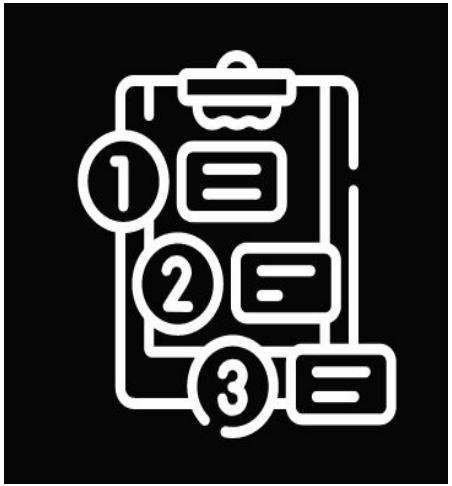
From A Stroll To A Sprint:
Expediting Submission And Ingest To Hathitrust For Locally Digitized Materials

Salwa Ismail, University of California, Berkeley
Lynne Grigsby, University of California, Berkeley
Paul Fogel, California Digital Library
Kathryn Stine, California Digital Library

October 29, 2020

Overview

- Where we were
- Our plans over summer - discussions
- Operations and process
- Pre-submission workflow
- Submission workflow
- How it all ties together
- Q & A



We were strolling...

- UC Berkeley has been submitting to HT from a few years
- Started as a service for students with disability
- Submission varied from 2 to 10+items per month
- 2018 library closed, shelf clearing approach

  - as time permitted, guillotined.
- Google Books Project

Warm up for our sprint...

- Fall 2019 - Fires and power outages
- Spring 2020 - COVID-9
- April 2020 - ETAS
- Planning for summer and esp. Fall
- Not offering curbside pickups, or ILL
- Semester was starting up
- e-Reserves using ETAS
- Conversations with HT started
    - optimize and sync our current workflows
    - not add extra overhead on HT
    - conform our workflows to HT requirements

We were forced to sprint…

- 2 to 10+ books 60+ books a week
- Bib #, barcode and OCLC #
- HT ETAS checking (versions, publishers, editions etc.)
- 400 books to date … more coming
- Change workflows and processes

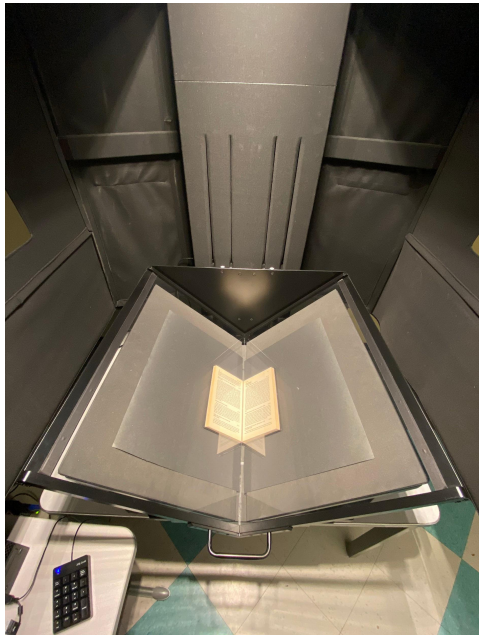
Moving Equipment

- Move 2 Atiz book scanners
- Separate rooms on same floor
- Close to book shelves

