The Great Migration: Moving First Generation Digital Texts to HathiTrust

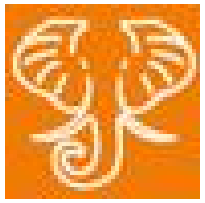
July 22, 2014 - Digital Preservation 2014

Lance Stuchell and Kat Hagedorn, University of Michigan Library

#digpres14

Overview

We've been migrating our older collections to HathiTrust for 5 years now (since 2009). Recently, we've been taking a closer look at correctly preserving this content. We've been doing that for the past 2 years (since early 2012).



#digpres14

Overview

DLXS: (Digital Library eXtension Service)

- Includes content from 1995 to present
- Images, texts, bibliographies, finding aids

HathiTrust:

- Consortium of research institutions and libraries
- Content includes material created through Google digitization project

Two repositories

Some things are similar:

- File formats (TIFF, JPEG2000), structure of repository, repository management staff
- Content to be migrated (printed volumes)

However:

- HathiTrust has stricter technical requirements
- HathiTrust has a more formal ingest process

Our team

Our team at U-M includes:

John Weise - Head of the Digital Library Production Service (DLPS)
Cory Snavely - Head of the Core Services (infrastructure) unit
Aaron Elkiss - Systems Programmer in Core Services
Chris Powell - Coordinator for Text Collections in DLPS
Matt LaChance - Data Processing Automation Programmer in DLPS
Lance Stuchell - Digital Preservation Librarian
Kat Hagedorn - Project Manager for Digital Projects in DLPS

The Weasley House effect

When we started, we were balancing pragmatism with best practices and a thoughtful approach.
We didn't have good validation tools, because they didn't exist.
We got smarter as we went along, and we got better tools to help us.
But it means we don't have consistency across all our collections.

The pain point

95% of our materials can be ingested easily - they pass current validation with little or no problems.
However, that remaining 5%...

It’s not really broken

It’s wrong to call these broken or unbroken - they don’t meet our standards of preservation as-is.

Problem	*Details*	*Solution*
Bitonal image resolution	Resolution value is incorrect or zero	Manually fix the images
Invalid but well-formed bitonal images	Images are viewable, but the image metadata is incorrect	Automation for those can batch fix, manual intervention for the rest
Unexpected contone image dimensions	Some images in a volume are suspiciously not a reasonable size, in relation to the other pages	Automate discovery mechanism, but in the end manually inspecting and annotating
“Funny” filenames	Filenames that met a previous spec don’t meet current spec	Automate discovery, and automatically fixed (pattern matching)
Bad/ambiguous dates	e.g., 4-5-98, 040506	Automate discovery and fix of patterns, but some will need manual fixing
Legitimate sequence skips	Skips in the numbering or nomenclature of a sequence (not the page)	Automate discovery and most fixes, some manually rename

It’s really broken

Problem	*Details*	*Solution*
Blank contone images	Contone is a blank image	Need to reload/rescan the volume
Contone images don’t match bitonals	Misaligned (sequences) of contones and bitonals within the volume	Manual work to discover the correct matches, then make those changes
Skipped pages	Illegitimate skips	Need to reload/rescan the volume
OCR but no images	OCR for certain sequences, but no matching images	After discovery, may need to rescan some images, but some images may not be needed
Character errors	e.g., em-dashes, control characters not recognizable by XML	After discovery, remove/replace those characters
Non-viewable, malformed images	Cannot open, view or discover the problems with certain images	“Cut bait” (rescan the entire volume)

The “note from mom” tool

We needed a way to indicate which volumes we were ingesting because a thoughtful analysis didn’t require a fix, but a method for noting our analysis prior to ingest. e.g.,

- legitimate printed volume errors (missing pages)
- those unexpected contone image dimensions (if they look good on manual inspection)

We built an ingest utility - aka the “note from mom” - to ingest these volumes.

“Final” result?

