U.S. Federal Documents in HathiTrust: A Collection Profile

March 2017
Valerie Glenn
valglenn@hathitrust.org

Collection Profile: U.S. Federal Documents in HathiTrust

The federal documents collection in HathiTrust has developed through mass digitization, opportunistic projects, and ad hoc contributions from members. We are beginning a new phase of focused collection development that builds upon this large base of federal documents. In late 2016 HathiTrust staff conducted an overall analysis of the current HathiTrust federal documents collection to inform collection development strategies, but also to investigate a variety of metrics based on the data available to us and to establish a baseline for reporting on the collection. There have been no previous attempts to conduct this type of analysis on a specific HathiTrust collection. This report describes our initial attempt at benchmarking our collection of federal government documents as it existed on September 1, 2016. In addition to the overall analysis, where we outline our findings and some data limitations, the report also describes a test we performed to determine comprehensiveness of digitized versions of a few key federal document titles within HathiTrust. We also highlight some areas we've identified for further exploration.

Scope & Identification of Federal Documents

The analysis is largely based on bibliographic data. It is important to note that the data in this report differs from the repository numbers reported regularly by HathiTrust and available via the Help page for the U.S. Federal Documents Program[1]. For this analysis, data on both full and limited view U.S. federal documents has been incorporated, in order to provide a richer picture of HathiTrust's federal documents content. These materials have been identified using the following criteria:

- The HathiTrust record matches an existing record in the U.S. Federal Documents Registry[2]; or
- The MARC 008 field contains an 'f' in character 28 and a 'u' in character 17; or
- The MARC 086 field contains a SuDoc number.

The analysis in this report incorporates the following pieces of information, drawn from the bibliographic records and from additional HathiTrust data:

- Rights attributes assigned to the digital object;
- HathiTrust members who deposited content;
- Digitization agent;

[1] https://www.hathitrust.org/help_usgovdocs. This number, reported on the 1st and 15th of each month, details the number of digital objects associated with a bibliographic record that has been coded as a US federal document after the bibliographic rights algorithm has been applied.
[2] https://www.hathitrust.org/usdocs_registry/

- The type of publication, as designated in the MARC record (ie, serial or monograph);
- The number of duplicate digital objects;
- Analysis of the corporate author field;
- Analysis of the publisher field;
- Geographic coverage;
- Publication date;
- Superintendent of Documents call number;
- Language;
- Usage data.

Findings

The resulting total number of bibliographic records in the analysis is 412,205, representing 970,315 digital objects.

Finding: Approximately 88% of federal documents are currently in full view and 12% in limited view.

Approximately 852,488 are fully viewable worldwide and 117,827 are limited view/search only. Potential reasons for an object to be limited view are:

- quasi-governmental author (ie, Smithsonian, Federal Reserve);
- the object is known or suspected to contain copyrighted material;
- the object contains personally identifiable information (PII);
- the place of publication is outside of the United States (ie, an embassy or military installation).

The HathiTrust rights algorithm defaults to an in-copyright determination when the MARC 008 field does not contain an ‘f’ in character 28. This may result in some federal documents being limited view in error.

View Status

Rights attributes are assigned to digital objects and dictate viewability. They can be assigned automatically, based on information in the bibliographic record, or manually, following review by a staff member.

Finding: Not all items meeting our definition of federal documents are available in full view.

The 970,315 digital objects in this profile have eighteen different rights attributes[3] assigned to them. The majority of these (838,993) are coded as public domain, but there are more than 100,000 objects coded as in-copyright. Two different rights attributes are indicative of privacy concerns: “Public Domain but access limited due to privacy concerns” (130) and “Available to nobody; blocked for all users” (66). 11,602 objects have the rights status “Undetermined.”

The variety of rights attributes may be indicative of out-of-scope, non-federal documents inadvertently included in this analysis, or other unknown factors. This is an area for further investigation.

Monograph/Serial Breakdown and Identification of Duplicates

Finding: 94% of the bibliographic records represent monographs vs 6% that represent serials. Monographs make up 56% of the digital objects, vs. 44% serials.

This analysis includes 387,766 monographic records representing 546,432 digital objects and 23,986 serial records representing 423,187 digital objects. The numbers were generated based on the presence of ‘m’ or ‘s’ in the eighth character (07 - bibliographic level) of the MARC record leader, then counting the number of objects associated with that record. 453 records (696 digital objects) did not include this data.

Finding: Limitations of data are too great for a definitive identification of duplicates.

