Practical Academic Librarianship: The International Journal of the SLA Academic Division
2(1):1-25, 2012 © The Author(s) http://journals.tdl.org/pal

A Comparison of HathiTrust and Google Books Using Federal Publications

Laura Sare
Texas A&M University

Abstract

This study compares the functionality of the Google Books and HathiTrust online repositories using federal government publications. From a random sample of more than 1,500 government documents, record overlap between the two databases was also examined. Functionality, such as ease of finding specific government publication titles, citation extraction, as well as the quality of metadata, in each repository is assessed in this article. Also addressed in this study is a look at the quality of bibliographic records for these repositories contained in the WorldCat database. This study found that Google Books and HathiTrust have unique strengths and weaknesses, as well as content and functionality overlap. The results of this study help to inform librarians which repository is best suited for federal document retrieval and patron needs.

Keywords: online repositories, federal publications, government documents, database functionality, Google Books, HathiTrust

Corresponding author: Laura Sare lsare@tamu.edu

Introduction

The development of Google Books, as well as the creation of the HathiTrust Repository through strategic partnership of major research institutions to digitize library collections, has enhanced the availability of government publications online. Formerly, federal publications could only be acquired at Federal Depository Libraries, but, through digitization efforts, these public domain materials are now available to anyone online. This is significant for academic librarians who work at institutions that are not members of the Federal Depository Library Program, or who have small or non-historical depository collections, because there is now more access to government information than ever before. Government publications are often the best sources for statistics and primary sources, subjects highly requested by academic library users. Therefore, this study compares the findability of federal government publications in both the Google Books and HathiTrust repositories, as well as coverage overlap, usability, functionality, and features such as book bags, citation exporting, the quality of catalog records in the Online Computer Library Center's (OCLC) WorldCat database (via the FirstSearch platform), and privacy policies in each of the repositories. A random sample of over 1500 federal government publication titles was used to conduct this comparison. This research will assist librarians in choosing the best database for answering their patrons' needs for historical government information online.

Literature Review

HathiTrust is a partnership of major research institutions to create a shared repository of digitized collections. HathiTrust began in 2008 as a collaboration of the thirteen member universities of the Committee on Institutional Cooperation (CIC), the University of California system, and the University of Virginia. Materials include items scanned under in-house initiatives by partner institutions, Google, and the Internet Archive (http://www.archive.org/). The initial focus of the partnership was to preserve and provide access to the digitized book and journal content held in partner library collections, including both copyrighted and public domain materials. HathiTrust provides "secure, reliable, long-term preservation for deposited materials" (HathiTrust, n.d.). The repository provides bibliographic metadata searching as well as a full-text search of all volumes in this "Digital Library" (public domain or copyrighted) and unrestricted access to digitized public domain materials. Access to copyrighted digital

Corresponding author: Laura Sare lsare@tamu.edu

materials is restricted to member institutions in limited circumstances such as for users with print disabilities (HathiTrust, 2011). Links to WorldCat.org, the free version of the WorldCat database, are included so that if inclined, users can find where the material is physically available in libraries. As of September 2011, the HathiTrust repository contains over nine million digitized volumes, of which approximately 27% are in the public domain (HathiTrust, n.d.).

The Google Books project began in 2004, and was originally known as Google Print. Working with the libraries of the University of Michigan, University of Oxford, Harvard University, Stanford University, and the New York Public Library, Google's goal was to assist publishers in making books and other offline materials searchable online, increasing the visibility of in- and out-of-print books (Google, 2004; Google 2011a). As the Google Books project added more partner libraries, the aim of the project changed to allow users to find relevant books by creating a, "virtual card catalog of all books in all languages" (Google, 2011b). Because of copyright restrictions, users may view a digitized book in its entirety only if it is in the public domain or with permission from the publisher. If the book is still within copyright, then readers may only view "Snippets"-brief extracts of the book. Similar to HathiTrust, links to WorldCat.org are included with each item to easily determine if the item is available through a local library. Links to purchase the book through Amazon, Barnes & Noble, and other vendors are also included.

Literature about the HathiTrust project has mostly been limited to announcements of new search features, or new library members, but a few scholarly articles have recently been written. O'Brien (2011) conducted a review of the availability of medical literature in HathiTrust. Christenson (2011) provides an analysis of how the HathiTrust cooperative handles issues such as long-term digital preservation, access and services, and identifies opportunities for libraries to collaborate.

There are numerous articles about Google's use in libraries in the scholarly literature. Most focus on Google or Google Scholar. Those articles written about Google Books mainly discuss the legal implications of the *Authors Guild et al. v. Google* lawsuit and the Google Books Settlement Agreement (Google Books, 2011). Baksik's (2006) article provides a good early overview of the initial copyright issues involved and analyzes the strengths and weaknesses of Google's arguments regarding fair use. An article by Martin (2008) compared Google Books' OCR-only searching and the enhanced searching capabilities of the Text Creation Partnership (TCP) project (a University of Michigan led effort to produce standardized, digitally encoded editions of early print books from commercial publishers) to demonstrate how library partnerships could be used to make Google Books search more useful to academic researchers.

Jones (2010) conducted a study to determine how the contents of Google Books' public domain literature compared to that of a major research library. Jones found that Google Books' coverage of latter 19th century publications was comparable to that of a research library and had the benefits of full-text indexing, but issues with poor digitization needed to be resolved before the Google Books could be fully relied upon by libraries (p. 86). Additional analyses of Google Books of interest to the government documents community include Blakeley's (2009) article comparing Google Books and the Internet Archive as a way to augment government document collections and Blakeley's (2008) blog post on effectively searching for public domain government document publications in Google Books published before 1923.

An article by Hawkins and York (2010), while primarily an introduction and overview of HathiTrust, mentions that HathiTrust provides full-text access to public domain works produced by the federal government, but that Google Books does not on federal publications created after 1923. Hawkins and York go on to state that, "HathiTrust asserts rights under U.S. copyright law that Google has not" (p.3).

Methodology

While the focus of this article is to compare the ease of discovery and record overlap of federal documents in Google Books and HathiTrust, the quality of these two repositories' bibliographic records in OCLC's WorldCat (FirstSearch version), privacy policies, basic and advanced searching as well as other features are also examined.

Selection of Random Sample Titles

To compare the two databases, a sample of titles was necessary. This study used a list of titles developed by Sare (2011) to determine the full-text availability of federal publications online, and applied the methodology in the same study, to generate a random sample of 1540 documents covering a span of roughly four decades (1943-1976). These titles were divided into four decades: 1.) to break the titles down into manageable groups for analysis, and 2.) to be able to identify any trends in digitization, such as coverage, that librarians should be aware of when recommending these repositories to users.

The random sample time frame (1943-1976) also corresponds to the *Monthly Catalog of U.S. Government Publications*' notations of documents the Government Printing Office distributed to depository libraries. This is important because it was necessary to know what publications were sent to depository libraries when creating a list of titles for the random sample, for depository libraries were providing their collections to Google for digitization. The *Monthly Catalog* lists most federal agency publications, not just those sent to depository libraries. This restriction was necessary to exclude publications listed in the *Monthly Catalog* that were not candidates for digitization.

Repository Search Strategy

Each government document was searched by its publication title in both the HathiTrust and Google Books repositories. For some documents, such as monographic series, the name of the series was searched if the publication could not be found under the individual publication title. For serials, an individual issue was searched for, usually in a volume. For example, a specific issue of the final bound version of the *Congressional Record*: X 86/1:105/pt.9. For searches in HathiTrust, data was recorded for three possible outcomes: 1) title not found, 2) document was available in full-text, or 3) if the document was available only in a "Limited" search-only view, meaning it had been digitized and the full-text indexed, but not viewable online due to copyright.

Similar to HathiTrust, data in Google Books was analyzed according to one of three possible outcomes: 1.) title not found, 2.) document available in full-text, or 3.) Google Books citation found. The Google Books citations had two possible outcomes: 1.) a "Snippet" view, where three small sections of the scanned page with the search terms highlighted displayed along with a thumbnail image of the cover, and 2.) a "No Preview Available" view, with only limited bibliographic information and no cover image. The titles were searched during the Fall of 2010 and repository functionality was analyzed in the Spring of 2011.

Results

Table 1 contains the results of the title searching in HathiTrust. Of the 1540 titles searched in HathiTrust, 436 were available, and 319 were available online as full-text. These ranged from 74 to 90 publications per decade, or 19% to 23% availability using the sample. HathiTrust had a total of 117 "Limited" (search-only) documents that are only viewable to users from the institution that submitted the publication for digitization. Some titles are designated as"Limited" (search-only) because Google handled the digitization of these specific documents, and Google catalogued them within copyright for those documents that were published after 1923 (H. Mercer, personal communication, September 9, 2011). Government documents digitized by other institutions have the copyright status correctly applied. Thus, the "Limited" (search-only) documents increase by decade, while the full-text decrease in Google Books.

