

# Massive Metadata Mashup

Roy Tennant

Senior Program Officer

OCLC Research



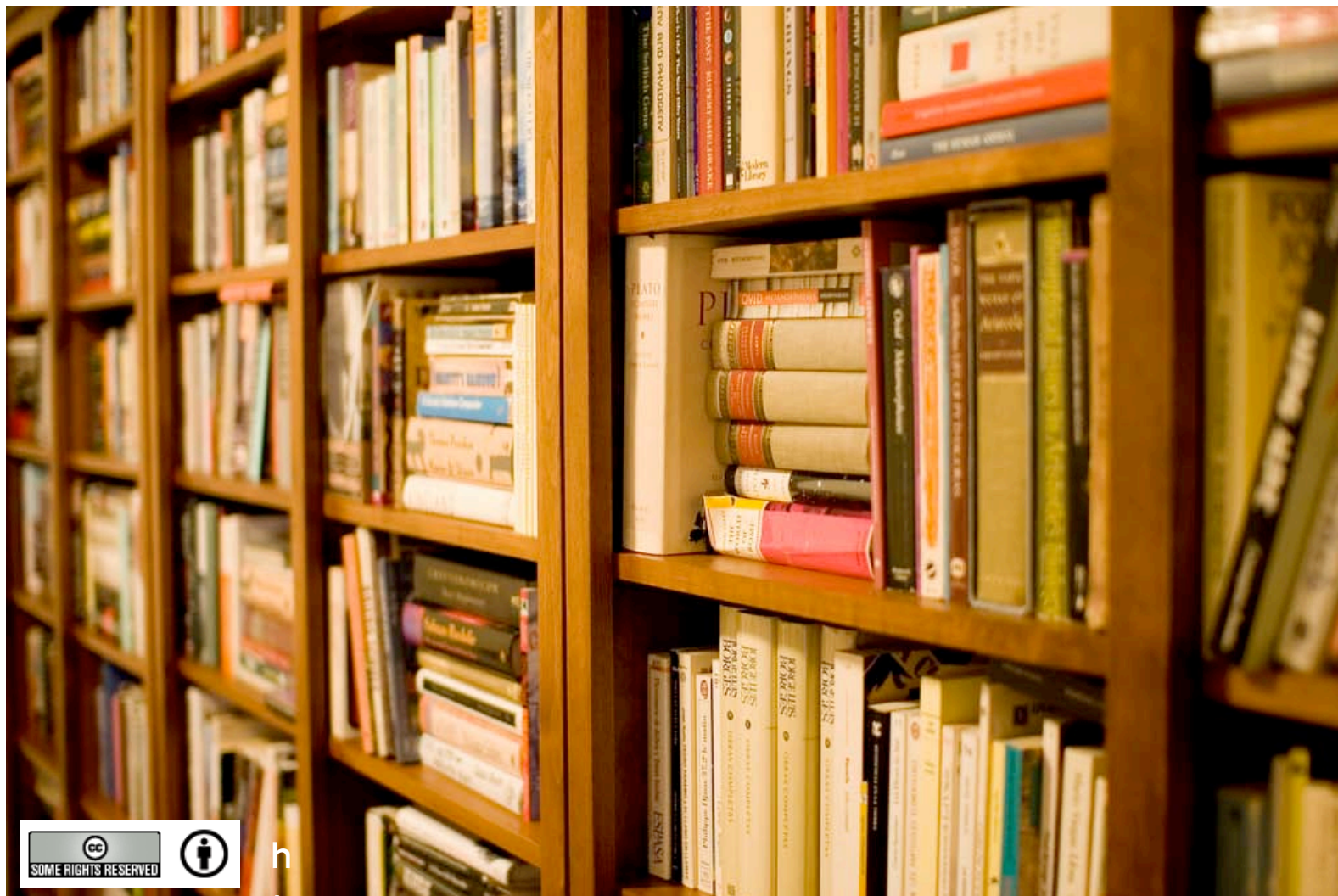
OCLC™

The world's libraries.  
Connected.

# So What is This All About?



# And Why Are You Doing This?





# HATHI TRUST

a shared digital repository

Catalog Search

Find

[About](#)[Access & Services](#)[Repository](#)[Partnership](#)[Updates](#)[About HathiTrust](#)[Mission & Goals](#)[Functional Objectives](#)[Governance](#)

## Welcome to the Shared Digital Future

### HathiTrust is a bold idea with big plans.

As a digital repository for the nation's great research libraries, HathiTrust ([listen to pronunciation](#)) brings together the immense collections of partner institutions.

HathiTrust was conceived as a collaboration of the thirteen universities of the Committee on Institutional Cooperation and the University of California system to establish a repository for these universities to archive and share their digitized collections. Partnership is open to all who share this grand vision.

