Submitted on: 15.06.2016

Detecting US Federal Documents to Expand Access

Mike Furlough,
HathiTrust, Ann Arbor, USA
E-mail address: furlough@hathitrust.org

Valerie Glenn,
HathiTrust, Ann Arbor, USA
E-mail address: valglenn@umich.edu

Abstract:

This paper reports on HathiTrust's Federal Documents Program, which facilitates collective action to create a comprehensive digital collection of United States government publications issued by the Government Printing Office and other agencies. Most government information is now produced and distributed digitally, but US research libraries, especially those that participate in the Federal Depository Library Program, hold large numbers of historical print publications that are difficult to discover, find, and use. In June 2016 HathiTrust held over 700,000 items identified as federal documents, but we know this to be only a fraction of what exists. Because of varied cataloging practices we have limited understanding of the number of federal documents at the title level, as well as the corresponding number of volumes, the number of pages, and their distribution across libraries in North America. All of these are important details necessary to plan comprehensive mass digitization of federal documents. A major component of HathiTrust's program has been the development of the US Federal Documents Registry, envisioned as a reliable inventory of items published at the expense of the US government. The methodology employed for the Registry's development includes extensive comparative bibliographic analysis, based upon more than 20 million records submitted by 40 libraries in response to a request from HathiTrust. This paper describes methods of de-duplication, relationship-detection, and record consolidation. While many potential use cases exist for such a registry, its primary role is as a tool for identification of materials to be digitized among HathiTrust member libraries and in partnership with other agencies and groups.

Keywords: US federal documents; metadata analysis; digitization

HathiTrust's US Federal Documents Program

HathiTrust is a cooperative organization of over 110 research libraries around the world with a mission to "contribute to research, scholarship, and the common good by collaboratively collecting, organizing, preserving, communicating, and sharing the record of human knowledge."[1] Central to this ambitious mission is HathiTrust's digital library, a digital preservation and access repository which in June 2016 held more than 14.5 million books, serials, and documents digitized from research library collections. The HathiTrust collection largely includes materials produced through mass digitization programs, particularly Google's well known book scanning project (first known as the Google Library project), but also projects led by Microsoft and by the Internet Archive. HathiTrust members also contribute items that they have digitized locally from their collections.

HathiTrust was founded in 2008 by the academic institutions that then formed the Committee on Institutional Cooperation and the University of California.[2] The early focus of HathiTrust was to develop cooperative infrastructure to ensure the long term preservation of mass digitized items and to develop new modes of digital access tailored for academic uses. This aggregation of millions of items from libraries affords an opportunity to support large scale cooperative programs that would be inconceivable at the local level. These include the HathiTrust Research Center (https://www.hathitrust.org/htrc), devoted to developing services for non-consumptive text and data mining, and a Shared Print Monograph Program (https://www.hathitrust.org/print_monograph_archiving), which will ensure the preservation of print collections that correspond to HathiTrust's digital collection.

In 2011, the membership of HathiTrust held a meeting organized to determine the future of the organization and there identified US Federal Documents as an area for strategic investment and cooperative activity.[3] Members endorsed a resolution to "expand and enhance access to U.S. federal publications" through the creation of a "comprehensive digital corpus of U.S. federal publications including those issued by GPO and other federal agencies."[4] Most government information is now produced and distributed digitally, but US research libraries, especially those that participate in the Federal Depository Library Program (FDLP), hold large numbers of historical print publications.[5] Nearly all US members of HathiTrust currently receive federal documents as part of the program and/or hold collections of federal documents acquired via other means. These collections occupy hundreds of miles of shelf space in their stacks and are notoriously challenging for general users to access due to complexities of publication history, cataloging, and format. The historic run of these print publications contains an enormous trove of information about US and international history, policy, economics, science, and law.

