

Depository Library Council Session  
Wednesday, April 6, 2011, 8:30-10 a.m.



# **Permanent Public Access to Digital Material**

Or...

# **TRAIL-blazing on the Digital Frontier**



## **Overview of current state of digital preservation**

- David Walls, Preservation Librarian, U.S. Government Printing Office (via Skype)
- Robin L. Haun-Mohamed, Director, Collection Management and Preservation, U.S. Government Printing Office



## **Permanent public access for born-digital web materials:**

### **Internet Archive and the TRAIL (Texas Records and Information Locator) Service**

- Lori Donovan, Partner Specialist, Web Archiving Services, Internet Archive
- Coby Condrey, Coordinator, Texas State Publications Depository Program, Texas State Library and Archives Commission



## **Permanent public access for digitized publications:**

### **Hathi Trust and the TRAIL (Technical Report Archive and Image Library) Project**

- Geoff Swindells, Government Information Department Head, Northwestern University Library
- Mel DeSart, Head, Engineering Library, University of Washington



## **Downstream uses of digital government information services:**

### **Using metadata to connect local users to remote digital collections**

- Chris Brown, Government Documents Librarian, University of Denver Penrose Library



U.S. GOVERNMENT PRINTING OFFICE | KEEPING AMERICA INFORMED

# Permanent Public Access GPO's strategies for managing and preserving government information

## Depository Library Conference San Antonio, Texas

April 6, 2011

Robin Haun-Mohamed

David Walls

# Our Information Landscape:

We have 150 years of paper-based legacy collections in FDLP libraries.

These collections are:

- Often inaccessible through modern search and discovery methods.
- Vulnerable to environmental disasters
- Potentially brittle due to acidic decay

# Our Information Landscape:

- Over 90% of current US Government Information is digital
- Subscriptions to the *Federal Register* in 1994, >20,000. In 2004, < than 3000.  
*Congressional Record* 1994 >20,000. In 2010, <5,000.
- Most of this information is disseminated via the Web
- Web-based information is temporary.
- ““Link Rot” & Legal Resources on the Web: A 2010 Analysis”  
<http://legalinfoarchive.org>

# Our Information Landscape:

- We have a 150 years of collective experience managing and preserving paper-based information.
- We have little more than 15 years of collective experience managing and preserving digital information.

# Mission: Permanent Public Access

- Aging paper collections
- Temporary web content
- A deluge of digital government information

Requires preservation strategies and initiatives

# A Digital Repository:

- **GPO's Federal Digital System, FDsys, a digital content management and repository of US government publications.**
- **Replaces "GPO Access"**
- **Extensible system architecture**
- **Extensive Metadata for searching**
- **Cryptographic hash authentication of all content. PKI on selected content.**
- **Designed around OAIS model to be a "trustworthy" digital repository**



# FDsys Content:

- 1. Deposited Content:**  
Federal Agencies deposit digital publications in FDsys.
- 2. Harvested Content:**  
GPO harvests and archives publications within the scope of the FDLP.
- 3. Digitized Content:**  
Selected FDLP legacy publications  
[www.digitizationguidelines.gov](http://www.digitizationguidelines.gov)

# Partnerships:

- **Identify maps, GIS data, databases, preservation file format copies, and other web publications not capable of being harvested and develop partnerships with agencies creating content.**
- **Reach out to agencies to encourage deposit of agency content with GPO**
- **Maintain networking relationships with agencies focused on sharing best practices**
- **Leverage LC partnership for web harvesting and digitization**
- **GPO is a NARA electronic affiliate.**

# Current and Future Strategies:

- **Work with FDLP community to establish priorities for preservation**
- **Reach out to FDsys user communities and potential user groups**
- **Develop cost models for FDsys sustainability**
- **Disaster Response and Recovery Plan for FDLP**
- **Strengthen our web harvesting effort through partnerships and technology**

# Issues:

- Harvesting interactive databases and web content
- PII (personally identifiable information)
- Authentication
- Need for multiple ingest models and metadata requirements for FDsys

# Permanent Public Access =

- A Trustworthy repository for digital assets
- FDL P partnerships to establish preservation priorities
- Technology tools and strategies for web harvesting
- Planning for disasters and how we will respond to them
- Partnerships with Federal agencies
- Networking and communication

# GPO

Robin Haun-Mohamed

Director, Collection Management and  
Preservation

[rhaunmohamed@gpo.gov](mailto:rhaunmohamed@gpo.gov)

David Walls

Preservation Librarian

[dwalls@gpo.gov](mailto:dwalls@gpo.gov)



# Government Documents 2.0: Archiving Government Information on the Web

April 2011

Lori Donovan  
Partner Specialist  
Internet Archive



# What is the Internet Archive?

- **We are a Digital Library**
- **Mission Statement:** Universal access to human knowledge
- Founded in 1996 by Brewster Kahle in San Francisco, California
- Officially designated a library by the state of California (2007)





# What the Archive Has:

## **Largest public web archive in existence**

- Approximately 4.5 petabytes of data
- Current archive is 200+ billion web pages, culled from 65+ million websites in over 40 languages
- Books and Texts
- Films and Videos
- Audio and the Spoken Word
- Still Images
- Software



# Archive-It

[www.archive-it.org](http://www.archive-it.org)

**First deployed in February 2006**

- Web based application that allows users to create, manage and preserve collections of web content
- Functions include: selection and scoping, harvesting, reports and analysis of captures, cataloging with metadata, full text search
- Archived content includes: text, html, video, audio, images, PDF, online newspapers, social networking and more...
- Includes hosting, access and storage (primary and back-up)
- Archived content available for viewing 24 hours after a crawl has completed



# The Tools Behind Archive-It

## Open Source Technology

primarily developed by Internet Archive, the open source community, and the IIPC

- **Heritrix:** web crawler - crawls and captures pages
- **Wayback Machine:** access tool for rendering and viewing pages. Displays archived web pages--surf the web as it was.
- **NutchWAX:** Open source search engine. Standard full-text search



# www.archive-it.org

Archive-It: Rice University Web Sites: Collection Management

Archiving the internet for future generations  
Collect it, manage it, search it..... ARCHIVE-IT

English

ARCHIVE-IT

Partners

FAQ

About Archive-It

Press Room

Contact Us

Partner Login

Enter Search Terms Here...

Select an Institution

...Or Search All Collections

GO

Advanced Search

## WELCOME TO ARCHIVE-IT

▶ **Attend a live online demo: April 5, 2011 11:30 AM PDT or April 19, 2011 11:30 AM PDT**

Archive-It, a subscription service from the Internet Archive, allows institutions to harvest and preserve collections of digital content and create Digital Archives. Through a user-friendly web interface, Archive-It partners can catalog, manage, and browse their archived collections. Collections are hosted at the Internet Archive data center and are accessible to the public with full-text search. [Learn more](#)

▶ As of today, Archive-It has collected **2,624,717,627** URLs for **1,423** public collections!

### Browse Our Partners

#### State Archives & Libraries

Choose a Partner

#### Colleges & Universities

Choose a Partner

#### Museums & Public Libraries

Choose a Partner

#### National Institutions

Choose a Partner

#### K-12 Program Schools

### Featured Collections



#### Japan Earthquake 2011

This collection depicts the events after the Earthquake and Tsunami in Japan in March 2011. Our partners at Virginia Tech: Crisis, Tragedy, and Recovery Network, Japan's National Diet Library, and Library of Congress have contributed websites for this collection.

### Browse Our Collections

- ▶ **Public Collections A-Z**
- ▶ **Arts & Humanities**
- ▶ **Blogs & Social Media**
- ▶ **Computers & Technology**
- ▶ **Government**
- ▶ **Spontaneous Events**
- ▶ **Politics & Elections**



# Why Archive Web Content?

- Construct an historical record of an institution or government agency's web presence over time
- Gather information and documents from the web to enhance and supplement traditional collections
- Capture "at risk" content that is not available in other formats
- Capture public reactions - tweets, blogs, comments
- Collaborate with other institutions and share research



# Who is Archiving Online Government Information?

- **20+ State Archives & Libraries:** Archive a range of government information, from county and state agencies to state officials and federal representatives.
- **University libraries:** Archive U.S. and international government information, often in regionally or topically based collections
- **Researchers:** Archive information on campaigns and elections, specific topics in state, local or federal government



# North Carolina State Archives & State Library of North Carolina

Purpose: archive state agency websites and publications

- Includes pages in a variety of formats: pdfs, text, images, audio, video and social networking sites
- Archive-It Partner since 2005 (pilot partner)



# North Carolina State Archives & State Library of North Carolina



Office of State Budget and Management  
*Balancing Needs - Improving Government*


[HOME](#) [BUDGET](#) [MANAGEMENT](#) [FACTS AND FIGURES](#) [ECONOMIC ANALYSIS](#) [OSBM LIBRARY](#) [ABOUT OSBM](#)

Home > Facts and Figures > Socioeconomic Data > Census and Survey Data

Saturday, April 02, 2011

## Facts and Figures Menu

- [+ Socioeconomic Data](#)
- [+ Budget and Management Data](#)
- [+ Performance Data](#)
- [■ Data Websites by Topic](#)
- [■ Other Statistical Resources](#)
- [■ About Facts and Figures](#)

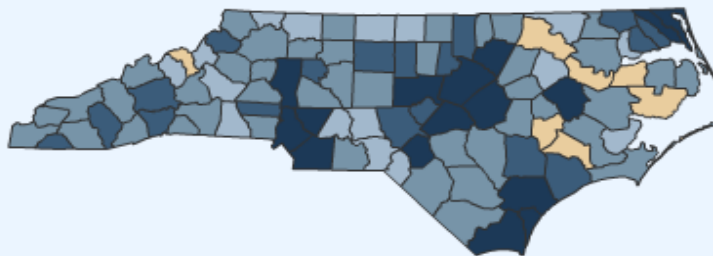
## Census and Survey Data

2010 CENSUS RESULTS

**North Carolina** STATE POPULATION: 9,535,483

POPULATION CHANGE BY COUNTY: 2000-2010

[LOSS](#) [0-5%](#) [5-15%](#) [15-25%](#) [25% +](#)



STATE POPULATION BY RACE  
NORTH CAROLINA: 2010

	PERCENT OF POPULATION	CHANGE 2000-2010
White alone	68.5%	12.5% ↑
Black or African American alone	21.5%	17.9% ↑
American Indian and Alaska Native alone	1.3%	22.7% ↑
Asian alone	2.2%	83.8% ↑
Native Hawaiian and Other Pacific Islander alone	0.1%	65.8% ↑
Some Other Race alone	4.3%	121.8% ↑
Two or More Races	2.2%	99.7% ↑

STATE POPULATION BY HISPANIC OR LATINO ORIGIN  
NORTH CAROLINA: 2010





# Stanford University, Social Sciences Resource Group

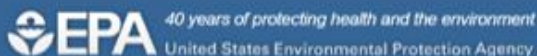
Purpose: Supports research, instruction and collection development within the social sciences

- Collections include California state/local, US Federal and International government information
- Topics range from CA Education to Congressional Research Service Reports, FOIA and US Foreign Policy



# Stanford University, Social Sciences Resource Group

You are viewing an archived web page, collected at the request of Stanford University, Social Sciences Resource Group using [Archive-It](#). This page was captured on 3:20:21 Feb 02, 2011, and is part of the [Fugitive US Agencies](#) collection. The information on this web page may be out of date. See [All versions](#) of this archived page. [Metadata](#)



Mobile | Español | 中文: 繁體版 | 中文: 簡體版 | Tiếng Việt | 한국어

LEARN THE ISSUES | SCIENCE & TECHNOLOGY | LAWS & REGULATIONS | ABOUT EPA | NEWSROOM

Advanced Search

A-Z Index

SEARCH

## Top 50 Green Power Partners

One corporation nearly doubles its green power usage.

