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

Submitted by the Government Documents Initiative Planning and Advisory Working Group to the HathiTrust Program Steering Committee October, 2014

EXECUTIVE SUMMARY

Charge to the Government Documents Initiative Planning and Advisory Working Group (GDIPAWG)

1. Develop and deliver the Initiative's overall strategy and plan, including scope and phasing of plan components, an operational model, resource requirements, and deliverables;
2. Recommend investments as needed to the HathiTrust Board of Governors by means of the Program Steering Committee (PSC) in support of the Initiative;
3. Advise the HathiTrust Board, by means of the PSC, on relevant policy issues; and
4. Serve in an advisory capacity and provide oversight to the project while in execution, including
    a. Reviewing and validating major tactical/operational plans;
    b. Supporting project communications to, and advocacy, engagement, and partnership development with the broader government documents community and others; and
    c. Monitoring the progress and accountabilities of the Initiative relative to stated milestones and objectives.

GDIPAWG Membership

Mark Sandler, Chair (CIC); Prue Adler (ARL); Ivy Anderson (CDL); Joni Blake (GWLA); Kirsten Clark (Univ. of Minnesota); Rick Clement (Univ. of New Mexico); Elizabeth Cowell (UC Santa Cruz); Michael Norman (Univ. of Illinois); Mark Phillips (Univ. of North Texas); Jon Rothman (Univ. of Michigan); Judy Russell (Univ. of Florida); Barbie Selby (Univ. of Virginia); Jeremy York (HathiTrust)

Project Status to Date

From the earliest years—2004 and onward—of the Google library scanning partnership, government documents were included in the items being shipped to scanning centers, and were thus included in the content flowing to HathiTrust. Beginning in 2009, the CIC Library Directors charged a CIC group to plan an initiative whereby the CIC libraries could work with Google to give focused attention to scanning U.S. federal documents, with the goal of creating a large and publicly accessible collection of digitized U.S. federal documents. As a result of these and other initiatives, there are now close to 570,000 U.S federal documents in HathiTrust, and the collection grows steadily, largely through the continued emphasis given to this content by the CIC.

- While HathiTrust houses nearly 570,000 documents volumes, various idiosyncrasies in how documents are classified, distributed, processed and bound makes it extremely difficult to determine the extent of the overall corpus against which the HathiTrust collection could be compared. To remedy this, HathiTrust has undertaken to build a registry of U.S. government documents dating back to 1789. Beginning in 1813, the Federal Depository Library Program (FDLP) was instituted to ensure public access by broadly disseminating government publications. The FDLP was ultimately expanded and brought under the management of the Government Printing Office in 1895. Estimates for the size of the overall corpus of federal documents currently range from 1.5 to 3 million volumes, but it is hoped that the Registry Initiative can narrow those estimates, giving partner libraries a more definitive goal as they attempt to build a "comprehensive" digital instantiation of the FDLP collection and other government publications.

Recommendations

In very general terms, the GDIPAWG believe that the "near-term recommendations" below can be pursued with limited additional investments at HathiTrust. The "Intermediate" and "Long-Term" recommendations, however, are likely to require additional cash expenditures and/or further investments in staff resources at HathiTrust, participating libraries, or both.

Near-Term Recommendations:

1) Continue to build a comprehensive registry of U.S. federal documents and compare to the known universe of cataloged library holdings distributed through the FDLP or otherwise acquired.
    a. Locate source libraries for content and facilitate its digitization
    b. Help to facilitate the processing of content once a potential source is identified
    c. Assist potential source libraries with efforts to estimate the overall costs and timelines for surfacing their documents holdings
    d. Help to facilitate print management once digital surrogates are secured
2) Provide, and regularly update, a descriptive analysis of government documents holdings already in HathiTrust.
    a. Maintain a running count of the items held
    b. Provide profiles of the collection by issuing agency, date, and topical or geographic focus
    c. Analyze commonly consulted serials for completeness and gaps
3) Encourage member libraries to continue to build the digital corpus by identifying available cataloged content for either sheetfed and non-destructive scanning.
    a. Encourage Google partner schools in the CIC and University of California system to continue to surface content for scanning
    b. HathiTrust members NOT partnering with Google should be encouraged to deposit federal documents files that were locally scanned or digitized by outside vendors.
    c. Focus initial attention on sourcing, digitizing and ingesting ALL post-1976 cataloged FDLP publications
    d. Either subsequently or simultaneously, source, digitize and ingest pre-'76 content for which catalog records are available, and holding locations can be identified
    e. Identify, prioritize and source "essential titles" for digitization and ingest
4) Enlist the support of documents librarians to analyze, organize, and promote the existing corpus.
    a. Near-term activities could be pursued in this arena. A formal communications plan to reach out to documents librarians and users could be an Intermediate-term activity.
5) Actively pursue partnerships with the Government Printing Office, publishing agencies, and national and other governmental libraries such as the National Agricultural Library, National Library of Medicine, the library of the US Geological Survey, etc.
6) Identify existing projects in which libraries, consortia, federal agencies, and other organizations are already undertaking the work of identifying, preserving, digitizing, hosting, organizing, advocating for, and describing federal documents.
    a. Build on the existing research done by HathiTrust staff at https://sites.google.com/a/umich.edu/hathitrust-government-documents/government-documents?pli=1
    b. Evaluate the overall outcomes and component aspects (identification, digitizing, cataloging, publicizing) of past and ongoing initiatives.
    c. Identify opportunities for HT to leverage these external efforts by partnering, facilitating access to content, identifying gaps that could be filled by HathiTrust,or otherwise accelerate progress toward the goal of building a comprehensive federal digital corpus.