The program was important to HathiTrust's membership for many reasons. Many members of HathiTrust had decided to prioritize digitization of federal documents early in their partnerships with Google and others, as well as via local and inter-institutional cooperative projects. Digitization of federal documents offered these libraries an opportunity to collaboratively address discovery and access challenges for federal documents that result from decades of legacy cataloging and metadata practices. Because US federal publications are by law not protected by copyright they can, for the most part, be digitized and made accessible online without concern for infringement or the need to seek permission from rights holders. (Exceptions can include third party inserts included in publications and publications from agencies such as the Smithsonian Institution.) Thus digitization of the documents enables full-text indexing and online viewing, thus reducing the need for physical access and making it possible to consider moving such collections to low-use storage facilities. Finally, for providing widespread comprehensive access to these materials fits squarely within the goals of the depository library program as well as HathiTrust's stated goal to create and "sustain a public good while at the same time defining a set of services that benefits member institutions."[6] Digitization of US federal documents thus serves both the libraries' immediate population of users and the general public.

The US Federal Documents Registry and its Methodology

In June 2016, HathiTrust counted more than 733,000 individual items in its digital collection as federal documents. However, it has been impossible to estimate the total number of documents in existence that were produced at federal expense. Estimates range from as low as 1.5 million to as many as 3 million. This poses a major challenge to creating a "comprehensive digital corpus of U.S. federal publications." In 2013 HathiTrust began the development of the US Federal Documents Registry, with the goal of identifying the full corpus of US federal documents, including their digitization status. The scope has been intentionally broad, and could ultimately include grant-funded or contract works, declassified materials, individual pieces of legislation (bills), administrative publications, and/or numerical data sets. The central purpose of the Federal Documents Registry is to define the full corpus and to inform collaborative digitization and collection building activities.

The project team has had to directly confront the the legions of metadata challenges presented by these federal documents. A HathiTrust advisory group for the Federal Documents program has summarized these as

> a) inaccuracies in government documents' status in cataloging records, b) metadata that inadequately represent the publications and their critical relationship to other resources, and c) differences in the cataloging policies and practices across of libraries contributing records to HathiTrust.[7]

Items can be cataloged variously as both a monograph and a serial. Changes in agency name and federal publication patterns over the last two centuries present obvious problems. Variations in local cataloging and collection practices more deeply complicate the challenge of identifying anything close to a comprehensive collection. Local library cataloging of federal publications has been selective and spotty over time. Libraries have also approached binding of government publications variously, so that the purported same "volume" from two different libraries are not identical in content. The GPO *Catalog of Government Publications* provides a critical source of metadata about federal documents, but it is known to be an incomplete representation of the universe of these materials, and does not offer data about local holdings, which is necessary to identify source documents for digitization. This recitation of metadata inconsistencies does not come close to being complete.

If the ultimate goal is to digitize items held by libraries, it is necessary to work from existing metadata describing the collections in these libraries. Clearly, a very broad record set was required to fulfill these challenging requirements. In the fall of 2013, HathiTrust issued an open call for records for US federal government documents.[8] Forty-three libraries responded, including non-HathiTrust members, submitting in total more than more than 26 million records. The majority of these contributing institutions were large public university libraries and members of the FDLP, including eleven regional depository libraries.[9] Those records, along with those from the HathiTrust repository, form the basis of the US Federal Documents Registry. The Registry is now updated daily with new and/or changed records from the HathiTrust repository, and weekly with the records from the GPO Catalog of Government

Publications (CGP). The public interface to the Registry, now in beta release, is available https://www.hathitrust.org/usdocs_registry. As of June 2016 there were 5.5 million records in the Registry, derived from the more than 26 million contributed "source" records.[10] These 5.5 million records contain an unknown number of duplicates and thus cannot yet be used to estimate the number of federal documents in existence.