- > [Read the press release](#)
- > [Learn about EPA's Green Power Partnership](#)
- > [Learn about the top 50 list](#)

Join Now and Position Your Organization for the Future



### Snow and Ice

Protect yourself and your family in winter storms.

### Popular Topics

- > Air ducts
- > Air pollution
- > Carbon monoxide
- > CFL cleanup
- > Climate change
- > eCycling
- > Fuel

### Info Where You Live



Find info about tribal areas

### Major Announcements

- 1/21** - EPA Grants E15 Fuel Waiver for Model Years 2001 - 2006 Cars and Light Trucks
- 1/19** - EPA and Chrysler to Take Latest Hybrid Technology from Lab to Street
- 1/18** - Administrator Jackson, SBA Administrator Mills Announce Launch of Water Technology Innovation Cluster
- 1/14** - EPA Grants Continue to Protect Beachgoers

- > All announcements | News on Twitter
- > All regional news releases

Administrator  
Lisa P. Jackson



Biography Schedule Twitter

are among Administrator Jackson's priorities. Learn more about all seven priorities for EPA's future.

Jobs at EPA



# Georgetown Law Library

Georgetown University Law Center

 **Georgetown Law Library**

Library Catalog | A - Z Index | Staff Directory

About the Library ▾

Databases ▾

Research Help ▾

For Faculty ▾

For Students ▾

[Home](#) » [Research](#) » Using the Federal Freedom of Information Act (FOIA)

Live Help  
Open

## Table of Contents

- › Introduction
- › Overview of the Federal FOIA
- › Making a FOIA Request
- › Handbooks & Treatises
- › OPEN Government Act of 2007
- › Administrative Materials & Regulations
- › Case Law
- › Advocacy Organizations

## Federal Freedom of Information Act Research Guide

### INTRODUCTION

This research guide discusses how to use the federal **Freedom of Information Act ("FOIA"), 5 U.S.C. § 552**, to obtain records from federal government agencies when those records have not been published in the **Federal Register** or otherwise distributed by the **Government Printing Office**. For example, if you are researching the history of a regulation and the public comments associated with that regulation weren't published in the **Federal Register**, you might be able to find them on the Internet in the agency's **Electronic FOIA Reading Room**. In the alternative, you might be able to obtain them by requesting them under the **Freedom of Information Act**.

This guide is intended only as an introduction to the **Federal Freedom of Information Act**. It provides citations to other guides and handbooks, as well as agency regulations, which may help you in drafting your own FOIA requests. This guide does **not** discuss state freedom of information laws.

Further Information:





Note that all 50 states and the District of Columbia have their own freedom of information statutes. The text of each of those statutes can be found in the appendices to Justin D. Franklin and Robert E. Bouchard, **Guidebook to the Freedom of Information and Privacy Acts** (2d ed. 1986-) [KF5753 .B68 1986].

### OVERVIEW OF THE FEDERAL FREEDOM OF INFORMATION ACT

#### 1. What Information Is Available Under the Freedom of Information Act?

The federal **Freedom of Information Act, 5 U.S.C. § 552**, which was originally enacted in 1966, provides for public access to many federal executive branch agency records. For purposes of the FOIA, the term "agency" . . . includes any executive department, military department, Government corporation, Government controlled corporation, or other establishment in the executive branch of the Government (including the Executive Office of the President), or any independent regulatory agency." **5 U.S.C. § 552(f)(1)**. **FOIA cannot be used to obtain records of Congress or the federal judiciary. It also cannot be used to obtain state records; each state has its own freedom of information statute.**

#### 2. Methods of Information Disclosure Under FOIA

 BOOKMARK     
About This

#### Related Resources

-  [Georgetown Law Guides](#)
-  [Databases](#)



# Overall Approaches and Challenges

- Where is relevant government information located on the web?
- Quality Review: review archived content thoroughly
- Collaborate: compare approaches and results with other Archive-It users
- Document detailed instructions, lessons learned, and best practices for other partners



# Thank you!

[www.archive-it.org](http://www.archive-it.org)

<http://www.facebook.com/ArchiveIt>

Lori Donovan

Partner Specialist, Internet Archive

lori@archive.org

# Texas Records and Information Locator

[www.tsl.state.tx.us/trail](http://www.tsl.state.tx.us/trail)

---

The TRAIL to a Texas Web Archive

Coby Condrey

State Publications Coordinator

Texas State Library and Archives Commission



# Desiderata

---

- Access to Online Content
  - Once a GILS, but Google does it better
- Preservation for Posterity
  - TSLAC mission + external pressure (TLA)
- Navigation / Directory / Finding Aid
  - List of Texas State Agencies



# Scope of the Collection

---

- Capture All Web-Accessible Content
- Three Primary Branches
  - Executive, Legislative & Judicial
- Some Quasi-Governmental Entities
  - State Bar
  - 'Resource' Authorities
- *Exclude* Institutions of Higher Ed. & Interstate Entities





# TRAIL Service Components

---

- Texas State Agency List
- Web Content Archive
  - Full-Text Search
  - Alternate Access via the 'Front Entrance' to Agency by Capture Date



# Coordination of Components

---

- Texas State Agency List
  - (a MySQL dB, rendered to web)
- Archive-It
  - (a web interface)



# Policy Guidelines: Agency List

---

- Check appropriations bill (& statutes)
- Check other reliable resources
  - Tx. Department of Information Resources
  - *Texas State Directory* (commercial publ.)
- Entity has impact statewide (or at least beyond a single county)
- Omit:
  - sub-entities
  - non-state
  - 'gateways'



# Policy Guidelines: Archive-It

---

- Appears on Agency List
- URL ends in .state.tx.us
- URL is out-of-scope, unusual
- Omit (or Disable):
  - sub-pages of another URL on List
  - non-state entities
  - defunct, abolished, absorbed

	A	B	C	D	E
1	Action	Code	Example		
2					
3	Add to TRAIL if entity represents a top-level administrative unit that:				
4	• is a state agency receiving an appropriation from the Legislature	A	See <a href="http://www.lbb.state.tx.us">www.lbb.state.tx.us</a> for appropri		
5	• was created by the Legislature to oversee an issue or to regulate a resource	B	Water authorities (rivers, aquifers, gr		
6	• is a multi-agency cooperative effort	C			
7	• provides a service under statutory authorization but operates as part of a university	D	Texas Cooperative Extension, etc.		
8					
9	Do NOT add to TRAIL if entity:				
10	• is a program or sub-division of another TRAIL entity	S			
11	• is not a state agency, or is a state-federal entity (omit esp. county & city government	N			
12	• is a gateway to a broad class of services (courts, health, e-govt)	G	Texas Courts Online, Texas Online		
13					
14	Add to Archive-It if entity:				
15	• meets criteria for addition to TRAIL	X			
16	• has a URL using the format [www.]~something~.state.tx.us	Y			
17	• has a URL that is out of scope or otherwise unreachable by crawl of primary domain	Z			
18					
19	Do NOT add to Archive-It if entity:				
20	• has a web site that is a sub-page of another TRAIL entity	O	Veterans Land Board		
21	• is not a state agency or state-federal entity (omit esp. county & city government enti	N			
22	• is defunct/abolished or has been absorbed by another TRAIL entity	N	[using "N" twice; it can mean "not" a		
23					
24					
25					



# Scope in Archive-It

---

- Omit calendars and other crawler traps
  - Archive-It provides expressions, help
- Extend crawl duration to maximum
  - Length and URLs captured
- Expand URLs that masquerade as separate domains
- Review after test crawls & regular crawls



# Crawls (Captures)

---

- Automated according to frequency
- Option to start manually




# Results

---

- Summary: start/stop, total size/URLs
- Host: URLs per seed
- Seed Status:  
verification, redirects, blocks
- Seed Source: number of URLs/seed
- File Types: URLs and Bytes per seed
- All downloadable as CSV





Welcome ccondrey [Help](#) [Settings](#) [Archive This!](#) [Templates](#) [Log Out](#)  
[Reset Password](#) | [English](#) [Español](#)

**Texas State Library and Archives Commission - Web Archive** **Partner Home**  
 Partner Since December 2006

**Current Subscription (started Dec 1 2010)**

Documents Crawled:	6,000,013
Subscription Document Budget:	12,000,000
Document Budget Used:	50.00%
Data Archived:	263.9 GB
Data Budget:	1,024 GB
Data Budget Used:	25.77%
Maximum Seed Count	300
Total Active Seeds:	238

**All Subscription Periods**

Documents Crawled:	33,634,526
Data Archived:	2,642.2 GB

**Getting Started**  
[Create New Collection](#)

Active Collections	Last Completed Crawl	Next Scheduled Crawl
<a href="#">TRAIL</a>	<a href="#">February 24, 2011 6:23:43 PM CST</a>	<a href="#">August 19, 2011 5:30:21 PM CDT</a>

**Welcome to Archive-It**

This home view gives you an overview of your account activity including subscription start date and budget.

To create a new collection, click the "create new collection" link from the "collections" drop down menu at the top of the screen.

To manage existing collections, select a collection from the "collections" drop down menu at the top of the screen. You can also get to your active collections by using the links under "active collections" at the bottom of the screen. Information about current or upcoming crawls is available under the "crawls" link at the top of the screen.

- [Learn more about getting started with Archive-It](#)
- [Frequently Asked Questions about Archive-It](#)
- [Glossary of Web Archiving Terms](#)

TRAIL Edit

Collection Management

Created: kevinmarsh Dec 15, 2006 11:10:55 AM Updated: ccondrey Sep 17, 2009 4:16:02 PM

[Activate] [Deactivate] [Mark Dormant]

Collection Management

- Add Seeds
Modify Crawl Scope
Edit Collection Metadata
Edit Document Metadata
View Reports

Seed Management

Table with columns: by seed state, by crawl frequency. Rows include All (277), Active (238), Inactive (39), Twice Daily (0), Daily (0), Weekly (0), Monthly (0), Bi-monthly (0), Quarterly (0), Semiannual (238), Annual (0), One-Time (0).

Crawling Activity

Table with columns: Frequency, Last Completed Crawl, Next Scheduled Crawl. Includes a 'Start Crawl Now' button.

Help

The Collection Management page allows you to make changes and manage your collection.

- Learn more about managing your collection
Learn more about starting One-Time crawls on demand

Frequently Asked Questions

- What is the difference between Active, Inactive and Dormant collections and seeds?
How do I un-schedule crawls?
How can I add more seeds to my collection?
How do I know how large my crawls will be?
How do I export the metadata I've added to my seeds and collections?

