Free Books:
Hathi Trust Cataloging Records

Roman S. Panchyshyn
Assistant Professor: Cataloging Librarian
Kent State University

OVGTSL

May 16. 2011

Introduction

- Goal--to review the project Kent State University Libraries undertook to load public domain records from Hathi Trust into KentLINK and OhioLINK Central
- Examine
  - Planning
  - Execution
  - Costs
  - Results
  - Future issues

The Elephant in the Catalog is …

Beginnings

- Late in 2009, OhioLINK Database Management and Standards committee (DMS) received a request from OhioLINK Reference and User Services to:
  - Ask DMS to create/adapt/load records for 5 openly available e-book collections: Hathi Trust (Google Book Project), Internet Archive, National Academies Press, Policy Archive, Project Gutenberg

Inspiration

- Also in 2009, Jeffrey Beall from the University of Denver published the following article”
  - Beall, J. (2009). Free Books: Loading Brief MARC Records for Open-Access Books in an Academic Library Catalog. *Cataloging & Classification Quarterly*, 47(5), 452-463.
- Article detailed how Auroria Library (UD) loaded 100,000 brief MARC records into their Innovative system for public domain materials from the Hathi Trust in 2008

DMS Approval

- In October 2009, DMS asked that Kent State investigate loading these groups of records
- Upon investigation, it was found that Hathi Trust public domain records and Project Gutenberg records had free MARC records that were easy to obtain
- Mike Kreyche, Systems Librarian, and I, picked up this task

MarcEdit

- MarcEdit, a cataloging utility tool developed and maintained by Terry Reese, was one of the main tools used for this project. See:

http://people.oregonstate.edu/~reeset/marcedit/html/index.php

- MarcEdit was useful for:
  - Harvesting MARC records for these projects
  - Batch editing MARC records in preparation for loading

Hathi Trust: Preliminary Test Results

- November 2009
  - KSU harvested 205,379 US public domain records, 88,534 other public domain records from Hathi
- Problems
  - Just as Beall experienced, we encountered poor MARC record quality and incomplete data
- According to Beall, data was stripped from the records to “fulfill the requirements of OCLC’s copyright on the records”. (Beall, p. 460)

Missing Data

- A review of the data revealed what was intentionally left out (MARC tags):
  - 100 & 600 $c and often $d
  - 130 & 240 uniform titles (all)
  - 245 $c
  - 300 $b and $c
  - 440, 490-830 (all)
  - 5XX (all)
  - 7XX (all)
- This presented significant issues for authority control and access

Sample Harvested Record

=LDR nam a22002291 4500

=005 19890623000000.0

=010 \\$a04034555

=035 \\$a(OCoLC)ocm03823852

=035 \\$asdr-wu6389926

=050 0\$aLB1541$b.H7

=100 1\$aHolton, Martha Adelaide.

=245 00$aIndustrial work for public schools,

=260 \\$aChicago,$aLondon [etc.]$bRand, McNally & company$c[1904]

=300 \\$a134 p.

=650 \0$aManual training

=856 4\$uhttp://hdl.handle.net/2027/mdp.39015059773377rpdxeContent

Decisions From First Test

- Pros
  - We can provide access to these materials at minimal cost to KentLINK and to Central
  - We can provide these records to other OhioLINK Libraries on demand if requested
- Cons
  - Poor record quality, major implications for quality of KentLINK and Central database
  - Major implications for authority control (cost) and for access (split files).

Outcomes

- OhioLINK DMS and KSU felt at that time that the quality of the Hathi records was too poor for the catalogs
- If we used the records, the cost of doing authority work for these records was estimated at $13,000 by Backstage. Made no sense to pay for authority control for records that lacked many access points

Finding Better Records

- Two ways to get records from Hathi Trust
  - OAI-PMH harvesting (poor quality records)
  - Bib API (Application protocol interface). Hathi had developed an API that returns bibliographic, copyright, and volume information from its catalog (including permanent URLs) when queried with a variety of standard identifiers (e.g., ISBN, LCCN, OCLC, etc.).
  - The API has controls to return brief or full bibliographic metadata.