Table 1: HathiTrust Results

	1940s (385 total)	1950s (385 total)	1960s (385 total)	1970s (385 total)
Number of Titles Found	98 (25%)	104 (27%)	114 (30%)	120 (31%)
Full Text	90 (23%)	78 (20%)	77 (20%)	74 (19%)
Limited View	8 (2%)	26 (7%)	37 (10%)	46 (12%)
Not Found	287 (75%)	281 (73%)	271 (70%)	265 (69%)

Table 2 shows the results of our search for government documents in Google Books. While Google Books has more total results of titles found, this is due to its combined "Snippet" and "No Preview Available" records. Only fourteen government publications were available in full-text online due to Google Books' policy of including federal publications with books in copyright that were published between 1923 and 1963 as they were scanning them. Google is, for the most part, not allowing full-text views because they fall within the orphan works copyright time frame (Orwant, 2008). Most of the documents found had "No Preview Available" views which does not provide the ability to search the full-text as the "Snippet" views do. Nineteen to 25% of the results were in "Snippet" form.

Table 2: Google Book Results

	1940s (385 total)	1950s (385 total)	1960s (385 total)	1970s (385 total)
Number of Titles Found	181 (47%)	187 (49%)	209 (54%)	232 (60%)
Full Text	4 (1%)	0 (0%)	4 (1%)	6 (1.5%)
No Preview Available	101 (26%)	102 (26%)	130 (34%)	128 (33%)
Snippet	76 (20%)	85 (22%)	75 (19%)	98 (25%)
Not Found	204 (53%)	198 (51%)	176 (46%)	153 (40%)

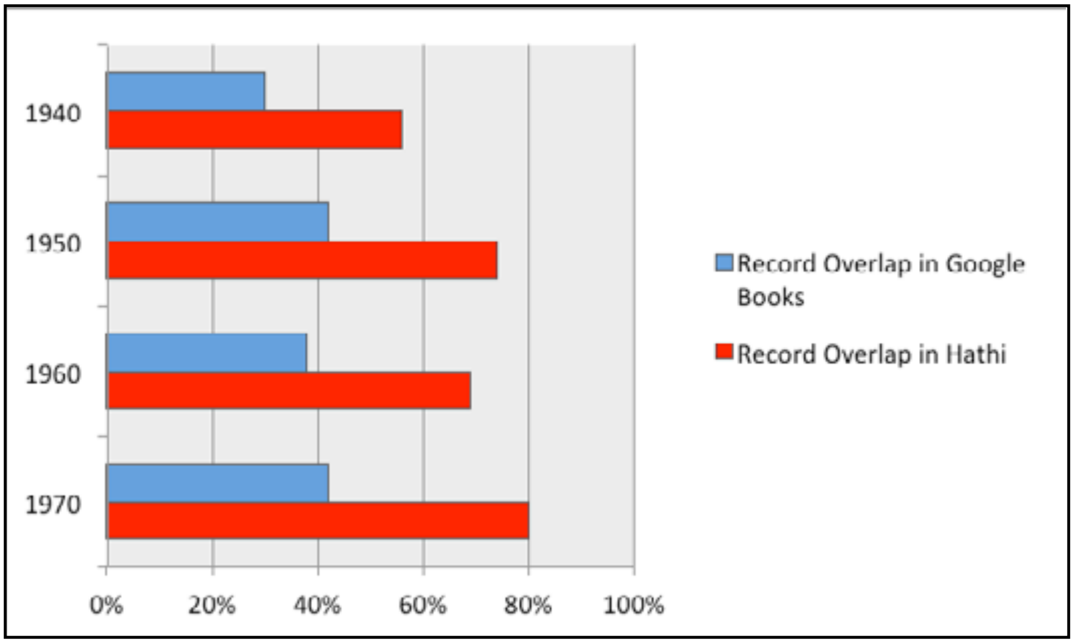
Table 3 shows the record overlap between Google Books and HathiTrust. A more comprehensive breakdown of record overlap by decade can be found in Appendices A and B. While the majority of the digital copies for both of these databases were scanned by Google, an overlap comparison was conducted to see how many titles were available in both databases. Surprisingly, there were only three documents for which the full-text was available online in both databases. And, Google Books contained 11 titles in full-text online for which HathiTrust did not even have a record.

Table 3: Overlap of Records in HathiTrust and Google Books

		Google Books				
HathiTrust		Full Text	Snippet	No Preview	No Record	Total
	Full Text	3	156	64	94	
	Limited	0	58	22	41	
	No Record	11	133	366	604	1540

Figure 1 below illustrates the percentage of government document record overlap by decade in the two repositories.. The record overlap shows that 56% to 80% of record citations in HathiTrust were also available in Google Books. Yet, only 30% to 42% of records found in Google Books also had records in HathiTrust; this is mostly due to Google Books containing "No Preview Available" and "Snippet" citations. Users wanting full-text for federal government

Figure 1: Percentage of Record Overlap between Google Books and HathiTrust



publications will be better served by searching in HathiTrust. However, to locate those publications not yet available in full-text, the Google Books "Snippet" view is useful because it displays the search terms within a portion of the digitized page, as well as stating how many times the terms appears in the document above the first "Snippet" image. The Google Books records are more helpful than HathiTrust's "Limited" view, since the "Limited" view only shows the user the number of times the search term appeared on certain pages, where as Google Books shows where within the document the search term is located.

Comparison of Database Functionality

The functionality for searching in HathiTrust and Google Books was also compared, as well as the usefulness of features in each repository. Features include the quality of WorldCat bibliographic records, basic and advanced searching, exported citation quality, and added functionality such as a word cloud feature. Privacy issues were also compared. At the time this article was written, HathiTrust had just begun to beta test their new WorldCat Local interface -that interface was not analyzed for this paper.

Monographic Searching

Searching for monographic titles was straightforward in both repositories. Both repositories have basic search functionality vis-à-vis a single search box, which worked well when the title searched for was unique. For more common words, putting the title in quotation marks for phrase searching narrowed the search results significantly in both databases. Both repositories have an advanced search that allows users to search specific fields (author, title, subject, publisher, year of publication, ISSN/ISBN), and a "limit to full text" option. The key difference between the advanced search functionality in Google Books is that this repository allows for Boolean searching without the user needing to know Boolean operators, while HathiTrust maintains a traditional Boolean search screen that requires the user to select the appropriate operator. Both Google Books and HathiTrust offer wildcard searching on single and multiple characters

In Google Books the only format limiter is a choice between books and magazines, while HathiTrust's limiter for format includes audio files, videos, and maps. HathiTrust also includes the additional search field of "Series Title". Table 4 shows a comparison of fields in the repositories' advanced searches.

Table 4: Advanced Search Fields Comparison

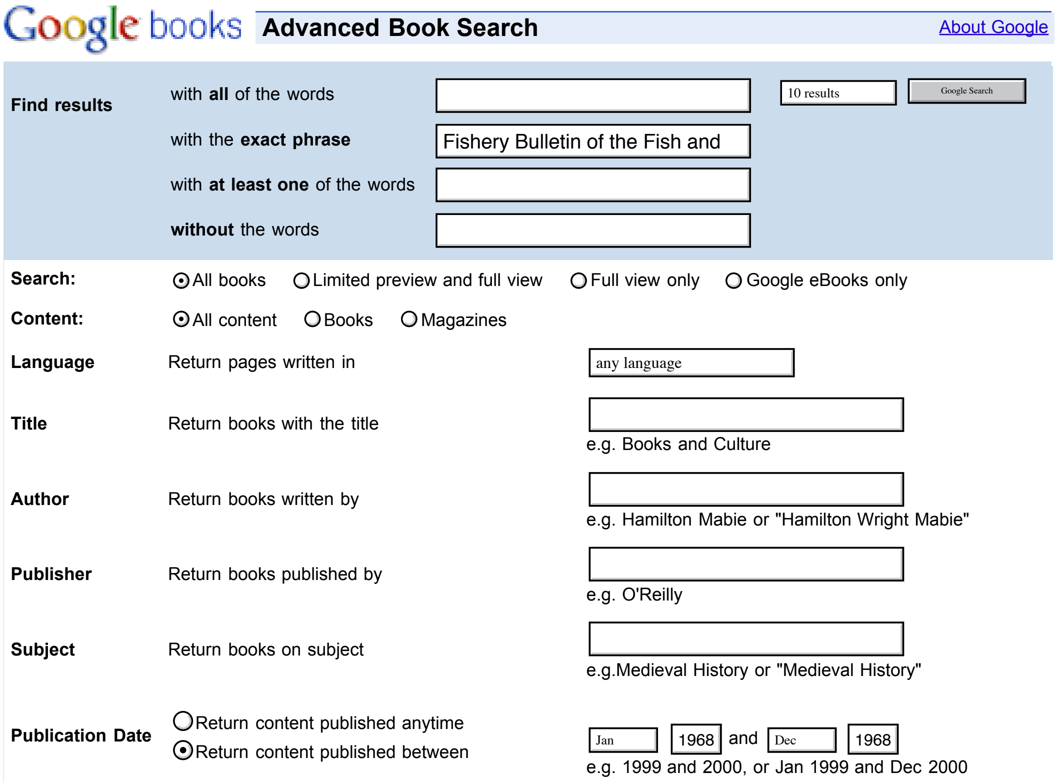
	Google Books	HathiTrust
Boolean	Implied but includes a NOT	AND, OR, no NOT
Title	Yes	Yes
Author	Yes	Yes
Subject	Yes	Yes
Publisher	Yes	Yes
Series Title	No	Yes
Date of Publication	No	Yes
ISBN/ISSN	Yes	Yes
Limit to full-text only	Yes	Yes
Year of Publication Date Range	Yes	Yes
Language	46	448
Format	Only Books or Magazines	29 Formats

Serials Searching

Because the searches conducted for this study were all on known titles, most of the search limiters were not applied. This is because the majority of government publications are not assigned ISBN or ISSN numbers, and some do not have individual authors. Instead, government publications designate the originating agency as author. In both databases, searching for serials can prove challenging. Users familiar with searching in full-text article databases who are accustomed to retrieving a journal title and scanning through specific years to individual issues within that year, may be frustrated with the lack of this search feature in both repositories.