### HathiTrust is a solution.

To prospective partners, HathiTrust offers leadership and reliability.


It provides a no-worry, pain-free solution to archiving vast amounts of digital content. You can rely on the expertise of other librarians and information technologists who understand your needs and who will address the issues of servers, storage, migration, and long-term preservation.

### You will shape HathiTrust.

Growing the world's largest library won't happen overnight. You've heard of other digital libraries. This one is different in concept and scale. Its greatest promise—and challenge—rests in defining how to serve researchers in

- [FAQ](#)
- [Press](#)
- [Papers, Reports & Presentations](#)
- [Contact](#)

## August Activities

- [Working Group Updates](#) 
- [Progress On Internet Archive Ingest](#)
- [UM Press Content To Enter HathiTrust With Links to Print-On-Demand](#)
- [HathiTrust Disaster Preparedness Report](#)
- [Prototype PageTurner Development](#)
- [New METS Profile](#)
- [Duplicate Volume Analysis](#)
- [Mobile Interface Update](#)
- [September Forecast](#)
- [Improved Indexing, New Hardware for Large-scale Search](#)
- [Collection Builder APIs](#)

[Read the full update >>](#)

# Current Partners

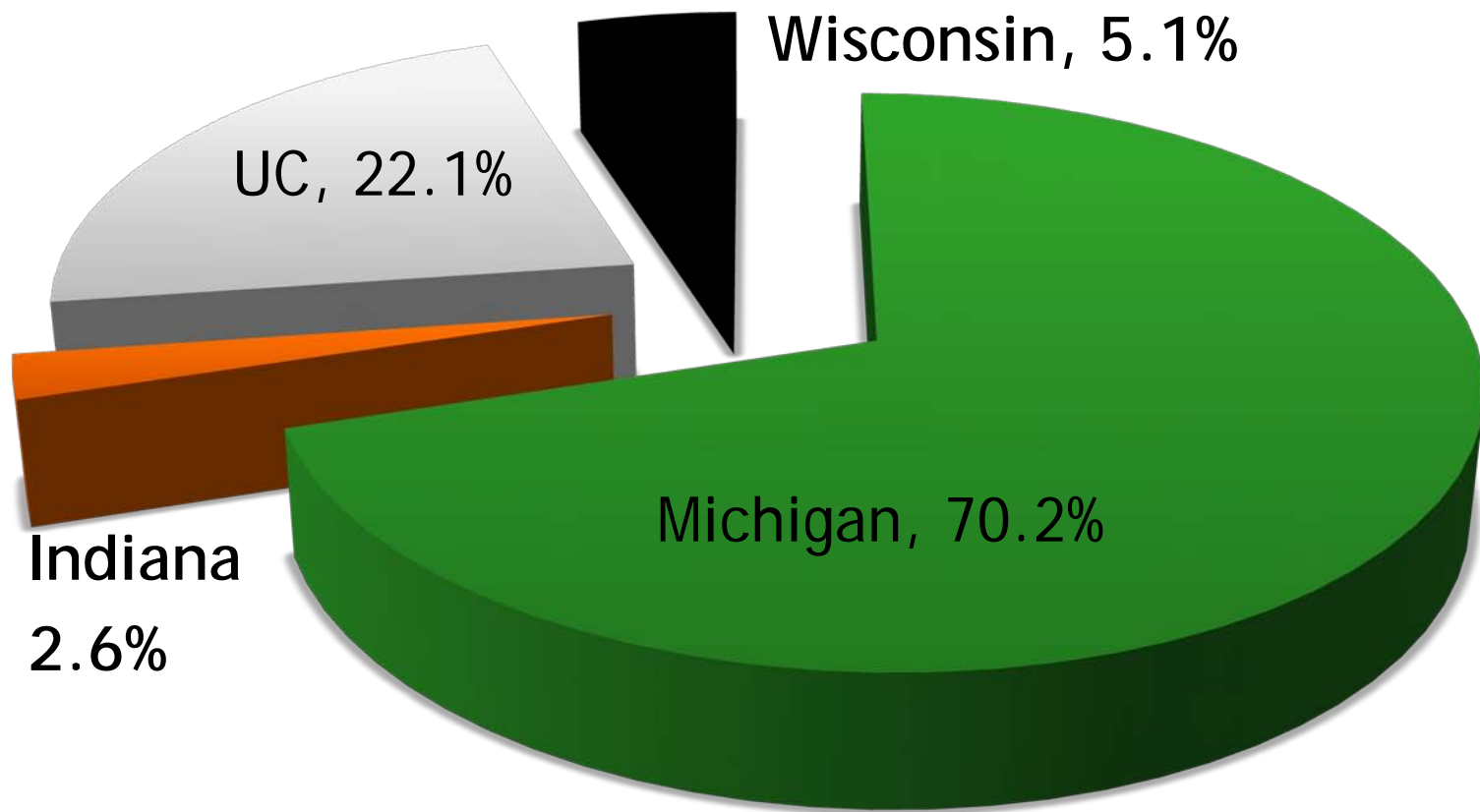


## Current Partners

HathiTrust membership currently consists of the member libraries of the [Committee on Institutional Cooperation\(CIC\)](#) the [University of California system](#), and the [University of Virginia](#), but is open to all interested research libraries. Current members include:

