Contributing to HathiTrust on a Smaller Scale

Suzanne Reinman, Juliana Nykolaiszyn, and Tabitha Carr

In 2021, the Oklahoma State University Library contributed titles from the Bureau of Indian Affairs in their regional depository library collection to the HathiTrust Digital Library. A coordinated effort, this allowed the university library to expand the mission and outreach of the Federal Depository Library Program and also disseminate information and further education as part of the responsibility of a land-grant institution. Contributing Bureau of Indian Affairs materials enhanced the OSU Library's role as a preservation partner with the Federal Depository Library Program, supported the state with the importance of Native American tribes in Oklahoma, and complemented the goals of the HathiTrust Federal Documents Collection Framework as a key agency to complete as to a comprehensive collection.

The Oklahoma State University Library (OSU) is a regional depository library for the Federal Depository Library Program and also a member institution of HathiTrust (HT), a "collaborative of academic and research libraries preserving over 17 million digitized items."[1] Working to comprehensively catalog its federal documents collection, second copies and other publications were retained for potential contribution to the HathiTrust database to fulfill in part its mission as a land-grant university in the advancement of knowledge.

The US Federal Documents Program at HathiTrust is one of six programs that furthers HathiTrust's mission and goals. The Program serves to "expand access to and preserve US federal publications."[2] Selecting "Browse Collections" from the HathiTrust main page, *US Federal Documents* is a collection in progress identified through the Federal Government Documents Registry, "a database of metadata intended to represent the comprehensive corpus of U.S. federal documents produced from 1789 to the present."[3] Other pages related to federal document collections include the HathiTrust US Federal Government Documents Program,[4] expanding access to and preserving US federal publications "through coordinated and collective action, expand and enhance digital access to U.S. federal publications including those issued by GPO and other federal agencies," through its goals of a comprehensive digital collection, enduring access, and community.

Project Background

As a member institution and a regional depository library, OSU considered it important to contribute federal publications not yet included in HathiTrust through its U.S Federal Documents Program, serving researchers worldwide. Contacting the Government Documents Registry Analyst at HathiTrust, the library inquired about contributing materials to the database. In 2019, HathiTrust included materials from about 55 members in the repository; most of the items in the database come from the Google scanning workflow. An overview of existing digitized content and the Hathi ingest process is presented at Getting Content into HathiTrust.[5] Although the bulk of the content is mass-digitized, HT is very interested in working with members to ingest locally-digitized material and they plan to work with members to provide additional support and guidance for this process in the coming years.

To align its contributions with the collections sought for the database, the OSU Library reviewed the HathiTrust Federal Documents Collection Framework,[6] also the US Federal Documents Collections outline.[7] Comprehensive runs of essential titles are listed as priorities as well as publications from specific agencies.

OSU serves as a Preservation Steward for the Federal Depository Library Program for the Bureau of Indian Affairs (BIA). The library chose to begin with this agency because it aligned with both institutions' priorities–US Federal Documents Collection. In addition to history of native peoples in Oklahoma and the tribes that were relocated to Indian Territory, Oklahoma is also a state with 35 currently federally recognized tribes and one of the highest percentages of Native

Americans as part of its population. Working with the Program Officer for Collections and Federal Documents and the Collection Services Librarian, the contact for digitization and ingest, the library sent their holdings for the BIA in 2019 for a comparison as to what the registry, updated daily, was able to ingest as to titles not yet in the HT database from OSU. Twenty-seven titles were selected for contribution from the comparison that were substantial in their content.

