Collaborative Librarianship

Volume 8 | Issue 3 Article 5

2016

Building a U.S. Federal Government Documents Collection in HathiTrust

Heather Christenson
HathiTrust, christeh@hathitrust.org

Building a U.S. Federal Government Documents Collection in HathiTrust

Heather Christenson (christeh@hathitrust.org)
Program Officer for Federal Documents and Collections, HathiTrust

Abstract

The HathiTrust Digital Library encompasses over 760,000 federal documents digitized from print. HathiTrust has recently begun to focus attention on further developing this collection via the U.S. Federal Documents Program. The program will leverage the power of HathiTrust infrastructure, services, and member contributions and will focus not only on collection building, but also on the enrichment of discovery and access for end users. This article provides history of HathiTrust's investment in federal documents, background on the program, a description of current goals and activities, and a brief look at the future.

Background

Launched in 2008, HathiTrust is well known as a collaborative digital library composed primarily of texts digitized from print. At this writing in late 2016, HathiTrust has over 120 member libraries and almost 15 million volumes in its shared collection.[1] HathiTrust offers a variety of services for users including catalog and full text search, distributed user support, and computational analysis via the HathiTrust Research Center. HathiTrust members participate in a shared governance structure that guides development and services for the shared collection. Via committees and working groups, members collaborate on important areas such as collection development, rights, quality, and metadata policy.

HathiTrust's collection is largely the result of mass digitization projects conducted since 2005 by U.S. research libraries in partnership with Google and, to a lesser extent, the Internet Archive. Mass digitization has been focused on a library or collection at a time rather than being selective at a finer level. Mass digitization of U.S. federal documents dates back to the beginnings of the Google Library Project, well before the founding of HathiTrust. Early partners with Google, especially the Big Ten Academic Alliance (BTAA) universities (then known as the Committee on Institutional Cooperation),[2] worked with Google to prioritize digitization of federal documents beginning in 2005, and were later joined by additional Google library partners such as the University of California and Cornell University. Other libraries, including the University of Florida and the Library of Congress, have partnered with the Internet Archive to digitize federal documents. Large digitization efforts have expanded in recent years to include federal agencies and collaborations such as the Center for Research Libraries' Technical Report Archive and Image Library (TRAIL).[3]

HathiTrust's initiative to create a U.S. federal documents collection dates from 2011, when members at the "Constitutional Convention", a gathering of the membership, approved a proposal to build on previous work and create a comprehensive collection of these materials.[4] Since then, HathiTrust has tackled the challenge of building this collection on a number of fronts: inventorying the universe of U.S. federal documents by building a database known as the U.S. Federal Documents Registry,[5] focusing on member deposit of mass-digitized federal documents into the repository, and convening a group of member library experts who articulated a strategy for federal documents[6] leading to the recent establishment of the HathiTrust U.S. Federal Documents Program in 2016.

The HathiTrust U.S. Federal Documents Program will leverage the power of HathiTrust infrastructure, services, and member contributions and will focus not only on collection building,

but also on enriching discovery and access for end users. HathiTrust has appointed a new advisory committee to consult with the Program Officer and ensure that program activities serve the interests and needs of the partnership. As the program develops, the focus will be on working within the library community to solve shared problems.

With a large and invested membership community, a growing digital collection, infrastructure for discovery, access, and preservation, along with the U.S. Federal Documents Registry database, HathiTrust is well-positioned as a locus of collaboration to improve digital access to U.S. federal documents.

HathiTrust's Investment in Federal Documents

By virtue of its membership, HathiTrust is committed to the inclusion of federal documents in its collections. Eighty-four HathiTrust member libraries also participate in the Federal Depository Library Program (FDLP),[7] and most other member libraries include federal documents in their collections. In addition to participation in the FDLP, many HathiTrust members also belong to consortia and organizations that have made significant contributions to the digital documents landscape. Among these are ASERL's (Association of Southeastern Research Libraries) Centers of Excellence for cataloging documents,[8] TRAIL's digitization program, BTAA's digitization progress in Google partnerships, and the University of California's FedDocArc project to archive print and digital versions of federal documents.[9] HathiTrust member libraries have played a role in all of these collaborative activities.