Of the 167,102 volumes in our text collections, we have validated and migrated 25,339 of them to HathiTrust (15%).

- 18 collections (mostly) ingested (20%), 31 next on the docket
- 41 cannot be currently migrated

It took us most of the past 2 years to validate and ingest these volumes.

Full disclosure

Categories that need special help:

- Some volumes we deemed okay to ingest, but they need to be revisited (reduced validation).
- Those collections with permission or ownership questions.
- Those collections that we built to contain contones we couldn't include in our text collections.
- Those collections that contain more highly-encoded volumes (level 4 TEI).

Some collections can't be migrated because they are licensed, don't have page images, etc.

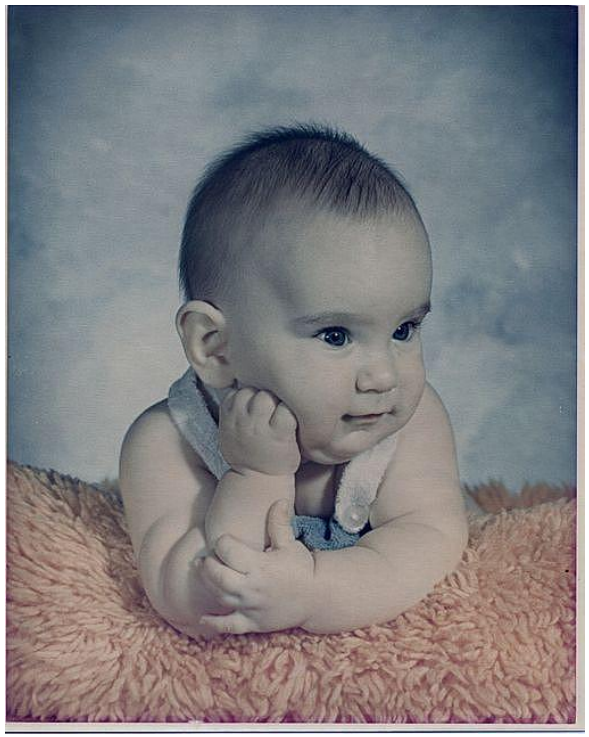
Hard, but intriguing

We have been:

- intrigued - by the extent of certain problems
- annoyed - by image reloads that seem to make no sense
- surprised - at the differences between latter years and today in terms of preservation tools and expertise

But in the end, this is fun! It just takes a lot of time, so we encourage you to start now.

Opportunity for reflection



Pensive_John by DiscourseMarker

DLXS collections were digitized with preservation in mind

Migration project offers an opportunity to evaluate how we did, where we have gone, and where we are now

#digpres14

How did we do?

- Small percentage of files are “broken,” non-assessable, or require rescanning
- Most errors happened in digitization and/or package creation - we’re not sure
  - Not caused by degrading in the repository
  - Not caused by anything inherent in the file formats
- Most errors discovered during migration

Way back in 1995...

- Content was created with long term preservation in mind, but…
  - Ingest validation was minimal
  - Relied on vendors creating correct content [pause for laughter]
  - Metadata that was to become engine of management at scale was inconsistent or incorrect

Road from then to now

- Shift from “hand-crafted” to mass digitization led to unprecedented scale
- Further reliance on tools facilitating automated processes like ingest (JHOVE, etc.)
- Increased use of standards (METS, PREMIS, adoption of OAIS and TRAC concepts)
- Effort to move away from anecdotal knowledge of issues with content

Are we better now?