Intermediate-Term Recommendations:

7) Hire a government documents project manager to ensure effective coordination of the activities of various planning groups and implementation teams, interact with potential partner institutions, and communicate with interested user communities.
8) Build on the general recommendations in this report by charging a series of working groups, subsidiary to the GDIPAWG, to address the general recommendations in this report.
    a. Review and assess the priority of these recommendations
    b. Develop the strategies, timelines and resources needed to operationalize the recommendations
    c. Provide benchmarks for measuring the progress of the various activities
9) Develop and/or coordinate a strategy to locate uncataloged/unrecorded library documents holdings.
    a. Convene a group to survey existing efforts, review alternative approaches, estimate timelines, and calculate costs
    b. Encourage the creation of a funding pool to support this work
    c. Evaluate partnering opportunities with bibliographic utilities (e.g., OCLC, GPO, MARCIVE) and commercial entities
    d. Implement several pilot projects to determine the efficacy and costs of a proposed approach to cataloging documents for the purpose of digitization
10) Reach out to potential partners—libraries, agencies and commercial vendors— with unique digital content to deposit.
    a. Develop a funding pool if cash payment is expected
    b. Work with CIC General Counsels and Google to determine the extent to which existing HathiTrust files can be shared in exchange for additional digital content
11) Undertake efforts to review the quality of and de-duplicate volumes in the existing corpus; enlisting the support of documents librarians where possible/feasible.
12) Enhance search and discovery options
    a. Implement SuDocs search and display capability
    b. Create links between legislatively and/or bibliographically related content
    c. Identify and expose relationships between superseded material and various versions/editions
    d. Otherwise enhance metadata to improve identification, discovery and use of federal documents

Long-Term Recommendations:

13) Expand coverage beyond federal government information distributed by GPO through the FDLP, including both born digital and print publications.
14) Enhance HathiTrust functionality to incorporate rich format government content, including maps, photos, time-based media, etc.
15) Review access policies for non-member downloading of government documents and other policy considerations.
16) Develop a communication plan for the HT government documents initiative.

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

Submitted by the Government Documents Initiative Planning and Advisory Working Group to the HathiTrust Program Steering Committee October, 2014

1. Background/History

In 2011, HathiTrust partners resolved to pursue an initiative "to facilitate collective action to create a comprehensive digital corpus of U.S. federal publications including those issued by GPO and other federal agencies."[1] The primary reasons for embarking on the initiative were the following:

- "Government publications provide historical context, inform policy, document critical trends, and reflect the evolution of graphic arts and publishing."
- "A comprehensive digital archive of US documents in digital form offers significant potential for research and opportunity to reduce costs of collection management and access in libraries. Safeguarding digital surrogates within a non-profit context will preserve the integrity of these valued resources. Strategies to enhance discovery offer models for access that may have applicability in other contexts."
- "Approximately 97% of new government publications available through the Program are disseminated electronically. With programs to convert legacy print collections to digital form comes the opportunity to develop a comprehensive, network-accessible, digital library of United States federal publication."[2]

1.1 The FDLP Collection

United States federal government publications are defined in the U.S. Code as "informational matter which is published as an individual document at Government expense, or as required by law" (44 U.S.C. 1901).[3] Government publications may be produced in a wide variety of formats including books, pamphlets, maps, microfiche, posters, puzzles, CDs, DVDs, floppy disks, cassettes, Web content or other born-digital format.

The US government began making publications available in 1789.[4] In 1813, the US Congress established the Federal Depository Library Program (FDLP) to provide public access to government publications at no cost.[5] The Government Printing Office (GPO) assumed responsibility for the program in 1895, and today approximately 1,230 libraries receive publications.

The total number of US federal publications that have been produced and distributed in one form or another by government agencies is nearly impossible to determine. The subset of federal documents distributed the Government Printing Office (GPO), the federal agency charged with printing and disseminating government publications, is more knowable, although surprisingly elusive in its own right. Since GPO disseminates its print outputs to libraries via the FDLP, there are longstanding retention commitments and decades of recordkeeping by both the suppliers and recipients of the content. Hence, here is an imperfect record of what has been produced, and therefore the reasonable expectation that systematic analysis could yield a number, and comprehensive listing of the items distributed through the FDLP. While that work proceeds, current estimates put the number at somewhere between 1.5 and 3 million documents.6

[1] Constitutional Convention Ballot Proposals. http://bit.ly/1q877Ra.
[2] Ibid.
[3] U.S. Government Printing Office, Federal Digital System (FDSys), http://www.gpo.gov/fdsys/pkg/USCODE-2009-title44/pdf/USCODE-2009-title44-chap19-sec1901.pdf
[4] http://memory.loc.gov/ll/llsl/001/0100/01920068.tif
[5] http://memory.loc.gov/ll/llsl/003/0100/01820140.tif
[6] John Butler, Meeting of the DPLA Content and Scope Workstream, February 28, 2013.

The number of documents that were distributed directly by the publishing agencies, and therefore not included in the FDLP, is more difficult to determine. Although they are not the initial focus of this project, these documents are within scope of the project based on the statement in the HathiTrust resolution: “...to create a comprehensive digital corpus of U.S. federal publications including those issued by GPO and other federal agencies.”

There are a large number of projects underway in the United States to digitize and make available US federal publications.7 Notably, through a targeted effort of the Committee on Institutional Cooperation in partnership with Google, and along with federal publications contributed by other participating libraries, HathiTrust has assembled a corpus of approximately one half million identified US federal publications (identified via bibliographic metadata). The number of government documents in the HathiTrust repository that are not identified as such is unknown.

1.2 Challenges

Discovery of government documents in HathiTrust is hampered by several factors including 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. Some issues that compound these fairly common issues in library cataloging for government publications in particular include the following:

- The publication structures for government publications are complicated and sometimes confusing. A single published volume may well have a monographic title and be considered part of multiple sets, series and/or serial publications. In addition, there are few definitive records on what was published and, even where these records exist, they are not always in agreement (e.g., we may have a volume in HathiTrust that an authoritative publication list(s) claims was never published).
- Many libraries have been ambivalent over time on whether to catalog government documents at all. When cataloging has taken place, it has frequently been selective, and libraries have made widely variant decisions: sometimes cataloging only selected analytic volumes, or only certain series; sometimes cataloging certain items as part of their general collection(s) with no indication that they are government documents, etc.
- Differences in collection management procedures at individual depository libraries mean that individually issued “documents,” “pieces,” or “items” can be bound and recorded differently from library to library such that “a volume” at any one library may not match a similarly labeled volume at any other. These differences in physical management (commonly referred to as “bound-withs”) add complexity to the analysis of collections when looking for duplicates and gaps.