Requirements

The original process for contributions as outlined by HT in 2019 was revised in 2020. Other institutions have described the earlier process.[8] The technical requirements did not change. The revised and current workflow is as follows. Working with the original process proved to be challenging; it is now much more streamlined and direct. Getting Content Into HathiTrust is part of the Digital Library page:

> Members may deposit digitized materials with HathiTrust for long-term preservation and access. These materials are stored in our repository and made available for search, display, and computational research, in addition to other uses as permitted by U.S. Copyright Law. We encourage all members to deposit material.[9]

The sections under Getting Content into HathiTrust include the following:

- Bibliographic Metadata Specifications, hathitrust.org /bib_specifications
- Bibliographic Rights Determination, hathitrust.org /bib_rights_determination
- Ingest Checklist, hathitrust.org/ingest_checklist
- Ingest Reports, hathitrust.org/ingest_reports
- Ingest Reports Description, hathitrust.org /ingest_reports_description
- Bibliographic Metadata Submission, hathitrust.org /bib_data_submission
- Ingest Tools, hathitrust.org/ingest_tools

Also see the Guidelines for Digital Object Deposit in the Policies section,[10] which includes four sections:

> The purpose of these guidelines is to facilitate deposit of digital content from a variety of sources into the HathiTrust repository. The guidelines contain an introduction to the HathiTrust Digital Library, a description of its guiding principles and design, a brief overview of the ingest process, and definitions, policies and procedures related to the ingest of digitized book and journal content and associated metadata.
>
> - Technical Requirements for Digitized Page Images Submitted to HathiTrust see https://bit .ly/2OS8byR
> - Submission Package Requirements for Digitized Content Submitted to HathiTrust https://bit .ly/2SB3ytK
> - Bibliographic Metadata Specifications see https:// bit.ly/2UMFUNE
> - Overview of Bibliographic Metadata Submission Process see https://bit.ly/39uD9Vq

The library reviewed the scanning and bibliographic record specifications, also submission requirements, *Bibliographic Metadata Specifications and Overview of Bibliographic Metadata Submission Process* above. Following the Ingest Overview in Getting Content into HathiTrust, the Digital Asset Submission Inventory form was completed to set up the content stream in the HathiTrust repository and signed by the dean of libraries, also the Administrative Coversheet to be used for setting up the configuration for loading the bibliographic data into Zephir, the bibliographic metadata management system for HT.

For the bibliographic records, a test file of a 10 percent sample of the bibliographic metadata to be submitted was required. Working with the metadata analyst at the California Digital Library, the process for the submission of bibliographic metadata to Zephir, is outlined at Bibliographic Metadata Submission,[11] and the specification for the records themselves is posted at Bibliographic Metadata Specifications.[12]

Scanning Workflow

HathiTrust outlines technical requirements when scanning materials for inclusion, such as specifications for image capture, resolution, color, format, and file naming conventions.[13] Many scanners have the ability to set up jobs with predefined settings within their proprietary software. Designing a HathiTrust specific preset is beneficial, and helps scanning operators understand key specifications without the need to consult documentation every time.

One of the technical barriers for OSU was identifying existing scanning equipment within the building to produce quality scans based on the technical requirements. Specifications include creating scans with a bitonal resolution of 600 pixels per inch (ppi) with CCITT Group 4 compression in TIFF format, or continuous tone images with a minimum resolution of

300 ppi in TIFF or JPEG2000 format. This, combined with the requirements to produce optical character recognition (OCR) for each scanned page via TXT files, object metadata as a YAML file, and an MD5 fixity check comprise the Submission Information Package (SIP) needed for transfer to HathiTrust.[14]

In breaking down the SIP further, producing OCR files for each scan can be a challenge. Institutions may turn to commercial products to assist in this process, but the need to produce one text file per image can be quite cumbersome. While open-source options like Tesseract or fee-based programs like Adobe Acrobat or ABBYY FineReader may be helpful to explore, at OSU, our scanner's proprietary software was able to generate the much-needed text file per image, based on presets set up during configuration.

Looking closer at the object metadata, while not difficult to generate, it does take some time to configure using a mix of CSV files and helpful Python code created by the HathiTrust community to generate the necessary YAML file. This file contains specific metadata information, including scanner type, DPI, compression, extensions, and page breakdowns. Helpful Github repositories to potentially use when setting up your own processes include Caruso's YAMLgenerator[15] or Tillman's YAML Generator for Digitized HathiTrust Submissions.[16] Each can be adapted to fit your needs, and also provides a great starting point with respect to understanding metadata requirements.