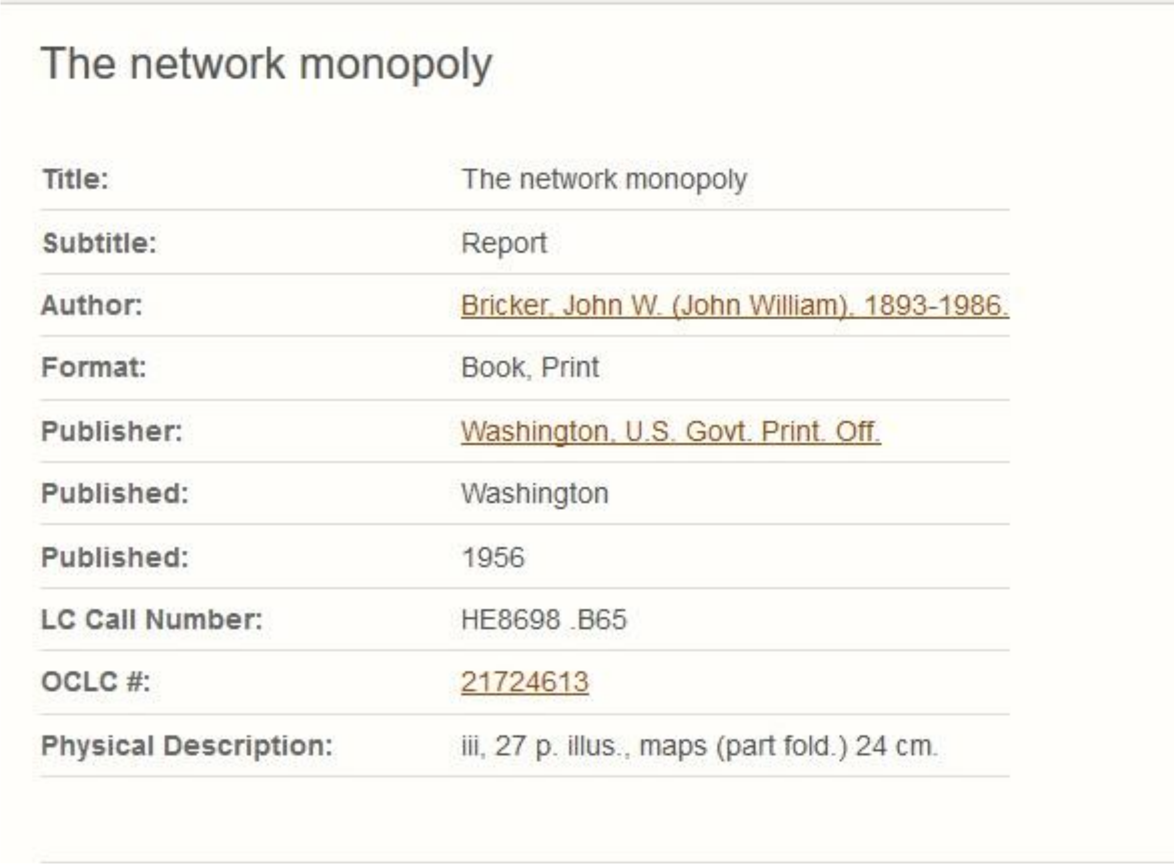
It was known from the beginning that the initial call for records would result in significant duplication among the records submitted, but it soon became clear that we would need to identify and filter records outside of the already broad scope. When classifying materials in the digitized collection, HathiTrust uses automated processes to analyze the MARC 008 field to determine if an item can be identified and treated as a US federal government document.[11] This is an imperfect method. Not all records for US federal documents are coded correctly, and null values in the 008 field have sometimes been truncated during previous processes, thus shifting the key identifying alphanumeric values in the field. This method can also incorrectly include other countries' federal documents produced in the United States and exclude US federal documents produced outside of the US (i.e., embassies; military bases). Yet it is the best possible method available for analyzing bibliographic records for our digital collections at a large scale.

However, in order to cast the widest possible net of potential documents, the call for records for the registry project deliberately did not specify using this method or any other to identify records for US federal government publications. As a result, out-of-scope records for state, non-US federal, and NGO publications as well as some commercially produced volumes were contributed, and filtering was required to winnow them out. Records for items that would seem clearly to be completely out of scope, such as video games and feature films, were also included in some sets.