2. Scoping the Initiative

While the Government Documents Initiative Planning and Advisory Working Group (GDIPAWG) recognizes the value of comprehensively digitizing the corpus of U.S. federal documents, it is recommended that the initial effort give priority to the body of material distributed in a print format by the Government Printing Office as part of the Federal Depository Library Program. It is understood that researchers and citizens would find value in content distributed in other formats, and disseminated through channels other than the FDLP. We would hope to move to encompass such content in due time, or sooner if it can be accomplished expeditiously. What is proposed here, however, is that primary attention be given to the widely disseminated content of the FDLP as this is both the most secure in terms of preservation and the most concerning to the HathiTrust membership and twelve hundred other depository libraries across the U.S.

2.1 Environmental Scan

The Constitutional Convention Ballot Proposal called for "HathiTrust, through coordinated and collective action, expand and enhance access to U. S. federal publications." Establishing the true current state of such access and current projects which are also intended to "expand and enhance" existing access is a crucial component of the Working Group's mission. Additionally, as no one organization can, or should, reasonably be expected to provide such comprehensive access it is essential that this task be accomplished "through coordinated and collective action." Undertaking an environmental scan of the current state of progress toward a comprehensive digital corpus of U.S. federal documents will enable HathiTrust and other interested organizations and agencies to leverage one another's efforts and collaboratively reach the goal.

[7] A list compiled by HathiTrust is available at http://bit.ly/1fRhcR0.

3. What HathiTrust Has Done So Far

3.1. The HathiTrust Initiative To-Date

While some HathiTrust partners (e.g., CIC member libraries) have engaged in specific projects to digitize government publications and deposit them in HathiTrust, the current initiative is the first partner-wide effort HathiTrust has undertaken. Prior to the formation of the GDIPAWG) in February 2014, work on the partner ballot initiative has proceeded in two areas: a) information gathering, including an effort to gather and analyze federal government publications in order to increase our understanding about the universe of government publications, and b) work to create a HathiTrust Government Documents Registry.

3.2 Information Gathering

HathiTrust staff assembled information about what US government publications are, where they can be found, how they are identified, and the current status of initiatives to catalog, disseminate, collect and house, and digitize, US government publications. The information is posted to Google website.[8]

In the fall of 2013, in conjunction with GWLA and ASERL, HathiTrust issued a broad call to US libraries, especially libraries participating in the FDLP program, for the submission of bibliographic records of government publications.[9] The purposes of the call were to gain a better understanding of the total corpus of US federal publications, and to try to determine the proportion of the total corpus held in print by libraries that has already been digitized. To this end, in February 2014, HathiTrust staff sent records from the 42 institutions that participated in the call to be analyzed by Google. A report about the volumes digitized and not digitized, and volumes were records were not good enough to make a determination, was expected earlier this year, but is still pending as of this writing. HathiTrust retains these records currently for further analysis, and for possible inclusion in the HathiTrust Government Documents Registry.

3.3 HathiTrust Government Documents Registry

In April 2013, HathiTrust hired a Government Documents Analyst, Valerie Glenn, to begin work to create a comprehensive registry of US federal publications.10 Work on the Registry to-date has included[11]:

- Development of a scoping statement;
- Definition of primary and secondary Registry audiences;
- Definition of project constraints and initial assumptions;
- Analysis of existing sources of metadata;
- Compilation of an increasingly comprehensive list of US federal agencies;[12]
- Manual efforts to identify relationships between items (based on metadata) and identify gaps in metadata coverage;
- Creation of a draft of functional requirements.[13]

[8] http://bit.ly/PCFTqR

[9] http://www.hathitrust.org/usgovdocs_call-for-records.

[10] http://www.hathitrust.org/usgovdocs_registry.

[11] All of these are available at http://www.hathitrust.org/usgovdocs_registry unless otherwise noted.

[12] http://bit.ly/1dr9WpI.

4. The Existing HathiTrust Collection

While there are challenges to accurately identifying U.S. federal documents in HathiTrust, the criteria[13] used for surfacing this class of content puts the current number of volumes at 568,291 (September, 2014); almost all of it is in the public domain. This collection grows daily because of a partnership between the CIC and Google to digitize federal documents and make them publicly accessible in HathiTrust. As of September, the CIC project had added 441,096 of the overall 568,291 items. The CIC content has largely been captured by sheetfed scanning, and thus differs from documents files captured non-destructively in the early years (~approx. 2005-2008) of the Google/Library partnership. Both Google engineers and CIC digital librarians believe that sheetfed capture yields somewhat denser images and, accordingly, more accurate optical character recognition.

With more than a half million federal documents already in HathiTrust, the GDIPAWG believes it is already a valuable resource for research, citizen awareness, preservation, and library collection management.
The GDIPAWG recommends the following actions to leverage the visibility and value of the existing corpus:

1. Develop a protocol to systematically check for file quality, replacing poorly scanned images and testing for differences in OCR accuracy between files created by non-destructive and sheetfed scanning methods. This protocol would build on current initiatives to identify file quality issues.
    a. General quality assurance testing could be carried out by HT on a sample of documents content
    b. Establish a path for crowd-sourcing further QA discovery and reporting.
2. Undertake a project to de-duplicate the existing collection, or to represent duplicate files in ways that can be readily interpreted by end-users.
3. Pull together series/serial volumes and fill in gaps with priority for the so-called “essential titles” as identified by the Government Printing Office and depository library community.
    a. HathiTrust might undertake analysis of the highest priority series/serials content
    b. The depository library community might be enlisted to assist with further efforts to identify and complete series/serials.
4. Draw upon the expertise of government documents librarians to update metadata to enhance metadata to improve identification, discovery and use of federal documents.
    a. Include series names, serial linking information, corporate authors, SuDocs classification, etc.
    b. Provide profiles of the collection by issuing agency, date, and topical or geographic focus.
5. Develop a marketing plan for the collection—to libraries, government agencies, and end- users—to facilitate maximum exposure and use.