- California Digital Library
- Indiana University
- Michigan State University
- Northwestern University
- The Ohio State University
- Penn State University
- Purdue University
- University of California Berkeley
- University of California Davis
- University of California Irvine
- University of California Los Angeles
- University of California Merced
- University of California Riverside
- University of California San Diego
- University of California San Francisco
- University of California Santa Barbara
- University of California Santa Cruz
- The University of Chicago
- University of Illinois
- University of Illinois at Chicago
- The University of Iowa
- University of Michigan
- University of Minnesota
- University of Wisconsin-Madison
- University of Virginia

# Current Contributing Partners (Dec. '09 Data)



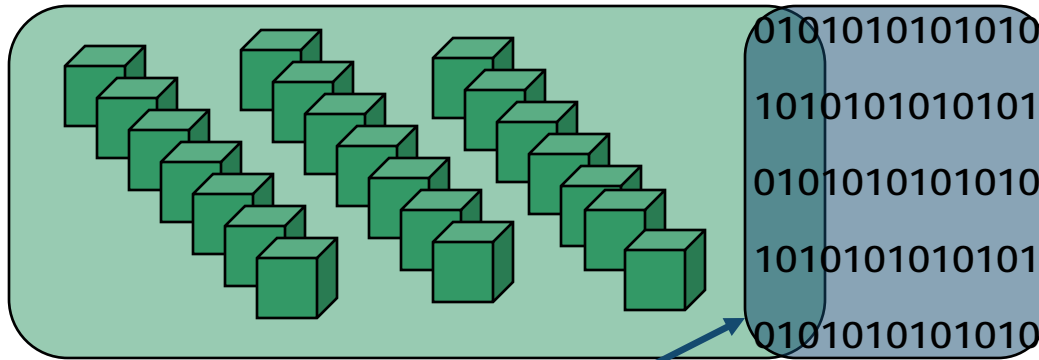


# OCCL Research “Cloud Library” Project



## Academic off-site storage

**25 years  
+70M vols.**



**15 months  
+5M vols.**

**HathiTrust**

***Will this intersection create new operational efficiencies?***

***For which libraries?***

***Under what conditions?***

***How soon and with what impact?***

Constance  
Malpas





Oh Crap

# How I Ended Up in the Middle of This

What on Earth was I thinking?



OCLC™

The world's libraries.  
Connected.





HATHI TRUST  
a shared digital repository

Catalog Search

Find

About

Access & Services

Repository

Partnership

Updates

Access To HathiTrust

Data Distribution & APIs

Datasets

Visualizations

## Hathifiles

Displaying *sites/www.hathitrust.org/files/hathifiles*.

Contains 33 files totaling 477.39 MB in size.

Name	Size
<a href="#">hathi_full_20090801.txt.gz</a>	218.46 MB
<a href="#">hathi_full_20090901.txt.gz</a>	232.36 MB
<a href="#">hathi_upd_20090801.txt.gz</a>	660.27 KB
<a href="#">hathi_upd_20090802.txt.gz</a>	1.25 MB
<a href="#">hathi_upd_20090803.txt.gz</a>	1.17 MB
<a href="#">hathi_upd_20090804.txt.gz</a>	1.17 MB
<a href="#">hathi_upd_20090805.txt.gz</a>	1.49 MB
<a href="#">hathi_upd_20090806.txt.gz</a>	1.98 MB
<a href="#">hathi_upd_20090807.txt.gz</a>	1.79 MB
<a href="#">hathi_upd_20090808.txt.gz</a>	1.38 MB
<a href="#">hathi_upd_20090809.txt.gz</a>	1.1 MB
<a href="#">hathi_upd_20090810.txt.gz</a>	1.13 MB
<a href="#">hathi_upd_20090811.txt.gz</a>	968.73 KB

- [HathiTrust Metadata Download](#)
- [HathiTrust Metadata Documentation](#)
- [HathiTrust Rights API](#)
- [HathiTrust Data API](#)

## FAQ

How does HathiTrust compare to Google Book Search?