Several important factors have driven the creation of a digital collection within HathiTrust. Over time, sizeable collections of documents have accumulated in libraries, taking up costly shelf space, and there has been a strong feeling from the libraries that documents are underused compared to their value. This state of affairs was described in a recent paper by Mike Furlough, HathiTrust's Executive Director:

> *These collections... 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.*[10]

To solve these challenges, the HathiTrust libraries have focused on digitization and aggregation to improve access and provide more flexibility to manage print collections.

HathiTrust currently includes over 760,000 digitized federal documents that will serve as a base for future expansion. Although this enormous collection has accumulated as a result of mass digitization projects, it has also grown from the inclusion of a large number of documents digitized in collaboration with TRAIL, and from individual libraries that have digitized their collections locally and deposited them into HathiTrust.

U.S. Federal Documents Registry

The collection continues to grow via mass digitization, but in order to reach the goal of comprehensiveness, more focused collection development will be necessary. Due to varying cataloging practices, the biggest challenge to building a comprehensive collection of federal documents is understanding the full spectrum of documents that exist. As described in the recent paper *Detecting US Federal Documents to Expand Access*, "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."[11]

The Registry database began with a set of over twenty million records contributed by forty libraries. It is intended to provide a full inventory of titles and volumes associated with those titles, and now includes 5.3 million records that have been consolidated via bibliographic analysis to de-duplicate and detect relationships. A primary use case for the Registry is to identify U.S. federal documents held in libraries but not yet digitized and deposited into the HathiTrust repository. The Registry holds promise for comparison of library holdings to HathiTrust and to the full inventory of federal documents, as well as support for HathiTrust's ability to create definitive collections. A user interface has been developed

for the Registry, enabling librarians or end users to search the database. Future Registry use cases and development are currently being evaluated, including those related to metadata remediation and enhancement.

Discovery and Access Benefits

The size of the collection and aggregation within HathiTrust provide significant benefits to libraries and end users. HathiTrust provides a number of end-user services across its entire collection including basic and advanced bibliographic search, full text search, a collection-building tool, and services for blind and print-disabled users. In particular, the ability to search the full text of over 750,000 federal documents is a major benefit that dramatically increases the research value of federal documents to users.

In keeping with its public service mission, HathiTrust takes a broad approach to access, opening up materials to the extent permitted by law. In general, the viewability of digital volumes within HathiTrust is based on bibliographic metadata:

> *All objects in the archive are either in the public domain, have the necessary permissions to support the level of access afforded, or are simply archived in such a way as to ensure an enduring copy of the content. HathiTrust only provides reading access to those publications where permitted by law or by the rights holder.*[12]

Most U.S. federal documents within HathiTrust are fully viewable by readers, with the exception of a small number of government entities whose publications are subject to copyright, such as the Smithsonian Institute.

HathiTrust is architected so that external services can incorporate and rely upon connections into its collection and services. Every item and every page in HathiTrust has a persistent URL, enabling reliable linking, either individually in web pages, blogs, or social media, or programmatically within services. Similarly, HathiTrust's "search widgets" allow a search box to be embedded in another location. HathiTrust also offers freely available data that may be incorporated into other tools and services such as library catalogs, discovery services, and link resolvers, enabling wider discovery. For example, via this data, the HathiTrust collection including federal documents, is surfaced within the Digital Public Library of America (DPLA),[13] alongside the DPLA primary source materials.

Current Goals and Activities

Collection Building

The U.S. Federal Documents Program is an endeavor to leverage the enormous collaboratively built mass-digitized HathiTrust corpus to develop a more specific collection. In order to build a comprehensive collection of U.S. federal documents, HathiTrust needs to both draw upon the mass-digitized collection and extend it. The program's primary collection development focus is digitized versions of U.S. federal documents distributed in print by the U.S. Government Publishing Office (GPO), but documents distributed by federal agencies outside the GPO will also be included. HathiTrust intends to expand the collection via digitization and deposit of already-digitized documents, and, later, inclusion of born-digital documents.

The program is beginning to tackle collection development by undertaking an analysis to profile the HathiTrust federal documents collection as it stands today. An overall goal of the project is to test the ability to differentiate and characterize the collection based on current state of bibliographic and other data within HathiTrust and the Registry. We are analyzing the collection on characteristics indicating provenance of the digital object (such as contributing institution, digitization agent), bibliographically determined characteristics ((such as agency, SuDoc (Superintendent of Documents) number, publication date, languages)), and usage data. The analysis will establish benchmarks for the collection, and enable HathiTrust staff to identify specific opportunities for collaborative collection building and refinement. Another intended outcome is to determine the best method for providing regular updates of the descriptive analysis of the collection, for the benefit of HathiTrust members and ultimately the greater library community.