5. Surfacing Additional Content for Digitization

Beyond the extant collection of government documents already archived in HathiTrust, there are a number of more or less clear and cost-effective paths for extending the holdings. These pathways, or strategies, include:

1. Identifying, digitizing and ingesting more cataloged government documents, whether issued prior to or after 1976
2. Surfacing uncataloged documents for processing, digitization, and ingest
3. Seeking out and ingesting “born-digital” and other already digitized content
4. Building capability in HathiTrust to ingest and manage non-print, non-book formats

[13] These criteria include the indicator “f” (federal) in the 086 MARC subfield. Since affixing this code is often overlooked, the actual number of government documents is likely higher than the reported 568,000.

5.1 Cataloged Content

A number of efforts are already afoot to surface additional government content for digitization:

- Google is analyzing the catalogs of CIC and University of California scanning partners for federal documents that have not yet been digitized. In recent months, Google updated its metrics for identifying federal documents, so that the lists offered to partnering libraries will be both more inclusive and more accurate.
- HathiTrust has gathered records from over forty member schools and has made those available to Google for analysis. The goal of this effort is to identify additional federal content that might not have already surfaced in other databases of records.
- In conjunction with the CIC digitization effort, the University of Illinois has done significant analysis of a subset of records identified as federal documents in the OCLC database. It is recommended that HathiTrust or member libraries fund an updated capture of records from OCLC to match against already digitized content.

To optimize the sourcing, digitization and ingest of cataloged documents, it is recommended that source libraries, digitization partners, and HathiTrust consider sequencing their work as follows:

1. Focus special attention on the post-1976 output of the FDLP for which comprehensive, or nearly comprehensive cataloging is believed to be available, either directly from GPO, through MARCIVE, or uploaded to OCLC and other bibliographic utilities.
    a. HathiTrust could track and report post-’76 items in queue for digitization from CIC members and University of California.
    b. To the extent that post-‘76 govdocs records and content are discovered beyond existing Google partners, determine the best path for adding this additional content.
        i. Use the records from elsewhere to expedite discovery and processing of content held but not cataloged in Google partner schools
        ii. Move the cataloged content for barcoding and Google scanning through existing partner schools
        iii. Create a pool of funds to support scanning through vendors other than Google
        iv. Consider incentives for non-HathiTrust members to scan and contribute uniquely held content.
2. Consider when and how to incorporate the significant number of pre-1976 documents for which catalog records are available in OCLC, but potential source libraries are difficult to identify because of a lack of local cataloging or holdings information.
    a. Review, adapt and adopt procedures for fast-track copy-cataloging developed at the University of Minnesota, University of Florida, et al.
    b. Seek out partner libraries willing to search for items cataloged elsewhere that might be held uncataloged on their own shelves.
    c. Develop standards and workflow for affixing a record to an item so as to facilitate digitization and subsequent ingest in HathiTrust.

5.2 Addressing Uncataloged Content

Identifying uncataloged candidates for digitization is the most challenging aspect of a project intended to create a comprehensive corpus of digitized federal documents. While the HathiTrust registry initiative and other reviews of bibliographies can confirm the publication of items, it will then prove a labor-intensive effort to match those known items against the uncataloged holdings of any particular library. Workflow options need to be evaluated for the following common situations:

1. A bibliographic record—brief or otherwise— is found for a yet to be digitized publication, but potential source libraries can’t be identified because they haven’t cataloged the item in question; or
2. A library undertakes shelf-reading of both cataloged and uncataloged documents, checking to determine if a digital surrogate already exists
    a. If a surrogate is found, and the print item is to be retained, some form of a bibliographic or holdings record should be added with a link to the electronic version.
    b. If the item in hand is to be withdrawn, a record leading users to the digital surrogate should be added.
    c. If no digital surrogate is discovered, a bibliographic or holding record should be added so the item can become a candidate for digitization.
    d. This process of shelf-reading could be divided across a number of libraries or concentrated in a

single library with a large, but largely uncataloged, collection.

3. Pilot efforts should be undertaken to assess the timeline and costs for cataloging documents as a check and refinement of data already available from local initiatives.

Cleaning up uncataloged documents collections will require considerable investments of effort and dollars; if not addressed, however, the problem will plague libraries in perpetuity, limiting options and adding costs for ongoing collection management. Creating a substantial digital collection of federal documents in HathiTrust does not, in and of itself, solve the problem for local libraries if they can't match their holdings to the Hathi collection. Thus, every individual depository library will still need to undertake a labor-intensive effort to match its holdings against the HathiTrust corpus since machine matches are impossible in the absence of local catalog records.
For this reason, library directors should recognize that user access may prove to be the most immediate benefit of building a comprehensive documents collection in HathiTrust, as opposed to providing an easy pathway for libraries to manage down their uncataloged print holdings.

5.3 Seeking and Ingesting Born-Digital and Other Already Digitized Content

A number of initiatives have been—or will be— undertaken by libraries, agencies or commercial interests to digitize government documents. These digitization projects are often limited in scope to subjects, geographic regions or agencies that are of specific interest to the institutions undertaking the digitization. Some of the larger programs to digitize government information have been undertaken by federal agencies such as the Library of Congress, which is working with the Government Printing Office to digitize the Bound (or Permanent) Congressional Record, having successfully completed previous digitization projects such as the U.S. Statutes at Large that are made accessible by FDSys.

