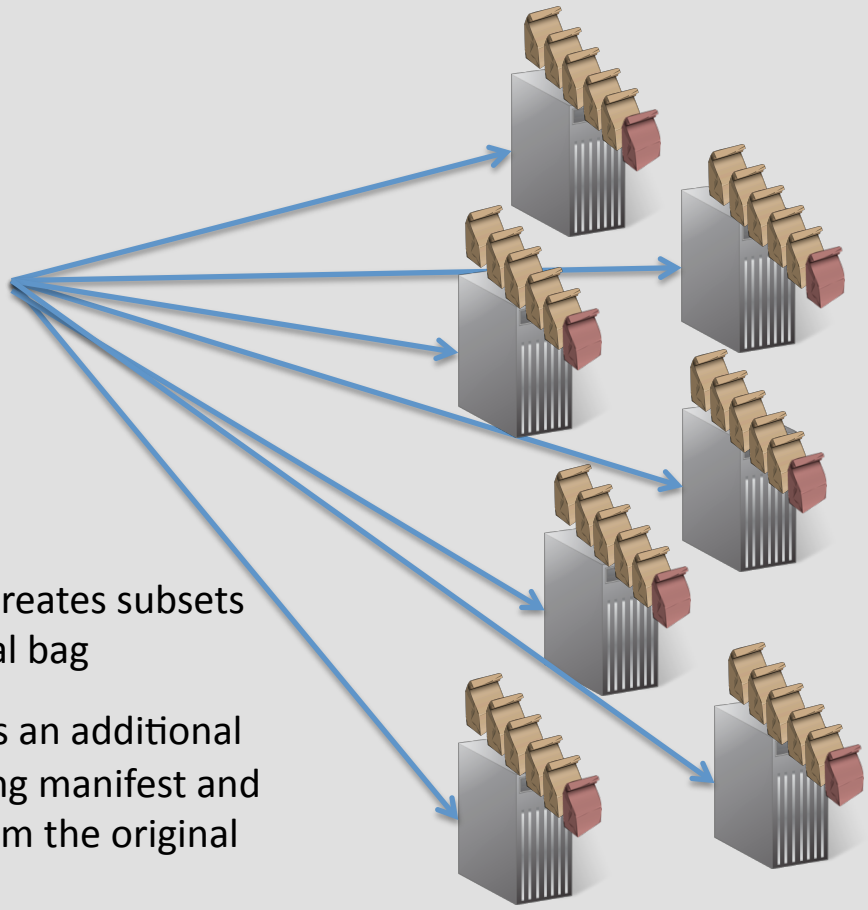
- For some systems, it can be helpful to break collections into manageable units in order to optimize checksum audit processes.
- The MetaArchive breaks collections into archival units (AUs) of 30GB.
- MetaArchive created a lightweight method of splitting and validating bags greater than 30GB and then reconstituting the original bag for export.

BagIt + Custom Scripts to Split



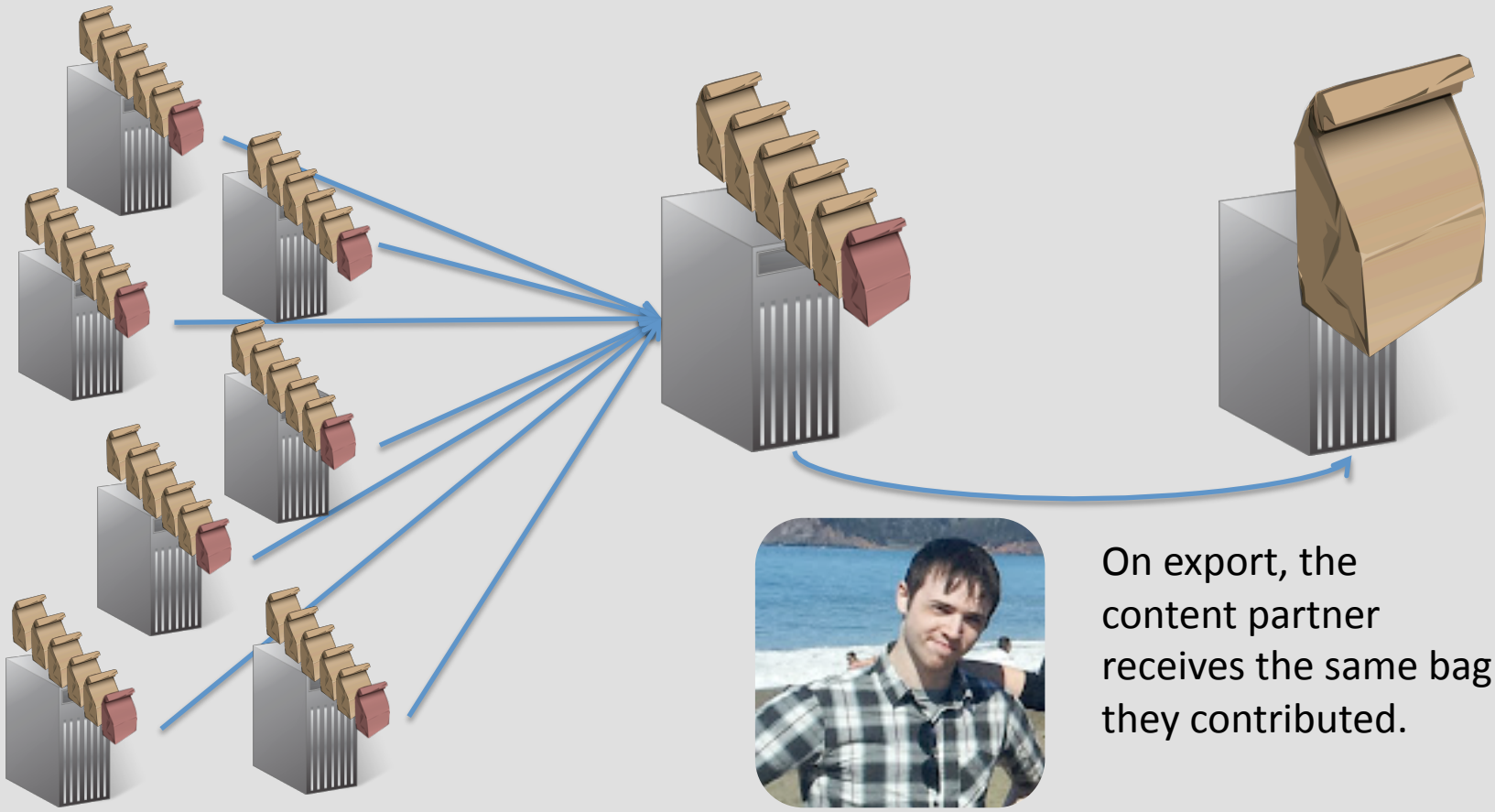
BagIt utility creates subsets of the original bag