Our efforts to detect duplicate records first focused on matching unique identifiers in the record. US federal documents, however, are not assigned a unique identifier when they are printed or published. The closest approximation available is the SuDoc number, but the SuDoc is unsuitable as a primary identifier for several reasons. Not all documents have been cataloged and added to OCLC by GPO (thus, the need for a Registry), nor do all libraries use the SuDoc number to arrange their collections. If a library has performed original cataloging on a government document, the SuDoc number may not have been added to the record. Thus we have used additional common identifiers as match points, including the OCLC number, LCCN (Library of Congress Control Number), and ISSN, as well as the SuDoc number. Roughly 75% of Registry records have a SuDoc number (4.1 million out of 5.5 million), and 80% of Registry records have an OCLC number (4.4 million out of 5.5 million). 23,362 Registry records have no identifiers.

Registry records represent individual items, as the aim is to identify physical volumes that have not yet been digitized. Therefore, the same bibliographic data will be associated with multiple Registry records. Upon receipt, source records were run through duplicate detection processes. Initial source records were clustered as duplicate (records for the same item), related (records for parts of the same series), or solo (unique). Records that have identical bibliographic data but variations in item description, also known as enumeration and chronology, are clustered as related. Records in duplicate clusters were assigned confidence scores ranging from 0 to 1, based on the number of common elements (i.e., title, publication date).

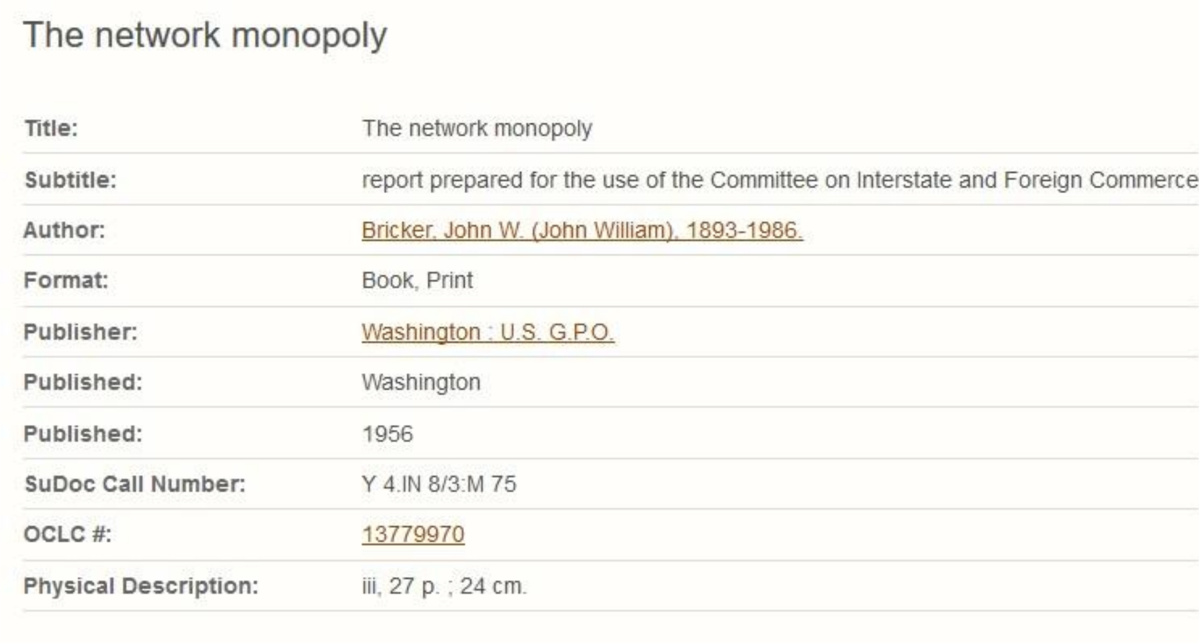
Confidence is based on the number of identical fields (i.e., OCLC number, publication date, title) that two records have in common. Record clusters with a confidence score of 0.8 or above indicate that multiple source records are likely describing the same item, and a Registry record is created. Records are not merged; rather, the Registry record is created by collating multiple underlying source records. The confidence level must be high to create a Registry record, in part to avoid false positives. Often, items have very common titles (e.g., Report) and similar SuDoc numbers, and it can be difficult to determine duplication based solely on information in the metadata records. Figures 1 and 2 below display two bibliographic descriptions of the same title, and demonstrate some of the basic challenges of detecting duplicates in the collection. Figure 1 includes a Library of Congress classification number, while the record in figure two does, and also displays variations in other identifiers, publisher, subtitle, and physical description.