Through the Federal Documents Program, HathiTrust will look for ways to coordinate digitization efforts to ensure that the resulting digitized documents continue to aggregate in HathiTrust. The decision to deposit digitized collections rests with individual HathiTrust member libraries, however the goal of a comprehensive federal documents collection is a powerful incentive to mobilize those libraries to fill in the gaps. HathiTrust also plans to seek additional paths to collaborate on large scale digitization projects, keeping in mind economy of scale and cost-effectiveness. Additionally, the program will seek to encourage and coordinate local digitization projects at HathiTrust member libraries.

In order to build a comprehensive collection, in the future we may also look for possibilities to incorporate already digitized and born-digital documents into the HathiTrust Digital Library. To this end, HathiTrust may investigate collaborative projects to incorporate digital documents held by federal agencies or other entities, as well as opportunities related to web-archived documents.

Shared Print

In addition to digital preservation, HathiTrust has initiated print preservation efforts by recently launching a Shared Print Program.[14] The initial phase of the Shared Print Program, focused on monographs, encompasses monographic federal documents. As the two programs progress, they will coordinate to address similar collection analysis challenges. We will also look for synergies between programs in areas such as building infrastructure for databases, services and tools for libraries, as well as in developing analysis techniques and practices.

In addition to focused collection development, there are other issues that deserve and need attention as we build the corpus; many of these are important for the entire HathiTrust collection but are of particular significance for federal documents.

Metadata Challenges

The limitations of existing federal documents metadata present the biggest challenge to targeted collection building, since comparisons and analysis must be based on bibliographic data. For federal documents in particular, cataloging has not been consistent over time or across institutions, and is missing entirely for many documents that were originally organized on the shelf by GPO-assigned SuDoc numbers. Compounding this problem is the fact that the libraries that were likely to catalog documents were also not likely to organize their materials by SuDoc numbers, so might not have coded the record as a federal document. For documents published after 1976, the prospects improve, as that is when the GPO began cataloging and sharing metadata via OCLC.[15]

HathiTrust staff who created the Registry have learned a great deal by bringing together the many millions of records thought to be federal documents and comparing them to distill into Registry records for each document. We will apply our expertise in bibliographic data patterns and analysis, gained in Registry work, to the process of deduping between collections and HathiTrust, and identifying gaps. Examples of this expertise include matching by identifiers and reconciling enumeration and chronology for serial records.

The Registry itself, and the metadata within, will also continue to be refined. We have identified a variety of use cases for the Registry such as outward-facing services or tools for libraries to use, or as a basis for enrichment of metadata within the HathiTrust digital library and HathiTrust Research Center. An assessment of the Registry, currently underway, will determine a path for potential future development.

Improving Discovery and Access

HathiTrust's current discovery environment offers a variety of services, but to a large extent, discovery and access within HathiTrust rely on the quality of existing metadata and digitization. Metadata-supported features in HathiTrust's user interface include catalog search, search facets, and listing of serials in results. However, as previously noted, metadata quality is a particular concern for federal documents more so than

for the broader HathiTrust collection. In addition to existing features, HathiTrust's current user interface has room for improvement to better support federal documents, and better metadata could pave the way. For example, a SuDoc search, or the ability to filter search results to include or exclude federal documents, would greatly improve discovery. Additionally, better authority control for names of federal agencies and other entities would make it easier for end users to zero in on specific document sets.

Metadata remediation is also likely to open up more federal documents to end users. Since viewability in HathiTrust is based on bibliographic data, inaccurate and incomplete data can result in a limited view for end users. It is estimated that approximately 93,000 federal documents are currently in limited view in HathiTrust. This set of documents may include a large number of items without enough cataloging to distinguish them as federal publications. HathiTrust has a distributed user support service that fields inquiries from HathiTrust end users. Anecdotally, an overwhelming majority of the user support issues regarding federal documents involve identification of items as federal documents where HathiTrust does not designate them so.

