Harvesting Government Documents into the Local Catalog: A New Model for Online Access

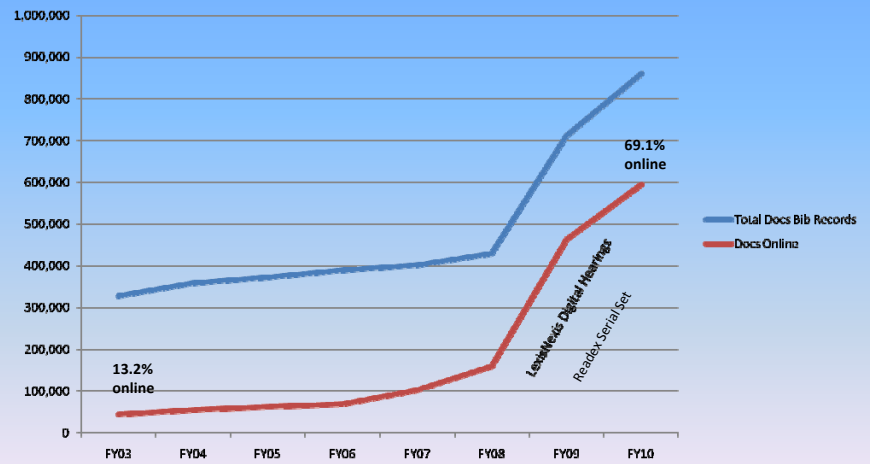
Christopher C. Brown
University of Denver, Penrose Library
(303) 871-3404
cbrown@du.edu

About University of Denver

- Depository since 1909
- Historically a 70-75% selective
- Now a 4.8% selective, but receive 100% of online cataloging
- Adding URLs to historic documents

UNIVERSITY OF DENVER
START FROM A HIGHER PLACE

We can't get enough online docs



University of Denver documents cataloging

Our Motivation to Harvest

- Our users are used to using electronic documents
- Our paper docs are almost entirely in storage
- We will be remodelling our library – totally displaced for at least 18 months

Hathi Trust Launch

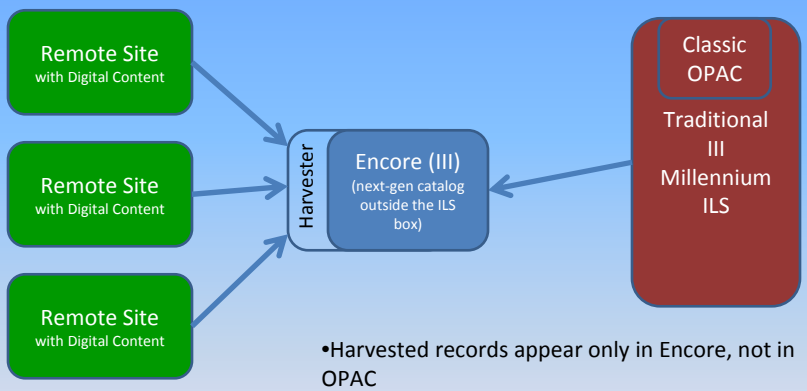
- October 13, 2008 – Hathi Trust launched
- 12-member Committee on Institutional Cooperation (CIC) + 11 libraries of the University of California system
- Currently 29 members
- University of Denver not a member

HATHI TRUST
a shared digital repository

OAI-PMH Harvesting

- http://www.openarchives.org/
- Promotes interoperability standards for dissemination of content
- Hathi Trust allows harvesting of its records
- Innovative Interface's Encore catalog allows for records to be harvested (with the purchase of a harvester connection)

Encore Model



- Harvested records appear only in Encore, not in OPAC
- ILS records displayed "live" in OPAC, but ~15 minute delay in Encore
- Harvested records update on a periodic schedule – in our case daily