Bookshelves

- Had plenty
- Needed lots of signage


ATIZ



Getting started

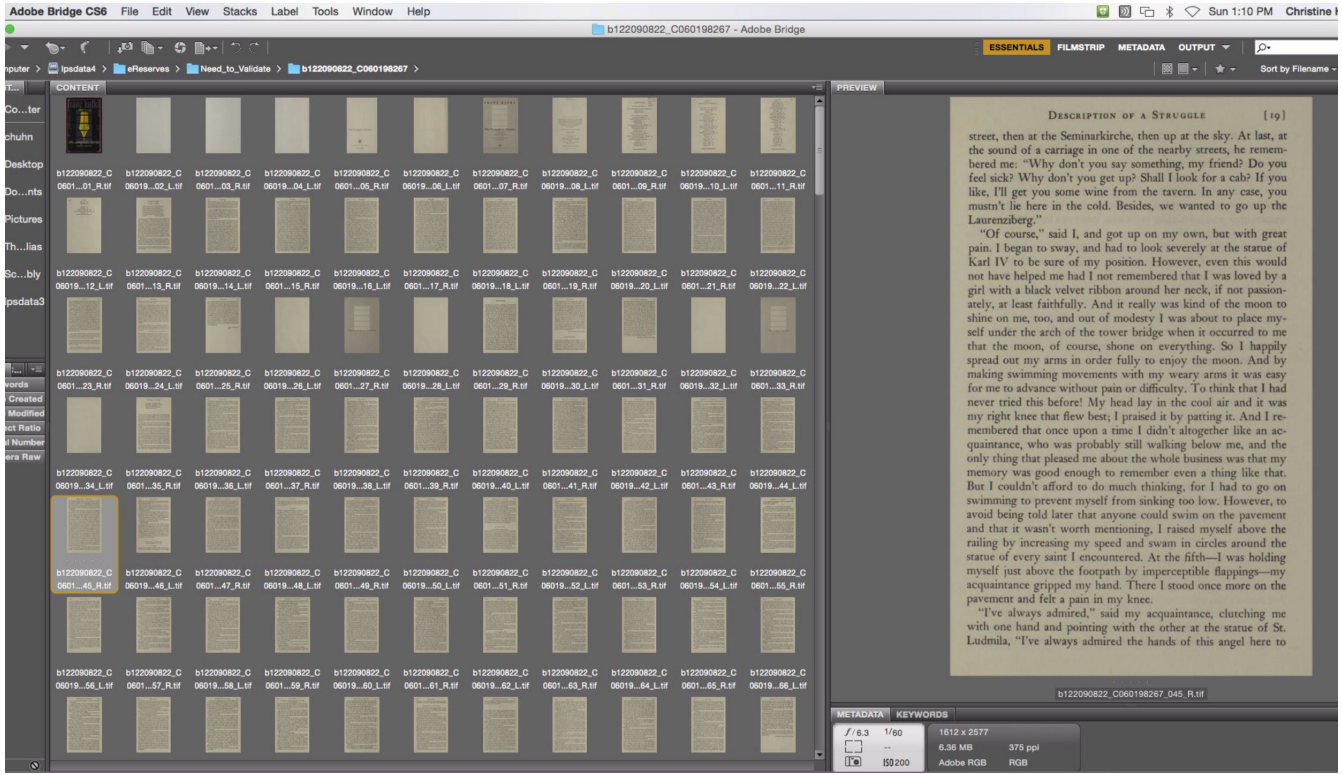
- Staffing
  - 4 Library Imaging Staff
  - 7 shared staff
- Scheduling -- all on site 2-3 days a week
- Training at a distance
- Paging staff managed separately
- We were all learning the process

Process

- Items received
  - Items triaged for binding, foldouts, size
    - Placed on appropriate shelf
- Spreadsheet updated
- Scanning staff pick up book take to scanner
- When finished scanning put on appropriate shelf
- Pick up next book...repeat

Quality Control

- Multiple steps
- Convert to tiff
- Crop
- Page counts
- Image quality
- 12-35 items batched together
- Passed to next step



Submission to HathiTrust

- UC Berkeley has been submitting to HathiTrust since 2016
- Since then we have submitted over 13,000 volumes consisting of over 3 million files
- e-Reserves required multiple changes to our process to handle smaller, faster submissions

Even more quality control

Final Validation Results

Location: Z:\eReserves\validating_20201015
Folders: 20
Files: 7377
Total size: 57.1 GB
Folder Errors: 0
File Errors: 0
Elapsed time: 50 minutes, 30 seconds

