Chronicles in Preservation
Preserving Digital News & Newspapers

IFLA 2013, Singapore

EDUCOPIA INSTITUTE

Chronicles in Preservation

- About: NEH grant-funded study (2011-2014)
- Objective: To study, document, and model data preparation and distributed digital preservation (DDP) 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

Why Digital Newspapers?

- At-risk and valuable scholarly content genre
- Success of the United States Newspaper Program (USNP) & National Digital Newspaper Program (NDNP) – cataloging, digitizing, archiving & providing access to public domain newspapers
- Success of research carried out by Center for Research Libraries (CRL) in the U.S.
- Digitized and born-digital newspaper collections have been created with a variety of
  - standards
  - metadata
  - data models
  - technologies

EDUCOPIA INSTITUTE

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?

EDUCOPIA INSTITUTE

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

Guidelines: Overview

- Present *essential* and *optimal* actions
  - Essential – The minimum to be considered preservation, requires limited resources
  - Optimal – Best preservation for objects, requires more resources
- Based on:
  - Interviews with publishers, libraries, and vendors
  - Project experiences
  - Standards (e.g. METS, NDNP, OAIS)
  - Community feedback
    - Draft is available for public review

EDUCOPIA INSTITUTE

Guidelines: Modules

- Inventorying Digital Newspapers for Preservation
  - How to record what content an organization has and how it is stored
- Format Management for Digital Newspapers
  - How to identify, validate, and migrate formats
- Metadata Packaging for Digital Newspapers
  - How to choose metadata formats, export metadata from repositories, and manage the storage of metadata
- Checksum Management for Digital Newspapers
  - How to generate and monitor fixity information
- Organizing Digital Newspapers for Preservation
  - How to structure folder hierarchies and names
- Packaging Digital Newspapers for Preservation
  - How to organize a collection for ingest into a digital preservation system

Guidelines: Sample Module - Inventorying

- A single collection might have had multiple curators, acquisition strategies, storage locations, and file formats.
- Inventories are essential to record this information, to understand the collection, and then to plan preservation action.

EDUCOPIA INSTITUTE

Guidelines: Sample Module - Inventorying

- Essential
  - Tools: File manager such as Windows Explorer or Finder for Mac
  - Information:
    - Newspaper titles
    - Number of files
    - File locations
    - File names
    - Inventory creation date
  - Container: Human-readable formats such as a document or spreadsheet

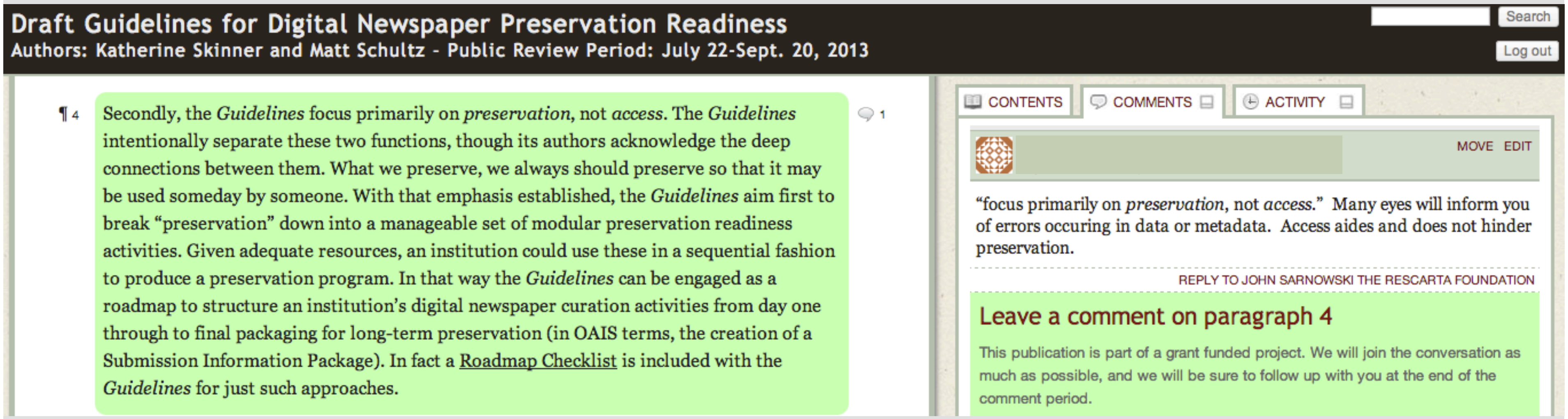
EDUCOPIA INSTITUTE

Guidelines: Sample Module - Inventorying

- Optimal
  - Tools: File manager and automated tools such as BagIt, PRONOM, or JHOVE
  - Information: Essential information and file formats, required application, checksums, and object identifiers
  - Container: Machine-readable formats such as a spreadsheet or database

EDUCOPIA INSTITUTE

Guidelines: Public Review



- http://publishing.educopia.org/chronicles
- We welcome comments and critique from the community to improve the Guidelines.

EDUCOPIA INSTITUTE

Comparative Analysis: Overview

- Three Distributed Digital Preservation (DDP) systems with 3 different infrastructures
  - MetaArchive – LOCKSS
  - Chronopolis – iRODS
  - UNT Coda – microservices
- Each library partner staged collections for DDP systems to document and analyze workflows with this type of content.