An ongoing working group, subsidiary to the GDIPAWG, should be convened to pursue opportunities to ingest federal documents already digitized outside the parameters of current efforts to build a documents corpus in HathiTrust. In some cases, potential partners might be seeking a preservation solution for already digitized content. In others, an exchange of content might prove mutually beneficial, or even co- investment in digitizing new bodies of content. And, paying for content—or records—should not be ruled out if that appears to be the most cost effective approach for the HathiTrust membership or library community writ large. Potential sources for already digitized content might include:

- GPO—current and retrospective output
- Library scanned content (e.g., House Un-American Activities Committee (HUAC), Public Papers of the Presidents (PPOTP), Foreign Relations of the United States (FRUS))
- Google partner content not already in HathiTrust—Stanford, Harvard, LC, etc.)
- Nonprofits (e.g., Law Library Microform Consortium (LLMC), Internet Archive)
- Federal agencies (e.g., USGS, National Agricultural Library, Department of Energy OSTI)
- Commercial providers (e.g., ProQuest, Readex, Alexander Street, Sage)

While it would be beneficial for HathiTrust to investigate the possibilities for integrating large existing digitization projects and collections into the HathiTrust corpus of government documents, it is recognized that the integration of some bodies of digitized content can prove difficult, and the associated costs of integration can be significant. Any such efforts involving partnerships or purchases will require prior due diligence and testing to assure that the quality and formatting of the content and associated metadata is compatible with existing HathiTrust collections.

5.4 Non-print, Non-book Formats

Government information is released and distributed in all manner of formats--maps, posters, photos, CDs, DVDs, microform, charts, datasets, etc. Some of these rich format materials are released individually, while others are distributed in conjunction with a traditional print volume. Since much of this content is distributed through the FDLP, its inclusions a logical extension for the scope of our efforts to build a comprehensive digital corpus. It is recognized, however, that each of these alternative formats creates its own unique challenges for digitization, storage and discovery. We recommend, therefore, that a working group be convened to explore the costs and benefits of tackling one or another of these enriched formats and consider some or all of the following:

- Determine the priority for completing already digitized textual content with undigitized supplemental material (i.e., fold-outs, maps, CDROMs, etc.)
- Digitization strategies or protocols for reformatting already digitized content

- Methods for capturing inserts, foldouts and inclusions in future digitization
- Necessary enhancements of the HathiTrust platform to support the retrieval and management of enriched content, such as maps, that are often distributed independent of traditional print volumes

6. Non-GPO/FDLP Government Information

A great deal of government information is disseminated through channels other than the Government Printing Office. To the extent that users would benefit from the integration of this content with GPO distributed content, and future generations would benefit from having these resources securely archived in a trusted repository, the HathiTrust membership might seek to dedvelop strategies for identifying, prioritizing and securing this content. Such strategies would likely include direct outreach to, and partnerships with, issuing government agencies that may be in a position to make available either or both digitized and print resources to extend the HathiTrust corpus. Many large depository libraries have substantial holdings of federal documents obtained from sources other than GPO. These documents, which are often called "fugitive documents" since they fall within the scope of the FDLP but were not distributed through the program, should be identified and assessed for inclusion in the HathiTrust corpus of U.S. federal documents.

7. Establishing Priorities for Digitization and Ingest

Since many options have been identified for extending the HathiTrust collection of U.S federal documents, a process needs to be established for setting priorities. It is recommended that the GDIPAWG work with the HathiTrust Collections Committee and Program Steering Committee to create a framework for prioritizing the work ahead. Even in advance of such consultation, some priorities seem obvious enough to recommend at the outset. These include:

1. Google partner libraries should continue to work with Google to digitize unique holdings.
2. Google partners that have not yet deposited documents content in HathiTrust should be encouraged to do so.
3. Analysis of the pre- and post-1976 cataloged collection should be carried out by HathiTrust/ Google/OCLC, and source libraries identified for content not already digitized.
4. Widely known and consulted series distributed by GPO—the so-called "essential titles"—should be analyzed in HathiTrust for quality and completeness. To the extent that gaps are identified, source libraries should be tapped to supply needed content for digitization.
5. HathiTrust member and non-member libraries with known bodies of unique digitized government content should be approached about their willingness to deposit.

Beyond these five fairly obvious—and relatively inexpensive—paths for extending the existing HathiTrust corpus, a strategy and funding model needs to be developed for the complex and more expensive workflow required to discover and process uncataloged content, identify government publications outside the Federal Depository Library Program, or to create content partnerships that might involve significant investments or complicated terms. Setting boundaries and priorities for these more complicated digitization and/or ingest opportunities will need more in-depth consideration from stakeholder groups within the HathiTrust membership. One means for sorting among these competing opportunities would be to establish a clear sense of the priority among the communities that the HathiTrust membership is seeking to serve, and the content that is most supportive of these respective communities (see Appendix 2). While it might not prove feasible to assign a higher or lower priority to different users or uses of a HathiTrust collection of government documents, creating linkages between documents digitization and other HathiTrust groups and initiatives would almost certainly aid in setting priorities and assessing the costs and benefits of different courses of action.

8. Intersections With Other HathiTrust Initiatives

While a targeted effort to create a *comprehensive corpus* of U.S. federal documents in HathiTrust is necessarily mindful of the special characteristics of this body of content and its mode of production and distribution, it lives within a larger body of content being preserved by HathiTrust on behalf of its member libraries. Accordingly, the GDIPAWG recognizes the importance of integrating the approaches taken to digitizing and managing government documents with the broader strategic goals and policies of HathiTrust. It is therefore important that the planning efforts around documents recognize and exploit the intersections with other HathiTrust initiatives, working groups and resources, including the Collections Committee, the Print Monographs Archive Planning Task Force, the Rights and Access Working Group, and the HathiTrust Research Center.

8.1 Collection Committee

The charge for HathiTrust Collections Committee (http://www.hathitrust.org/collections_committee_charge) intersects in numerous significant ways with the work of the GDIPAWG. As specified in its charge, the Collections Committee has general oversight responsibility for activities and policies affecting:

- Collection development Prioritization for new content types
- Collection management tools and analytics
- Impact of 2012 Recommendation not to de-duplicate holdings

The intersection between the federal documents collection and the Collection Committee requires agreement on the definition/scope of a comprehensive collection of federal documents, which the GDIPAWG will recommend, but the Collections Committee should affirm.