In Google Books, the best results were obtained by using the basic search function with the publication title and the year or the specific volume number. For example, a search for the September 1968, volume 66, issue number 3, of the quarterly publication, *Fishery Bulletin of the Fish and Wildlife Service*, using only the title and year 1968 as search terms retrieved volume 68 of this publication as the first search result. Yet, there was a link to more editions and from that link volume 66 was the fourth result. A basic search for search terms of the title and volume 66 had the same results as putting in the year. Conducting advanced searches was less successful in both repositories. Figure 2 below shows an attempt to find the issue by its publishing date, 1968 using the advanced search function. The "publication date" parameters were set between January and December, 1968, and the content field as "all content." This yielded five results, none of which were the correct title. Changing the content field to "magazine," which would seem logical considering that the title is a quarterly publication, returned no results at all. In order to retrieve this government document through an advanced search in Google Books, the publication date had to be limited to the date range of January 1967 to December 1968 because the Google Books record lists 1967 as the only year in that record, even though it is published up to 1969.

Figure 2: Advanced Search Attempt for a Serial in Google Example



Google books **Advanced Book Search** About Google

Find results

with **all** of the words

with the **exact phrase** Fishery Bulletin of the Fish and

with **at least one** of the words

without the words

10 results Google Search

Search: ● All books ○ Limited preview and full view ○ Full view only ○ Google eBooks only

Content: ● All content ○ Books ○ Magazines

Language Return pages written in any language

Title Return books with the title
e.g. Books and Culture

Author Return books written by
e.g. Hamilton Mabie or "Hamilton Wright Mabie"

Publisher Return books published by
e.g. O'Reilly

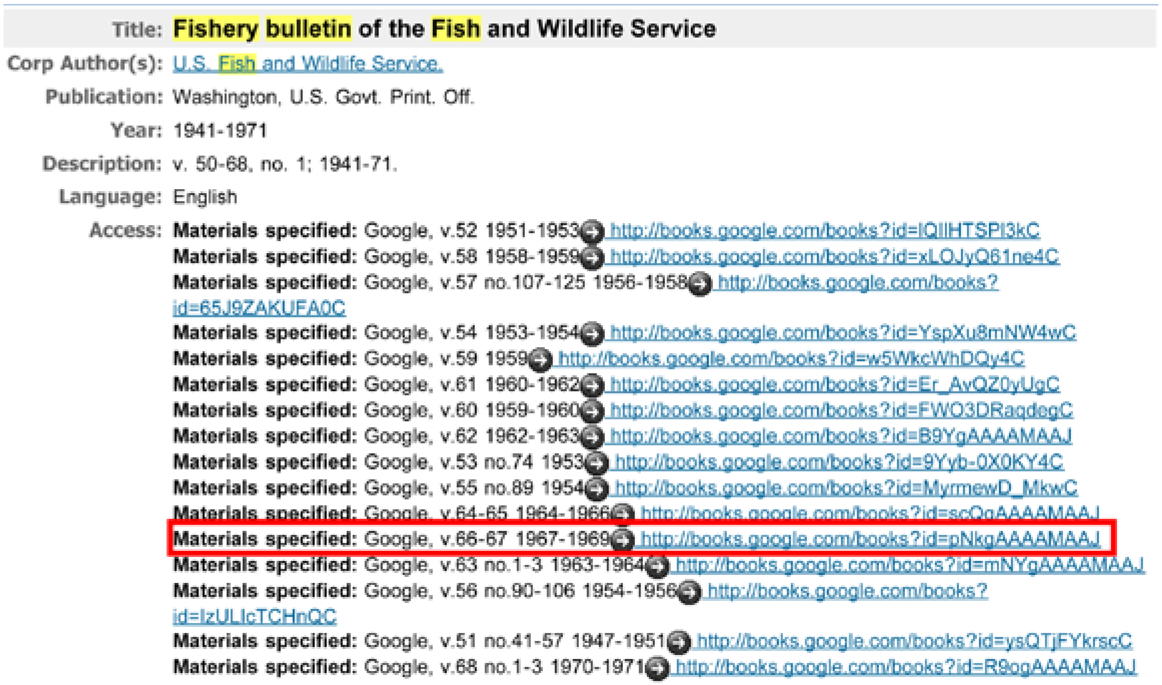
Subject Return books on subject
e.g.Medieval History or "Medieval History"

Publication Date ○ Return content published anytime
● Return content published between Jan 1968 and Dec 1968
e.g. 1999 and 2000, or Jan 1999 and Dec 2000

There were several instances where it was easier and faster to find the volume of a serial through the Google Books' WorldCat record than using either the Google Books' basic or advanced searches. Figure 3, shows the Google Books' record with a link that contains volume 66. From the Google Books' WorldCat records, there is no need to guess how large the date range needs to be to find the specified volume.

Contrary to how serial searching functions in Google Books, HathiTrust's basic search performed best for serials by just searching for the title of the serial without including

Figure 3: WorldCat Record with Sample Title Volume Highlighted



Title: Fishery bulletin of the Fish and Wildlife Service
Corp Author(s): U.S. Fish and Wildlife Service.
Publication: Washington, U.S. Govt. Print. Off.
Year: 1941-1971
Description: v. 50-68, no. 1; 1941-71.
Language: English
Access: **Materials specified:** Google, v.52 1951-1953 http://books.google.com/books?id=IQIIHTSPI3kC
Materials specified: Google, v.58 1958-1959 http://books.google.com/books?id=xLOJyQ61ne4C
Materials specified: Google, v.57 no.107-125 1956-1958 http://books.google.com/books?id=65J9ZAKUFA0C
Materials specified: Google, v.54 1953-1954 http://books.google.com/books?id=YspXu8mNW4wC
Materials specified: Google, v.59 1959 http://books.google.com/books?id=w5WkcWhDQy4C
Materials specified: Google, v.61 1960-1962 http://books.google.com/books?id=Er_AvQZ0yUgC
Materials specified: Google, v.60 1959-1960 http://books.google.com/books?id=FWO3DRaqdegC
Materials specified: Google, v.62 1962-1963 http://books.google.com/books?id=B9YgAAAAMAAJ
Materials specified: Google, v.53 no.74 1953 http://books.google.com/books?id=9Yyb-0X0KY4C
Materials specified: Google, v.55 no.89 1954 http://books.google.com/books?id=MyrmewD_MkwC
Materials specified: Google, v.64-65 1964-1966 http://books.google.com/books?id=scQgAAAAMAAJ
Materials specified: Google, v.66-67 1967-1969 http://books.google.com/books?id=pNkgAAAAMAAJ
Materials specified: Google, v.63 no.1-3 1963-1964 http://books.google.com/books?id=mNYgAAAAMAAJ
Materials specified: Google, v.56 no.90-106 1954-1956 http://books.google.com/books?id=IzULIcTCHnQC
Materials specified: Google, v.51 no.41-57 1947-1951 http://books.google.com/books?id=ysQTjFYkrscC
Materials specified: Google, v.68 no.1-3 1970-1971 http://books.google.com/books?id=R9ogAAAAMAAJ

publication years or volumes. Running a basic catalog search with only the title as the search terms for the same issue of the *Fishery Bulletin of the Fish and Wildlife Service* in the HathiTrust led to a link (first in the result list of fifteen) that led to a record similar to the Google Books' WorldCat record with a list of volumes from which a user could easily click on the desired years. HathiTrust also has a full-text search, and searching for the title as the only search terms provided 115,727 results. Fortunately, the link for volume 66 was the twelfth result on the first page. However, an advanced search for the same issue with the title in the "Title" field and setting the "Year of Publication" field for the same 1967-1968 range that worked in the Google Books' advanced search, did not return any results in HathiTrust. It was only when the "Year of Publication" field was changed to search "During or before" and the year 1941, which is the start year for this serial in the WorldCat and HathiTrust records, that the title did come up as the first result.

To summarize, Google Books has all of these serial volumes notated as "books" rather than "magazines" and in order to locate the right issue, the date ranges in the advanced search have to be broad to catch the years Google Books used. For HathiTrust, the publication date is not when the issue was produced, but taken from the MARC record field of when the serial began publishing under that specific title. This renders both the "Year of Publication" fields in the advanced search, as well as the "Date of Publication" limiter, in HathiTrust (found to the left of the list of results) useless when trying to narrow down to a specific issue. Users must find the serial record for that title in order to retrieve specific volumes and issues. The publication year limiter in both Google Books and HathiTrust are most useful to narrow results for monographs, not serial titles.