An attempt was made to report on duplicates in order to identify the number of objects in the HathiTrust federal documents corpus. However, enumeration and chronology information for serials and multi-part monographs was not included in that process, so the only information that is known is the number of unique items (353,416). The identification of duplicates is a challenging problem that is of wider collaborative interest across HathiTrust.

Contributors

Finding: Over fifty HathiTrust members have deposited federal documents.

Fifty-one different organizations have deposited content identified as a U.S. federal government document. Of those, thirty have deposited at least 1000 objects. Table 1 indicates the twenty largest contributing institutions - of those, nine are from the Big Ten Academic Alliance (formerly

[3] For more information about HathiTrust Access and Use statements, see https://www.hathitrust.org/access_use.

the CIC) and five are from the University of California System. Both the Big Ten and California have emphasized the digitization of federal documents in their partnership with Google.

Table 1. Top twenty contributing institutions, along with the number of federal documents deposited

Contributing Institution	**Number of Federal Documents Deposited**
University of Michigan	249066
University of Minnesota	125444
University of Illinois at Urbana–Champaign	102147
University of California, Northern Regional Library Facility	75387
Northwestern University	68518
Pennsylvania State University	48859
Technical Report Archive & Image Library (TRAIL)[4]	44437
Purdue University	43134
Cornell University	29898
Harvard University	23872
University of California, San Diego	21958
University of California, Riverside	12556
The Ohio State University	12442
Michigan State University	12418
University of California, Southern Regional Library Facility	10762
University of California, Santa Cruz	10715
University of Virginia	10253
Indiana University	9506
New York Public Library	7492
State University System of Florida	6960

[4] TRAIL is a project under the umbrella of the Center for Research Libraries. Not all libraries that participate in TRAIL are HathiTrust members. See https://www.crl.edu/programs/trail for more information.

Digitization Agent

Finding: Most federal documents in HathiTrust are a product of Google digitization, although documents originate from twenty digitization sources.

Based on information in the 974 $s of the MARC record, twenty organizations have digitized content identified as a U.S. federal government document. It is no surprise that Google is the leading digitization agent, digitizing 97.8% of the digital objects profiled in this report. However, of the fifty-one members that have deposited federal documents, seventeen have digitized items locally.

Table 2. Top two digitization agents, with volumes digitized and percentage of total.

Digitization Agent	Volumes Deposited	Percentage of Total
Google	948,486	97.8%
Internet Archive	17,356	1.8%

Bibliographic Analysis

Finding: Accurate characterization of the federal documents collection by author, publisher, and subject is difficult due to inconsistent cataloging.

Attempts to describe the collection based on corporate author (MARC field 110), publisher (MARC field 260 $b) and subject (MARC fields 650 and 651) have proven to be challenging for several reasons. Namely, each of these fields suffer from inconsistent cataloging practices involving abbreviations, word order, and punctuation, among other textual complications. Even when corporate authors or publishers have the same name, it is not absolutely certain they are the same entity (ie, The Commission).

Given the timeframe and resources allocated to this initial profile, we documented the challenges associated with this analysis so that we can explore solutions and re-analyze when feasible. Below is more detailed information about each field.

Corporate Author

Corporate author entries have been cataloged with differing levels of detail and accuracy over time. Subordinate units that may be the actual author are frequently omitted or, conversely, intervening departments are omitted and a subordinate unit is placed in the 110$a. The

evolution of the federal government's structure complicates this further, as the corporate author hierarchy can change even if the actual responsible office or department does not. Table 3 highlights some of the variations in corporate author for the same governmental author.

Table 3. Examples of the variation on corporate author in the 110 $a.

United States Commission on Civil Rights	U.S. Dept. of Commerce, Social and Economic Statistics Administration, Bureau of the Census	United States. Congress.
United States Civil Rights Commission	The Bureau	United States. Congress. Senate.
United States. Civil Rights Commission	Bureau of the Census	United States. Congress. Senate. Committee on Commerce, Science, and Transportation.
United States. Commission on Civil Rights		

Publisher

The Publisher field is another that suffers from a lack of uniformity of entries. The ballot initiative that established federal documents as a priority for HathiTrust encourages HathiTrust to "facilitate collective action to create a comprehensive digital corpus of U.S. federal publications including those issued by GPO and other federal agencies."[5] During this analysis more than forty-nine variations on "Government Printing Office" have been identified in the publisher field (260 $b), associated with 124,698 bibliographic records (30% of the whole). This means that 70% of records are represented by other publishers or no publishers (74,073 records do not include a publisher field; 41,572 publishers are represented among the 412,205 total bibliographic records).