Status	Count	Files	Size	Directory	Message
valid		486	3.2 GB	b103029266_C004756519	No warnings or errors found
valid		328	2.2 GB	b104281236_C068084141	No warnings or errors found
valid		282	2.6 GB	b105167538_C090038978	No warnings or errors found
valid		200	1.8 GB	b127102681_C093424348	No warnings or errors found
valid		298	2.1 GB	b129158276_C054755816	No warnings or errors found
valid		424	3.8 GB	b134328711_C110523403	No warnings or errors found
valid		360	2.5 GB	b135653083_C072630459	No warnings or errors found
valid		480	4.4 GB	b141076513_C058080787	No warnings or errors found
valid		472	3.7 GB	b142822589_C076322942	No warnings or errors found

And each tiff is checked

```
======================================================
b103029266_C004756519
======================================================
  valid   b103029266_C004756519_001_R.tif       (1962x3039, 375 ppi, 8 8 8 bits, LZW, 8 MB)
  valid   b103029266_C004756519_002_L.tif       (1564x2501, 375 ppi, 8 8 8 bits, LZW, 5 MB)
  valid   b103029266_C004756519_003_R.tif       (1564x2510, 375 ppi, 8 8 8 bits, LZW, 5 MB)
  valid   b103029266_C004756519_004_L.tif       (1564x2501, 375 ppi, 8 8 8 bits, LZW, 4 MB)
  valid   b103029266_C004756519_005_R.tif       (1564x2510, 375 ppi, 8 8 8 bits, LZW, 7 MB)
  valid   b103029266_C004756519_006_L.tif       (1564x2501, 375 ppi, 8 8 8 bits, LZW, 5 MB)
  valid   b103029266_C004756519_007_R.tif       (1564x2510, 375 ppi, 8 8 8 bits, LZW, 5 MB)
  valid   b103029266_C004756519_008_L.tif       (1564x2501, 375 ppi, 8 8 8 bits, LZW, 5 MB)
  valid   b103029266_C004756519_009_R.tif       (1564x2510, 375 ppi, 8 8 8 bits, LZW, 6 MB)
  valid   b103029266_C004756519_010_L.tif       (1564x2501, 375 ppi, 8 8 8 bits, LZW, 7 MB)
  valid   b103029266_C004756519_011_R.tif       (1564x2510, 375 ppi, 8 8 8 bits, LZW, 7 MB)
  valid   b103029266_C004756519_012_L.tif       (1564x2501, 375 ppi, 8 8 8 bits, LZW, 7 MB)
  valid   b103029266_C004756519_013_R.tif       (1564x2510, 375 ppi, 8 8 8 bits, LZW, 7 MB)
  valid   b103029266_C004756519_014_L.tif       (1564x2501, 375 ppi, 8 8 8 bits, LZW, 7 MB)
  valid   b103029266_C004756519_015_R.tif       (1564x2510, 375 ppi, 8 8 8 bits, LZW, 7 MB)
  valid   b103029266_C004756519_016_L.tif       (1564x2501, 375 ppi, 8 8 8 bits, LZW, 7 MB)
  valid   b103029266_C004756519_017_R.tif       (1564x2510, 375 ppi, 8 8 8 bits, LZW, 7 MB)
```

Creating the OCR

- 1 OCR file per image file
- Tesseract OCR recognition open source software
- Spin up multiple simultaneous processes
- e-Reserves books have tended to go relatively quickly
    - they are typically more plain-text heavy, newer and in better quality than the older books we have not previously
- For e-Reserves it takes 3 to 6 hours per batch
- Image files and OCR zipped together for submission

Metadata

1. File naming provides the bibliographic record number
2. We pull that MARC record from our local catalog
3. Convert to Hathi-compatible MARC XML

Ongoing Improvements

Constantly reviewing and looking for improvements

- Infrastructure changes to reduce the moving of files
- Adding LZW compression to reduce the size of tiffs without losing quality

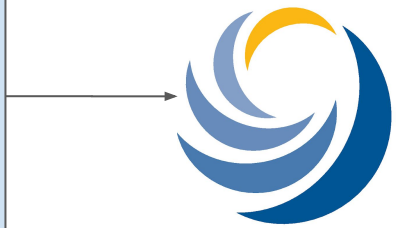
And some processes are still manual, such as,