The network monopoly

Title:	The network monopoly
Subtitle:	Report
Author:	Bricker, John W. (John William), 1893-1986.
Format:	Book, Print
Publisher:	Washington, U.S. Govt. Print. Off.
Published:	Washington
Published:	1956
LC Call Number:	HE8698 .B65
OCLC #:	21724613
Physical Description:	iii, 27 p. illus., maps (part fold.) 24 cm.

Figure 1: This record for The Network Monopoly includes an LC call number.



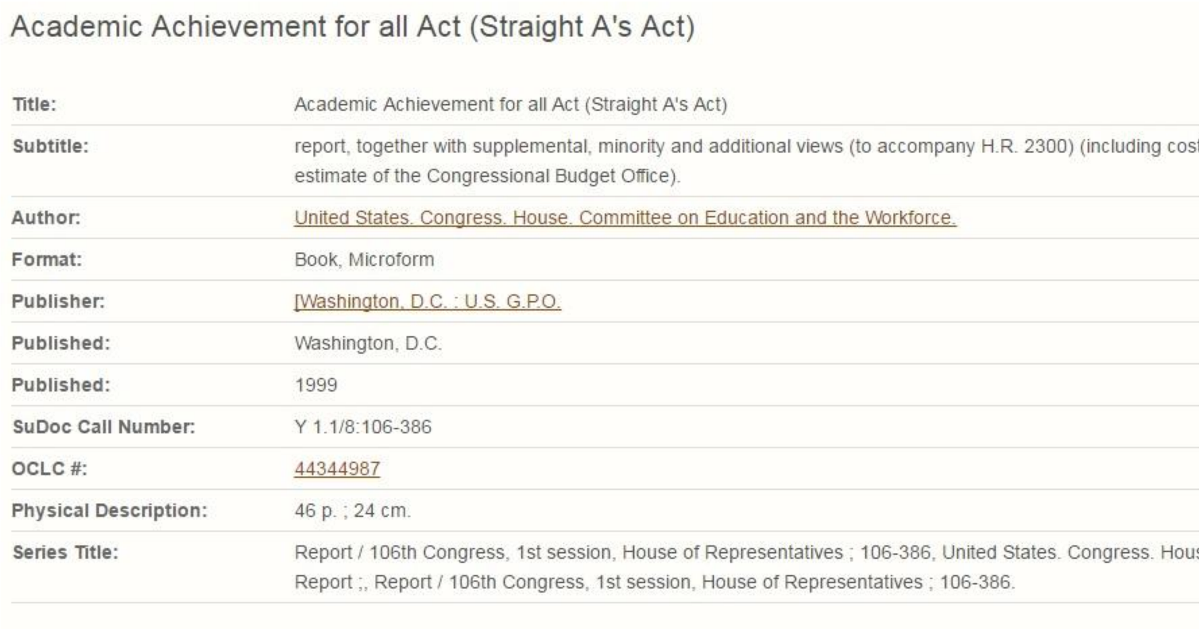
The network monopoly

Title:	The network monopoly
Subtitle:	report prepared for the use of the Committee on Interstate and Foreign Commerce
Author:	Bricker, John W. (John William), 1893-1986.
Format:	Book, Print
Publisher:	Washington : U.S. G.P.O.
Published:	Washington
Published:	1956
SuDoc Call Number:	Y 4.IN 8/3:M 75
OCLC #:	13779970
Physical Description:	iii, 27 p. ; 24 cm.

Figure 2: In contrast this record contains a SuDoc call number, as well as a different OCLC number, and other variations from the record shown in figure 1.

The lack of standardization of enumeration and chronology has been and continues to be one of the greatest challenges for detecting duplicates in the Registry. Different systems store the data in different fields, and libraries may use local specifications for recording information about a particular piece. Many US federal documents have been cataloged at different times as both serials and monographs, but the majority are serials. This makes inconsistent item description even more problematic. Some normalization of enumeration and chronology is performed upon record loading in the Registry. Addressing the “enum-chron” problem, which extends to serials records more generally, will be a continuing challenge for the Registry Project.