Finding the correct volume of a serial proved to be only half the challenge when searching for a specific issue in both repositories. Again, since depository libraries provided the documents to digitize, serial issues often bind together individual documents into a single bound item and were then digitized as a single volume. Therefore serial issues are not well-delineated in the resulting search. Thus, while the document is available online, it can harder to locate individual issues than from the print version (where you can flip through the volume to look for the thicker or colored covers). In the HathiTrust full-text view of a publication, above the image of the document is a drop down box, “Jump to Section”, and contains to the front cover, title page, table of contents, sections (numbered), bibliography, and index. For monograph titles, these links sometimes may/may not work, and the section links may direct to chapters within the book. Unfortunately, this system breaks down for serials and government documents.

Figure 4: HathiTrust Drop-box Example

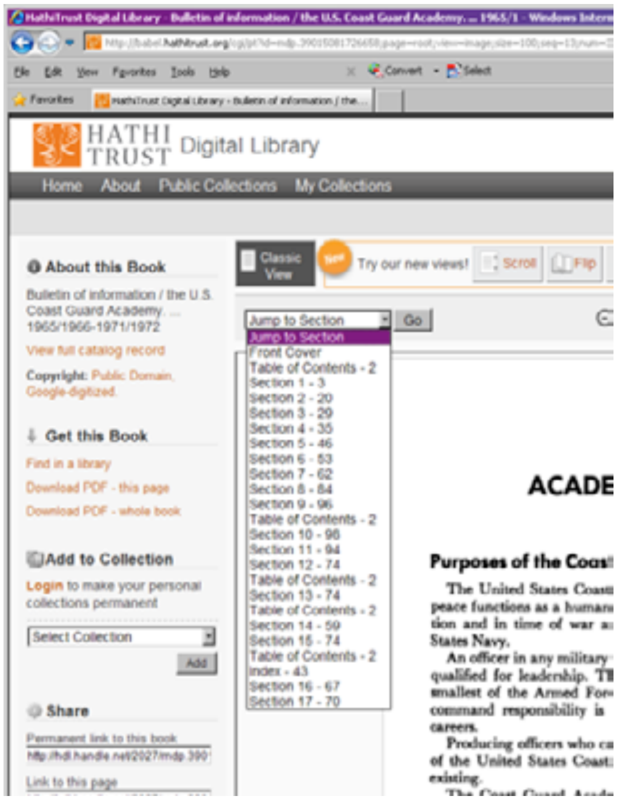
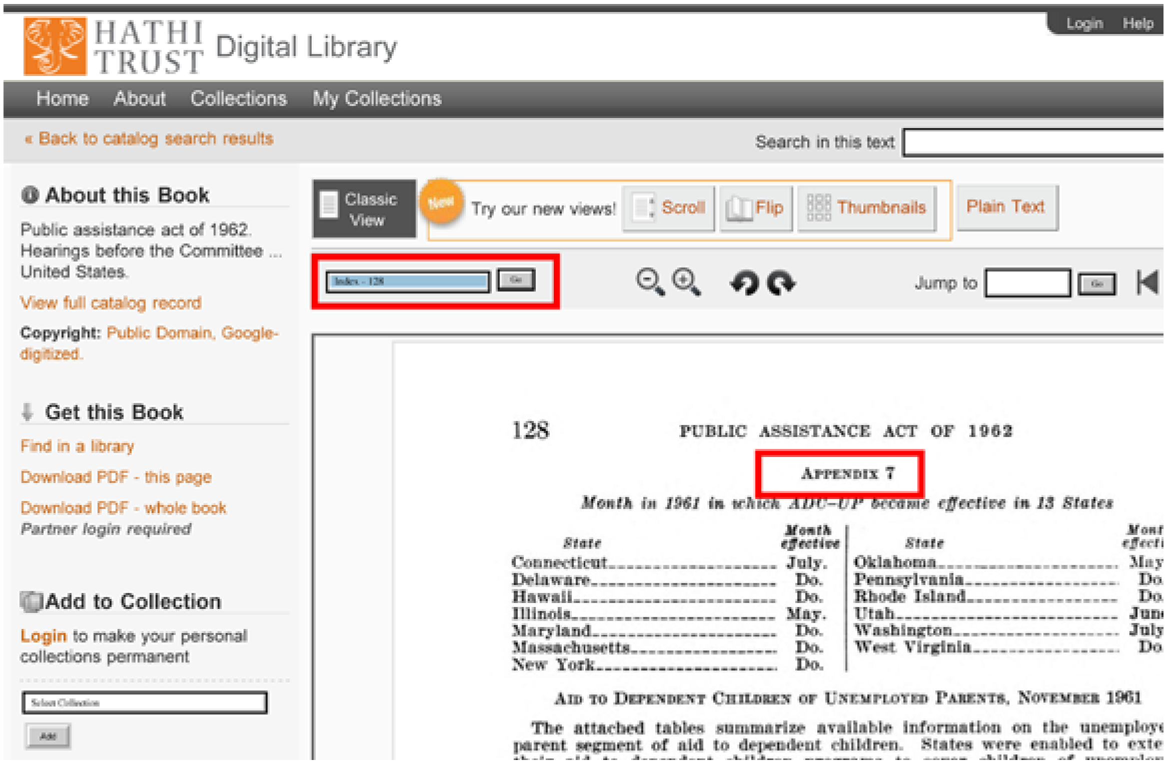


Figure 4 shows a serial in HathiTrust with several years bound together. Some attempt has been made to break out individual issues, but looking at the links for section 10, 11, and 12, table of contents links to few issues are missing (based on the pagination). For usability and accessibility purposes, it would be useful if these links were matched to individual issues or volumes, or, at least, include the dates or numeration of the issues included. Unfortunately the table of contents link is not helpful, for government publications in HathiTrust as many government documents lack a table of contents or have publication data on the front cover. It is difficult to determine where issues begin and end in each scanned volume when using the drop-box.

For other records, this study found that HathiTrust's content links are not accurate in some instances for government publications. In the hearing *Public Assistance Act of 1962*, there is an "Index" link to page 128, but the link actually goes to Appendix 7 of that document (Figure 5). The appendix starts on page 115, and there is no true index for this hearing. HathiTrust's "Search in this text" functionality proved useful to find specific issues and volumes of government documents in some instances, but often returns too many results to make finding individual issues easy. Part of the challenge with using its "Search in this text" search box for the author of this study was determining appropriate keywords to locate the individual issue. For example, searching for a December issue, the word December might not appear on the document itself, but rather as "Dec." or enumerated as the number 12. However, an effective alternative to locate individual issues is to use the "Thumbnails" view as seen in Figure 6 below. With this feature, the cover page for each issue is easily seen; the cover pages of the individual issues typically have more graphics and it is easier to scan visually for the graphical covers and count down to the specific issue.

Figure 5: Links Provided to a Bound Volume in HathiTrust Example



128 PUBLIC ASSISTANCE ACT OF 1962

APPENDIX 7

Month in 1961 in which ADC–UP became effective in 13 States

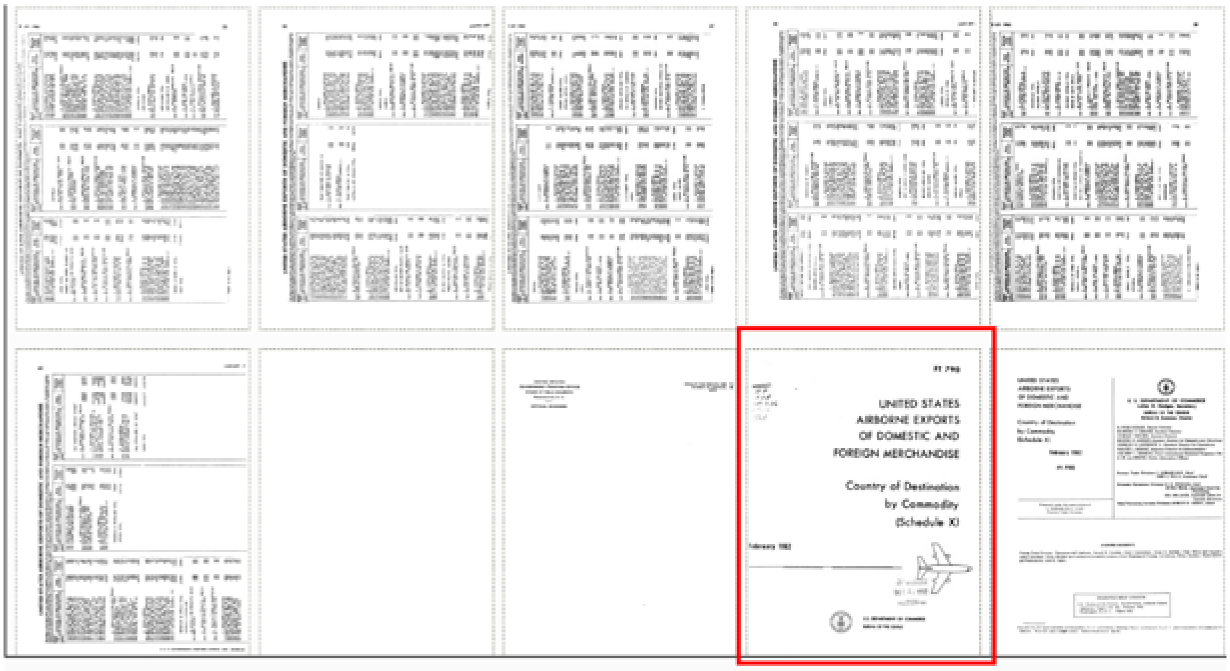
State	*Month effective*	*State*	*Mont effecti*
Connecticut	July.	Oklahoma	May
Delaware	Do.	Pennsylvania	Do
Hawaii	Do.	Rhode Island	Do
Illinois	May.	Utah	Jun
Maryland	Do.	Washington	July
Massachusetts	Do.	West Virginia	Do
New York	Do.		