Furthermore, as indicated in 5.4 above, it will not be possible to establish a comprehensive collection of federal documents without expanding the content types accepted for ingest into HathiTrust. Among the most prevalent non-book materials that would require new technical and policy consideration are maps, posters, newspapers and newsletters, pamphlets and flyers, teacher kits and a variety of loose-leaf services. Federal documents, even those that might be considered books and serials/journals, come in in almost every imaginable size and shape, from single sheets and broadsides to the extremely voluminous and oversized.

Some of the issues of concern to the GDIPAWG that intersect with the mandate of the Collections Committee include:

1. How will superseded items be denoted in the documents corpus as well as the larger corpus of HathiTrust content? While superseded items can be important for historical scholarship, we do not want to mislead users into thinking they are the latest versions upon which policy decisions or legal interpretations can be based. In the realm of government documents, slip laws are routinely superseded by the U.S. Statutes at Large, and the Statues are, in turn, superseded by integration into the U.S. Code. In other spheres, airport approach maps issued by the Federal Aviation Administration are routinely superseded by updated flight patterns.
    a. Is metadata tagging adequate, or are more prominent displays of the status of an item required?
    b. Should HathiTrust link superseded items with the current version of the law, regulation, map, or opinion currently in effect and vice versa?
    c. Should loose-leaf services be represented as individual sets of basic and supplementary issuances or can we help users better understand the ongoing nature of the updating of these publications?
    d. Since a document may not be superseded at the time of ingest into HathiTrust, but may well become superseded at a later date, what process could be put in place to monitor and accurately document the status of information over time?

2. Another important set of issues to be addressed with the Collections Committee is how to identify and manage documents that include personally identifiable information.
    a. Should the original scans be stored for future access with a redacted version made available for current access?
    b. Should the entire document be suppressed until sufficient time has passed so that personally identifiable information is not an issue?
    c. How would HathiTrust identify documents that contain personally identifiable information? Are there automated tools to assist with that process or would they need to be developed?

3. As is the case in other realms of publishing, there are frequently multiple versions of a single government document, often not clearly differentiated without careful analysis.
    a. Should the general HathiTrust policy not to de-duplicate versions be re-examined, either generally or for specific categories of government information?
    b. How would HathiTrust decide whether a specific volume being ingested is in fact a new version or new edition that requires distinctive metadata?

8.2 Print Monographs Archive Planning Task Force

It is necessary to coordinate with the Print Archive Planning Task Force (http://www.hathitrust.org/print_monographs_archive_charge) to ensure that government documents that fit within their definition of a monograph are included in their planning. It is likely that their charge should be expanded to ensure that there are identifiable print (and possibly other tangible format) federal documents that can be matched with their digital counterparts through a Distributed Print Documents Archive, whether or not those documents fit the current definition of a monograph. This is an expectation of many of the government documents professionals, but also of some deans/directors, GPO and other agency personnel, and probably many researchers, and it will provide assurance that we can, if necessary, repair or replace the digitized content in the future. We will certainly want common procedures and policies for a distributed print monograph archive and a distributed print government documents archive, whether this responsibility is assigned to the Print Archive Planning Task Force or the GDIPAWG.

8.3 Rights and Access Working Group

While most federal documents can safely be assumed to be in the public domain as a government work product, and hence not subject to copyright, this is not true for all federal documents distributed by GPO or otherwise disseminated. There are instances where copyrighted information has been incorporated into hearings or read into the Congressional Record, to mention two examples, and we will need clear policies about how such material will be treated. There are also contractor reports where it may not be easily determined whether the contract assigned the rights to the contractor or they were retained by the agency, and there are government agencies or quasi-governmental agencies that claim copyright in their work, such as the National Research Council, the Smithsonian Institution and the Library of Congress.

With the Smithsonian and the Library of Congress, this claim is often based on the co-mingling of private funds with government funds to produce the publication. Other agencies claim exemption in order to recover their costs. The CENDI Copyright Working Group has estimated that up to 15% of government publication is, in fact, covered by copyright restrictions[14].

In dealing with U.S. government documents, it will be necessary for HathiTrust to make a determination on the copyright status of individual federal publications. The GDIPAWG proposes to work with the HathiTrust Rights and Access Working Group (http://www.hathitrust.org/rights_and_access_charge) to develop procedures to expedite the review of documents when their copyright status is in question. The Groups may also work together to develop strategies and guidelines for negotiating unfettered access with agencies that assert copyright protections for their publications.

8.4 HathiTrust Research Center (HTRC)

Scholars from many disciplines have demonstrated an interest in applying the techniques of text mining and analysis to the corpus of U.S. federal documents. Whether combined with other publication types, or treated on their own, the record of congressional, executive branch and judicial output is a rich data source for longitudinal analysis. To facilitate this kind of research, the GDIPAWG proposes to work with the HTRC Executive Management Team (http://www.hathitrust.org/htrc_governance) to make sure that documents content is available and properly formatted for textual analysis.

[14] "U.S. Federal Government Information: Getting the Rights Right, CENDI Copyright Working Group, compiled by Bonnie Klein, April 13, 2012

9. Policy Considerations for Government Information

For the most part, the policies already in force with HathiTrust provide strong policies around access and use of government publications. For instance, copyright policy is explicit in mentioning that U.S. federal government documents are treated as public domain. While existing policies are generally supportive of encouraging public access to government information, there are areas where additional consideration could be given to the critical role of government information in our democracy, and the special relationship our libraries have forged with the government and citizenry to assure ongoing access to this content. Additional consideration should be given to the terms of access governing U.S. federal documents archived in HathiTrust, especially as these terms of access relate to the use of government publications by non-member individuals/institutions.