HathiTrust complements Google's massive undertaking to digitize the world's library collections. While both systems offer digitized books via the Internet, it is likely that HathiTrust will provide some content Google will not, such as digital collections unique to each institution, works from institutional repositories, and native born-digital materials.

HathiTrust also provides a new platform for the expert curation and consistent access long associated with research libraries. The trust and reliance developed over decades in providing essential print collections will extend to HathiTrust as a valued source for scholarly materials.

[More FAQ >>](#)

Hathitrust Metadata   www.hathitr...	+
Volume Identifier	<p>mdp.39015067674344</p> <p>This ID can be used to construct a handle or other link that provides access to the item. If you are providing access via the University of Michigan handle server, handles can be constructed as follows:  <a href="http://hdl.handle.net/2027/volume_identifier">http://hdl.handle.net/2027/volume_identifier</a>                      For example :  <a href="http://hdl.handle.net/2027/mdp.39015067674344">http://hdl.handle.net/2027/mdp.39015067674344</a></p>
Access	<p>Access code (values: allow or deny) mapped from the rights attribute using the same algorithm used in the <a href="#">page turner</a>.</p> <p>Notes on mapping for rights attributes where contextual user data would affect access:</p> <ol style="list-style-type: none"> <li>The opb (out of print and brittle) rights attribute is mapped to 'deny'.</li> <li>The umall (available to all University of Michigan affiliates) attribute is mapped to deny.</li> <li>The pdus (public domain if viewed in US) attribute is mapped to allow.</li> </ol>
Rights	<p>Rights attribute for this volume from the <a href="#">HathiTrust rights database</a>. Possible values are:</p> <p>pd .....public domain                      ic.....in copyright                      opb.....out of print and brittle (implies in copyright)                      orph.....copyright-orphaned (implies in copyright)                      umall.....available to UM affiliates and walk-in patrons (all campuses)</p> <p>world.....available to everyone in the world                      und.....undetermined copyright status                      pdus.....public domain only when viewed in the US</p>
University of Michigan record number	University of Michigan's record number for the associated bibliographic record.
Enumeration/Chronology	Enumeration and chronology data, if any, for this item
Source	Code identifying the source of the original from which this item was digitized. Currently the NUC code of the originating library is used for the code.
Source institution record number	Local bibliographic record number for the source institution identified in the source column.
OCLC numbers	OCLC number(s) for the bibliographic record. Where more than one number is listed, they are comma-delimited.

## 13 Elements

http://roytennant.com/proto/hathi/20090801/8/1777414.zhttp://roytenna...801/8/1777414.z

```
<record>
<identifier>mdp.39015018935323</identifier>
<access>deny</access>
<rights>ic</rights>
<recnum>002222646</recnum>
<enum></enum>
<source>MiU</source>
<sourcenum>002222646</sourcenum>
<oclc>15634978</oclc>
<isbn></isbn>
<issn></issn>
<lccn></lccn>
<title>Studies on Atlantic Canada : papers presented at the twent
<imprint>Dept. of Sociology and Anthropology, University of Princ
</record>
```

[ [Prototype Server Home](#) | [Hathi Trust Search Home](#) ]

## Hathi Trust Search

You are searching [brief records](#) for texts digitized by [Hathi Trust](#) institutions. You should also know about the [Hathi Trust Catalog](#) and the [experimental full-text search](#).

- |                                    |                                       |
|------------------------------------|---------------------------------------|
| ■ 1 April 2009: 2,779,517 records  | ■ 1 September 2009: 3,814,121 records |
| ■ 1 May 2009: 2,811,319 records    | ■ 1 October 2009: 3,976,132 records   |
| ■ 9 June 2009: 3,040,537 records   | ■ 1 November 2009: 4,537,138 records  |
| ■ 1 July 2009: 3,313,695 records   | ■ 1 December 2009: 4,699,669 records  |
| ■ 1 August 2009: 3,606,451 records | ■ 1 January 2010: 5,222,699 records   |

Search:  in the  index  Limit to public domain texts: ☐

---

*This is a prototype search service for which no promises are made. Contact [Roy Tennant](#) for comments, questions, or where to send the bottle of scotch.*



[ [Prototype Server Home](#) | [Hathi Trust Search Home](#) ]

## Hathi Trust Search

You are searching [brief records](#) for texts digitized by [Hathi Trust](#) institutions. You should also know about the [Hathi Trust Catalog](#) and the [experimental full-text search](#).

- 1 April 2009: 2,779,517 records
- 1 May 2009: 2,811,319 records
- 9 June 2009: 3,040,537 records
- 1 July 2009: 3,313,695 records
- 1 August 2009: 3,606,451 records

Your search for **prince edward island** found **76** items.