Reports  Started on or after  15 Record(s)

	Collection	Frequency	Status	Completed	Documents
<a href="#">View Report</a>	TRAIL	Semiannual	Finished (document limit)	Feb 24, 2011 6:23 PM	6,000,013
<a href="#">View Report</a>	TRAIL	Semiannual	Finished (document limit)	Aug 24, 2010 11:45 PM	6,999,914
<a href="#">View Report</a>	TRAIL	Semiannual	Stopped Unexpectedly	Jul 29, 2010 5:07 PM	5,705,231
<a href="#">View Report</a>	TRAIL	Semiannual	Finished (document limit)	Jan 31, 2010 1:31 PM	5,000,009
<a href="#">View Report</a>	TRAIL	Test	Finished (document limit)	Jan 12, 2010 7:54 AM	5,000,005
<a href="#">View Report</a>	TRAIL	Semiannual	Finished (time limit)	Jul 12, 2009 5:51 PM	4,154,286
<a href="#">View Report</a>	TRAIL	Annual	Finished (time limit)	Jul 6, 2009 6:00 PM	1,238,018
<a href="#">View Report</a>	TRAIL	Test	Finished (time limit)	Jan 19, 2009 6:04 PM	1,121,498
<a href="#">View Report</a>	TRAIL	Semiannual	Finished (time limit)	Jan 11, 2009 3:35 PM	3,172,371
<a href="#">View Report</a>	TRAIL	One-Time	Finished (time limit)	Oct 2, 2008 3:11 PM	247,791
<a href="#">View Report</a>	TRAIL	One-Time	Finished (time limit)	Sep 7, 2008 4:09 PM	354,955
<a href="#">View Report</a>	TRAIL	Semiannual	Finished (time limit)	Aug 26, 2008 5:35 PM	2,664,585
<a href="#">View Report</a>	TRAIL	Test	Finished	Jul 2, 2008 11:26 PM	1,187
<a href="#">View Report</a>	TRAIL	Annual	Stopped Unexpectedly	Jun 5, 2008 10:46 AM	3,222,549
<a href="#">View Report</a>	TRAIL	Annual	Finished (time limit)	Nov 24, 2007 6:53 PM	3,160,828

**Reports**

The Archive-It system will generate post crawl reports and information for every crawl instance. To view reports, click the "view report" link on the left side of the screen.

[Learn more about post crawl reports](#)



[TRAIL](#) **Crawl Report**  
**Semiannual (ID #20110219223022719)** 
 Started: February 19, 2011 4:30:22 PM  
 Completed: February 24, 2011 6:23:43 PM

[<< Back to Reports](#) [Scope-It Crawl Explorer](#)

Summary	Hosts	Seed Status	Seed Source	File Types	PDFs	Videos	QA
---------	-------	-------------	-------------	------------	------	--------	----

**Statistics**

Started	February 19, 2011 4:30:22 PM
Completed	February 24, 2011 6:23:43 PM
Status	Finished (document limit)
Average Doc Rate	13.79 urls/sec
Average KB Rate	1,677.0 KB/s
Total Documents Crawled*	6,000,013
Total Data Crawled	695.8 GB
New Documents Archived	6,000,013
New Data Archived*	263.9 GB

\* This number applies to your Archive-It account budget.

**Help on Reports**

Archive-It provides eight post crawl downloadable reports to assist partners in analyzing and understanding the data that has been archived. The reports can be downloaded as CSV files so they can be easily opened in Excel.

- [Learn more about reports](#)
- [Learn more about the host report](#)
- [Learn more about robots.txt](#)
- [Learn more about the QA Report](#)

<< Back to Reports

Scope-It Crawl Explorer

Summary Hosts Seed Status Seed Source File Types PDFs Videos QA

Hosts [Download](#)

The Hosts Report shows how many URLs were archived from each host, as well as the total amount of data for the collected documents. Other columns in this report provide more information about what was archived.

1. "URLs" refers to the number of documents crawled from each host. Click the number to view the 'URL Report' of exactly what URLs were crawled.
2. "New URLs" refers to documents that changed or were newly discovered since the previous crawl. Click the number to view the 'URL Report' of exactly what URLs were crawled.
3. "Queued" refers to the number of documents discovered but not crawled due to the crawl time limit.
4. "Robots.txt Blocked" refers to documents discovered but not crawled due to a robots.txt exclusion.
5. "Out of Scope" refers to documents that were discovered but not crawled as they were determined to be out of scope. The value in this column may be "n/a" before the report has been generated.

Click the number in each column to view more specific information. This information is available 24 hours after a crawl completes.

View only hosts containing  Filter Clear << First << Previous Next >> Last >> 1 through 100 of 70883

Host	URLs	Data	New URLs	New Data	Queued	Robots.txt Blocked	Out of Scope
info.sos.state.tx.us	<a href="#">561,799</a>	4.9 GB	<a href="#">70,835</a>	674.8 MB	<a href="#">299,986</a>	0	0
www.dshs.state.tx.us	<a href="#">539,474</a>	37.7 GB	<a href="#">477,896</a>	17.9 GB	<a href="#">2,058,268</a>	0	<a href="#">11,618</a>
facilityquality.dads.state.tx.us	<a href="#">478,646</a>	2.8 GB	<a href="#">474,976</a>	2.8 GB	<a href="#">7,267,892</a>	0	<a href="#">2</a>
www.window.state.tx.us	<a href="#">352,027</a>	9.6 GB	<a href="#">130,660</a>	3.6 GB	0	0	0
wit.twc.state.tx.us	<a href="#">336,085</a>	5.2 GB	<a href="#">334,666</a>	5.2 GB	<a href="#">3,829,752</a>	0	0
secure.sos.state.tx.us	<a href="#">229,709</a>	2.5 GB	<a href="#">229,635</a>	2.5 GB	<a href="#">65,132</a>	0	0
www.lcra.org	<a href="#">228,733</a>	8.6 GB	<a href="#">9,466</a>	2.2 GB	<a href="#">6,070</a>	0	0
www.oag.state.tx.us	<a href="#">211,034</a>	21.9 GB	<a href="#">47,394</a>	4.4 GB	0	<a href="#">856</a>	<a href="#">3</a>

downloadable report partners in analyzing understanding the archived. The report downloaded as CSV be easily opened

- [Learn more about](#)
- [Learn more about report](#)
- [Learn more about](#)
- [Learn more about](#)



# Access to Content

---

- TRAIL main page
  - Full-text search of harvested data
  - Simple search tips
- Advanced search
- Texas State Agency List




# TRAIL: Texas Records and Information Locator

Texas State Library and Archives Commission

- Agency Information
- Areas of General Interest
- Services to Librarians
- Services to Gov't Agencies
- Catalogs & Searches
- Our Publications
- News & Events
- 

Catalogs and Searches > TRAIL

- 
- Texas State Agency List
  - Search tips and help
  - About TRAIL
  - Send your comments

## Welcome to TRAIL

### Search TRAIL



*To search the archive, enter your key word(s) and click the "Submit Search" button.*

TRAIL searches and locates information collected in an archive of more than 180 Texas state agency web servers. After clicking the "Submit Search" button, you will leave this site and enter the Archive-It search results page. **To start a new search, please use your browser's "back" button to return to this page.**

The search engine locates pages in the archive that contain *all* your search terms; search results are ranked by relevance. Using specific or unique terms will yield better search results. Avoid searching ubiquitous words like "Texas," "governor," "state," etc. You may use [advanced search features](#) as well.

The search results page provides only one citation per web host address. The exact information you're seeking may be on a different page than the one cited in the results. Click on the **more from** link to see additional pages from a single web site that matched your term(s)



Archiving the internet for future generations  
Collect it, manage it, search it..... ARCHIVE-IT

English

Partners FAQ About Archive-It Press Room Contact Us Partner Login

tourism Select an Institution ...Or Search All Collections GO

[Advanced Search](#)

Collection: Texas State Library and Archives Commission - TRAIL RSS

[Basic Search](#) | [Advanced Search](#)

tourism  [Help with Search](#)

just this collection  all collections

Catalog Metadata Results: 1 result

[What's this?](#)

**Full Text Results**

Results 1 - 20 of about 356,231 for **tourism** found in 0.5 seconds 1 2 3 4 5 6 7 8 9 10 next» last»»

Texas Tourism  
Texas Tourism Sub Navigation News Releases New Attractions & Exhibits Texas Events Calendar Texas Tourism Images & Logos Media Assistance Texas Tourism Contacts Story Ideas Texas Trivia Links Información en Español Media Texas Tourism offers support services and resources for travel and tourism related



## Catalog Metadata Results: 1 result

[Hide](#)

[What's this?](#)

[TravelTex: Texas Travel and Tourism \[Governor's Office, Economic Development and Tourism\]](#)

Collection: [TRAIL](#)

Partner: [Texas State Library and Archives Commission](#)

<http://www.traveltex.com/>

## Full Text Results

Results 1 - 20 of about 356,231 for **tourism** found in 0.5 seconds

1 2 3 4 5 6 7 8 9 10 [next»](#) [last»»](#)

### [Texas Tourism](#)

Texas **Tourism** Sub Navigation News Releases New Attractions & Exhibits Texas Events Calendar Texas **Tourism** Images & Logos Media Assistance Texas **Tourism** Contacts Story Ideas Texas Trivia Links Información en Español Media Texas **Tourism** offers support services and resources for travel and **tourism** related media. In the Media section you will find **tourism** press releases, Texas travel images, Texas travel story ideas, contact information, and more. About Us Contact Us Privacy Policy © 2007 Office of the Governor, Economic Development and **Tourism**. All rights reserved....

text/html - 6.4 KB - crawled once Nov 20, 2007

<http://travel.state.tx.us/media.aspx> - [more results from travel.state.tx.us](#)

### [TCA Tool-Kit: Cultural Tourism](#)

TCA Tool-Kit: Cultural **Tourism** TCAnet | Texas Cultural & Arts Network | Texas Commission on the Arts Cultural **Tourism** In this section, you will find: an overview of cultural **tourism** basics; information on starting a cultural **tourism** program; steps for building a **tourism** message; ideas on how to tell your community's story for **tourism** purposes; guidelines for developing different types of tours; sample letters and agendas for your **tourism** committee; a **tourism** assessment survey; sample **tourism** messages; community surveys; funding and information resources; and more. Cultural **Tourism**: The Basics Building a **Tourism** Message Ten Easy Steps to Activate Cultural **Tourism** When Identifying Your... Resource Section Sample Letter to Cultural **Tourism** Steering Committee (PDF) Sample Agenda for...

text/html - 4.9 KB - crawled once Feb 16, 2007

<http://www.arts.state.tx.us/toolkit/tourism/index.asp> - [more results from arts.state.tx.us](#)



# Texas State Agency List

---

- Locator Record
  - directory information
  - homepage, contact, background
  - legal citations (statute, constitution, rules)
- Agency's Live Website
- Alternate Entry to Archive




# TRAIL: Texas Records and Information Locator

Texas State Library and Archives Commission

[Agency Information](#)
[Areas of General Interest](#)
[Services to Librarians](#)
[Services to Gov't Agencies](#)
[Catalogs & Searches](#)
[Our Publications](#)
[News & Events](#)


[Catalogs and Searches > TRAIL](#)



**Texas State Agency List**

**Search tips and help**









**About TRAIL**

**Send your comments**

### TRAIL List of Texas State Agencies

For each Texas state agency we have included a link to the TRAIL page that provides contact information, addresses, phone numbers and links to enabling legislation for that agency. We have also provided a link directly to the agency's Web site.

[A](#)[B](#)[C](#)[D](#)[E](#)[F](#)[G](#)[H](#)[I](#)[J](#)[K](#)[L](#)[M](#)[N](#)[O](#)[P](#)[Q](#)[R](#)[S](#)[T](#)[U](#)[V](#)[W](#)[X](#)[Y](#)[Z](#)[ALL](#)

Agency	TRAIL Page	Website	Archive
Accountancy, Board of Public			
Adjutant General's Dept.			
Administrative Hearings, Office of			
Affordable Housing Corporation			
Aging and Disability Services, Dept. of			
Agriculture, Dept. of			
AgriLife Extension Service, Texas			
AgriLife Research, Texas			



# TRAIL Web Archive (Texas State Library and Archives Commission)



Enter Web Address:  All  [Compare Archive Pages](#)

Searched for <http://www.governor.state.tx.us/> 14 Results [RSS](#)  
[Look up URL](#) in general Internet Archive web collection [Proxy Mode Help](#)

\* denotes when page was updated

## Search Results for Jan 1, 2005 - Dec 31, 2011

2005	2006	2007	2008	2009	2010	2011
0 pages	0 pages	2 pages	3 pages	3 pages	4 pages	2 pages
		<a href="#">Feb 16, 2007</a> *	<a href="#">Jun 2, 2008</a> *	<a href="#">Jan 6, 2009</a> *	<a href="#">Jan 25, 2010</a> *	<a href="#">Feb 19, 2011</a> *
		<a href="#">Nov 20, 2007</a> *	<a href="#">Aug 21, 2008</a> *	<a href="#">Jun 29, 2009</a> *	<a href="#">Jul 25, 2010</a> *	<a href="#">Feb 19, 2011</a>
			<a href="#">Sep 7, 2008</a> *	<a href="#">Jul 5, 2009</a> *	<a href="#">Aug 19, 2010</a> *	
					<a href="#">Aug 19, 2010</a>	

You are viewing an archived web page, collected at the request of Texas State Library and Archives Commission using [Archive-It](#). This page was captured on 23:35:18 Feb 19, 2011, and is part of the [TRAIL](#) collection. The information on this web page may be out of date. See [All versions](#) of this archived page.