PD = where docs generally live

ATTRIBUTES

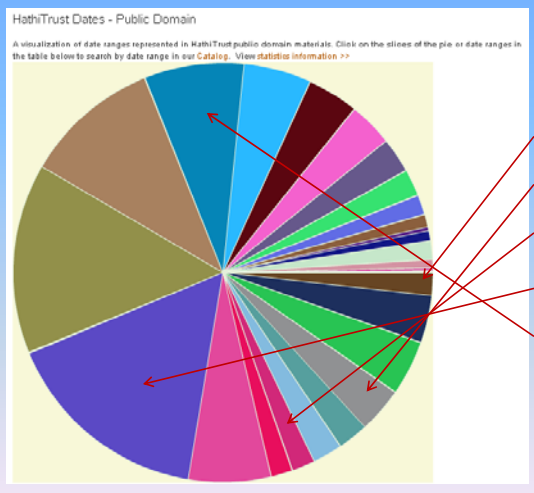
id	name	type	dscr
1	pd	copyright	public domain
2	ic	copyright	in-copyright
3	opb	copyright	out-of-print and brittle (implies in-copyright)
4	orph	copyright	copyright-orphaned (implies in-copyright)
5	und	copyright	undetermined copyright status
6	umall	access	available to UM affiliates and walk-in patrons (all campuses)
7	world	access	available to everyone in the world
8	nobody	access	available to nobody; blocked for all users
9	pdus	copyright	public domain only when viewed in the US

Hathi Trust Attributes
From: http://www.hathitrust.org/rights_database

PD vs. PDUS

- **Mass identification of copyright status based on bibliographically-derived information:** a) As texts are loaded, a set query in Mirlyn identifies those texts that are:US federal government documents, or
- published in the US prior to 1923, or
- published outside of the US before 1870
- These are treated as public domain (ATTRIBUTE name=pd) based on bibliographically-derived information (REASON name=bib). We do not restrict access to these materials. b) Those texts that do not meet these criteria (e.g,. US post-1923 and not a government document) are treated as in-copyright (i.e., ATTRIBUTE name=ic and REASON name=bib). c) An additional attribute is used to represent works published outside the United States between 1870 and 1923 because copyright status for these works depends on the location of the user. Works published outside the US prior to 1923 are in the public domain; however, due to the variations in copyright law in countries outside the US, it is estimated that 1870 is the earliest date works published in these countries may still be under copyright. Therefore, users accessing the volume from US IP addresses will have access to the works published outside the US between 1870 through 1923; however, users with non-US IP addresses will not (ATTRIBUTE name=pdus and REASON name=bib).

Public Domain Distribution



Date range	Count	Percent
2000-2009	13,781	1.79
1990-1999	28,092	3.65
1980-1989	32,431	4.22
1970-1979	26,931	3.50
1960-1969	18,190	2.37
1950-1959	17,703	2.30
1940-1949	13,752	1.79
1930-1939	12,654	1.65
1920-1929	49,465	6.43
1910-1919	123,504	16.06
1900-1909	112,681	14.66
1890-1899	80,528	10.47
1880-1889	59,672	7.76

Sampling Method

- I wanted to see how many government documents were in our Hathi Trust harvest
- Limit to Hathi Trust for a given year
- Examine first result on each page of 25 results (4% of results) [limitation: Encore only displays first 1,000 results]

