Chronicles in Preservation

Curation Practices for Born-Digital and Digitized Newspaper Collections

Katherine Skinner, Executive Director, Educopia Institute
Martin Halbert, Dean of Libraries, University of North Texas
Tyler Walters, Dean of Libraries, Virginia Tech

CNI 2012 Spring Membership Meeting
Baltimore, MD
April 3, 2012
Presentation Overview

- Chronicles Project background
- State of the Field report
- Early Findings
One day, through the primeval wood,
A calf walked home, as good calves should;
But made a trail all bent and askew,
A crooked trail, as all calves do.

Since then three hundred years have fled,
And, I infer, the calf is dead.
But still he left behind his trail,
And thereby hangs my moral tale.

The years passed on in swiftness fleet,
The road became a village street;
And this, before men were aware,
A city’s crowded thoroughfare;
And soon the central street was this,
Of a renowned metropolis;
And men two centuries and a half,
Trod the footsteps of that calf.

And from that day, o’er hill and glade,
Through those old woods a path was made,
And many men wound in and out,
And dodged and turned and bent about,
And uttered words of righteous wrath
Because ‘twas such a crooked path;
But still they followed — do not laugh —
The first migrations of that calf.

This forest path became a lane,
That bent, and turned, and turned again.
This crooked lane became a road,
Where many a poor horse with his load,
Toiled on beneath the burning sun,
And traveled some three miles in one.
And thus a century and a half,
They trod the footsteps of that calf.

Each day a hundred thousand men were led
By one calf near three centuries dead.
They follow still his crooked way,
And lose one hundred years a day,
For thus such reverence is lent
To well-established precedent.
Chronicles, an introduction

Educopia Institute-led partnership, comprised of the following:

**Preservation groups**
- MetaArchive (LOCKSS)
- Chronopolis (iRODS)
- University of North Texas (CODA)

**Content Curators**
- Penn State
- Virginia Tech
- University of Utah
- Georgia Tech
- Boston College
- Clemson University
- University of Kentucky

Funded by:

Skinner, Halbert, and Walters 2012
To study, document, and model the use of data preparation practices and distributed digital preservation frameworks to collaboratively preserve digitized and born-digital newspaper collections.
Three DDP frameworks

- MetaArchive
  - Founded 2004, 50+ members in 3 countries
  - Multi-node, wide distribution of content
- Chronopolis
  - 3-node system (SDSC, NCAR, UMIACS)
- CODA
  - Developing multi-node framework based on a micro-services approach
Two types of content

Born Digital

Digitized

Skinner, Halbert, and Walters 2012
Research Questions

- How can curators effectively and efficiently prepare their existing digitized and born-digital newspaper collections for preservation?
- How can curators ingest preservation-ready newspaper content into existing DDP solutions?
- What are the strengths and challenges of three leading DDP solutions when used to preserve digital newspaper content?
Project Deliverables

- Guidelines to Digital Preservation Readiness
- Interoperability Tools
- Comparative Analysis of DDP Frameworks
State of the Field

- Early findings based on the following surveys:
  - 2008 ETD Preservation Survey (VT-NDLTD)
  - 2009 Digital Preservation Needs Survey (NHPRC)
  - 2011 Digital Preservation SPEC Kit 325 (ARL)
  - 2011-12 Chronicles Survey (8 academic libraries)
ETD and NHPRC surveys

- Readiness is low. Desire is high.
  - >70% had NO preservation plan.
  - >25% were not even backing up
  - almost none engaged in active preservation

Skinner, Halbert, and Walters 2012
Threats identified by respondents

2008-2009 survey results

Skinner, Halbert, and Walters 2012
State of the field: 2011

- SPEC Kit #325: Digital Preservation (ARL)
  - Types of content
    - ~100% ETDs, images, special collections
  - 80% preserve some now; all but 4% plan to. Top barriers?
    - Lack of experienced staff
    - Lack of funding
    - Institutional policies and strategies