AID TO DEPENDENT CHILDREN OF UNEMPLOYED PARENTS, NOVEMBER 1961

The attached tables summarize available information on the unemploy parent segment of aid to dependent children. States were enabled to exte

Like HathiTrust, Google Books digitized multiple serial issues as a single file. There were only two serial titles available in full-text online from Google Books. In the first one, the volume was listed as containing issues 290 to 294, but only the 290 issue was digitized and viewable. The other serial title from Google Books that was available in full-text online was *United States Airborne Exports of Domestic and Foreign Merchandise: Commodity (Schedule X) by Country of Destination*. While this serial had all the issues from one year bound together, there are various ways to find a specific issue that include searching for the name of the month for the issue needed in the full-text search box which searches within the book, scrolling through the entire publication until the month was found, or, as in HathiTrust, utilize the thumbnail views of every page in the digitized document to find the covers of individual issues. Google Books also has a "Contents" drop-box, but that feature presented navigation problems similar to the "Contents" drop-box in HathiTrust when searching for government publications.

Figure 6: HathiTrust Thumbnails View Example



UNITED STATES
AIRBORNE EXPORTS
OF DOMESTIC AND
FOREIGN MERCHANDISE

Country of Destination
by Commodity
(Schedule X)

WorldCat Records

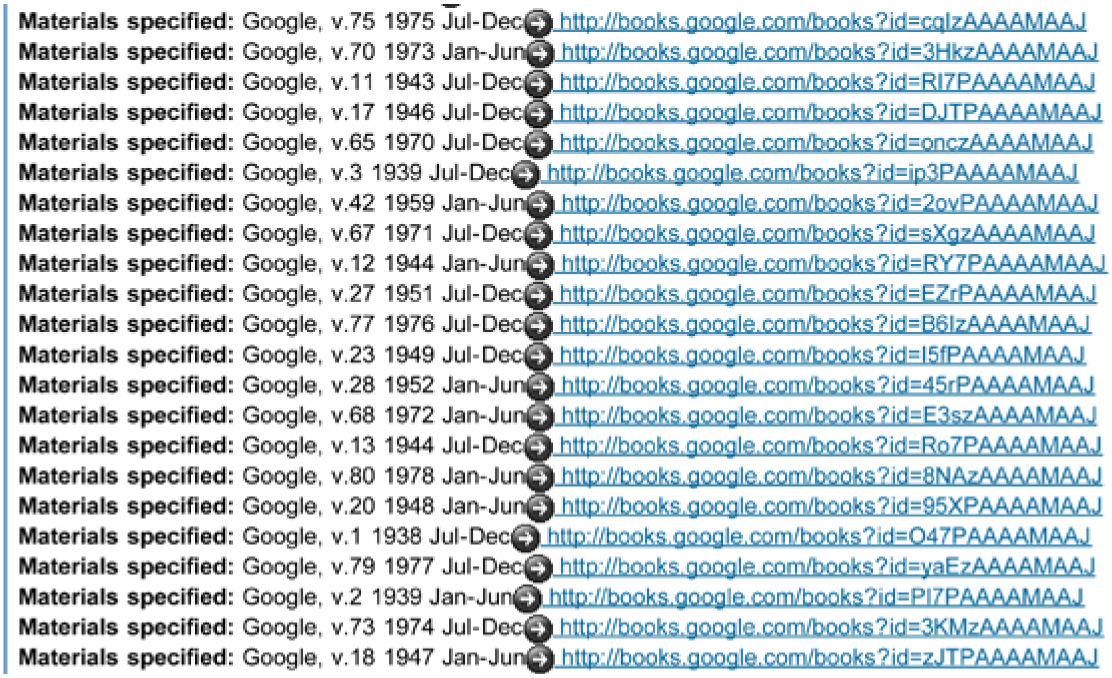
Both Google Books and HathiTrust have records for their resources catalogued in WorldCat. The quality of these records was analyzed to determine the feasibility of adding these records into a local OPAC. The HathiTrust records in WorldCat have bibliographic information similar in quality to the records for the same publication in print format.

Google Books records demonstrated two types of challenges with its bibliographic records in the WorldCat database. For most serial records, each volume has its own link in the serial record, yet these links are not listed in chronological order. As mentioned above, it is sometimes much easier to find a specific year or volume in Google Books via these WorldCat links, rather than using the Google Books search interface. However, by displaying the volume links in seemingly random order, Google Books' records makes navigating to a specific issue, especially for large sets, time consuming in WorldCat. Frequently the fastest way to reach a specific volume from these WorldCat records is to use the web browser's "Find" function to search for the year or volume. An example of this lack of chronological display appears in Figure 7 for the title *United States Customs Court Reports*.

By contrast, a majority of the HathiTrust online records have a single link from the WorldCat record to the serial record within their repository where the list of digitized volumes is in chronological order.

Another issue identified with Google Books' serial records in WorldCat is that each individual volume is listed on its own record, rather than including all the volumes together on a single serial record. One example shown in Figure 8 below is the *I.C.C. Practitioners' Journal*, which had 146 individual records in WorldCat. A similar situation occurred with the title, *The*

Figure 7: Google Books WorldCat Chronological Order Display



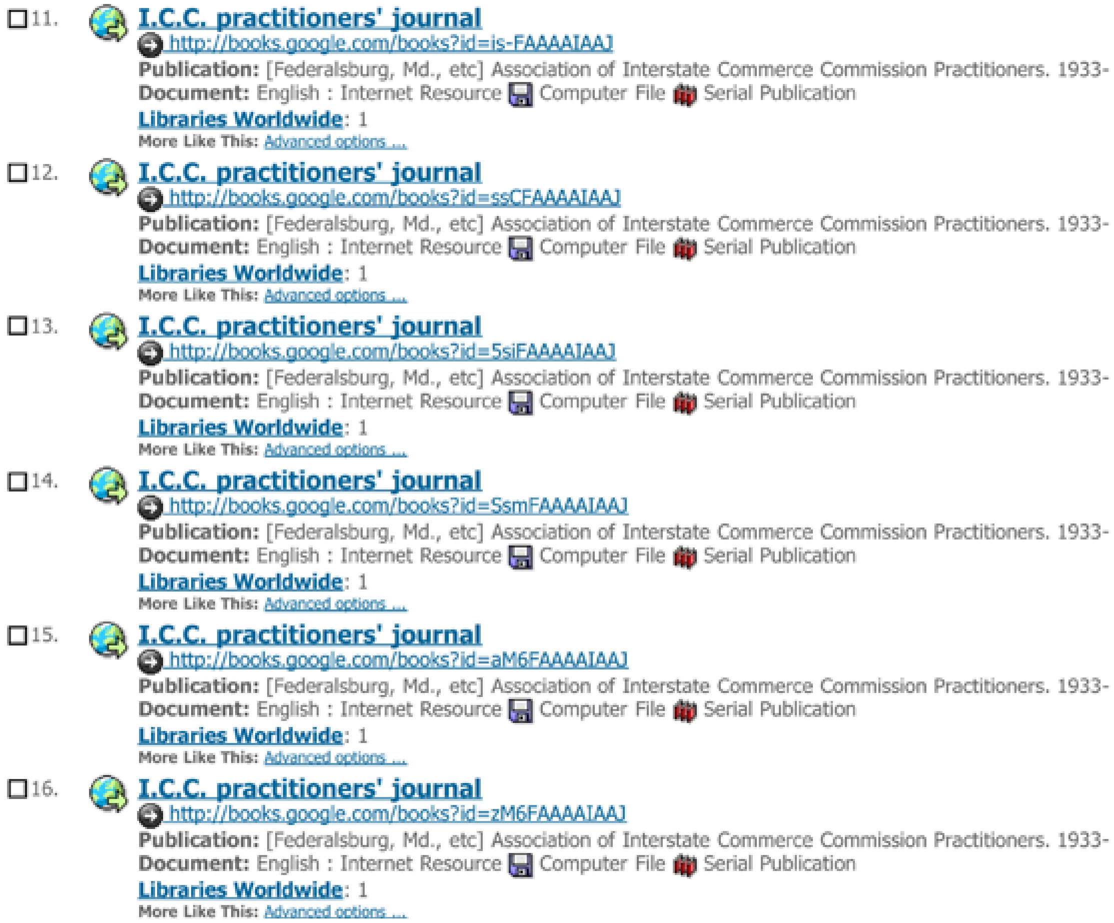
Materials specified: Google, v.75 1975 Jul-Dec http://books.google.com/books?id=cqlzAAAAMAAJ
Materials specified: Google, v.70 1973 Jan-Jun http://books.google.com/books?id=3HkzAAAAMAAJ
Materials specified: Google, v.11 1943 Jul-Dec http://books.google.com/books?id=RI7PAAAAMAAJ
Materials specified: Google, v.17 1946 Jul-Dec http://books.google.com/books?id=DJTPAAAAMAAJ
Materials specified: Google, v.65 1970 Jul-Dec http://books.google.com/books?id=onczAAAAMAAJ
Materials specified: Google, v.3 1939 Jul-Dec http://books.google.com/books?id=ip3PAAAAMAAJ
Materials specified: Google, v.42 1959 Jan-Jun http://books.google.com/books?id=2ovPAAAAMAAJ
Materials specified: Google, v.67 1971 Jul-Dec http://books.google.com/books?id=sXgzAAAAMAAJ
Materials specified: Google, v.12 1944 Jan-Jun http://books.google.com/books?id=RY7PAAAAMAAJ
Materials specified: Google, v.27 1951 Jul-Dec http://books.google.com/books?id=EZrPAAAAMAAJ
Materials specified: Google, v.77 1976 Jul-Dec http://books.google.com/books?id=B6IzAAAAMAAJ
Materials specified: Google, v.23 1949 Jul-Dec http://books.google.com/books?id=I5fPAAAAMAAJ
Materials specified: Google, v.28 1952 Jan-Jun http://books.google.com/books?id=45rPAAAAMAAJ
Materials specified: Google, v.68 1972 Jan-Jun http://books.google.com/books?id=E3szAAAAMAAJ
Materials specified: Google, v.13 1944 Jul-Dec http://books.google.com/books?id=Ro7PAAAAMAAJ
Materials specified: Google, v.80 1978 Jan-Jun http://books.google.com/books?id=8NAzAAAAMAAJ
Materials specified: Google, v.20 1948 Jan-Jun http://books.google.com/books?id=95XPAAAAMAAJ
Materials specified: Google, v.1 1938 Jul-Dec http://books.google.com/books?id=O47PAAAAMAAJ
Materials specified: Google, v.79 1977 Jul-Dec http://books.google.com/books?id=yaEzAAAAMAAJ
Materials specified: Google, v.2 1939 Jan-Jun http://books.google.com/books?id=PI7PAAAAMAAJ
Materials specified: Google, v.73 1974 Jul-Dec http://books.google.com/books?id=3KMzAAAAMAAJ
Materials specified: Google, v.18 1947 Jan-Jun http://books.google.com/books?id=zJTPAAAAMAAJ