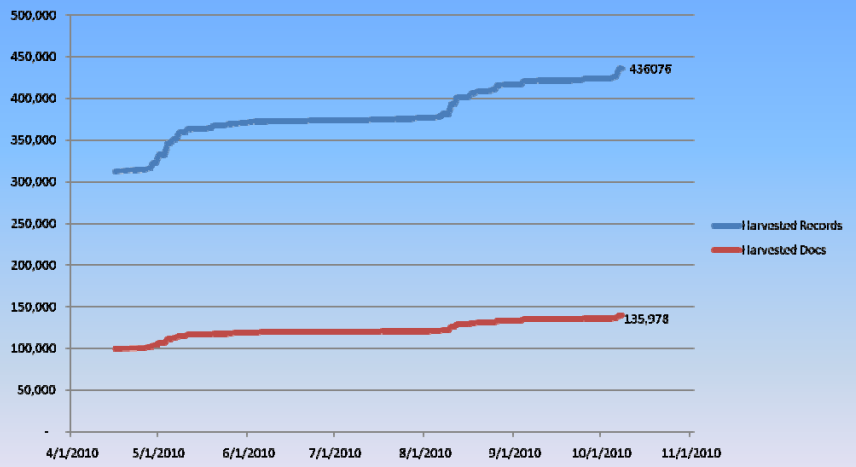
Harvesting Hathi Docs: The Stats

Date Range	Hathi Totals	Hathi All Pub Domain pdus + pd	Hathi pdus	DU pd Harvest	Docs Sampling	
2000-2009	457,620	13,781	726	13,055	13,055	100.00%
1990-1999	641,226	28,075	880	27,195	26,082	95.91%
1980-1989	654,837	32,411	1,204	31,207	30,226	96.86%
1970-1979	555,688	26,877	2,046	24,831	23,763	95.70%
1960-1969	478,498	18,066	1,987	16,079	6,124	38.09%
1950-1959	241,558	17,422	863	16,559	3,635	21.95%
1940-1949	160,670	13,377	600	12,777	2,976	23.29%
1930-1939	150,560	12,246	654	11,592	1,980	17.08%
1920-1929	141,441	48,103	27,108	20,995	1,130	5.38%
1910-1919	121,891	118,194	75,955	42,239	2,842	6.73%
1900-1909	122,945	103,171	70,900	32,271	1,529	4.74%
1890-1899	73,989	73,156	50,502	22,654	389	1.72%
1880-1889	54,481	53,969	38,928	15,041	444	2.95%
1870-1879	35,596	35,187	27,202	7,985	175	2.19%
1860-1869	26,253	26,049	2,273	23,776	134	0.56%
	3,917,253	**620,084**	**301,828**	**318,256**	**114,484**	**34.21%**

Statistics as of mid-September, 2010

The Docs Sampling columns show the estimated numbers of docs per year and the estimated percentage of docs per year from the Harvest

Hathi Harvest in Perspective



Tracking of daily harvesting since harvesting began, April 16, 2010 through October 8, 2010

Inclusion of Serials

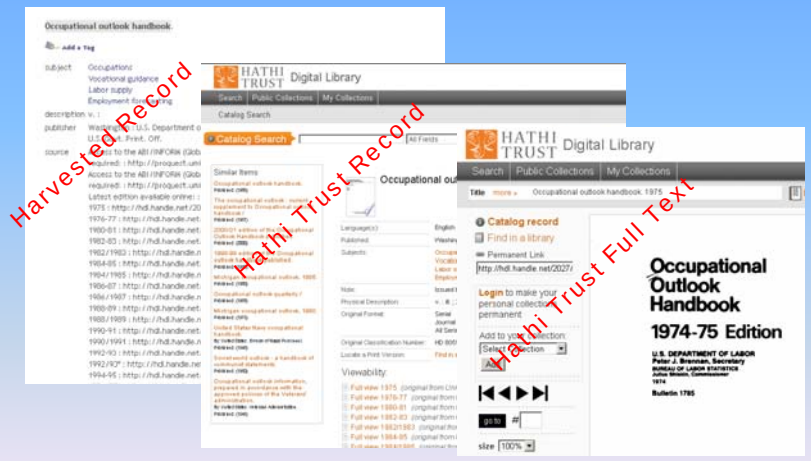
Bulletin of the Bureau of Standards.