> Once batch is completed, we send an email (with details) to Paul Fogel at the California Digital Library (CDL)

UC Print Digitization Coordination

<u>UC Locations</u>

- NRLF
- SRLF
- Berkeley
- Davis
- Irvine
- Riverside
- San Diego
- San Francisco
- Santa Barbara
- Santa Cruz
- UCLA



University of California

CDL

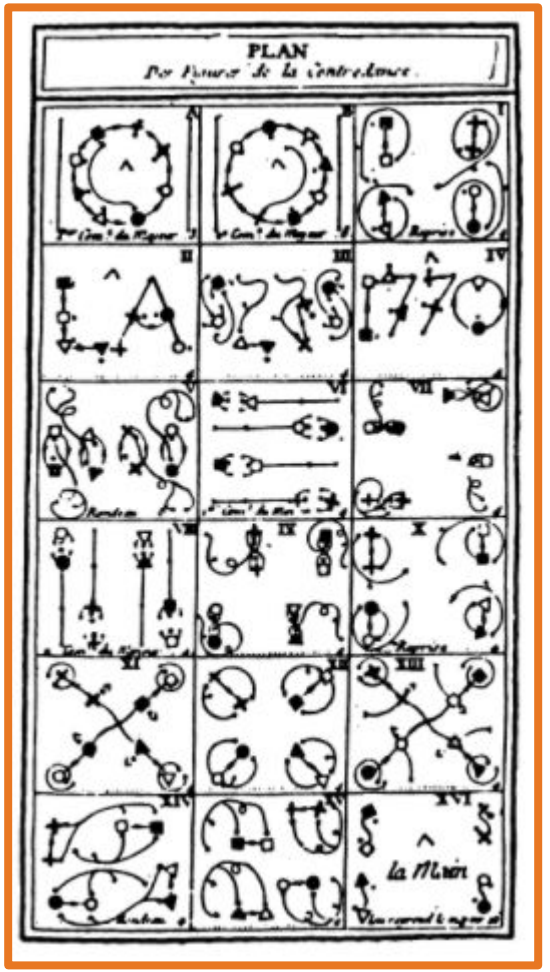
California Digital Library

HATHI TRUST

UC Berkeley Expedited HathiTrust Deposit Process

- Receipt of bibliographic records and object links from Berkeley
  - Typically on Friday or Monday
  - Content hosted on Berkeley servers
- Submission of records to Zephir (HathiTrust Metadata Management)
  - Upon receipt of files
- Confirmation of successful bibliographic record loading
  - Received the next 9:05am Pacific time
  - Upon successful load, data will be visible to HathiTrust systems the next 3am Pacific time
  - Load must occur no later than Tuesday morning to be available for the ingest systems
- Submission of request for content ingest
  - HT Ops staff performs ingest on Wednesdays
- Confirmation of successful content ingest
  - Typically on Thursday

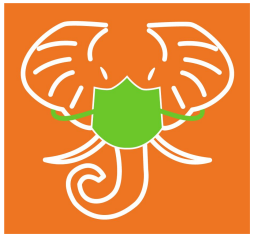
Shortening the turnaround from request to delivery relies on well-defined choreography



19th century French dance notations from (https://hdl.handle.net/2027/uc1.32106009453165?urlappend=%3Bseq=47)

Refining the Expedited HathiTrust Deposit Process

- Clarifying, routinizing, and optimizing roles and processes (e.g., quasi-daily → weekly)
- Benefiting from HathiTrust's predictable, and now more frequent processes implemented during ETAS
  - monthly print holdings db rebuild
  - early access for depositing institution
- Benefiting from existing local processes
  - Locally-digitized metadata and content stream workflows
  - Established handoffs and trouble-shooting approaches



Thank you!

- Please complete the survey
- Jump in on the conversation on the HathiTrust Community Slack: #commweek