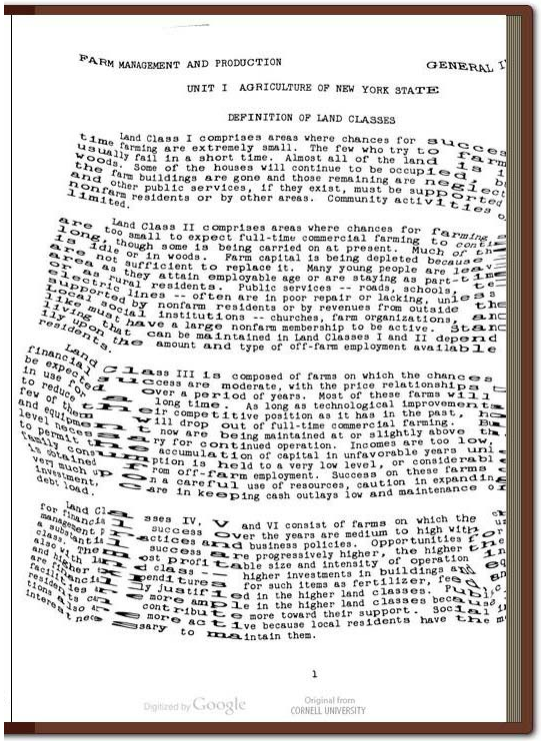
- Processes allows for consistent metadata creation and content validation at scale
- Use of standards accommodates large scale repository changes (METS & PREMIS uplift)
- Using PREMIS and local metadata for more documentation within the repository
- Content is much more consistent

Future issues: The outliers

outlier by Robert S. Donovan

- Necessity of documenting heterogeneous material increases along with scale
  - “Note from mom”
- How to document complicated special cases?
  - Shift from anecdotal knowledge is not complete

Future issues: tool dependance



FARM MANAGEMENT AND PRODUCTION GENERAL

UNIT I AGRICULTURE OF NEW YORK STATE

DEFINITION OF LAND CLASSES

Land Class I comprises areas where chances for success
time farming are extremely small. The few who try to farm
usually fail in a short time. Almost all of the land is i
woods. Some of the houses will continue to be occupied, b
the farm buildings are gone and those remaining are neglect
and other public services, if they exist, must be supported
nonfarm residents or by other areas. Community activities
limited.

Land Class II comprises areas where chances for farming
are too small to expect full-time commercial farming to contin
long, though some is being carried on at present. Much of th
is idle or in woods. Farm capital is being depleted because
area not sufficient to replace it. Many young people are leav
area as they attain employable age or are staying as part-time
electric rural residents. Public services -- roads, schools,
supported by nonfarm residents or by revenues from outside
lines -- often are in poor repair or lacking, unless
local social institutions -- churches, farm organizations, and
like must have a large nonfarm membership to be active. Stand
ing up that can be maintained in Land Classes I and II depend
residents. the amount and type of off-farm employment available

Land Class III is composed of farms on which the chances
financial success are moderate, with the price relationships
be especially over a period of years. Most of these farms will
in use reduced long time. As long as technological improvements
to reduce their competitive position as it has in the past, ho
few of them will drop out of full-time commercial farming. Ba
and equipment now are being maintained at or slightly above tha
level necessary for continued operation. Incomes are too low
to permit accumulation of capital in unfavorable years unl
adequate consumption is held to a very low level, or considerabl
usually obtained from off-farm employment. Success on these farms
is obtained by careful use of resources, caution in expanding
very much in keeping cash outlays low and maintenance o
investment,
debt load.