Add a Tag

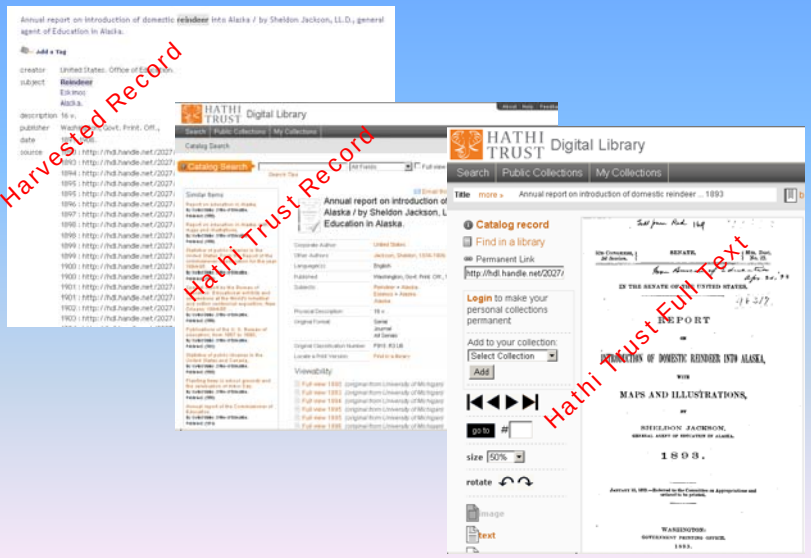
creator	United States. National Bureau of Standards.
subject	Science.
description	14 v.
publisher	Washington, Govt. Print. Off.
source	v.1 1905 : http://hdl.handle.net/2027/mdp.39015067123086 v.10 (1914) : http://hdl.handle.net/2027/uc1.b3505583 v.10 1914 : http://hdl.handle.net/2027/mdp.39015067122773 v.11 (1914-15) : http://hdl.handle.net/2027/uc1.b3505584 v.11 1915 : http://hdl.handle.net/2027/mdp.39015067122781 v.12 1915-1916 : http://hdl.handle.net/2027/mdp.39015067122799 v.13 (1916-17) : http://hdl.handle.net/2027/uc1.b3505586 v.13 1916-1917 : http://hdl.handle.net/2027/mdp.39015067122948 v.14 (1918-19) : http://hdl.handle.net/2027/uc1.b3505587 v.14 1918-1919 : http://hdl.handle.net/2027/mdp.39015018028434 v.2 (1906) : http://hdl.handle.net/2027/uc1.b3505575 v.2 1906 : http://hdl.handle.net/2027/mdp.39015006998218 v.3 1907 : http://hdl.handle.net/2027/mdp.39015067123094 v.4 (1907-08) : http://hdl.handle.net/2027/uc1.b3505577 v.4 1907-1908 : http://hdl.handle.net/2027/mdp.39015014125788 v.5 (1908-09) : http://hdl.handle.net/2027/uc1.b3505578 v.5 1908-1909 : http://hdl.handle.net/2027/mdp.39015018025125 v.6 (1909-10) : http://hdl.handle.net/2027/uc1.b3505579 v.6 1909-1910 : http://hdl.handle.net/2027/mdp.39015067122922 v.7 1911 : http://hdl.handle.net/2027/mdp.39015018028459 v.8 (1912-13) : http://hdl.handle.net/2027/uc1.b3505581 v.8 1912 : http://hdl.handle.net/2027/mdp.39015014125762 v.9 (1913) : http://hdl.handle.net/2027/uc1.b3505582 v.9 1913 : http://hdl.handle.net/2027/mdp.39015067122930

Although serial holdings do not sort properly, users can figure out what they need.

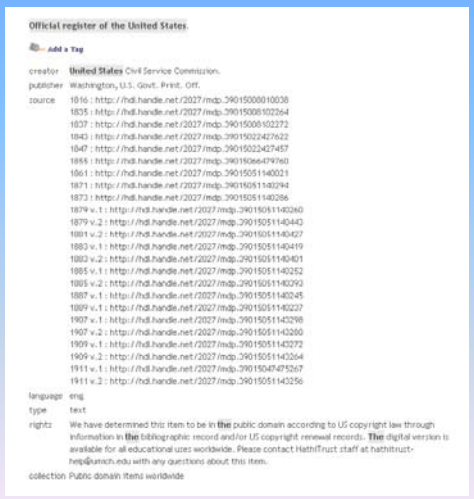
Access to Older Serials



And Very Old Serials



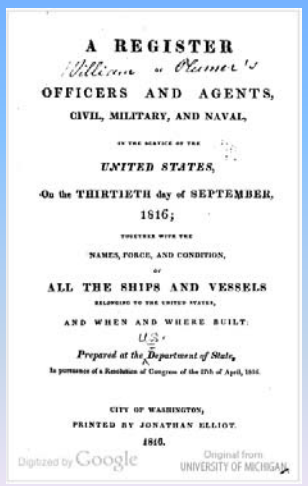
And Very, Very Old Serials



Official register of the United States.