Bib API

- Libraries could use Bib API to retrieve full records from the Hathi database
- Limited to 20 records per search
- One needs some sort of record ID number for exact match (ISBN, OCLC number, etc…)
- Records are returned in MARC-XML in JSON (Java Script Object Notation) format

Reexamine OAI-PMH data

- Each record retrieved via OAI-PMH had a unique header number & volume number in the 856 that could be used by the API to retrieve the full record
- KSU was able to use a PERL script to formulate an API search for each record using the following syntax **http://catalog.hathitrust.org/api/volumes/full/<id type>/<id value>.json**
- Id type = recordnumber id value = 000003198

Permission

- Use of the Bib API would allow us to get complete bib records from the Hathi catalog
- We requested permission from the University of Michigan to take records via API. They stated they did not have any current guidelines or quotas for the use of the API.
- As long as we did nothing to impact performance, we were free to use the API, but they asked us to limit search to one record at a time

Progress

- By April 2010, we began to download the Hathi records via API.
- By that time, the number of records in the public domain files had risen to over 471,000
- It took almost 4 weeks to retrieve good quality records from the Hathi catalog and convert them to MARC from MARC-XML

Individual Record Preparation

- Records needed to be prepared for
  - Loading into KentLINK following local standards
  - Loading into Central following DMS standards
- We held consultations with public services over display and indexing issues
- We presented DMS with lists of modifications we would make to the records for addition to Central
- Plan was to load them first into KentLINK, then later into Central, after any major issues resolved

Record Preparation

- KSU found it necessary to prep the records before loading into KentLINK
- KSU Cataloging Committee reviewed and approved the needed changes

Changes to Hathi Records

- OAI header number was paced in the 001, and a prefix, conforming to DMS standards, was added (e.g. 1MIU000003162)
- Prefix (OCoLC) was added to any records that contained an OCLC number for the print version in the 035
- Cat date at this point was set to blank to prevent harvesting for authority control

Changes …

- 006: type = m, file = d for every record
- 007 $a = c, $b = r for every record (minimum)
- Added GMD 245 $h[electronic resource]
- Added collocating field

  730 0 $a Hathi Trust collection ;$nMIU001724414

($n is unique identifier, the 001, used for indexing)

- Deleted $r and $x in 856, added “$z Connect to resource” as public note
- Changed Mat Type to “3” ebooks (later went back and changed all Bib lvl “s” to Mat type “s”

MarcEdit

- We trained a practicum student in the use of MarcEdit to save us some work
- Using MarcEdit and regular expressions, we were able to make all the changes to the Hathi records in batch
- We also cleaned up as many diacritics as possible
- Final count: 471,950 public domain records

Sample Authorities Test Results: 1XX Fields

Total Records	**1004**	
		Percentage
Records that have 1xx	**951**	
Matched on KentLINK	**367**	**39%**
Match see reference on KentLINK	**10**	**1%**
Matched a single heading on OCLC	**284**	**30%**
TOTAL THAT MATCHED ON OCLC OR KENTLINK	**661**	**70%**
Matched multiple authority records on OCLC	**50**	**5%**
No match against the OCLC Authority file	**200**	**21%**
Could not search because of invalid terms, diacritics, syntax, etc.	**40**	**4%**
TOTAL THAT DID NOT MATCH	**290**	**30%**

Authority Recommendations

- Sample test of 1004 records showed that 70% of records having a 1XX field had a match in the KentLINK or OCLC Authority File
- A no-match rate of 30% bolstered the argument to send these records out for authority control
- Asked authority vendors for quotes. Quote for Hathi records, plus Gutenberg, was over $17,000 by Backstage (cheapest). KSU was willing to foot this bill for OhioLINK.

Loading to KentLINK

- Began loading Hathi records to KentLINK only in June 2010 (Display code = “z”)
- Loading was complete by the end of July 2010
- Records were added in small batches to avoid indexing and performance issues
- DMS was informed at the August 2010 DMS meeting that our plan was to first send the records out for authority control, then load to Central when they return

Authority Processing

- Entire Hathi file, and Project Gutenberg, was sent to Backstage in Sept. 2010 for clean-up
- 505,276 bibliographic records were processed by Backstage:
  - 88% books
  - 6.6% computer files
  - 5% continuing resources
  - .4% other

Authority Work Summary

- 505K bibliographic records processed
- 99.9% of the records were changed in some way
- 510K access fields changed
- 488K title fields not under authority control changed
- 1.9 million miscellaneous changes
- 115K new authority records added to KentLINK
- 18K bib records automatically flipped by III when new authority records added
- Cost $18K