American Nautical Almanac For The Year…, published between 1882-1959. Interestingly, this title has an individual record for each volume in October 2010, however it had been whittled down to just nine records as of March 2011.

Both titles have an additional quality issue in that, while each volume has its own record with a single URL, there is no way to view in WorldCat's "List of Records", or the full record view, which individual volume the record is for until the user clicks on the link and views the record in Google Books. Only two titles for HathiTrust records in WorldCat had two different records for the same title from the random sample of documents.

WorldCat records are recommended to link patrons to full-text federal publications through their online catalogs, but this would involve some work in determining which records to download as the WorldCat records do not indicate if the links to either repository are leading to a full-text edition or a Google Books "Snippet" view or a HathiTrust "Limited" (search-only) view.

Figure 8: WorldCat Individual Volume Records Display



□ 11. I.C.C. practitioners' journal
http://books.google.com/books?id=is-FAAAAIAAJ
Publication: [Federalsburg, Md., etc] Association of Interstate Commerce Commission Practitioners. 1933-
Document: English : Internet Resource Computer File Serial Publication
Libraries Worldwide: 1
More Like This: Advanced options ...

□ 12. I.C.C. practitioners' journal
http://books.google.com/books?id=ssCFAAAAIAAJ
Publication: [Federalsburg, Md., etc] Association of Interstate Commerce Commission Practitioners. 1933-
Document: English : Internet Resource Computer File Serial Publication
Libraries Worldwide: 1
More Like This: Advanced options ...

□ 13. I.C.C. practitioners' journal
http://books.google.com/books?id=5siFAAAAIAAJ
Publication: [Federalsburg, Md., etc] Association of Interstate Commerce Commission Practitioners. 1933-
Document: English : Internet Resource Computer File Serial Publication
Libraries Worldwide: 1
More Like This: Advanced options ...

□ 14. I.C.C. practitioners' journal
http://books.google.com/books?id=SsmFAAAAIAAJ
Publication: [Federalsburg, Md., etc] Association of Interstate Commerce Commission Practitioners. 1933-
Document: English : Internet Resource Computer File Serial Publication
Libraries Worldwide: 1
More Like This: Advanced options ...

□ 15. I.C.C. practitioners' journal
http://books.google.com/books?id=aM6FAAAAIAAJ
Publication: [Federalsburg, Md., etc] Association of Interstate Commerce Commission Practitioners. 1933-
Document: English : Internet Resource Computer File Serial Publication
Libraries Worldwide: 1
More Like This: Advanced options ...

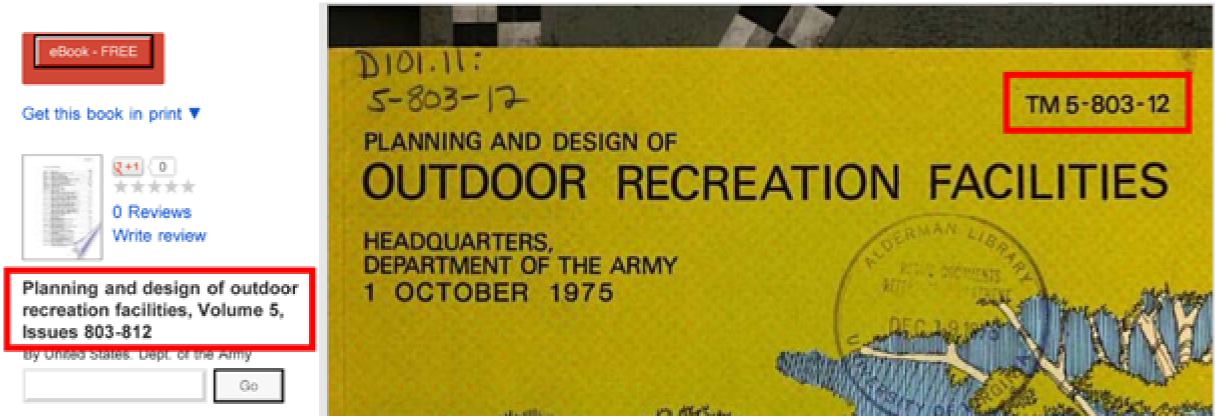
□ 16. I.C.C. practitioners' journal
http://books.google.com/books?id=zM6FAAAAIAAJ
Publication: [Federalsburg, Md., etc] Association of Interstate Commerce Commission Practitioners. 1933-
Document: English : Internet Resource Computer File Serial Publication
Libraries Worldwide: 1
More Like This: Advanced options ...

Digitization and Metadata Quality

While conducting this study, issues related to the quality of digitization and metadata surfaced. Irregularities mostly involved Google Books' records. Because Google relies on Optical Character Recognition (OCR) and computer processes to create metadata for their records, the monograph and serial templates used do not apply well to government publications. One of the most frequently recurring problems is Google Books' attempt to convert SuDoc numbers and agency document/publication identification numbers into volume and issue numbers. In the hearing, *Public Assistance Act of 1962*, the metadata incorrectly indicates that the title includes "Parts 96-98" -- derived from the latter half of this document's SuDoc number: Y 4.F 49:P 96/8. In another example, as shown in Figure 9, Google Books attempts to make the document's publication number into a volume and an issue number.

Another frequently found metadata issue was Google Books' provision of pseudo subject headings. A couple of examples are discussed here. For the document *Cases relating to Rules governing monthly reports of railway accidents (1922 revision)* and *Interpretations of rules governing monthly reports of railway accidents (1922 revision) issued in 1934*, a case reporter series; Google marked this as, "Crafts & Hobbies" in the Google Books record. However, the Google Books' WorldCat record has the correct subject heading of "Railroad accidents—United States." Another example of subject heading problems is illustrated by a record from the U.S. Naval Observatory's publication *The American Ephemeris and Nautical Almanac* which was marked as "Juvenile Nonfiction." Those searchers wanting to use authoritative subject headings are better served using HathiTrust or WorldCat than Google Books.

Figure 9: Google Books Metadata Example



Planning and design of outdoor recreation facilities, Volume 5, Issues 803-812
By United States. Dept. of the Army

Incorrectly assigned publication years are also an occasional metadata mistake. Google Books marked the *Regulations of the Civil Aeronautics Board* as 1980 in a "Snippet" view record, which is incorrect. This date was retrieved from the top page of this document but viewing this document in HathiTrust full-text shows that the first scanned page is actually a 1980 transmittal sheet for the 1977 (correct date) edition of the title. HathiTrust also had issues with this publication, and does not provide a year at all.