One interesting piece of information included in these records is the place of publication. As expected, Washington, DC, is the place of publication listed for the majority of records (274,243). Of the remaining 137,962 records, all fifty states are represented in place of publication and there are many locations outside of the United States. These may be indicators

[5] The ballot initiative from the 2011 Constitutional Convention is available at https://www.hathitrust.org/constitutional_convention2011_ballot_proposals#proposal4.

of regional agency or U.S. embassy publications, materials produced on military bases, or out-of-scope records erroneously included in the analysis.

Subject - Geographic Coverage

As a free text field, the Subject field is especially challenging to analyze, but staff did explore whether it was possible to extract geographic information from the MARC 651 field in order to provide a breakdown of coverage by geographic region. The analysis proved to be challenging in part because many entries begin with “United States;” also, not all geographic entries have state or country information [ie, Yellowstone National Park; Moon Surface].

A full analysis of the Subject field to characterize the federal documents collection by subject would be a major and complex project, and was not possible at this time.

Publication Date

Finding: A majority of the HathiTrust federal documents collection dates from the 1960s through mid-1990s.

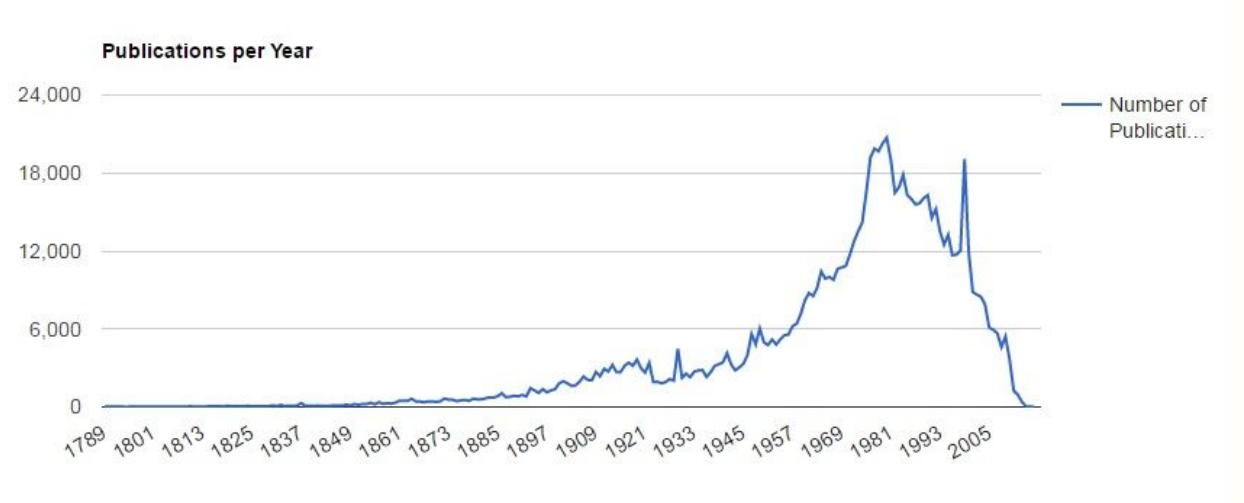
Information on publication date was pulled from the MARC 974$y reported in the HathiTrust holdings field. This was done so in order to try to more factually represent serial holdings across time. (Reports including data pulled from the MARC 008 or 260 subfield c fields may be flawed in that all objects associated with a serial record are reported as being produced in the first year that the serial was published.)

Among the 412,205 bibliographic records included in this profile, 7,521 are missing publication dates in the 008 and 260 $c fields. 91,208 of the digital objects are missing a 974 $y.

Image 1 shows the trend of publications produced per year. Publishing appears to peak in 1980 (20,721), then decrease steadily. The chart below shows a sharp increase in 1999 (19,087). This may not be a true spike, because the metadata for many documents is incomplete. The HathiTrust bibliographic rights algorithm[6] automatically assigns a publication date of 1999 if there is a ‘199u’ or ‘19uu’ in the end date field of the original MARC record. That automatic date assignment is reflected in the chart below.

[6] HathiTrust’s bibliographic rights determination algorithm can be found at https://www.hathitrust.org/bib_rights_determination.

Image 1. Trend of publications produced per year.