Office of the Governor  
Rick Perry

Contact RSS Feeds

Google Custom Search Search

HOME GOVERNOR PERRY FIRST LADY NEWS INITIATIVES ORGANIZATION CONTACT



### Small Businesses Drive Job Creation

Gov. Perry reiterated his commitment to strengthening Texas' competitive jobs climate, and renewed his call to make the small business tax cut passed last session permanent.

Read More

- 1 2 3 4 5

### RECENT NEWS

CONNECT & SHARE



# Statutory Modification

---

- Authorizing Law
  - Change is sometimes necessary
  - Be careful what you request (you might get it)
  - Be prepared to justify, defend
  - When needed, look at reinterpreting 'status quo'



# Rule Modifications

---

- Laborious process
  - Draft language
  - Secure management approval
  - Propose via publication in *Texas Register*
  - Respond to comments, revise if needed
  - Adopt by announcement in *TexReg*
- Ongoing process
  - Due to frequent changes in web

# Texas Administrative Code

## TITLE 13 CULTURAL RESOURCES

### PART 1 TEXAS STATE LIBRARY AND ARCHIVES COMMISSION

#### CHAPTER 3 STATE PUBLICATIONS DEPOSITORY PROGRAM

## Rules

- [§3.1](#) Definitions
- [§3.2](#) Standard Requirements for State Publications in All Formats
- [§3.3](#) Standard Deposit and Reporting Requirements for State Publications in Physical Formats
- [§3.4](#) Standard Deposit and Reporting Requirements for State Publications that are Internet Publications
- [§3.5](#) TRAIL Grant Database
- [§3.6](#) Standard Exemptions for State Publications in All Formats
- [§3.7](#) Special Exemptions
- [§3.8](#) State Publications Contact Person
- [§3.9](#) Designation and Termination of Depository Library Status for State Publications in Physical Formats
- [§3.10](#) Minimum Standards for Designated Depository Libraries for State Publications in Physical Formats
- [§3.11](#) Designation and Termination of Depository Library Status for State Publications Published as Internet





# Future Development Ideas

---

- Limit Crawls to Document Level
  - Highly Increases Seed Management
  - Increases Need for Agency Reporting
- Connect Catalog Records to Archive
  - Raise Awareness of Archive & Content
- Cataloging Online-Only Resources



# Challenges (Most Pressing)

---

- Incorporation of pre-Archive-It data
- Perpetually Moving Targets
  - Social Media
  - Proliferation of 'Disguised' Domains
  - Technology Advances Faster than Govt.
  - Transition to .texas.gov Domain Names
- Incomplete Harvests
- Look & Feel vs. Textual Content
- Resources: Funding & Staffing



# Thank You

---

- Questions, Comments & Suggestions Welcome ([trail@tsl.state.tx.us](mailto:trail@tsl.state.tx.us))
- Visit TRAIL at [www.tsl.state.tx.us/trail](http://www.tsl.state.tx.us/trail)
- Contact:
  - Coby Condrey
  - 512-463-5434
  - [ccondrey@tsl.state.tx.us](mailto:ccondrey@tsl.state.tx.us)

# HathiTrust Digital Library

Geoffrey Swindells  
Federal Depository Library Council Meeting  
06 April 2011



NORTHWESTERN  
UNIVERSITY

# HathiTrust

- Began in 2008 as a collaboration of the universities of the Committee on Institutional Cooperation (CIC), the University of California system, and the University of Virginia
- Currently comprises more than fifty partner libraries and consortia
- The mission of HathiTrust is to contribute to the common good by collecting, organizing, preserving, communicating, and sharing the record of human knowledge



# HathiTrust Digital Library

- Digital preservation repository
- Access platform
- Provides long-term preservation and access services for public domain and in copyright content from a variety of sources, including Google, the Internet Archive, Microsoft, and in-house partner institution initiatives



# Bulk ingest of federal publications

- Google Books Project
- CIC Federal Documents Digitization Project
- Technical Report Archive & Image Library (TRAIL)



# Types of materials

- Digitized books, journals, and manuscripts
- Future support for audio/visual
- Future support for born digital





## Standards-based

- University of Michigan digitization specifications (e.g. 600dpi ITU TIFF G4/300dpi JPEG2000)
- MARC/PREMIS/METS
- OASIS/TRAC

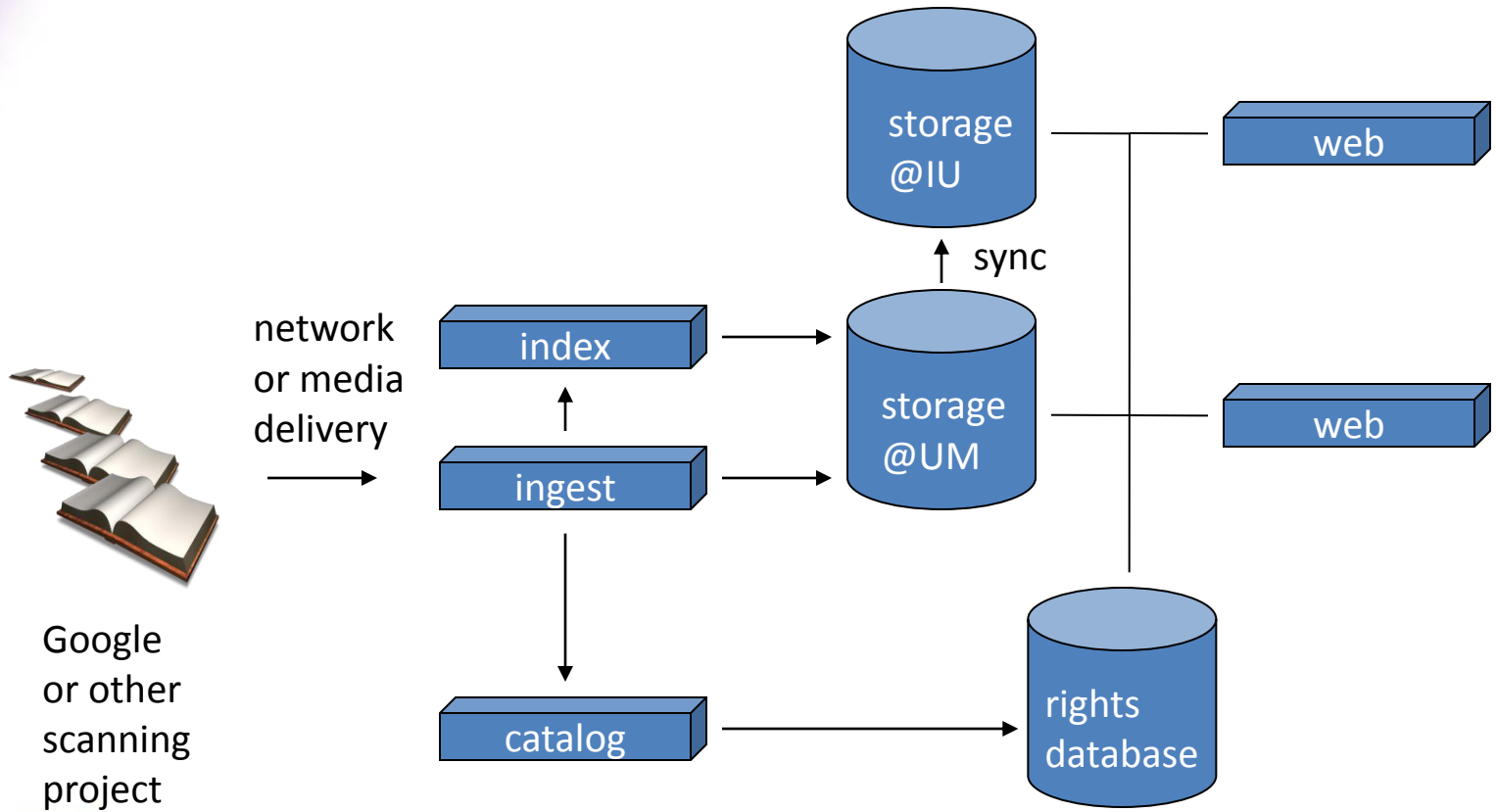


# TRUST = TRAC

- Trustworthy Repositories Audit and Certification
- Hathi Trust certified by CRL on March 30, 2011
- Objective measurement of trustworthiness
  - Organizational infrastructure
  - Digital object management
  - Technologies, technical infrastructure, and security
- Periodic review