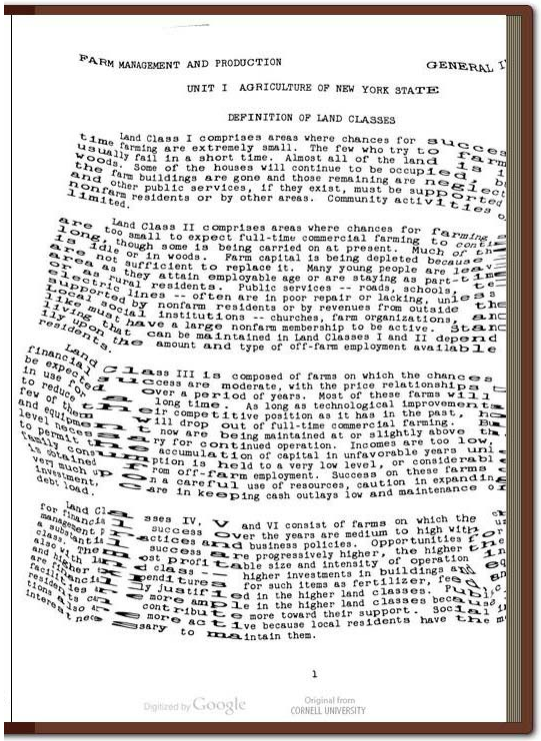
Land Classes IV, V and VI consist of farms on which the
for financial success over the years are medium to high with
management practices and business policies. Opportunities for
assistance success are progressively higher, the higher the
class. Land VI, the most profitable size and intensity of operation
and with the higher class -- higher investments in buildings and
are high expenditures for such items as fertilizer, feed
other than justified in the higher land classes. Public
facilities are more ample in the higher land classes because
residents also contribute more toward their support. Social
is can more active because local residents have the
interest necessary to maintain them.

1

- Some significant properties do not have perfect validation methods
- Scale and access demands make manual review impossible
- There are a variety of methods to detect and fix these after ingest, including re-scanning as a “last-resort”

#digpres14

Future issues: tool dependance



FARM MANAGEMENT AND PRODUCTION GENERAL

UNIT I AGRICULTURE OF NEW YORK STATE

DEFINITION OF LAND CLASSES

Land Class I comprises areas where chances for success in farming are extremely small. The few who try to farm usually fail in a short time. Almost all of the land is in woods. Some of the houses will continue to be occupied, but the farm buildings are gone and those remaining are neglected and other public services, if they exist, must be supported by nonfarm residents or by other areas. Community activities are limited.

Land Class II comprises areas where chances for farming are too small to expect full-time commercial farming to continue long, though some is being carried on at present. Much of the land is idle or in woods. Farm capital is being depleted because the area is not sufficient to replace it. Many young people are leaving as they attain employable age or are staying as part-time rural residents. Public services -- roads, schools, electric lines -- often are in poor repair or lacking, unless supported by nonfarm residents or by revenues from outside the area. Local social institutions -- churches, farm organizations, and like must have a large nonfarm membership to be active. Standing must be maintained in Land Classes I and II depending upon that can be maintained by nonfarm residents. The amount and type of off-farm employment available

Land Class III is composed of farms on which the chances for financial success are moderate, with the price relationships to be especially favorable over a period of years. Most of these farms will succeed in use for a long time. As long as technological improvements reduce the few of their competitive position as it has in the past, many will drop out of full-time commercial farming. Buildings and equipment now are being maintained at or slightly above the level necessary for continued operation. Incomes are too low to permit adequate accumulation of capital in unfavorable years unless consumption is held to a very low level, or considerable additional income from off-farm employment. Success on these farms is obtained by a careful use of resources, caution in expanding investment, and care in keeping cash outlays low and maintenance of debt load.

Land Classes IV, V and VI consist of farms on which the chances for financial success over the years are medium to high with management practices and business policies. Opportunities for success are progressively higher, the higher the class. The most profitable size and intensity of operation and the higher the class -- higher investments in buildings and equipment are justified in the higher land classes. Public expenditures for such items as fertilizer, feed, etc. are more ample in the higher land classes because residents are more active because local residents have the contribute more toward their support. Social interest necessary to maintain them.

1

- What happens in cases without rescanning as a “last resort”?
  - Born digital
  - File format migration
  - Degraded A/V material
- How does community prioritize tool development?

#digpres14

The punch line (for me)

Preservation is iterative, not static

- Formats have (so far) been very stable
- Metadata and supporting elements evolve
  - Tools to facilitate automated processes
  - METS and PREMIS uplifts
  - “note from mom,” etc.
- Constant decision making to balance real-world constraints with preservation ideals