EDUCOPIA INSTITUTE

Tools: Guiding Principles

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

EDUCOPIA INSTITUTE

Tools: BagIt

- Digital newspapers have a range of legacy collection structures & conventions
- BagIt is a file packaging format for storing and transferring data. The data model includes:
  - A data directory
  - A manifest inventory of the bag with checksums for all objects within
  - Metadata about the bag
- BagIt is an IETF Internet Draft
  - http://tools.ietf.org/html/draft-kunze-bagit-09
- Bagger
  - Java-based BagIt tool w/ GUI
  - Released 2012
  - Maintained by Library of Congress
  - http://sourceforge.net/projects/loc-xferutils/files/loc-bagger/
- bagit.py
  - Python-based BagIt tool
  - Released in 2010
  - Maintained by Ed Summers at the Library of Congress
  - https://github.com/edsu/bagit

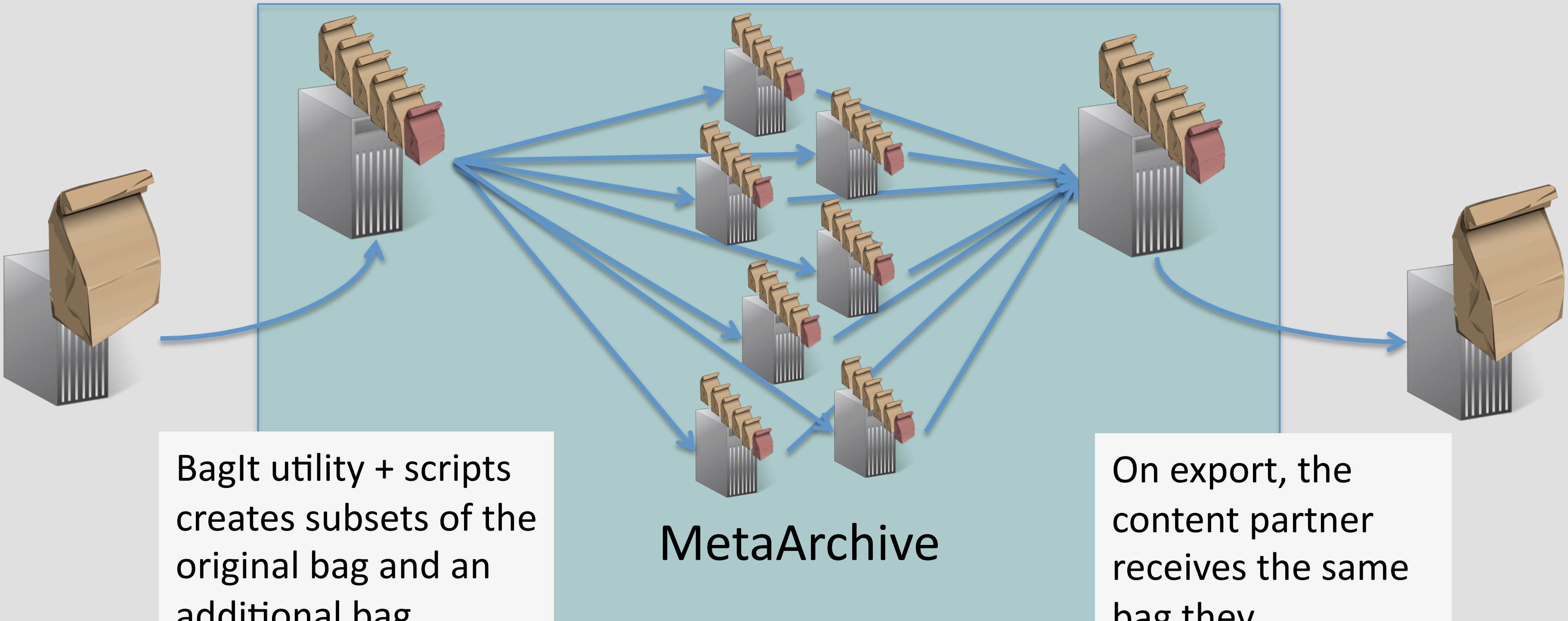
EDUCOPIA INSTITUTE

Tools: 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

EDUCOPIA INSTITUTE

Comparative Analysis: MetaArchive
BagIt + Custom Scripts to Split and Rebuild



BagIt utility + scripts creates subsets of the original bag and an additional bag containing manifest and metadata from the original bag

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

EDUCOPIA INSTITUTE

Tools: 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

EDUCOPIA INSTITUTE

Contacts & Links

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

EDUCOPIA INSTITUTE

Aligning National Approaches to Digital Preservation II

When: November 18-20, 2013

Where: Biblioteca de Catalunya (National Library of Catalonia), Barcelona, Spain

The Aligning National Approaches to Digital Preservation (ANADP) II Action Assembly will align digital preservation efforts internationally between communities—including national libraries, academic libraries, public libraries, research centers, archives, corporations, and funding agencies.

More Info: http://www.educopia.org/events/ANADPII

EDUCOPIA INSTITUTE