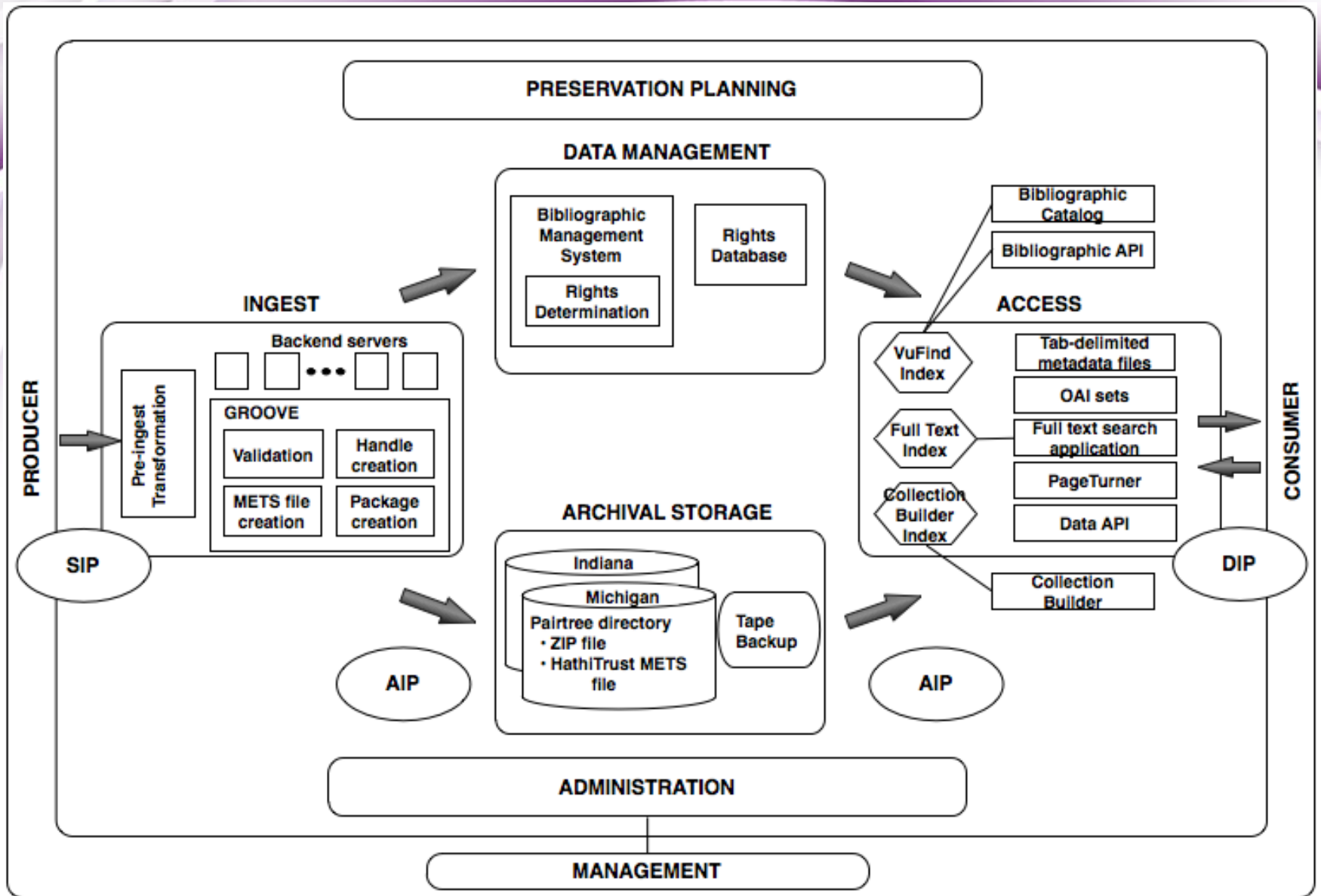


# Material and Data Flow



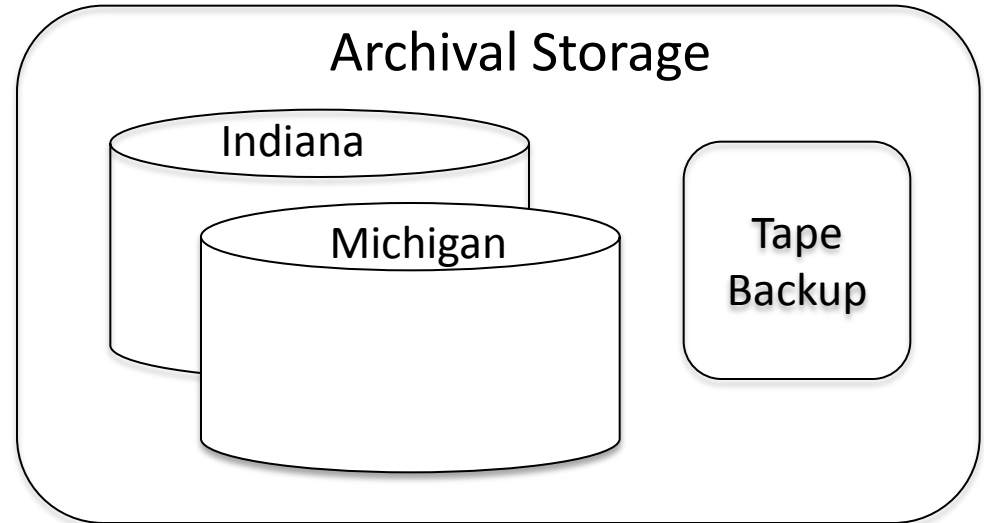
Google  
or other  
scanning  
project





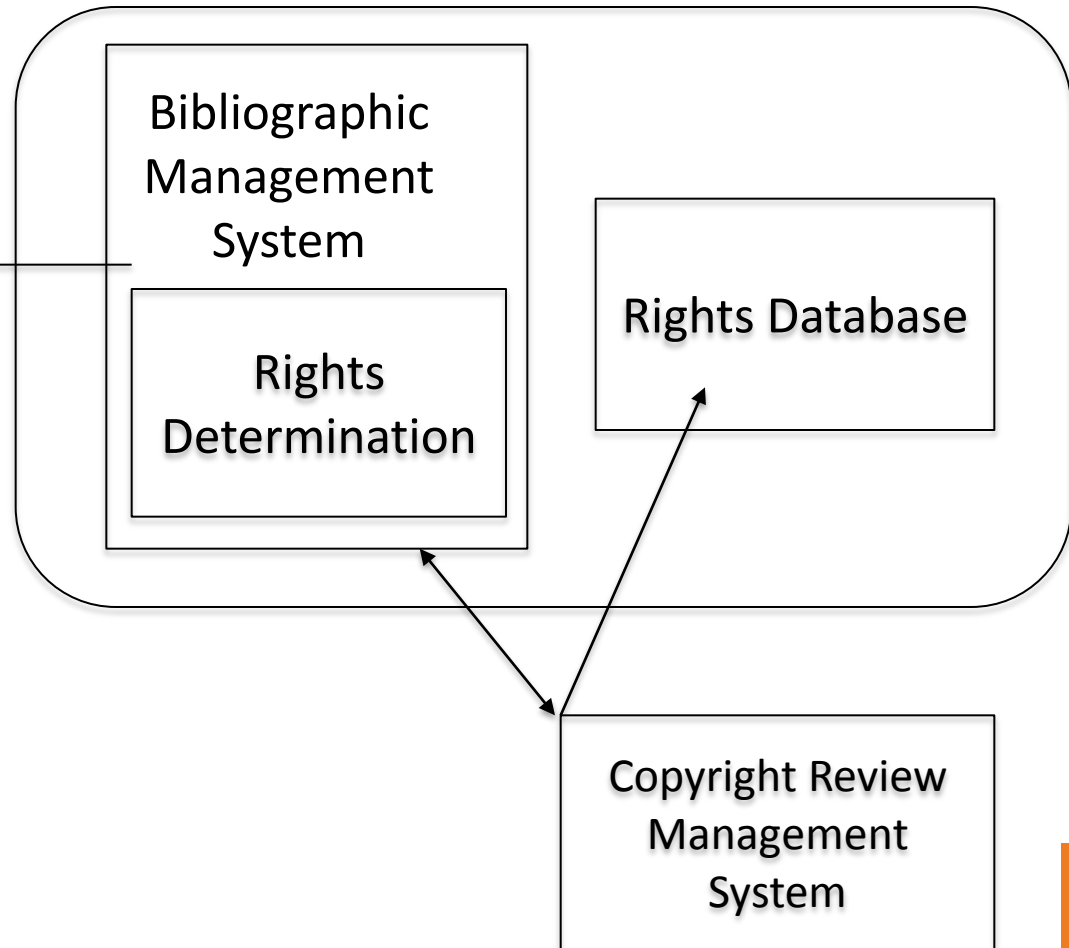
# Media & Architecture

- Isilon Systems
- Load balancing and failover
- Ingest at Michigan, replicated to Indiana
- Replacement on 3-4 year cycle



# Data Management

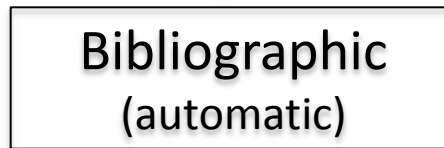
- Inventory
- Loading and updating records
- Duplicate detection and collation
- Solr indexes behind VuFind catalog
- Source of information for Access services
- Rights determination (automated and support for manual review)