Once all the scans, OCR files, and the YAML file have been generated, it is time to focus on creating an MD5 checksum to record fixity. Fixity refers to the overall integrity of a digital file. A fixity check, in this case, in the form of a checksum, helps verify if the file has been changed or altered in the transfer process. While there are many types of checksums, the MD5 hash is commonly used for this purpose, and is included as an option in many fixity programs. While there are many tools on the market to record fixity, one free option is ExactFile.[17] ExactFile is easy to use, cost effective, and provides numerous hash options, such as MD5. Once the checksum is complete, this is also included in your SIP to HathiTrust. It is at this point when you compress all items in a ZIP file named accordingly, usually with the item's barcode. According to HathiTrust, the package should contain the following:

- An image file for each page with proper naming conventions (.tif or .jp2)
- A plain text OCR file for each page (.txt) or coordinate OCR for each page (.html)
- A metadata file (.yml)
- A checksum file for the SIP (.md5)

For those concerned if everything is completed properly, HathiTrust also provides a Submission Information Package Validator18 to check ZIP files. Once done, completed ZIP files are then ready for the final step, transmission to HathiTrust for ingestion.

Bibliographic Records and Metadata

The Bibliographic Metadata Specifications for HathiTrust are outlined in their digital library. The documents coordinator, working with members of the cataloging department, was responsible for the metadata process and utilized Alma, FileZilla, MarcEdit, and Outlook as tools for this process. The steps for the metadata follow.

The first step in providing useable metadata was to create an itemized set of physical items, working from a spreadsheet containing barcodes of each item scanned in the first column under the header "Barcode" and formatted as a text column (format as text so numbers are not converted to scientific notation). It is also possible to work from a list of OCLC numbers as long as the text file has the header "035 field" as the first coleumn. See images 1 and 2 for examples of our spreadsheets for creating a set of records.

Once the spreadsheet is created, return to Alma and navigate to "Add set," select "itemized," and set the content type as "Physical Items." This is preparation for the upload of the spreadsheet file. There is a space to upload the file and this is where you will upload a spreadsheet like those above and finish creating the set. After the set is created, confirm the count of set members matches the items on the spreadsheet uploaded. If it does, then proceed to step two.

Export the file from the library's management system (Alma at OSU). A publishing profile and job are required for this. Check that the profile includes the holding and items information and full bibliographic information. Once the file is published, download it from the local ftp server. Filezilla was used to facilitate this process.

Using MarcEdit, open the file in MarcEdit-MARC tools and convert the file to .mrk using the MARC Breaker function. Open the .mrk file and remove all 9xx fields except for the 955 field with the barcode of the specific item you are including in the Alma set. The example below highlights what this will look like for a single record. Other records follow in the same format after a single line break (see image 3).

If there are multiple items that are associated with one bibliographic record (for example a multi-volume set), use a separate copy of the record for each item; the only difference between each will be the 955. Only include one 955 per record. MarcEdit has a find (ctrl-f) function that will allow you to

	A
1	BARCODE
2	36135019278318
3	36135022320206
4	36135000058950
5	36135022319687
6	36135003970912
7	36135019278250
8	36135003971050
9	36135022330585
10	36135022329900
11	36135022121976
12	36135010825919

or

	A
1	035 field
2	(OCoLC)320199034
3	(OCoLC)320220628
4	(OCoLC)05630081
5	(OCoLC)13961262
6	(OCoLC)811619226
7	(OCoLC)2788400
8	(OCoLC)42141115
9	(OCoLC)38525738
10	(OCoLC)263684196
11	(OCoLC)947057456
12	(OCoLC)822030432

Images 1 and 2. Examples of OSU Library's records spreadsheets.

locate all instances of a field and see if multiple versions of that field exist in one record. Image 4 shows an example of what that search could look like.

The results above show us that many records have multiple 955 fields (often because a 955 field containing the call number as well as barcode was added). By navigating to each additional occurrence, the records are able to be edited for compliance. HathiTrust asks submitters to also remove 035 $9, 019, and 035 $z. All records must include LDR (000), 001, 008, 035 $a (OCoLC), 040 $c, 245, and 300 $a. Removing and adding fields, except for the 955, can be accomplished using a normalization rule.