Both databases contain instances where the scanned document was blurred or had the image of the scanner's hand in the PDF. This seldom occurred, and both databases have feedback mechanisms to report problems of blurred text or missing pages.

Comparison of Exporting Citation Functionality

Citation exporting is an important feature for all databases, and both Google Books and HathiTrust provide this service but in a limited manner. Google Books has the ability to export to BiBTex, EndNote, and RefMan bibliographic citation management software. While HathiTrust has a "Cite this" function that returns a formated citation for the document in APA and MLA formats, and only exports to EndNote (with a caution that serial citations may be incomplete). It would be a useful addition if HathiTrust could also provide a Chicago/Turabian format option for historians who frequently use older government documents that typically cite with this format. In both Google Books and HathiTrust, users are limited to exporting records one at a time, since the export citation option appears only on the individual title or catalog record.

Since EndNote was the only common citation export format in both databases, a comparison of downloaded citations was conducted on twenty documents from the main sample. The HathiTrust exported citations were more accurate, with the exception of its "Author" field. HathiTrust usually listed "United States" as the author, which is technically correct for most documents, but it should also provide users with the parent agency as well. In contrast, Google Books' citations included the publishing agency, e.g. "Census Bureau", along with "United States" in the author field. However, HathiTrust consistently had data in the city field, which Google Books lacks. HathiTrust was also more likely to categorize a government serial publication as a "Reference Type" or "Journal Article," as opposed to Google Books classifying them as a "Book." And, sometimes Google Books displays the volume information for serials in the "Notes" field.

Formatting for the publication year in serial records was incomplete in HathiTrust generated citations. It exports the start year of the serial from the MARC record into Endnote's year field. This is correct for a serial catalog record, but users must edit the record to include the specific volume. The MARC record 500 field appears in EndNote's notes field from exported HathiTrust citations. Both repositories provided title, publisher, and URL data correctly and consistently in the random sample of exported citations.

WorldCat allows users to export citations to EndNote, Refworks, or as a basic text file. A look at the same titles exported to EndNote showed that WorldCat citations were not an improvement since it identifies the serial records as "generic." It does correctly identify monographs as books and the government agency as publisher. WorldCat also provides a "Cite this Item" function that formats for all of the major citation styles , i.e., APA, Harvard, MLA, and Turabian.

Additional Functionality

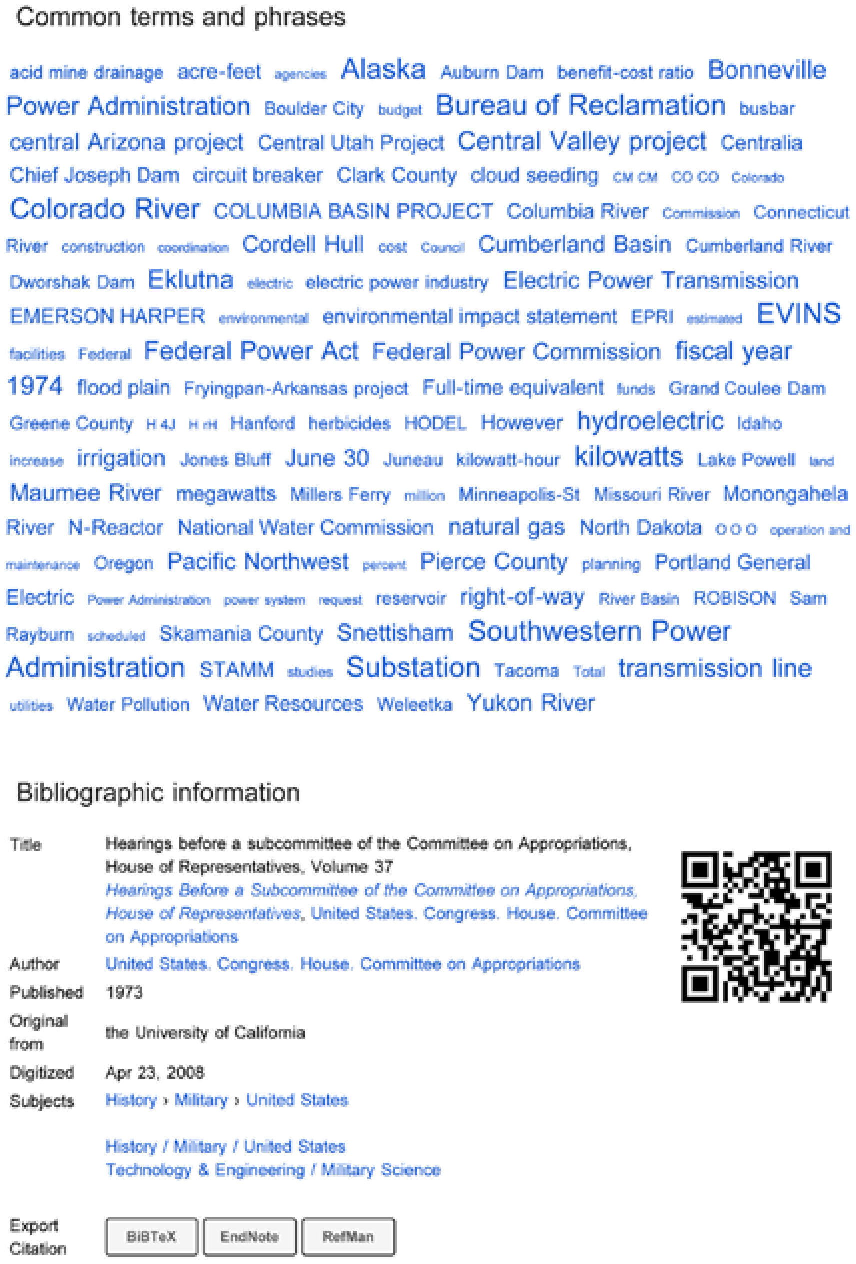
Google Books provides additional functionality for documents available only in its "Snippet" view. The first is a "Common terms and phrases" feature that creates a word cloud from the text within the document (Figure 10). This word cloud feature provides a useful discovery tool within the repository. It is more useful than HathiTrust's "Limited" records display, which simply shows how many times the search term appeared on each page.

Occasionally Google Books also provided a map feature in the "Snippet" view, called "Places mentioned in this book." This displays a Google Map with location pointers on the places mentioned, as well as a link that will launch Google Earth. This usually occurred when street addresses appeared in the digitized text. Google Books also has a QR code in the "Snippet" view that provides a simple URL to the book record being viewed. HathiTrust does not have any comparable features at this time.

Google Books also has a feature called "Add to My Library" where citations can be added to create a list. The list of titles added to the "My Library" function are publicly viewable by default, but can be marked as private.

HathiTrust has a "Share" feature that allows users to either share the permanent link of the document, or a specific page in the publication. Like Google Books, HathiTrust also has an "Add to Collection" function that allows the end-user to add records to a previously created collection or to create a new collection. There is also a "Collections" function where users can make their own collections available to be viewed by other users, or create a private list. Several

Figure 10: Additional Features in Google Books Example



Common terms and phrases

acid mine drainage acre-feet agencies Alaska Auburn Dam benefit-cost ratio Bonneville Power Administration Boulder City budget Bureau of Reclamation busbar central Arizona project Central Utah Project Central Valley project Centralia Chief Joseph Dam circuit breaker Clark County cloud seeding CM CM CO CO Colorado Colorado River COLUMBIA BASIN PROJECT Columbia River Commission Connecticut River construction coordination Cordell Hull cost Council Cumberland Basin Cumberland River Dworshak Dam Eklutna electric electric power industry Electric Power Transmission EMERSON HARPER environmental environmental impact statement EPRI estimated EVINS facilities Federal Federal Power Act Federal Power Commission fiscal year 1974 flood plain Fryingpan-Arkansas project Full-time equivalent funds Grand Coulee Dam Greene County H 4J H rH Hanford herbicides HODEL However hydroelectric Idaho increase irrigation Jones Bluff June 30 Juneau kilowatt-hour kilowatts Lake Powell land Maumee River megawatts Millers Ferry million Minneapolis-St Missouri River Monongahela River N-Reactor National Water Commission natural gas North Dakota O O O operation and maintenance Oregon Pacific Northwest percent Pierce County planning Portland General Electric Power Administration power system request reservoir right-of-way River Basin ROBISON Sam Rayburn scheduled Skamania County Snettisham Southwestern Power Administration STAMM studies Substation Tacoma Total transmission line utilities Water Pollution Water Resources Weleetka Yukon River

Bibliographic information

Title	Hearings before a subcommittee of the Committee on Appropriations, House of Representatives, Volume 37 *Hearings Before a Subcommittee of the Committee on Appropriations, House of Representatives*, United States. Congress. House. Committee on Appropriations
Author	United States. Congress. House. Committee on Appropriations
Published	1973
Original from	the University of California
Digitized	Apr 23, 2008
Subjects	History › Military › United States History / Military / United States Technology & Engineering / Military Science
Export Citation	BiBTeX EndNote RefMan

examples of these user-created lists already in Google Books pertaining to federal publications include:"96th Congress Documents," "NASA Technical Reports," and "Nuclear Reactor Library."