# Rights Database

---

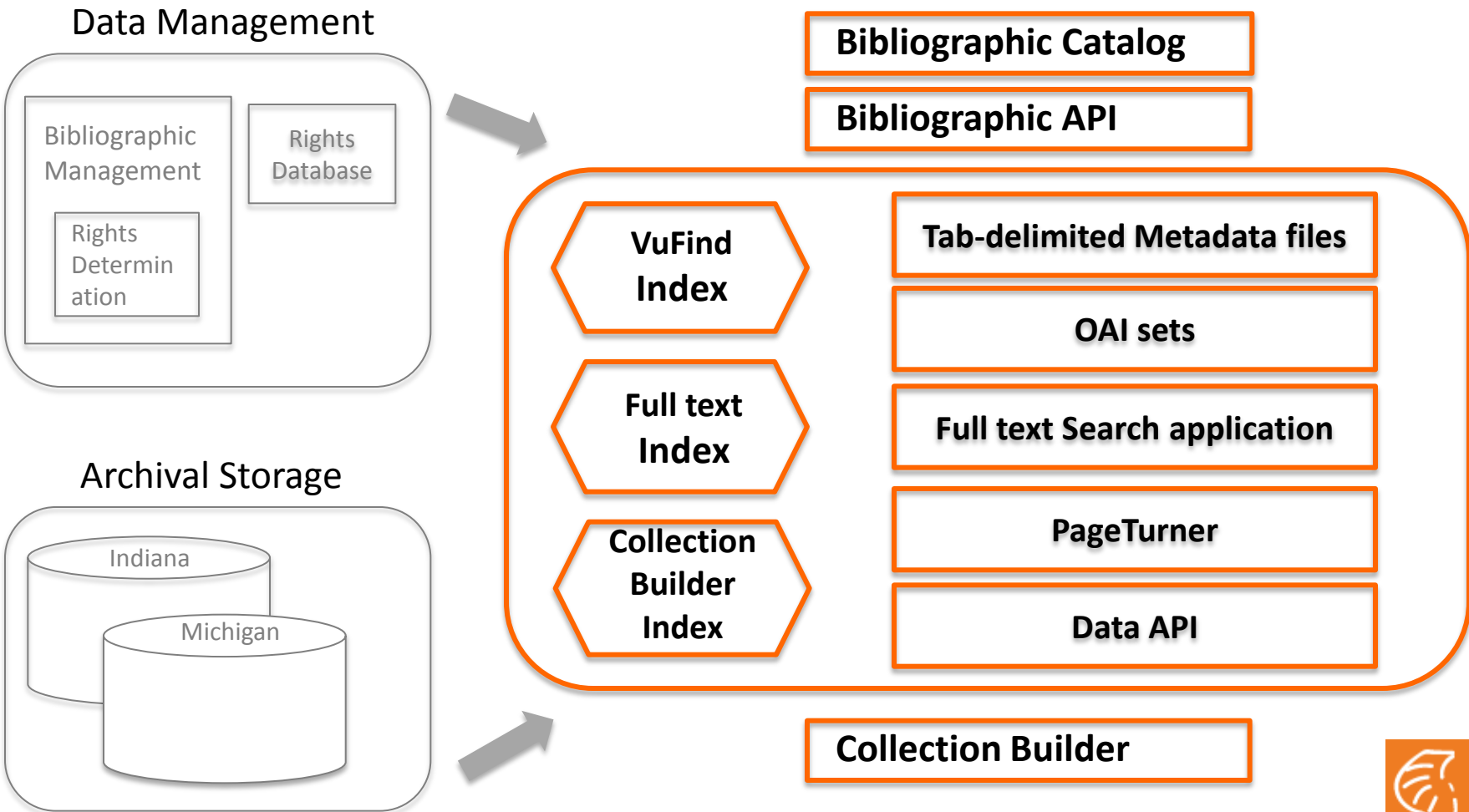
- System of precedence



- 9 attributes
- 11 reason codes



# Access





# **Blazing a TRAIL: Digitizing and Preserving Legacy U.S. Government Technical Reports**

**Mel DeSart**  
**Head, Engineering Library**  
**University of Washington**

**Depository Library Council Meeting**  
**06 April 2011 – San Antonio, TX**

# Charge

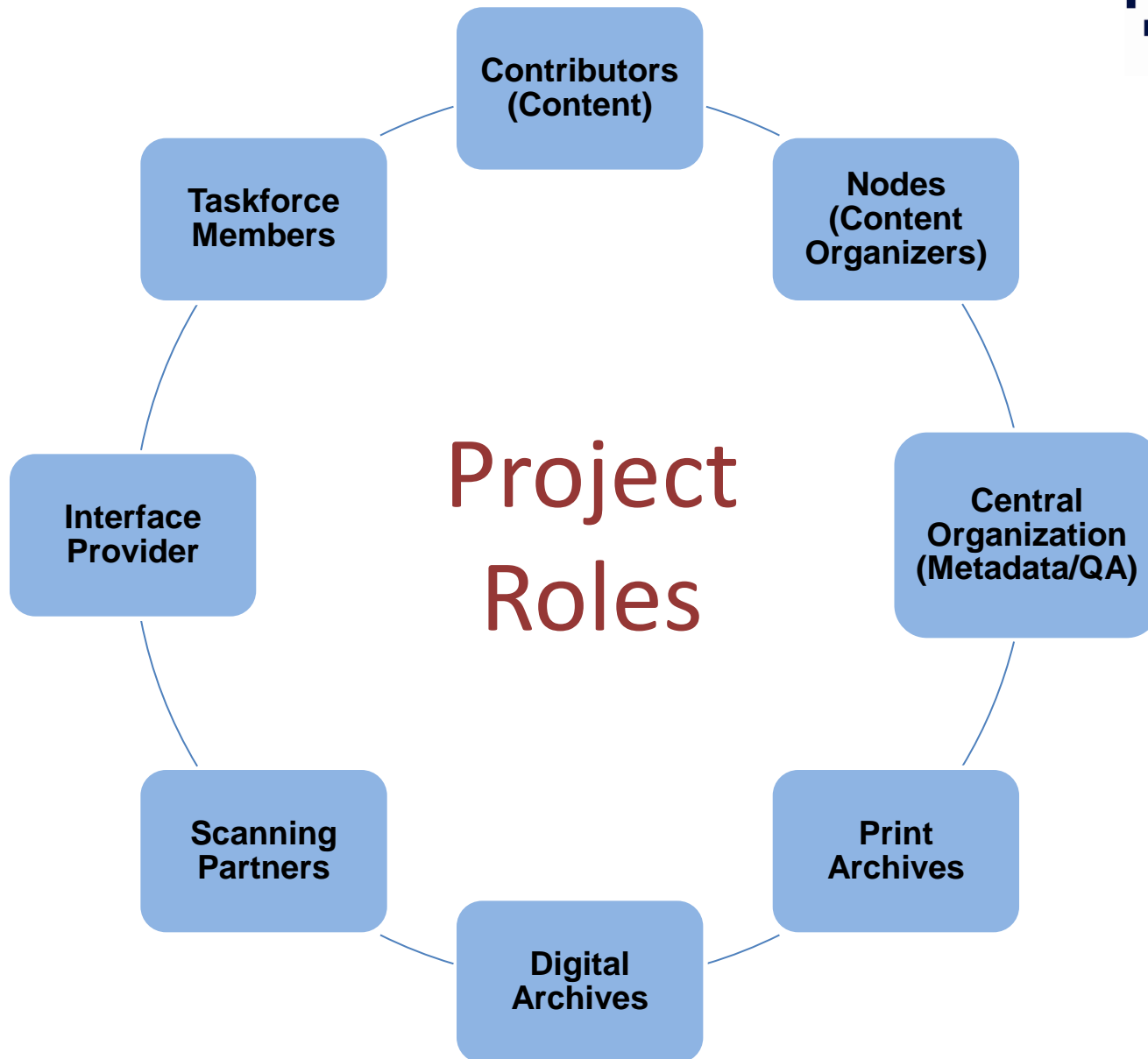
Develop a collaborative project with the Center for Research Libraries to identify, digitize, archive, and provide persistent and unrestricted access to federal technical reports issued prior to 1975.

# History 1

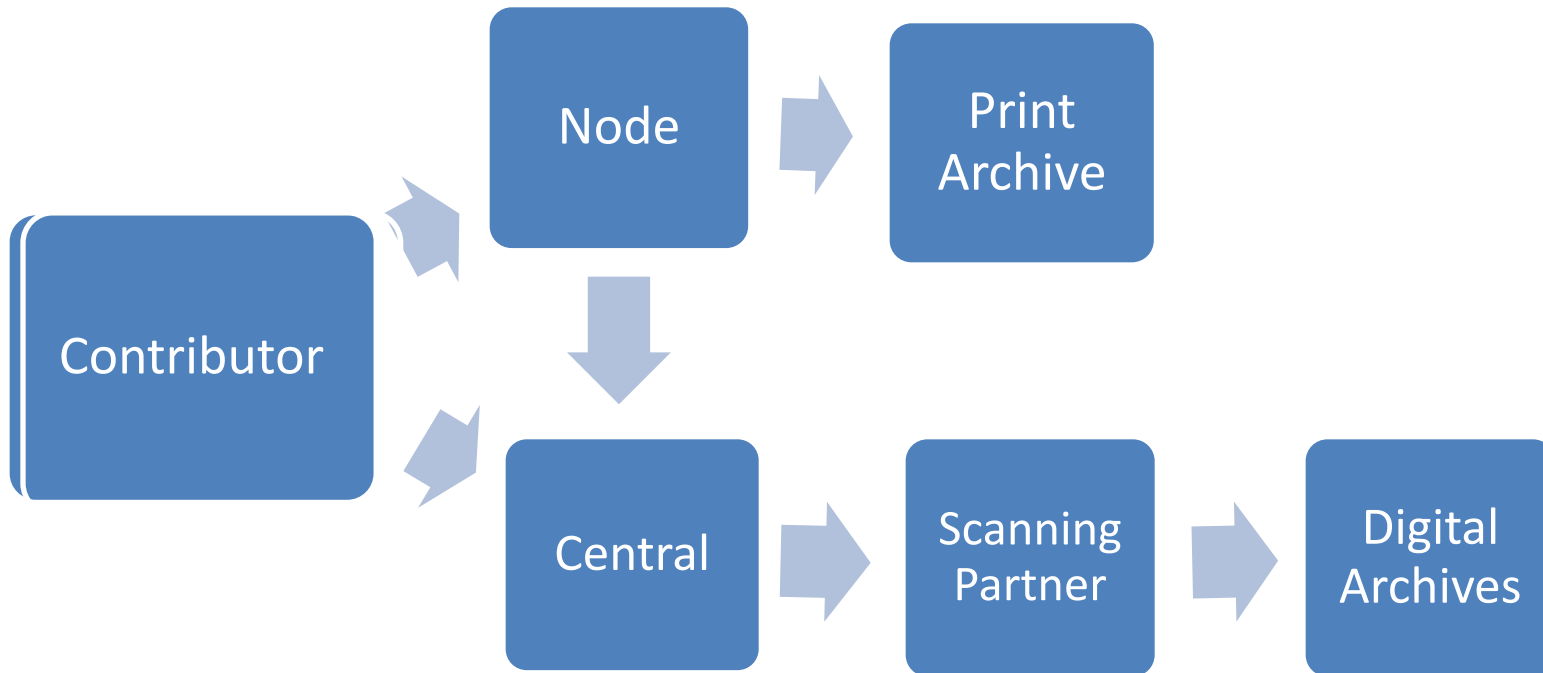
- Idea began with Maliaca Oxnam at the University of Arizona – grew out of frustration of engineering librarians at how difficult it was, especially for end users, to find individual technical reports
- Initial support from the Center for Research Libraries (CRL) and the University of Arizona
- Accepted as a Greater Western Library Alliance (GWLA) digital initiative
- TRAIL project formed - 2006
- Realized we needed greater government document librarian input – added government document librarians to the group
- Pilot site created at the University of Hawaii -- Manoa
- Funding an issue – almost all scanning initially had to be paid for – also cataloging costs

# History 2

- Enter the University of Michigan, Google, and the Hathi Trust – 85% of scanning now done at no cost to the project
- 15% of scanning still to vendor - that content now stored at the University of North Texas
- Early 2010, TRAIL moves entirely to CRL as newest Global Resource Network – now a member organization
- New set of funding issues – with library budgets depressed, fewer schools than expected became members – less funding currently than when a GWLA initiative
- 2010 – domain name acquired and new user interface developed
- Late 2010, new user interface debuts – [www.technicalreports.org](http://www.technicalreports.org)



# Process



# Accomplishments

- Completed an inventory of all defunct federal agencies and report series that were issued by those agencies
- 17,575 page-views of the pilot site in the first year (~ 200 reports available)
- Over 23,000 items cataloged & scanned; over 2.2 M pages scanned
- Establishing collection sets of MARC records in OCLC
- Digital archives at Michigan / Hathi Trust and at the University of North Texas
- Print Archive (Oklahoma State)
- Inventory control for all documents
- Award winning project



COLLABORATIONS

TRAIL

About GRN

GRN Forums

GRN Reports

Global Resources  
Newsletter

AFRINUL

CIFNAL

DSAL

GNARP

HRADP

ICON

LARRP

TRAIL

+ About TRAIL

Contact TRAIL


+ Current Activities

Member List

+ Working Groups



TECHNICAL REPORT  
ARCHIVE & IMAGE LIBRARY

 [Invitation to join TRAIL](#)

The **Technical Report Archive & Image Library (TRAIL)** is an initiative led by the University of Arizona in collaboration with CRL and other interested agencies to identify, digitize, archive, and provide access to federal technical reports issued prior to 1975. The TRAIL project began under the auspices of the [Greater Western Library Alliance](#).

Technical reports communicate research progress in technology and science; they deliver information for technical development to industry and research institutions contributing to the continued growth of science and technology. These highly detailed reports contain valuable information serving specialized audiences of researchers. While availability to more recent (1994–current) technical report literature has greatly improved with Internet access, legacy technical report documents remain elusive to researchers. Most large research libraries across the country have sizeable collections of federally funded technical research reports—frequently a million or more ranging from several pages to several hundred pages.

An example of some report series digitized includes:



Catalog Search Options »

QUICK LINKS ▲

RECENT TRAIL NEWS

- Dec 10 2010  
TRAIL Search Interface Now Available
- Dec 1 2010  
TRAIL Project Information Session at ALA Midwinter 2011
- May 24 2010  
TRAIL Poster at ALA 2010 Annual
- May 24 2010  
TRAIL Receives 2010 Documents to the People Award

[More News »](#)



Search U.S. government technical reports issued primarily prior to 1975 and digitized by the TRAIL Working Groups.