Good quality records were now available for Central

Reloading

- In late Sept. 2010, the Hathi records returned from Backstage and were overlaid locally using the .b number (Innovative record number)
- Cat date was reset/backdated to 05/31/10
- No holdings were set on OCLC
- Late Oct. 2010, after conversations with OhioLINK, we began loading the Hathi records to Central, slowly, to not affect performance
- Nov. 2010, all Hathi records completed loading at Central and OhioLINK libraries were offered a copy of the file on demand

Sample of Complete Hathi Record

- =LDR 00984nam 22002891 4500
- =001 1MIU001145362
- =005 20100902081840.0
- =006 m\\\\\\\\d\\\\\\\\
- =007 cr\bn\---auaua
- =008 890623s1886\\\\enk\\\\\\\\\\\000\0\eng\d
- =035 \\$a(OCoLC)ocm00503093
- =040 \\$cCarP$dUtOrBLW
- =043 \\$ae-uk---
- =050 00$aJN508$b.G7
- =100 1\$aGneist, Rudolph,$d1816-1895.
- =245 04$aThe English Parliament in its transformations through a thousand years$h[electronic resource]
- =260 \\$aLondon :$bH. Grevel & co.,$c1886.
- =300 \\$axxvii, 380 p.$c23 cm.
- =538 \\$aMode of access: Internet.
- =610 10$aGreat Britain.$bParliament$xHistory.
- =700 1\$aShee, Richard Jenery,$d1826-$etr.
- =730 0\$aHathi Trust collection ;$nMIU001145362.
- =856 40$uhttp://hdl.handle.net/2027/mdp.39015009315097$rpdus$zConnect to resource

Project Evaluation

- We fulfilled, over time, and with some expense, DMS's request to add public domain MARC records for Hathi Trust materials to the Central catalog
- We felt the records were of acceptable quality for use by the consortium
- We obtained valuable experience working with large data sets using MarcEdit
- We are now in the process of evaluating KSU's and OhioLINK's decision to load these records

Issues with the Hathi Trust Project

- Following is a recap of issues that have been raised about this project over the last 6 months.
- Some of these issues will have a significant impact on the future of this project
- They will be covered in no specific order

Display in Title Indexing

- Display of title searches under ”Hathi Trust” in KentLINK and Central varies.
- Central displays both the 730 and 245, allowing users to identify the item by title
- KentLINK displays the 730 only. User cannot differentiate by title.
- KSU is considering modifying local display to bring it in line with Central.
- Also we are considering moving the 730 to an unindexed field

Public Services Comments

- Some public services librarians feel that the addition of 471,000 Hathi records has made the database too cluttered.
- There are still issues with the quality of some of the MARC records. One example is that it is difficult to identify originals vs. reproductions, since the dates in many of the records are not consistent
- The question has been asked:Should the records just be displayed at Central only, and not locally?

Weeding

- Since many of the records contain OCLC numbers for the print editions, Hathi records can be compared against local collections to identify duplicates and potential weeding candidates.
- KSU has a project underway to analyze this overlap. This could also be done on a consortial level.

Future Updates

- Public domain materials in Hathi are growing exponentially.
- We loaded 471,000 records. There are now over 800,000 items in public domain.
- Do we continue to add new MARC records to the catalogs, and absorb the costs?
- Should we just point users to the Hathi database interface on the website?
- Will the addition of a discovery layer make this project irrelevant?

Usage Statistics

- At this time, both locally and centrally, we are not able to obtain accurate usage statistics for Hathi Trust materials.
- This makes it even harder to justify the expense for the project

Future of the Hathi Trust

- Hathi Trust itself is undergoing organizational changes. It is restructuring to a membership-based organization.
- Will they continue to support free access to public domain materials?
- Will they continue to provide API access to MARC records?
- Will their partnership with OCLC impact availability of records?
- Will URLs remain static or change?
- The Google Book project status is still in court