Quality of digitization is also an important factor for access, and for discovery, since indexed text is derived via optical character recognition (OCR) from the images. HathiTrust has taken steps to address quality issues by forming several new groups. In September 2016 we launched a new Quality Assurance and Standards Working Group that will develop strategies, processes and techniques to make scalable improvements of "digital surrogate fidelity at the item/object level."[16] A new focused user support group will handle digitization and digital object composition errors such as missing pages. This work will have benefits for the full digital library collection and for the HathiTrust Research Center corpus, both encompassing federal documents.

Conclusion

HathiTrust is a community of libraries with shared needs and interests, pooling their resources for shared collections and services. By virtue of serving the interests of its membership, HathiTrust also creates a public good in making millions of digitized volumes available to end users. In a landscape where new U.S. federal information now originates primarily in digital form, many of the retrospective print publications remain to be digitized, and libraries continue to serve an essential role in collectively preserving and providing access to federal documents. HathiTrust members have designated federal documents as strategically important and look to HathiTrust to preserve both print and digital versions comprehensively, through this transition, and beyond.

HathiTrust's development of shared digital and print collections give member libraries a basis to make collection decisions, and set the stage for documents to be discoverable in whatever context is needed. In addition, computational access to a collection of federal documents provides new opportunities for research and learning for scholarly purposes, and also may improve access within the digital library if research results in richer metadata or other new access methods.

A venue for digital versions of federal documents that exists outside of the government, and is developed and maintained collaboratively by libraries, is in itself an important public resource. Through its partners, HathiTrust has developed a valuable complement to the Federal Depository Library Program, expanding access beyond print distribution, and helping to realize the goal of wider access to federal information. The HathiTrust U.S. Federal Documents Program's goal of a comprehensive collection is extremely ambitious, but ultimately will ensure that federal documents will be digitally preserved and made accessible to library patrons, online users, and citizens.

[1] "Statistics and Visualizations," HathiTrust, accessed October 21, 2016, https://www.hathitrust.org/statistics_visualizations

[2] "Member Universities," Big Ten Academic Alliance, accessed October 21, 2016, https://www.btaa.org/about/member-universities

[3] "About TRAIL," Center for Research Libraries, accessed October 25, 2016, https://www.crl.edu/grn/trail/about-trail

[4] "Constitutional Convention Ballot Proposals," HathiTrust, accessed October 21, 2016, https://www.hathitrust.org/constitutional_convention2011_ballot_proposals#proposal4

[5] "Creating a Registry of U.S. Federal Government Documents," HathiTrust, accessed October 21, 2016, https://www.hathitrust.org/usgovdocs_registry

[6] "Government Documents Initiative Planning and Advisory Group Charge," HathiTrust, accessed October 25, 2016, https://www.hathitrust.org/usgovdocs_planning_charge

[7] "Federal Depository Library Program," U.S. Government Publishing Office, accessed October 25, 2016, https://www.gpo.gov/libraries/

[8] "Collaborative Federal Depository Program (CFDP): ASERL's Plan for Managing FDLP Collections in the Southeast," Association of Southeastern Research Libraries, accessed November 14, 2016, http://www.aserl.org/programs/govdoc/

[9] "UC Federal Documents Archive Report," University of California Libraries, accessed November 14, 2016, http://libraries.universityofcalifornia.edu/content/uc-federal-documents-archive-report

[10] Mike Furlough and Valerie Glenn, "Detecting US Federal Documents to Expand Access" (paper presented at the IFLA World Library and Information Congress 2016. Wednesday, June 29, 2016)

[11] Furlough and Glenn, "Detecting US Federal Documents to Expand Access"

[12] "Copyright," HathiTrust, accessed October 25, 2016, https://www.hathitrust.org/copyright

[13] "About," Digital Public Library of America, accessed November 14, 2016, https://dp.la/info/

[14] "Shared Print Program," HathiTrust, accessed November 14, 2016, https://www.hathitrust.org/shared_print_program

[15] "GPO Historic Shelflist," U.S. Government Publishing Office, accessed November 14, 2016, https://www.fdlp.gov/project-list/gpo-historic-shelflist

[16] "HathiTrust Quality Assurance and Standards Working Group," HathiTrust, accessed November 14, 2016, https://www.hathitrust.org/qaswg_charge