Skinner, Halbert, and Walters 2012
Chronicles Project Survey

- **Type**
  - NDNP: 18; non-NDNP: 459; born digital: 19

- **Image formats**
  - TIFF, JP2, PDF, HTML, TXT, XML

- **Metadata formats**
  - METS/ALTO, MIX, MODS, PREMIS

- **OCR formats**
  - METS, ALTO, PDF, Abbyy, XML, PRIME OCR.pro
Chronicles Project Survey (cont)

- Object identifier schemes
  - Fedora PID, Handles, Veridian and CONTENTdm custom URLs, ARKs
  - All but two are internal to the repository system

- Validation
  - \( \frac{1}{2} \) use JHOVE at least for some content

- Versioning
  - Only one institution
Chronicles Project Survey – Findings (cont.)

- Access and storage systems
  - Access: local, hosted, open, & proprietary
    - e.g., Fedora, Dspace, Olive, Veridian, CODA, web-server
  - Masters: e.g., SAN, tape, hard-drive

- Preferred ingest mechanisms
  - Secure FTP or “Frisbee-net”
VA Tech Case Study

- VA Tech - starting with the essential
  - Well entrenched in the calf-path
  - “diverse and un-normalized legacy” collections
  - the “born-digital dilemma” institution
  - extensive Data Wrangling experience
  - Hosting e-news since 1997
    - HTML 4.0, PDF 1.1
    - Metadata?
  - Outside NDNP recommendations

Skinner, Halbert, and Walters 2012
WELCOME TO CLUB BIO MOUNTAIN LAKE
BIOLOGICAL STATION IS A SCIENTIFIC HAVEN
FOR STUDENTS AND TEACHERS

Call it Club Bio.

Nestled within the woods of Salt Pond Mountain, the Mountain Lake Biological Station looks almost like a rustic resort with its small cabins and casually clad inhabitants. Most of the cabins are named after Southern scientists, however, and research - not a dip in the lake - is of primary concern.

For nearly seven decades, the biological station has served as a scientific haven for students and teachers from throughout the
SLUG: print


MODIFIED:
BY: TIME: STATUS: 
06/04/08 News 7 at 11 

NATASHA ON CAM NATASHA ON CAM 

GRAPHIC=Clinton quitting Good evening I'm Natasha

Ryan.

01 11 Campaign 2008 P pkg Clinton, quitti 3104 --
MODIFIED: MOD BY: ED. BTAPE TIMECODE WR BACKTIMEREAD:SOT: TOTALEP VIDEO -
06/04/2008 1alawson mmo 1381:44 0:25 1:23 1:48 11 Ca-- NATASHA ON CAM NATASHA ON CAM

GRAPHIC=HOLD Developments in the Democratic

Race for President -- Barack Obama has
more than the needed number of
delegates for his party's nomination...
What strategies help to improve and optimize newspaper digitization workflows?

Avoiding the calf-path requires a willingness to re-examine workflow and impose discipline.

Normalization is required for all incoming content – including newspapers.

Digitizing and preserving to current standards, using local flavors.

Builds off NDNP foundations.

Skinner, Halbert, and Walters 2012
State of the field example: UNT Portal to Texas History

- Relatively large scale and streamlined state digitization project (2.5M files, 186K serials/titles, now used 275K times/month)
- Digitizes content from 220 libraries and museums across Texas
- Strong ties to state educational groups and learning standards
- Much of the portal was created through NDNP funding streams
- Part of the much larger UNT Digital Library
- Micro-services modular system architecture based on open standards
Challenges and Early Findings

- Back-up vs. preservation
- Adoption of existing standards
  - e.g., OCR, metadata
- Lack of standards
  - e.g., file structures, naming identifier schemes
- Diverse array of expectations
  - very institution-specific
- Versioning processes will be necessary
  - e.g., for growing, changing, and/or remediated projects

Skinner, Halbert, and Walters 2012
Contact

Martin Halbert
martin.halbert@unt.edu

Katherine Skinner
katherine.skinner@metaarchive.org

Tyler Walters
tyler.walters@vatech.edu