Add a Tag

creator United States Civil Service Commission.
publisher Washington, U.S. Govt. Print. Off.
source 1816 : http://hdl.handle.net/2027/mdp.39015008610038
1835 : http://hdl.handle.net/2027/mdp.39015008102264
1837 : http://hdl.handle.net/2027/mdp.39015008102272
1843 : http://hdl.handle.net/2027/mdp.39015022427622
1847 : http://hdl.handle.net/2027/mdp.39015022427457
1855 : http://hdl.handle.net/2027/mdp.39015066479760
1861 : http://hdl.handle.net/2027/mdp.39015051140021
1871 : http://hdl.handle.net/2027/mdp.39015051140294
1873 : http://hdl.handle.net/2027/mdp.39015051140286
1879 v.1 : http://hdl.handle.net/2027/mdp.39015051140260
1879 v.2 : http://hdl.handle.net/2027/mdp.39015051140443
1881 v.2 : http://hdl.handle.net/2027/mdp.39015051140427
1883 v.1 : http://hdl.handle.net/2027/mdp.39015051140419
1883 v.2 : http://hdl.handle.net/2027/mdp.39015051140401
1885 v.1 : http://hdl.handle.net/2027/mdp.39015051140252
1885 v.2 : http://hdl.handle.net/2027/mdp.39015051140292
1887 v.1 : http://hdl.handle.net/2027/mdp.39015051140245
1889 v.1 : http://hdl.handle.net/2027/mdp.39015051140227
1907 v.1 : http://hdl.handle.net/2027/mdp.39015051143298
1907 v.2 : http://hdl.handle.net/2027/mdp.39015051143280
1909 v.1 : http://hdl.handle.net/2027/mdp.39015051143272
1909 v.2 : http://hdl.handle.net/2027/mdp.39015051143264
1911 v.1 : http://hdl.handle.net/2027/mdp.39015047475267
1911 v.2 : http://hdl.handle.net/2027/mdp.39015051143256
language eng
type text
rights We have determined this item to be in the public domain according to US copyright law through information in the bibliographic record and/or US copyright renewal records. The digital version is available for all educational uses worldwide. Please contact HathiTrust staff at hathitrust-help@umich.edu with any questions about this item.
collection Public domain items worldwide



A REGISTER
of
OFFICERS AND AGENTS,
CIVIL, MILITARY, AND NAVAL,
IN THE SERVICE OF THE
UNITED STATES,
On the THIRTIETH day of SEPTEMBER,
1816;
TOGETHER WITH THE
NAMES, FORCE, AND CONDITION,
OF
ALL THE SHIPS AND VESSELS
BELONGING TO THE UNITED STATES,
AND WHEN AND WHERE BUILT:
U.S.
Prepared at the Department of State,
In pursuance of a Resolution of Congress of the 27th of April, 1816.
CITY OF WASHINGTON,
PRINTED BY JONATHAN ELLIOT.
1816.
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

Multivolume Works

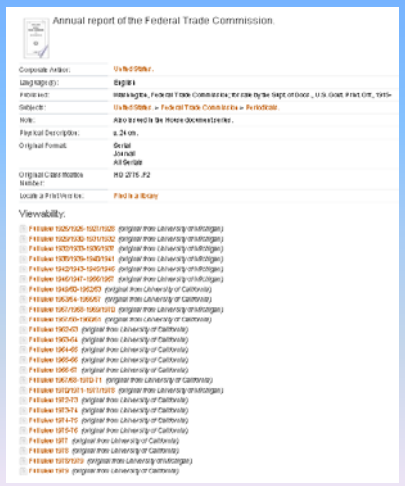
Foreign operations appropriations for 1964 : hearings before a subcommittee of the Committee on Appropriations, House of Representatives, Eighty-seventh Congress, second session.

Add a Tag

creator United States. Congress. House. Committee on Appropriations.
subject Economic assistance, American.
description 4 pts. ;
publisher Washington : U.S. Govt. Print. Off.,
date 1963.
source v.44 (pt.2) : http://hdl.handle.net/2027/uc1.b4291542
v.45 (pt.3-4) : http://hdl.handle.net/2027/uc1.b4291543
language eng
type text
rights We have determined this item to be in the public domain according to US copyright law through information in the bibliographic record and/or US copyright renewal records. The digital version is available for all educational uses worldwide. Please contact HathiTrust staff at hathitrust-help@umich.edu with any questions about this item.
collection Public domain items worldwide