Search:  in the  index  Limit to public domain texts: ☐

1. [Studies on Atlantic Canada : papers presented at the twentieth annual meeting of the Atlantic Association of Sociologists and Anthropologists at the University of Prince Edward Island, March 15-17, 1985 / edited by Satadal Dasgupta ; with contributions by William Buxton ... \[et al.\] - raw record - explanation of elements](#)  
Dept. of Sociology and Anthropology, University of Prince Edward Island, 1985.  
*Rights: In Copyright*
2. [St. Lawrence pilot; with the coast of Quebec, New Brunswick, Prince Edward Island, Nova Scotia, Cape Breton Island, and Newfoundland bordering on the Gulf of St. Lawrence and the River St. Lawrence to Quebec, including the banks of Newfoundland and the approaches to the gulf by Cabot Strait, the Strait of Belle Isle, and the Strait of Canso. - raw record - explanation of elements](#)  
[R. Duhamel, Queen's printer] 1963 [c1964]  
*Rights: In Copyright*
3. [Gulf of St. Lawrence pilot, with the coast of Quebec, New Brunswick, Prince Edward Island, Nova Scotia, Cape Breton Island and Newfoundland, bordering on the Gulf of St. Lawrence, including the Banks of Newfoundland and the approaches to the Gulf by Cabot Strait, the Strait of Belle Isle and the Strait of Canso. - raw record - explanation of elements](#)  
Canadian Hydrographic Service, Marine Sciences Branch, Dept. of energy, Mines, and Resources [1968]  
*Rights: In Copyright*
4. [Three centuries and the island; a historical geography of settlement and agriculture in Prince Edward Island, Canada. - raw record - explanation of elements](#)  
University of Toronto Press [1959]  
*Rights: In Copyright*