Superintendent of Documents (SuDoc) call number

Finding: Of the approximately 64% of records with a SuDoc number, three of the top ten SuDoc stems represent Congressional materials.

Many physical U.S. federal documents collections are arranged by Superintendent of Documents Classification and are not cataloged. Thus, one request from users (in focus groups for the Federal Documents Registry[7] and in the report issued by the Government Documents Initiative Planning and Advisory Group[8]) has been to add a search by SuDoc number to the HathiTrust Catalog.

During this analysis it was determined that 36% of bibliographic records (148,642) do not include a SuDoc number, which means that incorporating a SuDoc search into the Catalog could be misleading to users as they may assume that a search retrieving zero results means that the publication is not in HathiTrust.

Table 4 includes a list of the top ten SuDoc numbers assigned to records in the analysis. Out of the 263,563 records that include SuDoc numbers, 78,556 have stems that represent congressional hearings. In fact, three of the top ten SuDoc stems assigned to these records represent congressional materials. Given that many source libraries have contributed their hearings to Google for digitization, it's not too surprising that Y 4 is the most popular SuDoc stem. What is surprising is the gap between the first and second most popular SuDoc stems:

[7] A summary of focus group comments can be found at http://bit.ly/16QeM3i.
[8] The report of the Government Documents Initiative Planning and Advisory Group is available at https://www.hathitrust.org/documents/HathiTrustGDIPAWGwhitepaper.pdf.

the Y 3 stem, representing congressional commissions and independent agencies, has been assigned to over 57,000 fewer records (20,868).

Table 4. Top ten SuDoc stems included in the analysis.

SuDoc Stem	Agency Represented	Number of Records
Y 4.	Congressional Committee Hearings	78,556
Y 3.	Congressional Commissions and Independent Agencies	20,868
A 13.	United States. Forest Service	12,536
NAS 1.	United States. National Aeronautics and Space Administration	9,876
I 28.	United States. Bureau of Mines	9,273
Y 1.	Congress.	9,129
C 3.	United States. Bureau of the Census	6,919
HE 20.	United States. Office of Public Health and Science [and subordinate offices]	5,921
GA 1.	United States. General Accounting Office. / United States. Government Accountability Office	5,822
D 101.	United States. Department of the Army	4,792

Language

Finding: 147 languages are represented, raising questions.

U.S. federal documents are produced in multiple languages for a variety of reasons. English is the primary language used, but there are 147 languages represented in this collection profile - twenty-seven languages have at least twenty-five associated bibliographic records. While many of these are, in fact, U.S. documents, an investigation of records associated with certain languages may reveal records for materials that are not. For example, Ancient Greek is the language coded for fifty-six records. And while we can't rule out a Smithsonian or Library of Congress publication in that language, it's likely that many of these are out of scope.

The data on publication language was pulled from the MARC 008 (characters 35-37) field and the MARC 041 field (subfields a, d, e, j). 737 records were missing language information.

Table 5. Top five languages represented, with number of bibliographic records.

Language	**Number of Bibliographic Records**
English	401,806
Spanish	3,582
French	2,585
German	2,037
Russian	1,101

Usage

One metric that we are beginning to explore is the usage of those objects identified as federal documents in the HathiTrust Digital Library. We are still determining the best method for identifying the most requested objects, but we have identified the ten objects with the most unique sessions, based on PageTurner log data[9].

[9] PageTurner is the application that displays digital page images and OCR and allows users to download digital content in other formats. Usage was based on analysis of PageTurner logs from October 2015 to October 2016.

Table 6. Top ten most accessed federal documents based on PageTurner log data.

Title	Number of Unique Sessions (10/2015-10/2016)	URL
Library of Congress catalogs: 1976 V.4	47793	https://babel.hathitrust.org/cgi/pt?id=mdp.39015082933030
Annual report of the Commissioner of Patents for ... 1916	43958	https://babel.hathitrust.org/cgi/pt?id=njp.32101049919598
Smithsonian physical tables.Smithsonian Institution	42413	https://babel.hathitrust.org/cgi/pt?id=mdp.39015002910647
History of military mobilization in the United States Army	19993	https://babel.hathitrust.org/cgi/pt?id=uiug.30112012299092
The strike at Lawrence, Mass. Hearings before the Committee	16157	https://babel.hathitrust.org/cgi/pt?id=nyp.33433031320033
List of pensioners on the roll	11350	https://babel.hathitrust.org/cgi/pt?id=mdp.39015012089846
Catalog of copyright entries. n.s. pt.3 v.40 no.2 1945 Music	10114	https://babel.hathitrust.org/cgi/pt?id=mdp.39015077982075
Catalog of copyright entries. Ser.3 pt.11B v.4-6 1950-1952 Labels	8545	https://babel.hathitrust.org/cgi/pt?id=mdp.39015084451015
A short guide to New Zealand	7948	https://babel.hathitrust.org/cgi/pt?id=uiug.30112101024682
Library of Congress catalogs: 1978 v.8	6561	https://babel.hathitrust.org/cgi/pt?id=mdp.39015082940548