Privacy Issues

User privacy is an issue in these repositories, which is different from typical library vendor databases because both repositories are freely accessible to anyone online. There are features within each that are viewable to anyone, and in the case of Google, some information is sent to third parties. Patrons using library resources expect their privacy to be protected, and librarians directing users to resources that track online searching behavior need to take user privacy into account. Google received many calls to provide greater privacy protection during the Google Books settlement agreement discussion in 2009 (Center for Democracy and Technology, 2009; Electronic Frontier Foundation, 2009; Kravets, 2009). When users purchase books through Google's bookstore, some user information is shared with third parties. Publishers receive sales information, but not personal information, and this data is connected to a user's Google Account. Google specifically states, however, "Unless you are logged into your Google Account, your activity on Google Books will not be associated with your Google Account" (Google, 2010). This means that activity will be tracked and associated if a user is logged into their Google account when using Google Books. Activity is also tracked when users log into Google Books to use features such as "My Library" or to purchase books. When users add books to "My Library" in Google Books, they must mark these lists as "public" in order to share links with others. Reviews, in contrast, are automatically made public, but may be deleted (Google, n.d.). Pre-court settlement FAQ's for Google Books mention that Google Books was considering selling institutional subscriptions that would enable subscribers to view more materials online and would use Shibboleth, a technology that authenticates a user as a subscribing institution without identifying the individual user (Google Data, 2009).

HathiTrust's privacy policy states that it only logs transactions tied to individuals for a limited period of time, that these logs are used for trouble-shooting and problem resolution, and it states that no personal information will be shared with third parties. Additionally, it states that when a problem is resolved the logging information is destroyed. Authentication for users from the partner institutions uses Shibboleth. HathiTrust utilizes Google Analytics, which uses a cookie and transmits IP addresses to Google. HathiTrust's privacy website explains how users can opt out of this by turning off cookies or by using the Google Analytics Opt-out Browser Add-on (HathiTrust, n.d.b).

Conclusion

Since most users want access to full-text, HathiTrust offers the best database for finding government documents after 1923. Users concerned with privacy issues may prefer HathiTrust or want to use Google Books while logged out of their Google account. Those familiar with the Google eBookstore or who want the added functionality of data visualization to read and provide reviews may prefer Google Books. Regarding record overlap, HathiTrust had a greater percentage of publication records also available in Google Books, but with fewer records overall, while Google Books had records for more government documents than HathiTrust, and therefore a smaller overlap range. These results show that if a user cannot find a federal document in HathiTrust, Google Books might have a "Snippet" view record for that document and that record may provide more information for users to determine if the document is one that would be useful to them. However, caution is also advised for Google Books' records, as metadata mistakes on the full-text and "Snippet" view records should be taken into account by librarians and users alike.

Both repositories' citation exportation capabilities were disappointing. Users should be warned that the information exported is often incomplete. Users should follow the citation style manuals for government publications rather than rely on the computer-generated citations. The simple search feature for both repositories worked best when it was unclear whether the document sought after was a monograph or a serial, as the problems with how the bound volumes and issues were digitized may negate the ability to narrow by date to find specific issues. Serial searching is easier in HathiTrust when finding individual volumes. Sometimes the only way to find a specific Google Books volume was through the link via WorldCat, but finding an individual issue within a volume was easiest in both repositories using the thumbnail view feature. It would be helpful to researchers if both repositories, instead of presenting serials in the monograph templates, created a way that makes accessing individual issues easier.

Since these online book repositories are still relatively new, the interfaces and services are evolving. Digitization and quality control of materials is ongoing. Further research is needed to re-evaluate the ability and quality of federal publications over time. Additional research is necessary to assess the availability of federal publications through HathiTrust and Google Books in discovery layer interfaces (software capable of searching multiple catalog and article databases at once) such as Primo, Summon, Encore, and others.

References

Baksik, C. (2006). Fair use or exploitation? The Google Book search controversy. *Portal: Libraries and the Academy*, 6(4), 399-415. doi: 10.1353/pla.2006.0047

Blakeley, R. (2009). What was lost, now is found: Using Google Books and Internet Archive to enhance a government documents collection with digital documents. *Dttp: Documents to the People*, 37(3), 26-9. Retrieved from http://wikis.ala.org/godort/images/0/0e/DttP_37n3_web.pdf

Blakeley, R. (2008, March 4). Creating gov doc 'libraries' in Google Books. *Free Government Information*. [Web log post]. Retrieved from http://freegovinfo.info/node/1689

Center for Democracy & Technology. (2009, September 9). *CDT files brief urging privacy safeguards for Google Books*. Retrieved from http://www.cdt.org/policy/cdt-files-brief-urging-privacy-safeguards-google-books

Christenson, H. (2011). HathiTrust: A research library at web scale. *Library Resources & Technical Services*, 55(2), 93-102.

Electronic Frontier Foundation. (2009, July). *Google Book Search settlement and reader privacy*. Retrieved from http://www.eff.org/issues/privacy/google-book-search-settlement

Google. (2011a). *About Google books*. Retrieved from http://www.google.com/googlebooks/about.html

Google. (2011b). *Google Books library project - An enhanced card catalog of the world's books*. Retrieved from http://www.google.com/googlebooks/library.html

Google. (2010, December 6). *Google Books privacy policy*. Retrieved from http://www.google.com/googlebooks/privacy.html

Google. (2004, Dec 14). *Google checks out library books*. Retrieved from http://www.google.com/press/pressrel/print_library.html

Google. (n.d.). *My library FAQ*. Retrieved from http://books.google.com/support/bin/answer.py?hl=en&answer=75375

Google Books. (2011). *Google Books Settlement*. Retrieved from http://www.googlebooksettlement.com/agreement.html

Google Data. (2009, July 23). *The Google Books settlement and privacy: frequently asked questions*. Retrieved from http://googledata.org/google-books/the-google-books-settlement-and-privacy-frequently-asked-questions/

HathiTrust. (2011, September 9). *Update on August 2011 activities*. Retrieved from http://www.hathitrust.org/updates_august2011

HathiTrust. (n.d.). *Features and benefits.* Retrieved from http://www.hathitrust.org/features_benefits

HathiTrust. (n.d.). *Home*. Retrieved from http://www.hathitrust.org/

HathiTrust. (n.d.). *Privacy*. Retrieved from http://www.hathitrust.org/privacy

Hawkins, K.S. & York, J. (2010). *The HathiTrust shared digital repository for preserving and providing access to the world's print culture*. Retrieved from http://www.hathitrust.org/documents/HathiTrust-Rumyantsev-201003-en.pdf

Jones, E. (2010). Google Books as a general research collection. *Library Resources & Technical Services*, 54(2), 77-89.

Kravets, D. (2009, September 8). Google Book plan hits privacy snag. *Wired*, Retrieved from http://www.wired.com/threatlevel/2009/09/google-books-plan-hits-privacy-snag/

Martin, S. (2008). To Google or not to Google, that is the question: Supplementing Google Book search to make it more useful for scholarship. *Journal of Library Administration*, 47(1/2), 141-150. doi: 10.1080/01930820802111025

O'Brien, K. (2011). HathiTrust. *Journal of the Medical Library Association*, 99(2), 177-8. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3066577/

Online Computer Library Center (OCLC). (2010, April 5). *OCLC adding records to WorldCat for Google Books Library Project and HathiTrust Digital Library collections*. Retrieved from http://www.oclc.org/news/releases/2010/201019.htm

Orwant, J. (2008, June 23). U.S. copyright renewal records available for download. *Inside Google Books.* [Web log post]. Retrieved from http://booksearch.blogspot.com/2008/06/us-copyright-renewal-records-available.html

Sare, L. (2011). Availability of legacy documents online. *Internet Reference Services Quarterly*, 16(1-2), 55-66. doi: 10.1080/10875301.2011.582824

Appendix A: Overlap of Records by Decade

	1940	1950	1960	1970
Full Text in both HathiTrust and Google Books	0	0	2	1
Full Text in HathiTrust and No Record in Google Books	36	21	22	15
Full Text in HathiTrust and only Snippet view in Google Books	38	42	32	44
Full Text in HathiTrust and only No Preview Available view in Google Books	15	15	20	14
Limited view in HathiTrust and Full Text in Google Books	0	0	0	0
Limited view in HathiTrust and No Reord in Google Books	5	16	12	8
Limited view in HathiTrust and Snippet view in Google Books	2	9	20	27
Limited view in HathiTrust and No Preview Available view in Google Books	2	4	5	11
No record in HathiTrust and Full Text in Google Books	4	0	2	5

Appendix B: Number and Percentage of Record Overlap by Decade

	1940	1950	1960	1970
Record Overlap in HathiTrust	55 of 98 (56%)	79 of 107 (74%)	79 of 114 (69%)	97 of 120 (80%)
Record Overlap in Google Books	55 of 181 (30%)	79 of 187 (42%)	79 of 209 (38%)	97 of 232 (42%)