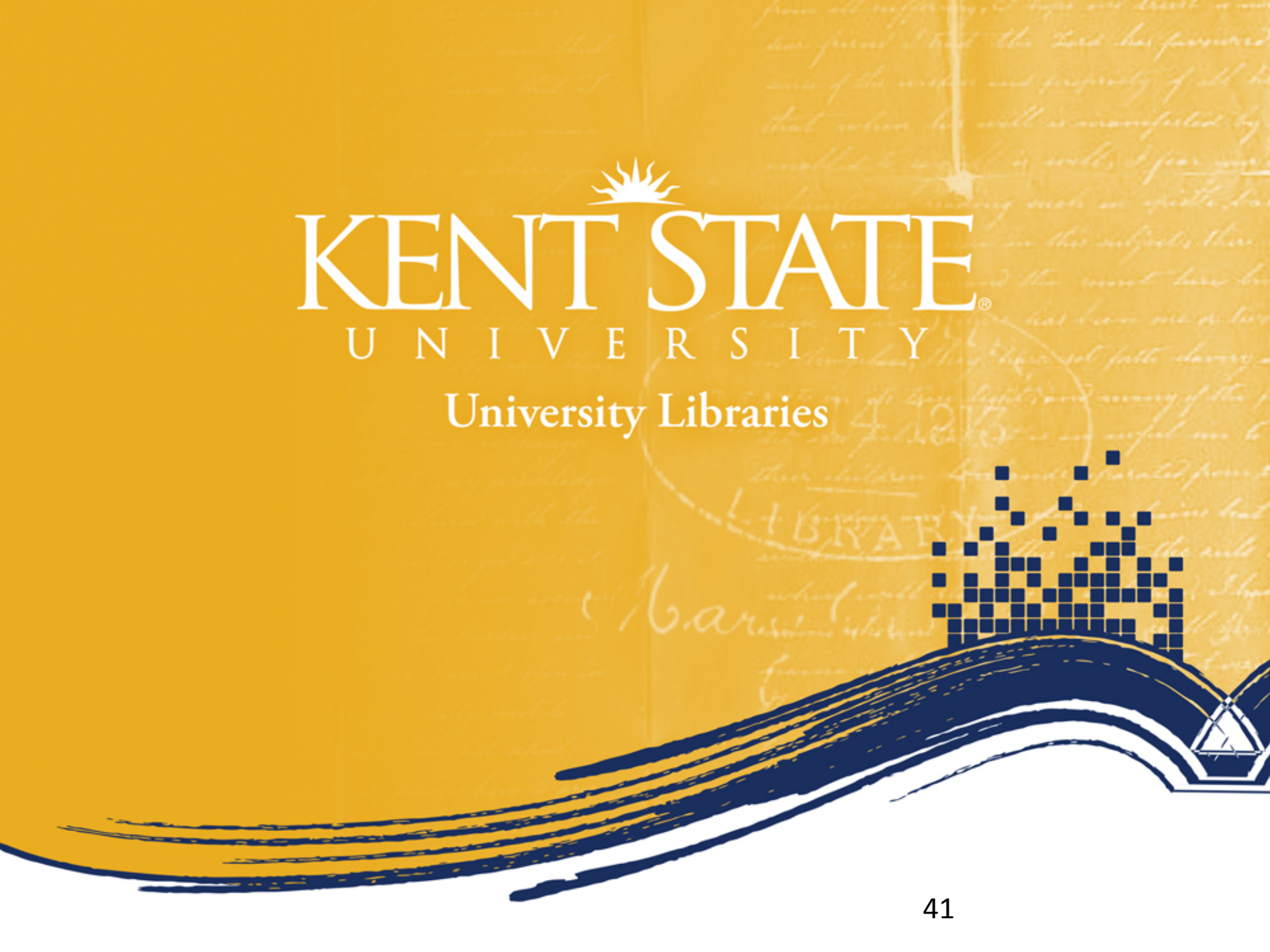
Conclusions

- Future of the Hathi project very much up in air
- Should we, or can we, be selective as to which records we should load?
- If these types of materials are excluded from local or consortial catalogs, what does that bode for the future of the catalog?
- Obvious need to find a way to effectively measure the use of Hathi Trust records and their impact on scholarship.

Questions

Contact Information

Roman S. Panchyshyn MLIS
Catalog Librarian, Assistant Professor
Kent State University
Tel: 330-672-1699
E-mail: rpanchys@kent.edu


KENT STATE
UNIVERSITY
University Libraries