A good verification step is to check that the number of each required field matches the number of items (one bibliographic record per item record) that you are submitting to HathiTrust. This may be done using Find All and comparing the results count of each field. After this is confirmed, run MarcEdit Marc validation report (ctrl-M) and validate headings. Next, compile the saved .mrk file into MARC and close the resulting .mrc file. Next, use MarcEdit Marc Tools to convert the .mrc file to MARC21XML. The resulting .xml file will be the final file and the one that is shared with HathiTrust.

Upload the files to Zephir using an FTP tool (such as Filezilla) using the naming convention provided by HathiTrust. If a second file is uploaded on the same date, make sure that it has a unique distinguishing identifier at the end. Files cannot have the same name as a previously loaded file. Once the file is uploaded to Zephir, send a notification email and make sure to check the FTPS space for run reports or error files later that week. If there are errors, re-submit the corrected records

```
=LDR  01182cam a2200301Ma 4500
=001  9937481184602681
=005  20190530112940.0
=008  901108s1900\\\\dcu\\\\\\\\\\000\0\eng\d
=035  \\$a(OCoLC)320199034
=035  \\$a(OCoLC)ocn320199034
=040  \\$aCLU$beng$cCLU$dOCLCO$dOCLCQ$dOCLCA
=049  \\$aOKSU
=110  1\$aUnited States.$bBureau of Indian Affairs.
=245  12$aA new era for the American Indians :$bpresident Nixon sets new Indian policies and goals /$cBureau of Indian Affairs.
=260  \\$aWashington, D.C. :$bThe Bureau,$c[19--]
=300  \\$a12 pages
=336  \\$atext$btxt$2rdacontent
=337  \\$aunmediated$bn$2rdamedia
=338  \\$avolume$bnc$2rdacarrier
=500  \\$aCover title.
=500  \\$a"The president's message to the congress of the United States on the America Indians, July 8, 1970."
=583  \\$apermanently retained$fFederal Information Preservation Network Preservation Copy of Record
=948  \\$hHELD BY OKS - 2 OTHER HOLDINGS
=955  \\$a36135022330320
=960  \\$aSPM$1(OCoLC)ocn320199034$29937481184602681$70
=994  \\$aZ0$bOKS
=852  \\$b10049267$822337971770002681
```

Image 3. Example of a single record.

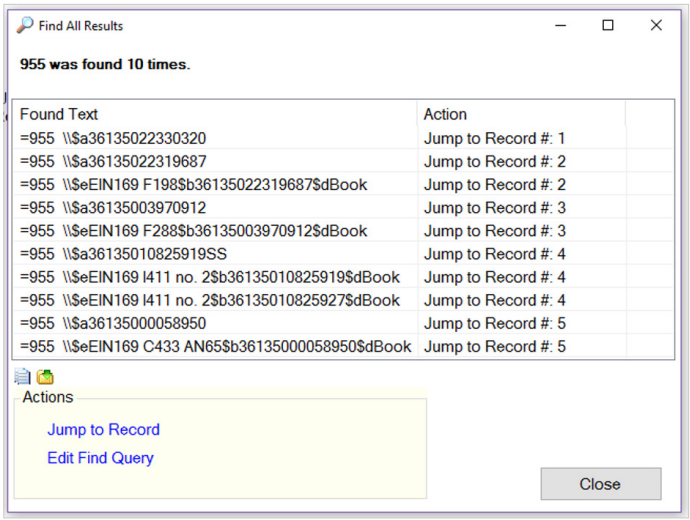

Image 4. Example of MarcEdit find results of 955 fields.

following the same process and notification as before updating the date in the filename.

Submission of Records and Staffing

A box folder at the University of Michigan was shared for the uploading of the digital object packages. HathiTrust is now using DropBox. Correspondence with feedback@issues.hathitrust.org included items ready for deposit and that they were successfully submitted. Materials were available in the database shortly after submission.

The Government Documents department managed the project with the Digital Resources and Discovery Services department navigating HathiTrust's scanning requirements. Staff with a background in cataloging and metadata are also necessary to work with the HT bibliographic metadata specifications. The months with COVID extended the process. The staff at HathiTrust were essential in their guidance in addition to the documentation.

