Against the Grain

Volume 29 | Issue 6 Article 8

December 2017

The HathiTrust Federal Documents Program-Towards a Digital U.S. Federal Documents Library at Scale

Heather Christenson
HathiTrust, christeh@umich.edu

The HathiTrust Federal Documents Program: Towards a Digital U.S. Federal Documents Library at Scale

by **Heather Christenson** (Program Officer for Federal Documents & Collections, HathiTrust) <christeh@umich.edu>

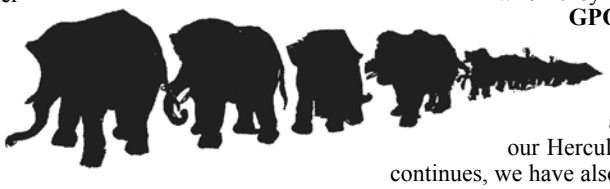
HathiTrust is a collaborative organization founded in 2008 as a solution to the need of a group of libraries to preserve and provide access to large amounts of digital content produced via mass digitization partnerships with **Google** and the **Internet Archive**. As **HathiTrust** has grown to over one hundred and twenty member libraries, the digital library has evolved to encompass mass-digitized volumes contributed by the many additional members who have joined since its founding, as well as locally-digitized and some born-digital volumes. By way of its emergence from libraries themselves, **HathiTrust** is rooted in an attitude of service to end users, continuous improvement, and use of standards, and is attuned to the needs of libraries. **HathiTrust** runs one of the few repositories certified as a Trusted Digital Repository under the **Center for Research Libraries'** Trustworthy Repositories Audit Certification (TRAC) process. Over the nine years of its existence, **HathiTrust** has developed robust access services at scale, including full text and catalog search, flexible viewing functions, and metadata download features. Additionally, **HathiTrust** hosts a copy of the digital library in the **HathiTrust** Research Center, which provides infrastructure, software tools, and services for computational access and research.

As of this writing, **HathiTrust** offers close to 15.8 million digital volumes for use, with close to 6 million fully viewable volumes in the public domain. Items in **HathiTrust** are the result of hundreds of years of library stewardship. Libraries selected and held the volumes, cataloged them, maintained them as they were accessed and used, transferred the cataloging to machine readable records, created digital versions of the volumes (whether on their own or in partnership with commercial entities such as **Google**), and now store these versions in **HathiTrust**, in a powerful digital aggregation that would not have been possible without a collective investment over time.

HathiTrust U.S. Federal Documents Program

U.S. federal documents within **HathiTrust** largely result from a particular formal stewardship program, the **Government Publishing Office (GPO)'s Federal Depository Library Program (FDLP)**. The **FDLP's** mission is "to ensure that the American public has access to its Government's information," and currently has 1,143 participating libraries.[1] Eighty-four **HathiTrust** member libraries participate in the **FDLP** as depository libraries, so this shared interest in federal documents is reflected in **HathiTrust's** priorities.

HathiTrust now includes about 1,045,000 items identified as U.S. federal documents, possibly the largest existing publicly-accessible digital collection of these materials. As of this writing, forty-nine member libraries have contributed digitized federal documents to **HathiTrust**. This total includes contributions from **FDLP** collections as well as documents from collections developed to meet needs and purposes outside of the **FDLP**. The **HathiTrust** digital collection brings all these documents together in a large-scale aggregation of federal documents that reflects the scale and scope of **FDLP** collections, but also draws richness from the inclusion of topically-focused documents collections.



The **HathiTrust Federal Documents Program** seeks to build the digital collection, and enrich discovery and access for end users. This relatively new program is a result of many years of investment and effort amongst staff and the community of member libraries. In 2011, at the **HathiTrust** "Constitutional Convention," members enthusiastically approved a proposal to build a comprehensive digital collection of U.S. federal documents in **HathiTrust**. Since then, **HathiTrust** staff have developed the U.S. Federal Documents Registry database, intended to provide an inventory of all known published federal documents. During the same time period, a collaborative working group of members assessed the many challenges and opportunities presented by an aggregate digital federal documents collection, and articulated a set of strategic priorities that led **HathiTrust** to establish this new program.

The Program positions **HathiTrust** to make progress on a number of fronts: in particular, print preservation, digital collection building, and enrichment of discovery and access for end users. By aggregating digitized federal documents in a collectively managed digital library, libraries can solve some longstanding issues and accommodate new kinds of uses. Over time, large collections of print federal documents have accumulated on library shelves, and for some libraries, maintenance costs persist while documents are underused. A related challenge is that depository library collections may not be well represented in some library discovery environments, since many libraries had historically chosen to minimally catalog documents, especially those published prior to 1976, which is when **GPO** began sharing metadata in the Catalog of Government Publications.