### KEYWORD SEARCH

Enter your search term(s):

[\[advanced search\]](#)

Examples: Bureau of Mines; Information Circular; Smith, John

[About TRAIL](#) | [FAQ](#) | [Join TRAIL](#) | [Contact Us](#)

Search developed and maintained by the University of Washington Libraries  
Copyright 2010

Title  contains  AND

Title  contains  AND

Title  contains  AND

Title  contains

[\[basic search\]](#)

[About TRAIL](#) | [FAQ](#) | [Join TRAIL](#) | [Contact Us](#)

Search developed and maintained by the University of Washington Libraries  
Copyright 2010


## Synthetic fuel from coal for supersonic aircraft


<b>Author:</b>	Schlesinger, Martin D.
<b>Additional Authors:</b>	Hiteshue, Raymond W.
<b>Year:</b>	1961
<b>Document Type:</b>	BMRA
<b>Issuing Agency:</b>	U.S. Dept. of the Interior, Bureau of Mines,
<b>SUDOC:</b>	I 28.23:5902
<b>Series:</b>	Report of investigations / United States Department of the Interior, Bureau of Mines ;
<b>Report Number:</b>	5902
<b>Subjects:</b>	Coal Jet planes--Fuel
<b>Link:</b>	<a href="http://hdl.handle.net/2027/mdp.39015078529917">http://hdl.handle.net/2027/mdp.39015078529917</a>

[< New search](#)

 [Catalog record](#)

 [Find in a library](#)

 [Buy a copy](#)

 [Permanent Link](#)

<http://hdl.handle.net/202>

**Login** to make your personal collections permanent

Add to your collection:

Select Collection



go to #

size

rotate  

 [image](#)

 [text](#)

 [1-page PDF](#)

 [Partners login](#)

# SYNTHETIC FUEL FROM COAL FOR SUPERSONIC AIRCRAFT

By M. D. Schlesinger and R. W. Hiteshue

\* \* \* \* \* report of investigations 5902

Login to make your personal collections permanent

Add to your collection:

Select Collection

Add



go to #

size 100%

rotate

image

text

PDF 1-page PDF

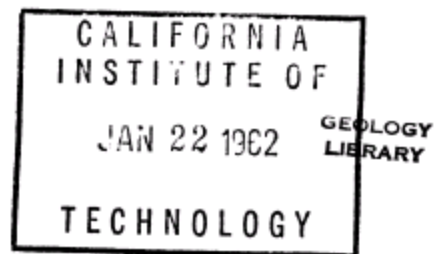
PDF Partners login for full PDF

Contents:

Front Cover	
Title Page	
Table of Contents	
Section 1	1
Section 2	7
Section 3	13

SYNTHETIC FUEL FROM COAL FOR SUPERSONIC AIRCRAFT

By M. D. Schlesinger and R. W. Hiteshue



# Useful Info

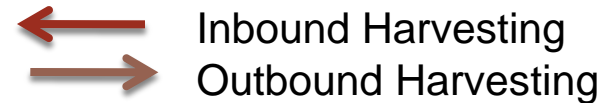
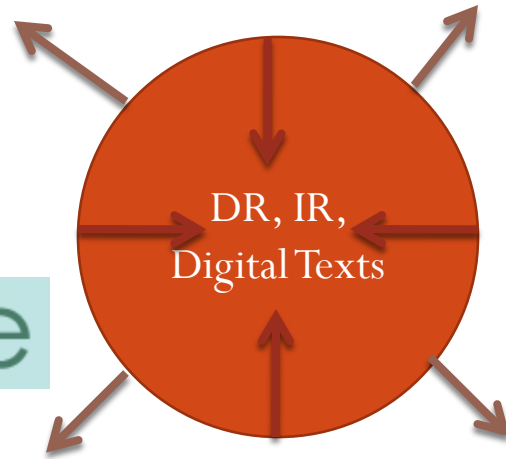
- ANY institution can join TRAIL – don't have to be a member of CRL
- Institutions that are project members have a vote on content, project direction and priorities, can serve on project committees and task forces, etc.
- Non-members can still contribute content, either for digitization or for the print archive, if needed
- TRAIL project site - [www.crl.edu/grn/trail](http://www.crl.edu/grn/trail)
- User interface – [www.technicalreports.org](http://www.technicalreports.org)

Thank you! Any questions???

# Downstream Uses of Digital Government Information Services: Using Metadata to Connect Local Users to Remote Digital Collections

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

**DLC San Antonio, TX**  
**April 6, 2011**



This presentation will show how Encore harvesting can be used to mitigate a space problem in a library, substituting online access for the need for physical access to the collection. The government documents collection will be the primary focus.

Note: Encore is the next-generation catalog interface produced by Innovative Interfaces, Inc.



# Collection Downsizing?

Cloud-sourcing Research Collections: Managing Print in the Mass-digitized Library Environment

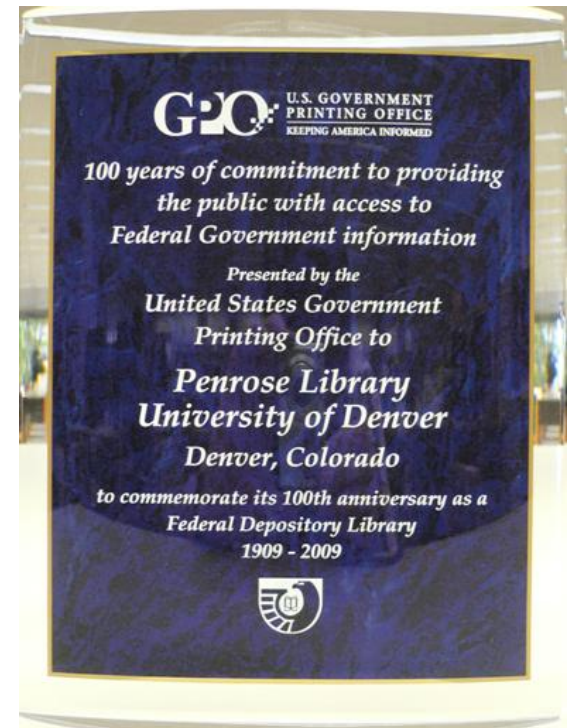
---

available, further reducing the need for print inventory. Among titles classified as government documents in the HathiTrust Digital Library, nearly 80% are designated as public domain content. *One can easily imagine that many academic libraries will choose to downsize local document collections in favor of online versions; for such institutions, the Hathi preservation services could provide a compelling and cost-effective alternative to local print archiving.* Even those libraries that choose to maintain their status as selective depositories could achieve significant cost savings by transferring physical copies of the government publications replicated in the HathiTrust Digital Library to high-density storage facilities.

Malpas, Constance. 2011. *Cloud-sourcing Research Collections: Managing Print in the Mass-digitized Library Environment*. Dublin, Ohio: OCLC Research. <http://www.oclc.org/research/publications/library/2011/2011-01.pdf>.

# 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
- Currently 100% of our paper documents are in storage
- We are remodeling our library. Under the remodeling plan, all docs will remain in remote storage.



# Partial Solution: Using Encore for Outbound Harvesting

- All documents off-site
- Our users are accustomed to using electronic documents
- Need to divert attention away from physical collection holdings
- Encore harvesting of Hathi Trust can do this

# 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](http://www.hathitrust.org/rights_database)

# 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]

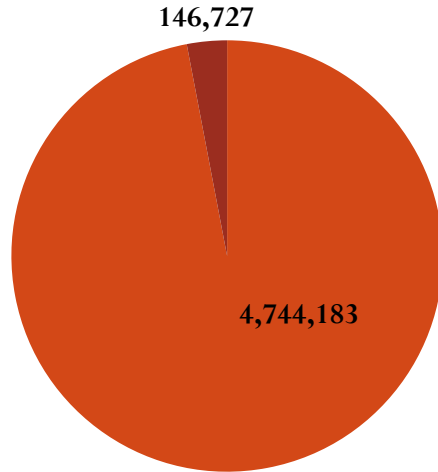
# Harvesting Hathi Docs: The Stats

Date Range	Hathi Totals	Hathi All Pub Domain			Docs Sampling	
		pdus + pd	Hathi pdus	DU pd Harvest		
2000-2009	505,682	14,140	726	13,369	13,340	99.78%
1990-1999	709,214	29,163	880	28,164	26,662	94.67%
1980-1989	723,657	33,753	1,204	32,321	31,370	97.06%
1970-1979	631,110	28,633	2,046	26,189	25,607	97.78%
1960-1969	546,914	21,244	1,987	18,991	7,668	40.38%
1950-1959	281,615	20,861	863	19,893	3,888	19.54%
1940-1949	184,755	17,096	600	16,253	3,771	23.21%
1930-1939	175,103	16,237	654	15,317	2,600	16.97%
1920-1929	175,226	66,563	27,108	28,854	1,529	5.30%
1910-1919	175,148	169,923	75,955	61,230	4,124	6.73%
1900-1909	179,018	153,284	70,900	47,999	2,265	4.72%
1890-1899	112,295	110,605	50,502	34,742	596	1.72%
1880-1889	83,950	82,809	38,928	23,855	699	2.93%
1870-1879	58,624	57,826	27,202	17,751	319	1.80%
1860-1869	50,907	50,337	2,273	45,790	248	0.54%
	<b>4,593,218</b>	<b>872,474</b>	<b>301,828</b>	<b>430,718</b>	<b>124,686</b>	<b>28.95%</b>

Statistics as of mid-March, 2011

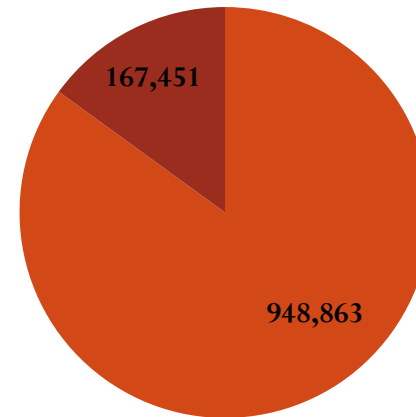
The Docs Sampling columns show the estimated numbers of docs per year and the estimated percentage of docs per year from the Harvest

# Malpas: Docs about 3% of Hathi Total and 15% of Public Domain

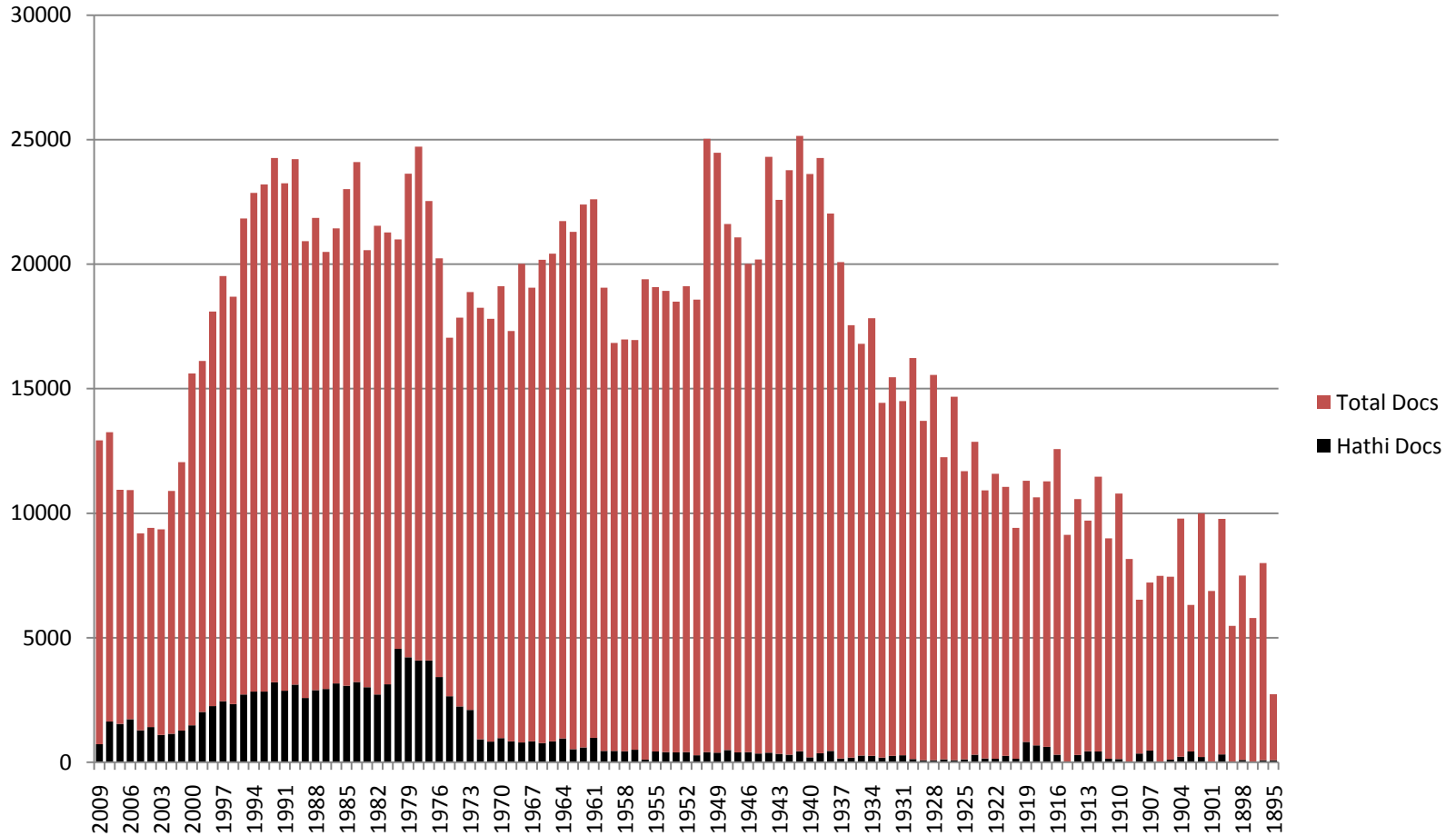


GovDocs: 3% overall

GovDocs: 15% of Public Domain



# Hathi Docs Usage in Proportion to Docs Distribution



Sources: 1895-1976 data: Monthly Catalog, 1895-1976 (ProQuest); 1976 onward data: CGP