Conclusion

A technical and labor intensive but valuable project, a combined departmental effort and melding of skill sets made this

contribution possible. Depository libraries nationwide have comparable collections but making them digitally available through a national database serving international communities expands the mission and outreach of the Federal Depository Library Program. It also disseminates information and furthers education as part of the responsibility of a land-grant institution. Contributing Bureau of Indian Affairs materials enhanced the OSU Library's role as a Preservation partner with the Federal Depository Library Program, supported the state with the importance of Native American tribes in Oklahoma, and complemented the goals of the HathiTrust Federal Documents Collection Framework as an agency to complete as to a comprehensive collection. Future contributions will be considered based on the HathiTrust Federal Documents Collection Framework.

Suzanne Reinman (suzanne.reinman@okstate.edu), Documents Librarian, Oklahoma State University. **Juliana Nykolaiszyn** (juliana.nykolaiszyn@okstate.edu), Head, Digital Resources and Discovery Services, Oklahoma State University. **Tabitha Carr** (tabitha.manners@okstate.edu), Government Documents Coordinator, Oklahoma State University

Notes

1. "HathiTrust Digital Library," HathiTrust, accessed February 10, 2022, http://hathitrust.org/.
2. "Welcome to HathiTrust!" HathiTrust, accessed February 10, 2022, https://www.hathitrust.org/about.
3. "United States Government Documents Registry," HathiTrust, accessed February 10, 2022, https://hathitrust.org/usdocs_registry.
4. "HathiTrust U.S. Federal Government Documents Program," HathiTrust, accessed February 10, 2022, https://hathitrust.org/usgovdocs.
5. "Getting Content into HathiTrust," HathiTrust, accessed February 10, 2022, https://www.hathitrust.org/ingest.
6. "HathiTrust Federal Documents Collection Framework," HathiTrust, accessed February 10, 2022, https://hathitrust.org/hathitrust-federal-documents-collection-framework.
7. "U.S. Federal Documents Collections," HathiTrust, accessed February 10, 2022, https://hathitrust.org/u-s-federal-documents-collections.
8. Kyle R. Rimkus and Kirk M. Hess, "HathiTrust Ingest of Locally Managed Content: A Case Study from the University of Illinois at Urbana-Champaign," *Code4Lib Journal* no. 25 (2014), https://journal.code4lib.org/articles/9703.
9. "Our Digital Library," "US Federal Documents Collections," HathiTrust, accessed February 10, 2022, https://hathitrust.org/digital_library.
10. "Guidelines for Digital Deposit," HathiTrust, accessed February 10, 2022, https://hathitrust.org/deposit_guidelines.
11. "Bibliographic Metadata Submission, Overview of Bibliographic Metadata Submission Process," HathiTrust, accessed February 4, 2022, https://hathitrust.org/bib_data_submission.
12. "Bibliographic Metadata Specifications," HathiTrust, accessed February 10, 2022, https://www.hathitrust.org/bib_specifications.
13. "Technical Requirements for Digitized Page Images Submitted to HathiTrust: A Guide for HathiTrust Members," HathiTrust, accessed February 4, 2022, https://www.hathitrust.org/technical-requirements-digitized-page-images-submitted-to-hathitrust.
14. "Submission Package Requirements for Digitized Content Submitted to HathiTrust," Hathitrust, accessed February 4, 2022, https://www.hathitrust.org/submission-package-requirements-digitized-content-submitted-to-hathitrust#two-two.
15. Moriah Caruso, "YAMLgenerator," accessed February 4, 2022, https://github.com/moriahcaruso/HathiTrustYAMLgenerator.
16. Ruth Tillman, "YAML Generator for Digitized HathiTrust Submissions," accessed February 4, 2022, https://github.com/ruthtillman/yaml-generator-for-hathitrust.
17. Brandon Staggs, "ExactFile," accessed February 4, 2022, https://www.exactfile.com/.
18. "Ingest Tools," HathiTrust, accessed February 4, 2022, https://www.hathitrust.org/ingest_tools.