Comprehensiveness Analysis

The comprehensiveness analysis is different than the rest of this report's analysis because it incorporates metadata from the U.S. Federal Documents Registry that is not in HathiTrust.

The goal of the Registry is to define the comprehensive corpus of U.S. federal documents, in order to determine what is left to be digitized and deposited into HathiTrust. The Registry is a

work in progress and does contain some duplication of titles, so some of the comprehensiveness numbers may be overestimated. A "comprehensiveness determination" comparison between the Registry and the digital collection was conducted for six titles and one agency. The titles were chosen because they are popular titles, some have been cataloged in multiple ways (ie, serial vs monograph), and it's clear that these titles are not completely represented in HathiTrust.

As a part of this process, HathiTrust records for these objects were compared to Registry records. Records for specific titles were identified by SuDoc stem and OCLC number, and specifications were developed for normalizing the enumeration and chronology of those titles. Records for the Civil Rights Commission publications were identified only by SuDoc stem. Given the available data regarding the percentage of records lacking a SuDoc, it is likely that not all records have been identified for comparison.

Table 7. List of Titles, along with percentage of comprehensiveness. HathiTrust data is current as of September 1, 2016. Registry data is current as of November 7, 2016.

Title	**# of objects in HathiTrust**	**# of records in the Registry**	**Percentage of Comprehensiveness**
Congressional Record [bound][10]	486	16,383	2.97%
Statistical Abstract of the United States	132	229	57.64%
United States Reports	1,138	2,519	45.18%
Foreign Relations of the United States	1,020	6,830	14.93%
Congressional Serial Set	13,855	43,187	32.08%
Economic Report of the President	32	146	21.92%
Publications of the Civil Rights Commission	870	3,078	28.27%

[10] The bound edition of the *Congressional Record* was chosen instead of the Daily Edition of the same title, partly because some libraries only retain the bound edition.

Conclusions

This initial analysis has given us a new window into our federal documents collection. It provides a benchmark for future analysis, and points to avenues for further exploration of the data. The analysis process has also helped to identify areas of potential collaboration between HathiTrust program areas and operations.

The variable quality of federal documents metadata has long been recognized and, not surprisingly, our analysis reconfirms that data quality will be an ongoing issue in describing the collection and identifying gaps that need to be filled. Given the problems with corporate author and the number of records lacking a SuDoc number, determining the completeness of a particular agency's collection will be a challenge. Our effort to determine the comprehensiveness of Civil Rights Commission materials is a first step in identifying how this work may be expanded to other agencies.

Other data limitations that have been identified include:

- Incomplete bibliographic data (records missing publisher, language, etc.)
- Enumeration and chronology - attempting to determine duplication among serials and multi-part monographs can be challenging due to varying library practices
- Local practice regarding cataloging (ie, cataloging a publication as a monograph vs. a serial; enumeration and chronology data recorded), binding decisions

HathiTrust staff have developed strategies to help address these limitations; namely, the development of specifications for enumeration and chronology of certain titles. So far this approach has proved successful in reducing the overall number of records, and we intend to look for ways to apply similar strategies to this set of data challenges.

Areas for Further Exploration:

While cognizant of the challenges, our goal is an accurate picture of the collection that enables targeted collection-building through digitization and contributions of digitized documents, and also can enable richer more informed discovery and access. Some activities that we will undertake:

- Begin regular reporting of the more authoritative metrics in the report

- Continue to pursue and develop additional strategies for overcoming the limitations of data, such as:
    - Further investigations into corporate author and publisher, and best ways to reconcile, via authority records or other methods
    - Continue to develop specifications for enumeration and chronology
- Continue to monitor comprehensiveness of selected titles and identify additional agencies/publications for which to determine comprehensiveness
- Investigate HathiTrust's process for the assignment of rights attributes as it relates to federal documents. Collaborate with HathiTrust staff to improve and scale processes and develop mechanisms to open up more federal documents