Because the Registry has identified the “work,” rather than the “manifestation” as the primary unit to record, we have also reduced the number of duplicate records by grouping item records regardless of format (see figure 3 for an example). This is done based on information in the MARC 776 field, so that print, microfiche, and electronic versions of the same title are clustered together as one item in the Registry



Academic Achievement for all Act (Straight A's Act)

Title:	Academic Achievement for all Act (Straight A's Act)
Subtitle:	report, together with supplemental, minority and additional views (to accompany H.R. 2300) (including cos estimate of the Congressional Budget Office).
Author:	United States. Congress. House. Committee on Education and the Workforce.
Format:	Book, Microform
Publisher:	[Washington, D.C. : U.S. G.P.O.
Published:	Washington, D.C.
Published:	1999
SuDoc Call Number:	Y 1.1/8:106-386
OCLC #:	44344987
Physical Description:	46 p. ; 24 cm.
Series Title:	Report / 106th Congress, 1st session, House of Representatives ; 106-386, United States. Congress. Hou Report ;, Report / 106th Congress, 1st session, House of Representatives ; 106-386.

Figure 3. Registry record for the title "Academic Achievement for all Act (Straight A's Act) in both print and microform formats.

Current Work and Near-term Program Development

The primary use case for the Registry is to identify US federal documents that are held by libraries but not yet digitized and deposited into the HathiTrust repository. Although record clustering and duplicate detection is ongoing, in spring 2016 the project team began the initial analysis necessary to identify items for digitization. The Registry contains roughly 4.7 million records without HathiTrust IDs to indicate that a digital copy of the work exists in the HathiTrust repository. Given that there are still duplicate records within that large set, several methods will be used to develop the gap detection and sourcing process. The ultimate goal is to be able to generate target "pick lists" that could be used by libraries to identify items on the shelf that could be digitized. The lists would include information such as OCLC number, title, publication date, enumeration and chronology, as well as source library. The goal is to ensure that both HathiTrust and the source libraries have enough information to identify the physical item.

Initial tests have focused on specific serial titles published between 1930 and 2000. Focusing on specific serial titles allows for refinement of enumeration and chronology specifications, and allows staff to attempt to identify a publication pattern. With the *Federal Register*, which is produced daily, the project team is investigating the creation of placeholder records that have enumeration and chronology describing each individual issue in the series, and relying less heavily on contributed records, which may have enumeration and chronology for multiple issues bound together. Placeholder records would serve to fill gaps in the metadata, ensuring that the Registry represents a comprehensive picture of the US federal documents corpus. "Item" in the Registry can be an individual issue or volume, or it can a be a bound volume containing multiple publications. Specifications for titles such as the *Congressional Record*, *United States Reports*, and the *Statutes at Large* are currently being developed, in an attempt to further eliminate duplication based on enumeration and chronology.

Because the Registry project itself is based at the University of Michigan Library, the initial list of "missing" items, i.e., items not yet digitized, will be compared with the University of Michigan's collection. However, additional sets of metadata will be brought to bear in this and future analysis, including the HathiTrust print holdings database. HathiTrust collects from all members an annual report of holdings data at the level of the volume. This data is used for various purposes, including shared print analysis, service provision, and fees. This annual extract reports what the library claims to hold in its physical collection, not what has been contributed to the digital collection. Therefore this data can be used to identify which libraries hold items identified as undigitized.[12] Project staff have recently begun an overlap comparison between data in the Registry data and the HathiTrust print holdings database. Currently staff are focusing on records with OCLC numbers marked as government documents in the print holdings database that do not appear in the Registry. These methods offer a promising method of sourcing materials for digitization, but were still in the earliest phases of development when this paper was drafted, and thus no data on the results of these analyses can be provided here.