HathiTrust launched a Shared Print Program last year with a goal to secure retention commitments for print monograph items that have digital counterparts within **HathiTrust**. Out of 4.8 million monograph titles committed to the Shared Print Program so far, over 222,000 are monographic federal documents committed for retention in at least one **HathiTrust** member library. Later rounds of shared print planning will seek to secure commitments for non-monographic documents. Currently, items committed to the Shared Print Program must be lendable to all **HathiTrust** members, which may be problematic for some depository libraries that have committed to retain print documents within the **GPO's Federal Information Preservation Network (FIPNet)** program and are required to designate those copies as non-circulating. This policy may be revisited in a later phase.

Collection Building

HathiTrust intends to build a comprehensive digital collection of U.S. federal publications distributed by **GPO** and other agencies. This is an ambitious endeavor, given the challenges of identifying every document originating from a government entity, and publication practices have not necessarily been standard across governmental units or time. Despite a long history of libraries, **GPO**, other agencies, and commercial entities producing catalog records, there is no one place to go for a record of every U.S. federal document ever published. **HathiTrust's** answer has been the development of the U.S. Federal Documents Registry Database, known as "the Registry." To build the Registry, **HathiTrust** solicited over twenty million bibliographic records from forty libraries, and has spent several years consolidating them via bibliographic analysis to de-duplicate and detect relationships, resulting in a database of around 5.3 million records. The database is enriched by regular updates of metadata provided by **HathiTrust** member libraries when they deposit digital documents, and also includes **GPO** metadata. Recently, we have incorporated the **Library of Congress** Name Authority File in order to more reliably identify agency authors.[2] The comprehensiveness and accuracy of the Registry is improving, and as our Herculean quest to identify the full set of documents continues, we have also begun to make use of the Registry to identify gaps in the **HathiTrust** federal documents collection.

Another piece of the comprehensive collection challenge has been gaining an understanding of what we have accumulated already: users have access to over a million U.S. federal documents within **HathiTrust** as a result of mass digitization and aggregation, but what exactly is this collection? We took a deep dive into collection analysis and published the results in early 2017. Not surprisingly, publication dates largely followed a curve mimicking overall government publishing, and a rich variety of subject matter is present with a wide distribution of agency

continued on page 24

The HathiTrust Federal Documents Program ...
from page 23

authors. This "Collection Profile"[3] snapshot enabled us to establish the baseline on which we are building the federal documents collection.

Based on this analysis we have set specific collection development priorities in consultation with the **HathiTrust Federal Documents Advisory Group**.[4] The priorities were chosen after considering a number of criteria including recommendations of **HathiTrust** working groups, widely known and consulted series distributed by **GPO**, titles commonly held by **HathiTrust** member libraries in print, synergies with the broader **HathiTrust** collection, synergies with other large collaborative endeavors related to federal documents, potential for **HathiTrust Research Center** use, and finally (and importantly), compelling and broader general interest for both member libraries and end users.

Discovery and Access

HathiTrust offers users the ability to build curated collections via its Collection Builder tool, which we have used to establish a U.S. Federal Documents Collection that will be curated and maintained by staff. This new collection provides end users with a way to filter searches to *only* include federal documents. As we build new subsets of federal documents, we are adding searchable collections for those as well. For example, we have created a collection of *Statistical Abstract of the United States*, an annual compendium of U.S. statistics beloved by librarians for providing the most commonly asked-for statistics all in one place, with references to more in-depth sources. *Statistical Abstract* had been published by the government since 1878 but was discontinued in 2012 when the government program that produced it was eliminated, although a commercial version is now produced by **ProQuest**. The digital surrogates in the **HathiTrust** collection are created with access in mind, but will be preserved for the long term, as well.

HathiTrust's full text search feature solves some classic federal documents discovery problems, for example, locating items of interest out of over fourteen thousand volumes of published federal reports and Congressional documents commonly known as the "Serial Set," or unexpectedly uncovering federal documents in the course of a broad search. Digital federal documents are freed from shelf order, and can be accessed or grouped topically, by date, with or without non-government works, or in any number of other flexible ways depending on user needs. Within the **HathiTrust Research Center,** scholars may look through the lens of federal documents over time and across agencies, and see paths of evolution for government, politics, social issues, health issues, culture, and more via computational analysis. Additionally, federal documents collections can be imported into the **HathiTrust Research Center** environment as worksets for computational analysis.