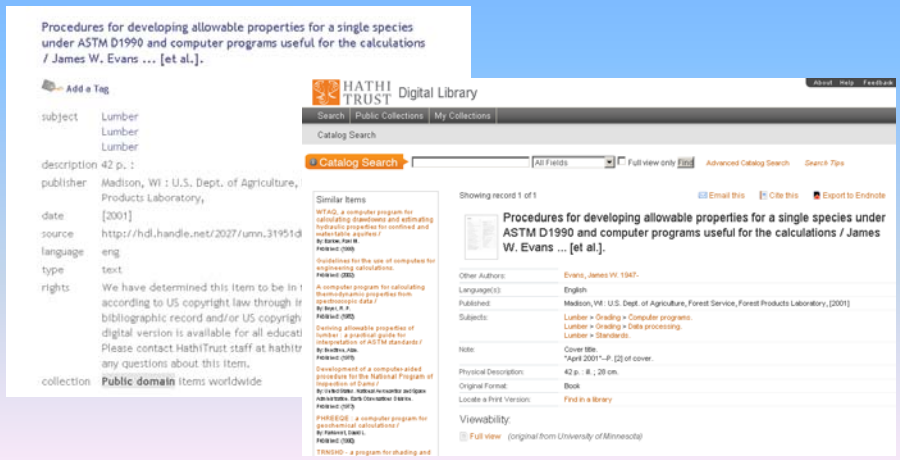
Community Tags
Add a Tag

Duplicate Holdings



U. Of Michigan and U. of California holdings both show in this record

Now, the Bad News:
Records are Stripped Down
"Lumber, Lumber, Lumber"



Harvested Record from our Catalog

<< Back to results

Add to list

Access online

Long hard road : NCO experiences in Afghanistan and Iraq.

Add a Tag

subject	United States. Iraq War, 2003 Iraq War, 2003 Iraq War, 2003 History, 21st Century History, 21st Century History, 21st Century Military Personnel Military Personnel Military Personnel War War War United States. United States. Afghan War, 2001- Iraq War, 2003-
description	195 p. :
publisher	Fort Bliss, Tex. : U.S. Army Sergeants Major Academy,
date	[2007]
source	http://hdl.handle.net/2027/mdp.39015075635493 http://hdl.handle.net/2027/umn.31951d02383016c
language	eng
type	text
rights	We have determined this item to be in the public domain according to US copyright law through information in the bibliographic record and/or US copyright renewal records. The digital version is available for all educational uses worldwide. Please contact HathiTrust staff at hathitrust-help@umich.edu with any questions about this item.
collection	Public domain items worldwide

Community Tags

Add a Tag

Notice the multiple duplications of subject headings

Original Record in Hathi Trust

Long hard road : NCO experiences in Afghanistan and Iraq.

Language(s):	English
Published:	Fort Bliss, Tex. : U.S. Army Sergeants Major Academy, [2007]
Subjects:	United States. > Army. > Noncommissioned Officer Corps. Iraq War, 2003- > Afghanistan > Personal Narratives. Iraq War, 2003- > Iraq > Personal Narratives. Iraq War, 2003- > United States > Personal Narratives. History, 21st Century > Afghanistan > Personal Narratives. History, 21st Century > Iraq > Personal Narratives. History, 21st Century > United States > Personal Narratives. Military Personnel > Afghanistan > Personal Narratives. Military Personnel > Iraq > Personal Narratives. Military Personnel > United States > Personal Narratives. War > Afghanistan > Personal Narratives. War > Iraq > Personal Narratives. War > United States > Personal Narratives. United States. > Army. > Noncommissioned Officer Corps > History. United States. > Army > Non-commissioned officers > History. Afghan War, 2001- > Personal narratives, American. Iraq War, 2003- > Personal narratives, American.
Note:	"October 2007." Shipping list no.: 2008-0072-P.
Physical Description:	195 p. : ill. (some col.), maps ; 22 cm.
Original Format:	Book
Original Classification Number:	DS 371.413 .L86 2007
Locate a Print Version:	Find in a library

Same record, but subject heading subfields are present

Stripped-Out Fields

008 fixed field data

650 subfields other than “a”

500 notes
5xx shipping list info

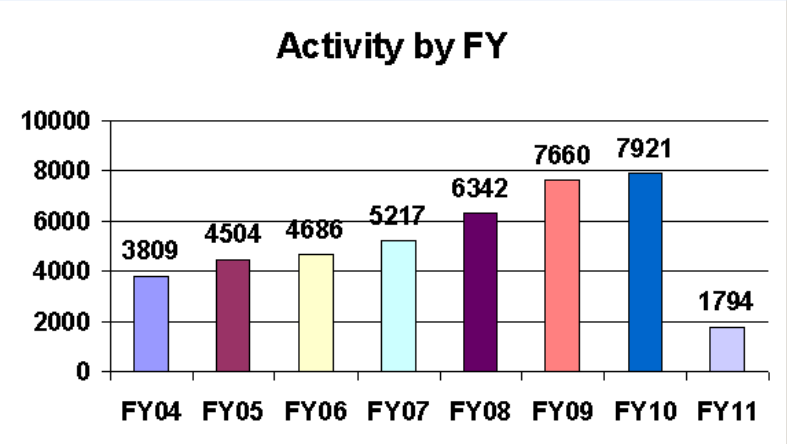
300 subfields after “a”

086 SuDocs number

Long hard road : NCO experiences in Afghanistan and Iraq.