While this work is underway, HathiTrust is developing general plan for coordinating digitization among its member libraries. Historically HathiTrust has not undertaken a coordinating or planning role for digitization among its members, so this will require new planning and policy. Among the issues still needing confirmation are the following:

- *Identifying source holding libraries.* Although many libraries may hold items that need to be digitized, it may be that libraries currently involved in mass digitization activities would be approached first. Other criteria for a source library could include the strengths and profile of that library's collection and whether it can withdraw and scan items destructively.
- *Publication date.* Because we have more comprehensive metadata/records for items published after 1976, it is probable that it will be easier to locate undigitized items.
- *Prioritizing materials cataloged as monographs versus serials*. In theory, monographs will present fewer issues with enumeration and chronology, which should mean less challenging duplicate detection.

The project team also continues to focus on increasing the comprehensiveness and reliability of the Registry. Other possible near-term activities could include gap detection in Registry metadata, with the eventual hope of creating or adding records for items not yet in the Registry, whether they be serials or monographs. The team will also explore other sources for metadata contribution of current records.

The HathiTrust US Federal Documents Program will leverage the Registry for a number of activities including characterizing HathiTrust's large and growing collection of federal documents, and engaging libraries in collaborative digitization strategies and operations. HathiTrust also intends to assess the potential of the Registry to provide services beyond the current public interface. During initial planning for the Registry project, several additional use cases were identified through consultation with focus groups of interested stakeholders. These additional use cases include: print collection management decisions based on what is available full view in the HathiTrust repository; collection development based on agency, title, and/or subject; and metadata creation or enhancement.[13] These are clearly valuable uses, but we have deferred work on them until we have accomplished significant digitization of missing materials.

Although a comprehensive itemization of every federal document ever published remains an ideal, the HathiTrust US Federal Documents Registry is getting much closer. The lessons learned from aggregating this very large set of federal documents data, and in conducting in-depth analysis of the data, will be put to good use as the Registry becomes actionable for HathiTrust and the library community.

Acknowledgments

We wish to acknowledge the work of the other members of the Federal Documents Registry project team: Bill Dueber, Jon Rothman, Josh Steverman, and Martin Warin. We are also thankful for the comments and editorial guidance provided by Heather Christenson, who joined HathiTrust as the Program Officer for Federal Documents and Collections just as this paper was being finalized.

References

1. The mission of HathiTrust is defined in its organizational bylaws, found online at https://www.hathitrust.org/bylaws.

2. All ten universities within the University of California are included here. The members of the Committee on Institutional Cooperation in 2008 were the University of Chicago, the University of Illinois, Indiana University, the University of Iowa, the University of Michigan, Michigan State University, the University of Minnesota, Northwestern University, Ohio State University, the Pennsylvania State University, Purdue University, and the University of Wisconsin-Madison.

3. This meeting, known as the Constitutional Convention, is documented at https://www.hathitrust.org/constitutional_convention2011.

4. The text is quoted from the original proposal presented at the Convention, found online at https://www.hathitrust.org/constitutional_convention2011_ballot_proposals#proposal4.

5. Founded in 1813, the FDLP (http://www.fdlp.gov) became the responsibility of the Government Printing Office (GPO) in 1895, and now includes over 1,100 members.

6. Quoted from the HathiTrust Bylaws.

7. Government Documents Initiative Planning and Advisory Working Group. “A Status Report and Set of Recommendations for Continued Action on Building a Comprehensive Collection of U.S. Government Documents in the HathiTrust Digital Library,” page 5, October, 2014. The rest of the paragraph summarizes observations found on the same page of the report.

8. The text of the Call for Records, along with an FAQ, is available at https://www.hathitrust.org/usgovdocs_call-for-records.

9. A complete list of contributing institutions is available at https://www.hathitrust.org/usdocs_registry/about.

10. An estimated 5 million of the 26 million contributed records were excluded, as the information contained in those records is too minimal to interpret.

11. HathiTrust performs this analysis on all items deposited into its repository in order to determine the copyright and view status of the work. It is documented at https://www.hathitrust.org/bib_rights_determination.

12 . Specifications for the print holdings data collected by HathiTrust are found at https://www.hathitrust.org/print_holdings.

13. These additional use cases are documented at https://docs.google.com/document/d/18fR-lpoTGbBpsXHFnSQspiQcuMHhW4OzKpvZnz3UfA0/edit.