9.1 Expansion of full document downloading for public domain documents to non-members

Access policies, as listed in the Copyright Policy, center around recognition of IP address detection, user authentication, and geography detection along with copyright status to determine level of use. Expansion of this policy to include the document type, such as government publications indicators, could allow for more refined ability to identify documents that could be automatically available as a full document download.

As outlined under the Privacy Policy, non-affiliates can create a University of Michigan Friends Account which allows for creation of permanent collections in the Collection Builder. What is not currently allowed with Friends Accounts is the ability to download an entire document. This is a big concern of the government documents community especially in discussions around equivalent access to online publications in comparison to print.

Currently, there are workarounds available such as contacting a member library to download documents and provide access through resources such as Dropbox. This, however, would not be sustainable as the HathiTrust government documents collection, and its potential audience, continues to grow.

9.2 Expansion of search capabilities for discovering and retrieving government documents

While not a specific policy as outlined on the HathiTrust website, discussions should occur around expansion of searching mechanisms specifically for government documents. Librarians and federal agencies are likely to want to search HathiTrust—and compare holdings with HathiTrust—by SuDocs number. It would also be useful to create series links for government publications so that, for example, reports to Congress that are issued as part of the Serial Set can be retrieved or identified as such.

Finally, attorneys and academics carrying out legislative histories would be best served if forms of bills were to be linked by HathiTrust metadata. We'd expect that some of these enhanced discovery and retrieval features could be accomplished through managed crowdsourcing, especially by enlisting the expertise of documents librarians and documents catalogers across the country.

9.3 Privacy

Since many government documents contain sensitive information about individuals, companies, or organizations, some policies will need to be developed to protect the privacy rights of citizens. This issue is being addressed more generally by the Collections Committee (see 8.1.2 above), so the GDIPAWG could inform them if there are particular issues likely to arise with government documents that are not being addressed in the overall HathiTrust privacy policies. On a related matter, there are occasions where a branch of government, GPO, or an issuing agency will ask libraries to remove an item from circulation. There should be some understanding in advance of how Hathi would respond to such requests.

10. Public Relations and Outreach

The essence of a successful digital project is that it helps people accomplish their information-seeking goals. For that to happen, users need to know what is being made available, and the creators of a collection need to know who will be using it and toward what end. This requires that robust channels of two-way communication need to be established between HathiTrust and the many and varied constituencies for government information. Some of this communication could be facilitated by the Government Printing Office, and the affiliated community of FDLP documents librarians that has a long history of representing a complex body of content to a diverse user community. Engaging the documents specialists working in the HathiTrust member libraries seems like the surest way to develop a more comprehensive program of communication with the government, library community, and the public at large.

Given the base of government content already in the corpus, and the prospects for continued digitization, there is a big opportunity for documents librarians at HathiTrust member libraries to take greater control of the shape of this project. Most immediately, the GDIPAWG recommends calling upon them to assess the utility of the existing content, requesting their input and help for how the content could be better organized and promoted to user communities--including libraries--across the country. Securing the support of a knowledgeable cadre of content specialists will not only help to attract users to the collection, but also mitigate some of pitfalls that might otherwise trip up these users as they begin to engage with the content.

Any marketing program needs to identify the targets and purpose of the proposed communication. The following among the primary constituencies for accessing U.S. federal documents in HathiTrust:

- Academic faculty and researchers
- Academic publishers
- Attorneys
- Federal Depository libraries, both regional and selective
- General public/interested citizenry
- Government Printing Office
- Information technology leaders
- Issuing government agencies
- Non-depository libraries, including, but not limited to, public and law libraries, whose patrons need/use this information
- Open access advocates
- Students (post-secondary undergraduate and graduate students, but some high school groups as well)

This list could be longer, but the main point is that each of these constituencies has specific needs for this content, so an associated communication program should be tailored to the particular needs of the respective user communities, both apprising them of the content being made available and soliciting their feedback around the utility of the resources being offered. A communication plan should be developed to address the dual needs for promotion and feedback, and volunteer or paid staff should be monitoring the implementation of the program.

11. Recommendations

Table 1 below summarizes the high-level recommendations embedded throughout this report. Items 1-5 could be pursued—at least in limited ways—without significant new member investments or undue demands placed on existing HathiTrust staff. The GDIPAWG is aware of a number of national initiatives in the arena of government documents that could advance public access to this content. Forging partnerships with these libraries, agencies and organizations could significantly reduce the demands placed on HathiTrust staff or HathiTrust member libraries to carry out all of this work on their own. That said, even reaching out to widely disparate groups, or attempting to negotiate terms with them, could prove quite time consuming and burdensome, so it is likely that additional resources will need to be committed to this effort. Item #6 below raises the prospect of adding dedicated staff to coordinate planning and operations that would increase the body of federal content in HathiTrust. Items 7-12 then address what the GDIPAWG consider to be the more labor and resource intensive aspects of a more proactive program to approach comprehensive reformatting of documents holdings across HathiTrust member libraries.

Table 1. Strategic Priorities for building, organizing, supporting and promoting the digital government documents corpus in HathiTrust

Recommendation	Timing	Lead	Comments
1. Continue to build a comprehensive registry of the FDLP corpus and compare to the known universe of library holdings	Immediate	HathiTrust Staff	
2. Provide, and regularly update, a descriptive analysis of government documents holdings already in HathiTrust, including efforts to assess and report on the quality of scans and OCR.	Immediate	HathiTrust Staff and docs librarians	
3. Encourage member libraries to continue to build the digital corpus by identifying available cataloged content for either sheetfed and non-destructive scanning	Ongoing	CIC and CDL/UC	
4. Enlist the support of documents librarians to analyze, organize, and promote the existing corpus	Immediate and ongoing	HathiTrust staff, Collections Committee, and documents librarians	
5. Pursue partnerships with the Government Printing Office, publishing agencies, and national and other governmental libraries	Immediate and ongoing	GDIPAWG, Hathi staff, PSC	
6. Gather data about projects in which libraries, consortia, federal agencies, and other organizations are undertaking the work of identifying, digitizing, hosting, organizing, and preserving federal documents.	Immediate	GDIPAWG	
7. Hire a government documents project manager to coordinate activities related to the HT documents initiative	By February 2015	HathiTrust staff, PSC, HT Board and GDIPAWG	
8. Charge a series of working groups, subsidiary to the GDIPAWG, to address and operationalize the general recommendations in this report.	Immediate thru June 2015	HathiStaff, GDIPAWG members and additional experts	
9. Develop and/or coordinate a strategy to locate uncataloged/unrecorded library holdings	By January 2015	GDIPAWG. specialized working group, and HathiTrust staff	
10. Reach out to potential partners—libraries, GPO, agencies and commercial vendors— with unique digital content to deposit.	2015	HathiTrust staff, Executive Director and Program Steering Committee	
11. Enlist the support of documents librarians to review the quality of existing files and deduplicate volumes when appropriate.	2015 and ongoing	HathiTrust staff	
12. Enhance search and discovery options such as	2015	HathiTrust staff	