Script creates an additional bag containing manifest and metadata from the original bag



EDUCOPIA INSTITUTE

…and to Rebuild



On export, the content partner receives the same bag they contributed.

EDUCOPIA INSTITUTE

Preservation Metadata for Objects

- Preservation metadata standards and specifications (METS/PREMIS) can be costly to implement
- Curators need lightweight and bulk applications to create and manage preservation metadata

- DAITSS Format Description Service
  - Web app that links DROID and JHOVE to create PREMIS
  - Released in 2009
  - https://github.com/daitss/describe

- UNT PREMIS Event Service
  - Web service to detect and log object events in an associated PREMIS file.
  - Available in 2014

Format Description Service

```
cd $BAG/data                                    # Enter the bag's "data" directory
find . -type f | while read line; do            # Go through each file found under this location
        mkdir -p "$BAG/premis/`dirname "$line"`" # Make a directory for the output file
        EXT=${line##*.}                         # Get the file's extension
         # Send the file to the description service (using its HTTP API) and save the output to a file
        curl -F "document=@$line" -F "extension=$EXT" $DESCRIBE_URL/description > "$BAG/premis/$line.xml"
done
```

EDUCOPIA INSTITUTE

PREMIS Event Service

Events

- <premis:event xmlns:premis="info:lc/xmlns/premis-v2">
- <premis:eventType>
- http://purl.org/net/meta/vocabularies/preservationEvents/#MigrateSuccess
- </premis:eventType>
- <premis:linkingAgentIdentifier>
- <premis:linkingAgentIdentifierValue>
- http://metaarchive.org/agent/metaMigrateSuccess
- </premis:linkingAgentIdentifierValue>
- <premis:linkingAgentIdentifierType>
- http://purl.org/net/meta/vocabularies/identifier-qualifiers/#URL
- </premis:linkingAgentIdentifierType>
- </premis:linkingAgentIdentifier>
- <premis:eventIdentifier>
- <premis:eventIdentifierType>
- http://purl.org/net/meta/vocabularies/identifier-qualifiers/#UUID
- </premis:eventIdentifierType>
- <premis:eventIdentifierValue>
- e8ee3b1a8c9e4a5daf0a1e0446383d90
- </premis:eventIdentifierValue>
- </premis:eventIdentifier>

Agents

- <?xml version="1.0"?>
- <premis:agent xmlns:premis="info:lc/xmlns/premis-v2">
- <premis:agentIdentifier>
- <premis:agentIdentifierValue>
- MigrateSuccess
- </premis:agentIdentifierValue>
- <premis:agentIdentifierType>
- FDsys:agent
- </premis:agentIdentifierType>
- </premis:agentIdentifier>
- <premis:agentName>
- http://metaarchive.org/agent/metaMigrateSuccess
- </premis:agentName>
- <premis:agentType>
- softw
- </premis:agentType>
- </premis:agent>

Meeting NDSA Metadata Levels

Green indicates fulfilled metadata requirements, red indicates metadata requirements not in scope

	Level 1	Level 2	Level 3	Level 4
Storage				
Fixity	BagIt			
Security				
Metadata	BagIt	BagIt/ Event Service		
Formats		Format ID		

Level 1 Fixity: Check or create fixity

Level 1 Metadata: Store object manifest

Level 2 Metadata: Administrative and transformative metadata

Level 2 Formats: Inventory file formats

EDUCOPIA INSTITUTE

Contacts & Links

- Matt Schultz (Program Manager, MetaArchive)
  matt.schultz@metaarchive.org
- Nick Krabbenhoeft (Project Manager, Educopia)
  nick@metaarchive.org
- Project URL: www.metaarchive.org/neh
- BagIt: http://sourceforge.net/projects/loc-xferutils/
- Description Service: http://description.fcla.edu/
- NDSA Levels: http://bit.ly/ndsa_levels

EDUCOPIA INSTITUTE