You're  
effing  
kidding me

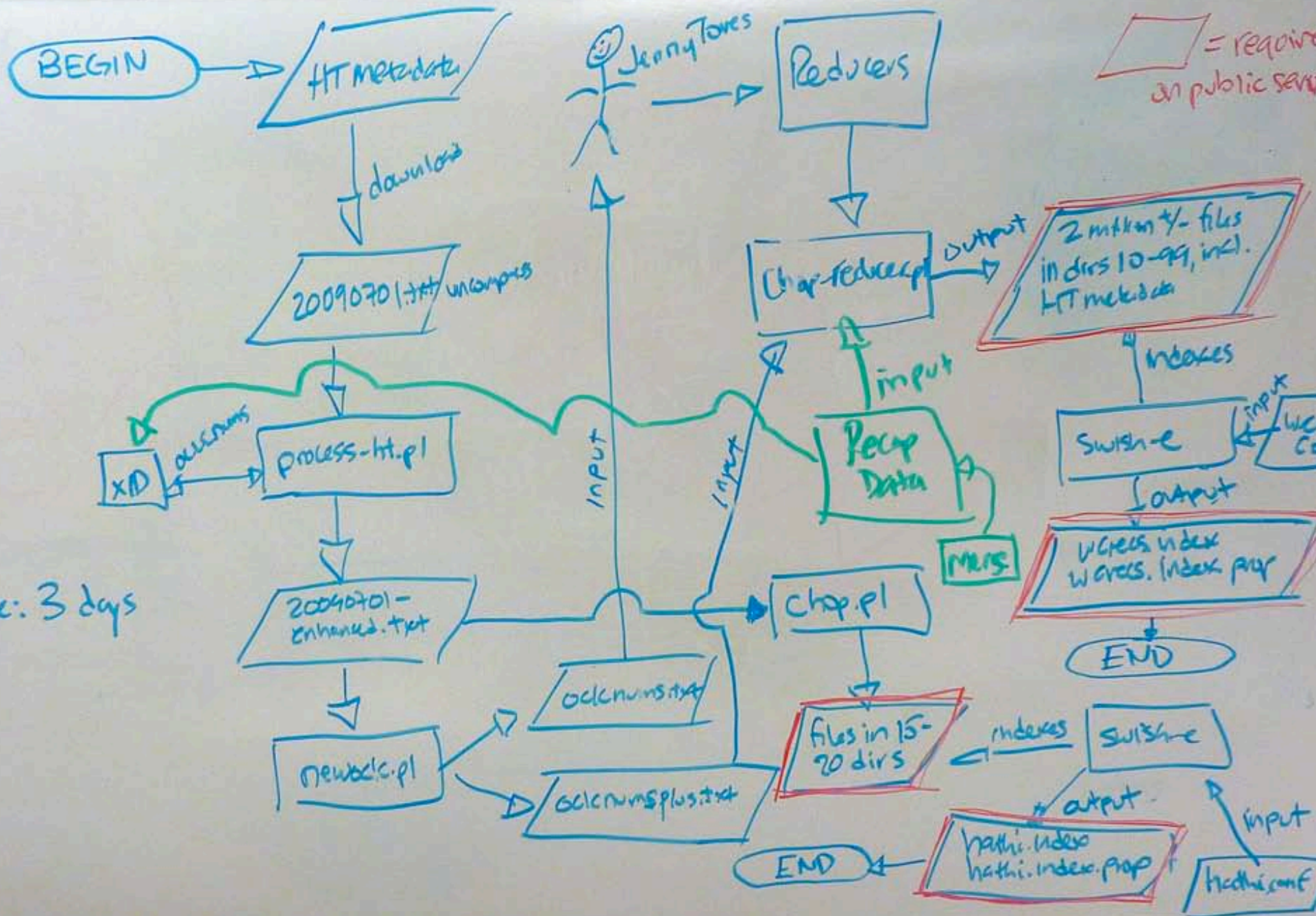
# The Process We Developed

What on Earth were we thinking?



OCLC™

The world's libraries.  
Connected.





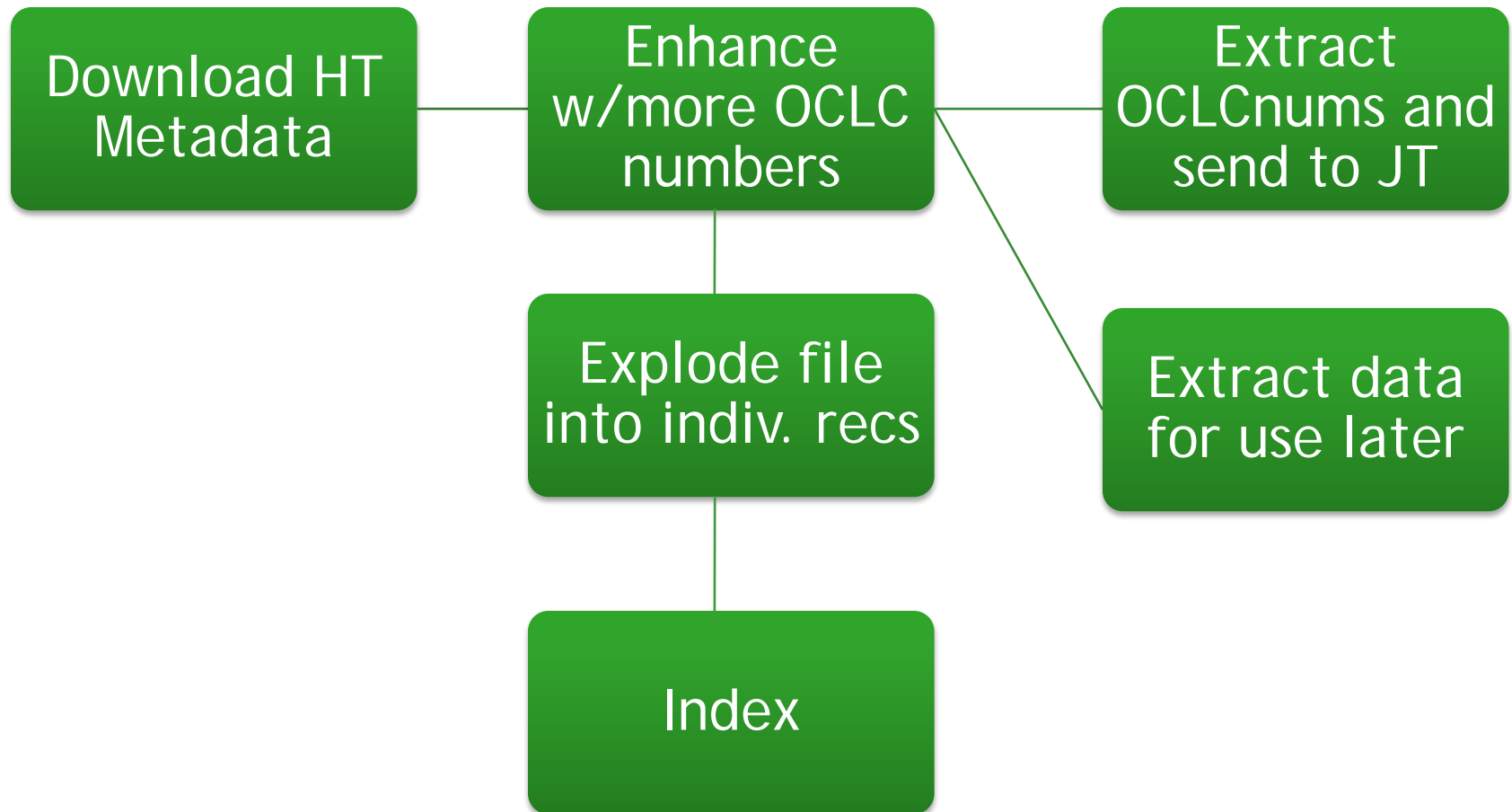




```
Working on /Data/byMonth/July/reducers/reducer106.cdf
Working on /Data/byMonth/July/reducers/reducer107.cdf
Working on /Data/byMonth/July/reducers/reducer108.cdf
Working on /Data/byMonth/July/reducers/reducer109.cdf
Working on /Data/byMonth/July/reducers/reducer110.cdf
Working on /Data/byMonth/July/reducers/reducer111.cdf
Working on /Data/byMonth/July/reducers/reducer112.cdf
Working on /Data/byMonth/July/reducers/reducer113.cdf
Working on /Data/byMonth/July/reducers/reducer114.cdf
Working on /Data/byMonth/July/reducers/reducer115.cdf
Working on /Data/byMonth/July/reducers/reducer116.cdf
Working on /Data/byMonth/July/reducers/reducer117.cdf
Working on /Data/byMonth/July/reducers/reducer118.cdf
Working on /Data/byMonth/July/reducers/reducer119.cdf
Working on /Data/byMonth/July/reducers/reducer11.cdf
Working on /Data/byMonth/July/reducers/reducer120.cdf
Working on /Data/byMonth/July/reducers/reducer121.cdf
Working on /Data/byMonth/July/reducers/reducer122.cdf
Working on /Data/byMonth/July/reducers/reducer123.cdf
Working on /Data/byMonth/July/reducers/reducer124.cdf
Working on /Data/byMonth/July/reducers/reducer125.cdf
Working on /Data/byMonth/July/reducers/reducer126.cdf
Working on /Data/byMonth/July/reducers/reducer127.cdf
```