We are pursuing a number of avenues to provide a better experience for end users. As librarians well know, one of the biggest barriers to better discovery is metadata quality, and this is especially true for federal documents. Access depends on documents being available to users for reading and download, and **HathiTrust** relies heavily on metadata to determine the rights status of publications. Federal documents are largely in the public domain with a few exceptions, but inaccurate and incomplete bibliographic metadata can result in the interface providing only limited view for end users. Metadata remediation has the potential to open up viewability for many federal documents. To improve both discovery and access, we are exploring targeted metadata remediation in collaboration with our member libraries and the staff of **Zephir**, **HathiTrust's** metadata management system,[5] as well as continuing to enrich our Registry database and exploring automated metadata creation projects in partnership with scholars.

HathiTrust is also focused on quality of its digital objects, and a member-led working group is currently developing a schema to characterize quality for end users and use cases. **HathiTrust** has an active community that is very interested in quality and reports to us on it. Over the last six years, our User Support Working Group has received and successfully resolved over 2600 quality issues. The Federal Documents Program is planning user experience research to better understand specific needs for discovery and access to the documents within the **HathiTrust Digital Library**, and to understand user experience problems inherent to documents that can be addressed through improved interface design. We are also exploring needs related to federal documents content sets and analysis within the **HathiTrust Research Center**.

Looking Ahead

We have our work cut out for us in the near future, with a priority to intentionally develop the digital federal documents collection and services in order to realize the value of this tremendous community asset. We will continue filling collection gaps through digitization of print, and are launching collaborative projects to do so. Since our goal is comprehensiveness, in the coming year we plan to investigate possibilities for incorporation of born-digital and web-archived federal documents into **HathiTrust**. The range of possibilities and quality of experience for end users will continue to improve not only as we grow our overall federal documents collection, but also as we delineate specific collections for access in the both the **HathiTrust Digital Library and Research Center**, and as we improve metadata, assess quality, and ensure that federal documents are available in full view. **HathiTrust** has a relatively small staff and large ambitions, so our success will depend on working collaboratively across our membership and with the broader library community.

Endnotes

1. **U.S. Government Publishing Office,** *LSCM FY 2016 Year in Review*, accessed November, 25, 2017. *https://www.fdlp.gov/file-repository/about-the-fdlp/lscm-year-in-review/2843-lscm-fy2016*
2. **Joshua Steverman**, "Problems with Authority," *Library Tech Talk* (blog), last modified August 10, 2017. *https://www.lib.umich.edu/blogs/library-tech-talk/problems-authority*
3. **Heather Christenson**, "Federal Documents in HathiTrust: A Look at Our Collective Collection," Perspectives From HathiTrust (blog), last modified March 20, 2017. *https://www.hathitrust.org/blogs/perspectives-from-hathitrust/federal_documents_collective_collection*
4. **HathiTrust**, "HathiTrust Federal Documents Collection Framework," accessed September 8, 2017. *https://www.hathitrust.org/hathitrust-federal-documents-collection-framework*
5. **HathiTrust**, "Zephir, the HathiTrust Metadata Management System," accessed September 8, 2017. *https://www.hathitrust.org/zephir*

Rumors
from page 16

in the areas such as library systems, cataloging (especially in the context of special library collections) and on collection development issues associated with autism. On a lighter note — he has a cooking blog and a somewhat over-the-top obsession with squirrels and cats (talk about diverse!). *http://www.against-the-grain.com/2017/10/atg-the-podcast-episode-with-corey-seeman/*

The past several weeks have seen **Michael Paul Pelikan** swept into an increasing tempo of medical interventions related to his ongoing back saga that has eclipsed much of his normal activity. **Michael** was appalled to discover, amidst examinations, imaging, treatments, and injections, that he missed the deadline for his Dec/Jan column! Shame on him! Moving right along, **Michael** just underwent surgery to fuse his right ilium to his sacrum, his fourth surgery on his lower back. OUCH!!!! He says he will be down for a minimum of six weeks, and perhaps longer based on his recovery time. **Michael** says that ***Against the Grain*** and his "Antidisambiguation" column, have been a marvelous outlet and a source of considerable satisfaction for him as well as for all of us including Yr. Ed. **Michael** recalls a luncheon in Anchorage many, many years ago, for the Editor in Chief of one of the two important newspapers in town who was going on a leave of absence, and was being feted and roasted by colleagues and friends. In his comments after all

continued on page 43