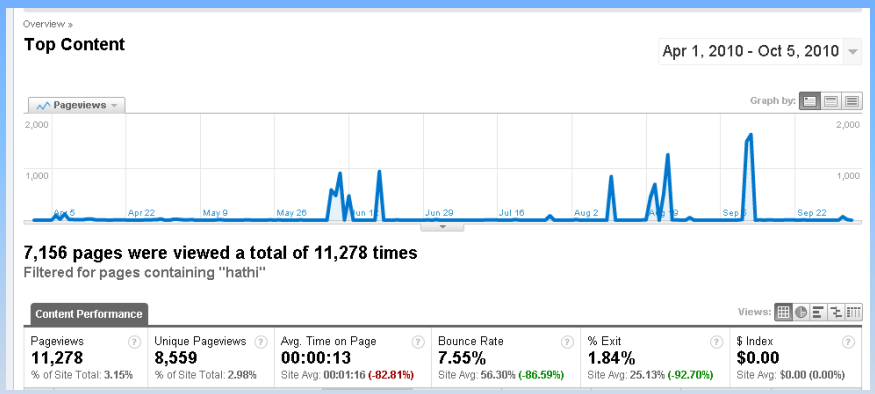
Language(s):	~~English~~
Published:	Fort Bliss, Tex. : U.S. Army Sergeants Major Academy, [2007]
Subjects:	United States. > ~~Army.~~ > ~~Noncommissioned Officer Corps.~~ Iraq War, 2003- > ~~Afghanistan~~ > ~~Personal Narratives.~~ Iraq War, 2003- > ~~Iraq~~ > ~~Personal Narratives.~~ Iraq War, 2003- > ~~United States~~ > ~~Personal Narratives.~~ History, 21st Century > ~~Afghanistan~~ > ~~Personal Narratives.~~ History, 21st Century > ~~Iraq~~ > ~~Personal Narratives.~~ History, 21st Century > ~~United States~~ > ~~Personal Narratives.~~ Military Personnel > ~~Afghanistan~~ > ~~Personal Narratives.~~ Military Personnel > ~~Iraq~~ > ~~Personal Narratives.~~ Military Personnel > ~~United States~~ > ~~Personal Narratives.~~ War > ~~Afghanistan~~ > ~~Personal Narratives.~~ War > ~~Iraq~~ > ~~Personal Narratives.~~ War > ~~United States~~ > ~~Personal Narratives.~~ United States. > ~~Army.~~ > ~~Noncommissioned Officer Corps~~ > ~~History.~~ United States. > ~~Army~~ > ~~Non-commissioned officers~~ > ~~History.~~ Afghan War, 2001- > ~~Personal narratives, American.~~ Iraq War, 2003- > ~~Personal narratives, American.~~
Note:	~~“October 2007.”~~ ~~Shipping list no.: 2008-0072-P.~~
Physical Description:	195 p.,~~: ill. (some col.), maps ; 22 cm.~~
Original Format:	~~Book~~
Original Classification Number:	~~DS 371.413 .L66 2007~~
Locate a Print Version:	~~Find in a library~~

Use Stats for Regular Online Docs



Represents clickthroughs from the catalog record to individual government documents over 7+ years.

Use Stats for Hathi Trust?



Statistics from Google Analytics

- Statistics for all Hathi Trust records accessed, not just documents
- Spikes in usage are docs librarian (my) testing, not real users

Conclusions

- Hathi Trust records are freely available and are easy to harvest
- The harvested records are stripped-down and inadequate, providing too few access points and inadequate descriptions
- The content is superb, contain monographic and serial documents holdings over a span of about 150 years
- Overall the project is worth having in our Encore catalog, especially since our legacy documents will all be in storage during our renovation