SuDocs search and display and linking items in series.			
13. Expand coverage beyond GPO distributed content to include both born digital and print federal publications released directly by agencies.	By January 2016		
14. Enhance HT functionality to incorporate rich format government content distributed by GPO including maps, photos, time-based media, etc.	By July 2016	HathiTrust staff	
15. Review access policies for non-member access to government documents.	By July 2015	HathiTrust PSC and Board	
16. Develop a communication plan to enlist the support of GPO and documents librarians and to draw users to the corpus.	By July 2015	GDIPAWG and specialized subgroup, HathiTrust staff	

Appendix I: HathiTrust partners that are Federal Depository Libraries

HathiTrust Members	**Depository Status**
Allegheny College	Selective
Arizona State University	Selective
Baylor University	Selective
Boston College	Selective
Brandeis University	Selective
Brown University	Selective
Colby College	Selective
Columbia University	Selective
Cornell University	Selective
Dartmouth College	Selective
Duke University	Selective
Florida A&M University*	Selective
Florida Atlantic University*	Selective
Florida State University	Selective
Florida International University*	Selective
Harvard University	Selective
Indiana University	Selective
Iowa State University	Selective
Johns Hopkins University	Selective
Kansas State University	Selective
Library of Congress	Selective
Massachusetts Institute of Technology	Selective
Michigan State University	Selective
New College of Florida*	Selective
New York Public Library	Selective
New York University	Selective

North Carolina State University	Selective
Northwestern University	Selective
Ohio State University	Selective
Pennsylvania State University	Selective
Princeton University	Selective
Purdue University	Selective
Stanford University	Selective
Syracuse University	Selective
Temple University	Selective
Texas A&M University	Selective
Tufts University	Selective
University of Alabama	Regional
University of Arizona	Selective
University of California, Berkeley	Selective
University of California, Davis	Selective
University of California, Irvine	Selective
University of California, Los Angeles	Selective
University of California, Merced	Selective
University of California, Riverside	Selective
University of California, San Diego	Selective
University of California, San Francisco	Selective
University of California, Santa Cruz	Selective
University of California, Santa Barbara	Selective
University of Central Florida	Selective
University of Chicago	Selective
University of Connecticut	Selective
University of Delaware	Selective
University of Florida	Regional
University of Houston	Selective
University of Illinois, Chicago	Selective
University of Illinois, Urbana Champaign	Selective

University of Iowa	Regional
University of Kansas	Regional
University of Maryland	Regional
University of Massachusetts, Amherst	Selective
University of Miami	Selective
University of Michigan	Selective
University of Minnesota	Regional
University of Missouri	Regional
University of Nebraska-Lincoln	Regional
University of North Carolina, Chapel Hill	Regional
University of North Florida	Selective
University of Notre Dame	Selective
University of Oklahoma	Selective
University of Pennsylvania	Selective
University of Pittsburgh	Selective
University of South Florida	Selective
University of Tennessee, Knoxville	Selective
University of Texas System	Selective
University of Utah	Selective
University of Vermont	Selective
University of Virginia	Regional
University of Washington	Selective
University of Wisconsin-Madison	Regional
Utah State University	Regional
Vanderbilt University	Selective
Virginia Tech	Selective
Wake Forest University	Selective
Washington University, St. Louis	Selective
Yale University	Selective

HathiTrust members that are not federal depository libraries:

- Boston University
- California Digital Library (though a number of the UC Libraries are depositories as noted above)
- Carnegie Mellon University
- Emory University
- Getty Research Institute
- Lafayette College
- McGill University
- Universidad Complutense de Madrid
- University of Alberta
- University of British Columbia
- University of Calgary
- University of California, Santa Barbara
- University of California, Santa Cruz
- University of Queensland
- Florida Gulf Coast University
- University of West Florida

Appendix II: User Communities and Needs For Government Documents

User Group	Content Preferences	Discovery Strategy
Undergraduates	Historical (major events); public policy analysis;scientific and technical reports; legislative history	Known title (course assignment); keyword or subject searching
Seasoned scholars (faculty, researchers, graduate students)	Political science, public policy, history, demography, foreign relations, natural resources,	Narrow topical (advanced search capabilities),
Attorneys	Codes and Regulations, Case Law	Legislative histories; legal citation
Librarians	Collection Management and Research Assistance	SuDoc# and known title to advanced searching capabilities
General Public	Current government information; legal information	News citations; core series; taxation; social benefits
Federal employees	Agency documents	Advanced search capabilities; chronology and keywords; author; narrow subject
Business/tech sector	CFR; standards; regulations;	citation information

While this table is in no way intended as a complete representation of all of the user communities that rely upon government documents, it does direct our attention to the fact that there are different constituencies for the content, with differing goals and different strategies for tapping the content to fulfill those goals.
Accordingly, as HathiTrust sets priorities for adding content or building discovery tools, it should be understood that some groups will derive additional benefits while others—implicitly or explicitly—are being asked to defer optimal service and content delivery. This understanding might prove helpful in staging and shaping outreach to identifiable constituencies.