# % Docs in HathiTrust (est.)

Year	Hathi Docs	Docs Total	% Hathi Docs	Year	Hathi Docs	Docs Total	% Hathi Docs	Year	Hathi Docs	Docs Total	% Hathi Docs
2009	724	12198	5.9%	1969	847	16471	5.1%	1929	80	13630	0.6%
2008	1639	11609	14.1%	1968	799	19209	4.2%	1928	82	15469	0.5%
2007	1546	9389	16.5%	1967	837	18219	4.6%	1927	118	12136	1.0%
2006	1732	9196	18.8%	1966	771	19400	4.0%	1926	77	14603	0.5%
2005	1281	7911	16.2%	1965	841	19577	4.3%	1925	116	11572	1.0%
2004	1409	7996	17.6%	1964	963	20760	4.6%	1924	308	12563	2.5%
2003	1103	8252	13.4%	1963	525	20776	2.5%	1923	155	10758	1.4%
2002	1146	9753	11.7%	1962	591	21804	2.7%	1922	146	11439	1.3%
2001	1269	10777	11.8%	1961	984	21621	4.6%	1921	251	10808	2.3%
2000	1483	14132	10.5%	1960	452	18602	2.4%	1920	142	9263	1.5%
1999	2019	14093	14.3%	1959	459	16382	2.8%	1919	816	10488	7.8%
1998	2252	15846	14.2%	1958	448	16530	2.7%	1918	673	9964	6.8%
1997	2455	17071	14.4%	1957	502	16451	3.1%	1917	623	10654	5.8%
1996	2330	16361	14.2%	1956	102	19289	0.5%	1916	306	12271	2.5%
1995	2727	19107	14.3%	1955	428	18644	2.3%	1915	0	9127	0.0%
1994	2833	20030	14.1%	1954	405	18524	2.2%	1914	292	10270	2.8%
1993	2836	20367	13.9%	1953	394	18099	2.2%	1913	439	9258	4.7%
1992	3206	21056	15.2%	1952	413	18701	2.2%	1912	428	11033	3.9%
1991	2871	20372	14.1%	1951	285	18295	1.6%	1911	136	8855	1.5%
1990	3122	21092	14.8%	1950	406	24632	1.6%	1910	135	10654	1.3%
1989	2577	18349	14.0%	1949	389	24084	1.6%	1909	0	8167	0.0%
1988	2892	18964	15.3%	1948	486	21130	2.3%	1908	352	6174	5.7%
1987	2941	17546	16.8%	1947	402	20675	1.9%	1907	467	6745	6.9%
1986	3164	18267	17.3%	1946	412	19597	2.1%	1906	0	7485	0.0%
1985	3075	19942	15.4%	1945	350	19834	1.8%	1905	111	7341	1.5%
1984	3222	20878	15.4%	1944	383	23925	1.6%	1904	221	9566	2.3%
1983	3009	17550	17.1%	1943	336	22241	1.5%	1903	432	5890	7.3%
1982	2718	18828	14.4%	1942	305	23464	1.3%	1902	214	9784	2.2%
1981	3133	18143	17.3%	1941	446	24702	1.8%	1901	0	6879	0.0%
1980	4552	16443	27.7%	1940	198	23427	0.8%	1900	310	9464	3.3%
1979	4212	19415	21.7%	1939	363	23895	1.5%	1899	0	5475	0.0%
1978	4090	20624	19.8%	1938	453	21573	2.1%	1898	92	7399	1.2%
1977	4090	18439	22.2%	1937	157	19922	0.8%	1897	0	5793	0.0%
1976	3424	16809	20.4%	1936	193	17353	1.1%	1896	87	7907	1.1%
1975	2655	14387	18.5%	1935	266	16533	1.6%	1895	85	2643	3.2%
1974	2247	15601	14.4%	1934	261	17568	1.5%				
1973	2104	16774	12.5%	1933	183	14254	1.3%				
1972	917	17333	5.3%	1932	262	15200	1.7%				
1971	824	16986	4.8%	1931	283	14214	2.0%				
1970	970	18142	5.3%	1930	128	16098	0.8%				

# Hathi Docs Links Provide Access to Docs in Storage

**Title** Current population reports. Series P-70, Household economic studies.  
**Publ Info** Washington, D.C. : U.S. Dept. of Commerce, Bureau of the Census : For sale by the Supt. of Docs., U.S. G.P.O., 1984-

The links below are for electronic versions of this publication:  
[Access online via Hathi Trust](#)  
[Access online via Census Bureau](#)

LOCATION	CALL #	STATUS
Internet	<a href="#">C 3.186:P-70/2/</a>	ONLINE

**Description** v. ; 28 cm.  
**Frequency** Quarterly  
**Pub History** No. 1-  
**Note(S)** "Average monthly data from the Survey of  
 Latest issue consulted: no. 70 (published in  
 Continued by an online version.

**Title** Bureau of Standards journal of research / Department of Commerce, Bureau of Standards.  
**Publ Info** Washington : The Bureau : For sale by the Supt. of Docs., U.S. G.P.O., 1928-

The links below are for electronic versions of this publication:  
[Access online version via Hathi Trust](#)

LOCATION	CALL #	STATUS
<i>Location</i>	<b>PASCAL Off-Site</b> QC1 .U52	
<i>Lib. Has</i>	v.1 (July 1928)- v.12 (June 1934)	
<b>PASCAL Off-Site</b>	<b>QC1 .U52</b> v.1 1928 July-Dec.	AVAILABLE
<b>PASCAL Off-Site</b>	<b>QC1 .U52</b> v.2 1929 Jan-Jun	AVAILABLE
<b>PASCAL Off-Site</b>	<b>QC1 .U52</b> v.3 1929 July-Dec	AVAILABLE
<b>PASCAL Off-Site</b>	<b>QC1 .U52</b> v.4 1930 Jan-Jun	AVAILABLE
<b>PASCAL Off-Site</b>	<b>QC1 .U52</b> v.5 1930 July-Dec	AVAILABLE
<b>PASCAL Off-Site</b>	<b>QC1 .U52</b> v.6 1931 Jan-Jun	AVAILABLE
<b>PASCAL Off-Site</b>	<b>QC1 .U52</b> v.7 1931 July-Dec	AVAILABLE
<b>PASCAL Off-Site</b>	<b>QC1 .U52</b> v.8 1932 Jan-Jun	AVAILABLE
<b>PASCAL Off-Site</b>	<b>QC1 .U52</b> v.9 1932 July-Dec	AVAILABLE
<b>PASCAL Off-Site</b>	<b>QC1 .U52</b> v.10 1933 Jan-Jun	AVAILABLE

[View additional copies or search for a specific volume/copy](#)

**Description** 12 v. : ill. ; 24 cm.  
**Frequency** Monthly  
**Pub History** Vol. 1, no. 1 (July 1928)-v. 12, no. 6 (June 1934).  
**Indexed In** Chemical abstracts 0009-2258

**Corp Author** [United States Commission on Civil Rights.](#)  
**Title** Indian tribes : a continuing quest for survival : a report / of the United States Commission on Civil Rights.  
**Publ Info** Washington, D.C. : The Commission : For sale by the Supt. of Docs., U.S. G.P.O., 1981.

The links below are for electronic versions of this publication:  
[Access online version via Hathi Trust](#)

LOCATION	CALL #	STATUS
<b>PASCAL Off-Site</b>	<a href="#">CR 1.2:In 2/7</a>	AVAILABLE
Internet	<a href="#">CR 1.2:In 2/7</a>	ONLINE

**Description** xi, 192 p. : ill. ; 26 cm.  
**Note(S)** "June 1981."  
 Includes bibliographical references.  
**Note** Access online version via Hathi Trust: <http://babel.hathitrust.org/cgi/pt?id=umn.31951d00823960q>

# Stripped-Out Fields

## Long hard road : NCO experiences in Afghanistan and Iraq.

008 fixed field data

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](#)

650 subfields other than "a"

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](#)

500 notes

5xx shipping list info

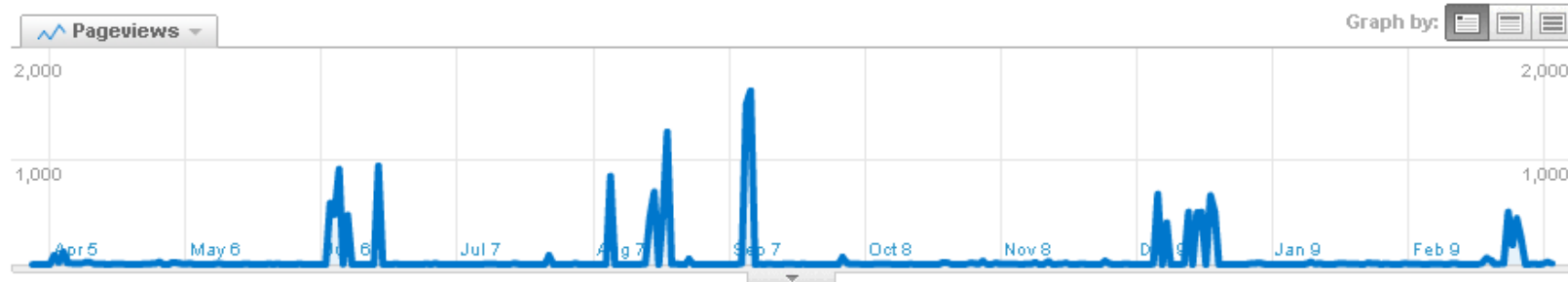
300 subfields after "a"

086 SuDocs number

# Use Stats for Hathi Trust?

Top Content

Apr 1, 2010 - Mar 14, 2011



**7,910 pages were viewed a total of 17,130 times**

Filtered for pages containing "hathi"

Content Performance

Views:

Pageviews <b>17,130</b> % of Site Total: 1.65%	Unique Pageviews <b>10,986</b> % of Site Total: 1.85%	Avg. Time on Page <b>00:00:13</b> Site Avg: 00:00:58 (-77.12%)	Bounce Rate <b>6.29%</b> Site Avg: 31.49% (-80.01%)	% Exit <b>1.87%</b> Site Avg: 19.55% (-90.42%)	\$ Index <b>\$0.00</b> Site Avg: \$0.00 (0.00%)
---	--	---	--	---	--

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

- Documents content in HathiTrust can provide a suitable surrogate for a limited subset of documents, but not a wholesale replacement.
- HathiTrust documents can be used as surrogates for selected titles, especially larger serial runs. But it is difficult at this time to isolate those titles.
- HathiTrust is definitely worth harvesting into local catalogs or other digital repositories.