# Part I





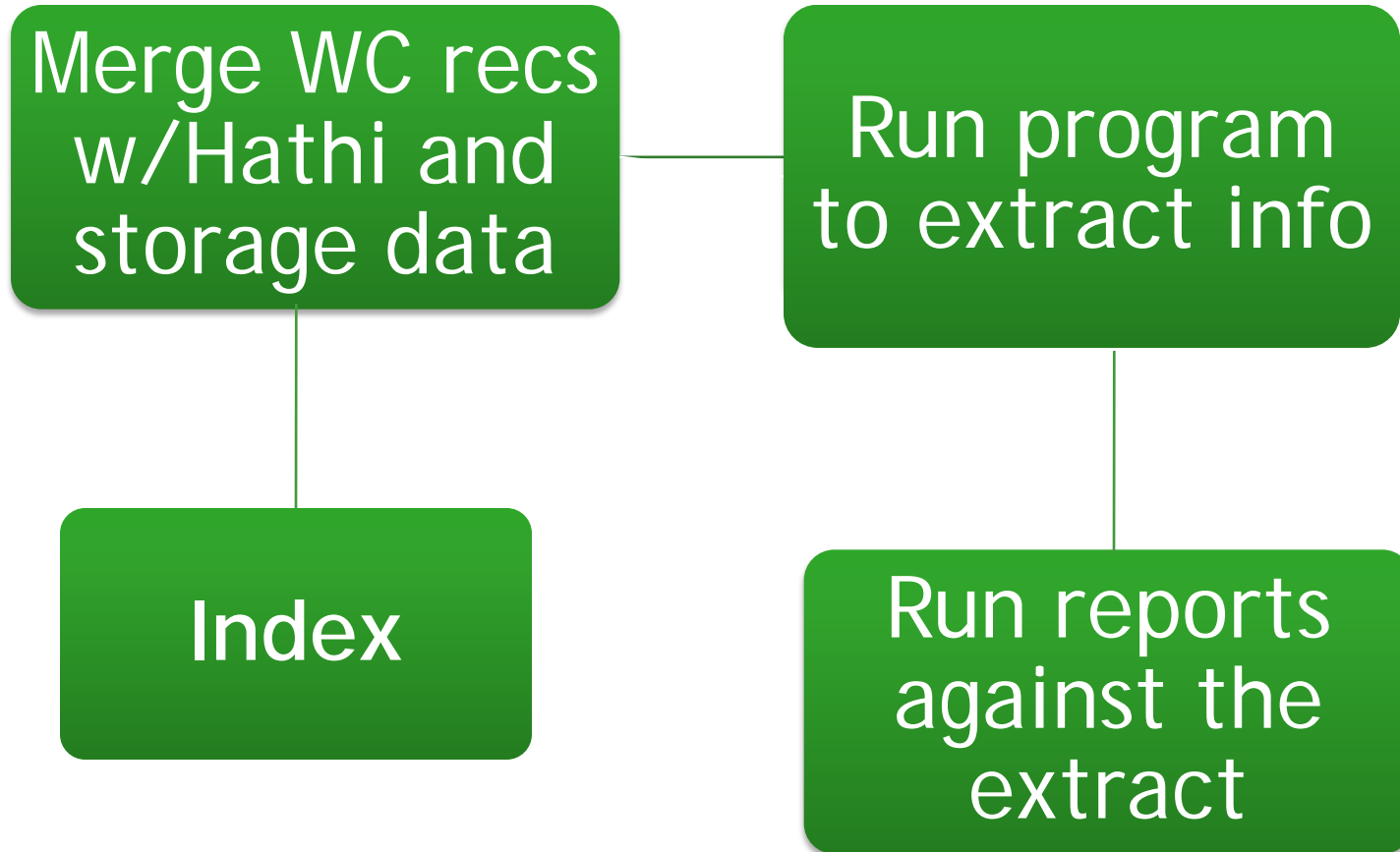
# Part II

Merge WC recs  
w/Hathi and  
storage data

Run program  
to extract info

Index

Run reports  
against the  
extract



# Tools Used

- Perl for text processing and user interface
- XML on the filesystem for data
- Swish-e for indexing [swish-e.org]
- XSLT for XML rendering
- xsltproc for XSLT parsing

# Data Structure - Original

```
<UniqueKey>394608</UniqueKey>
<primaryDocType>bks</primaryDocType>
<primaryDocType2>bks</primaryDocType2>
<primaryLanguage>eng</primaryLanguage>
<catalogingLanguage>eng</catalogingLanguage>
<libraryCountRange>20</libraryCountRange>
<year>1944</year>
<libraryCount>284</libraryCount>
<ulInstCount>1</ulInstCount>
<coOpProgsIndicator>y</coOpProgsIndicator>
<workId>1536352</workId>
<authorSort>BRUNER JEROME S (JEROME SEYMOUR)</authorSort>
<titleSort>MANDATE FROM THE PEOPLE</titleSort>
- <recordData>
  - <hathi>
    <oclcnum>394608</oclcnum>
    <rights>ic</rights>
    <rightsBinary>ic</rightsBinary>
    <source>MiU</source>
    <identifier>mdp.39015001794356 </identifier>
  </hathi>
  - <CDFFRec>
    <a>cam I </a>
    <c001>394608</c001>
    <c008>720829s1944 nyuab 000 0 eng </c008>
    - <v010 i1=" " i2=" ">
      - <sa>
        <d> 44040141 </d>
      </sa>
```

Hathi metadata

MARC bib



# Data Structure — Updated



```
<UniqueKey>394608</UniqueKey>
<primaryDocType>bks</primaryDocType>
<primaryDocType2>bks</primaryDocType2>
<primaryLanguage>eng</primaryLanguage>
<catalogingLanguage>eng</catalogingLanguage>
<libraryCountRange>20</libraryCountRange>
<year>1944</year>
<libraryCount>284</libraryCount>
<ulInstCount>1</ulInstCount>
<coOpProgsIndicator>y</coOpProgsIndicator>
<workId>1536352</workId>
<authorSort>BRUNER JEROME S (JEROME SEYMOUR)</authorSort>
<titleSort>MANDATE FROM THE PEOPLE</titleSort>
- <recordData>
  - <hathi>
    <oclcnum>394608</oclcnum>
    <rights>ic</rights>
    <rightsBinary>ic</rightsBinary>
    <source>MiU</source>
    <identifier>mdp.39015001794356 </identifier>
  </hathi>
```

Hathi metadata

```
<recap>
<oclcnum> 394608</oclcnum>
<custcode>CU PB AU</customercode>
<custoccur>3</custoccur>
</recap>
```

ReCAP metadata

# Data Extract Format



oclcnum::100000051

year::1970

doctype::url

librarycount::1

rights::ic

language::eng

libcountrange::11

division::99000000

category::99100000

subject::99100100

source::WU

identifier::wu.89105673826

# Sample Report (fragment)



August 2009 (2,279,499 records)

Doctype :

art	210
bks	2198974
...	
sco	2372
ser	77693
url	1134
vis	8

Year :

0000	4237
0001	1
0200	1
1000	1923
1047	1
1125	1
.....	
2005	28118
2006	24363
2007	17479
2008	4570
2009	190
2010	2
2300	1
6546	1
9999	8

Language :

abk	1
ace	25
ady	4
afr	238



Hmm...

# What We've Discovered So Far

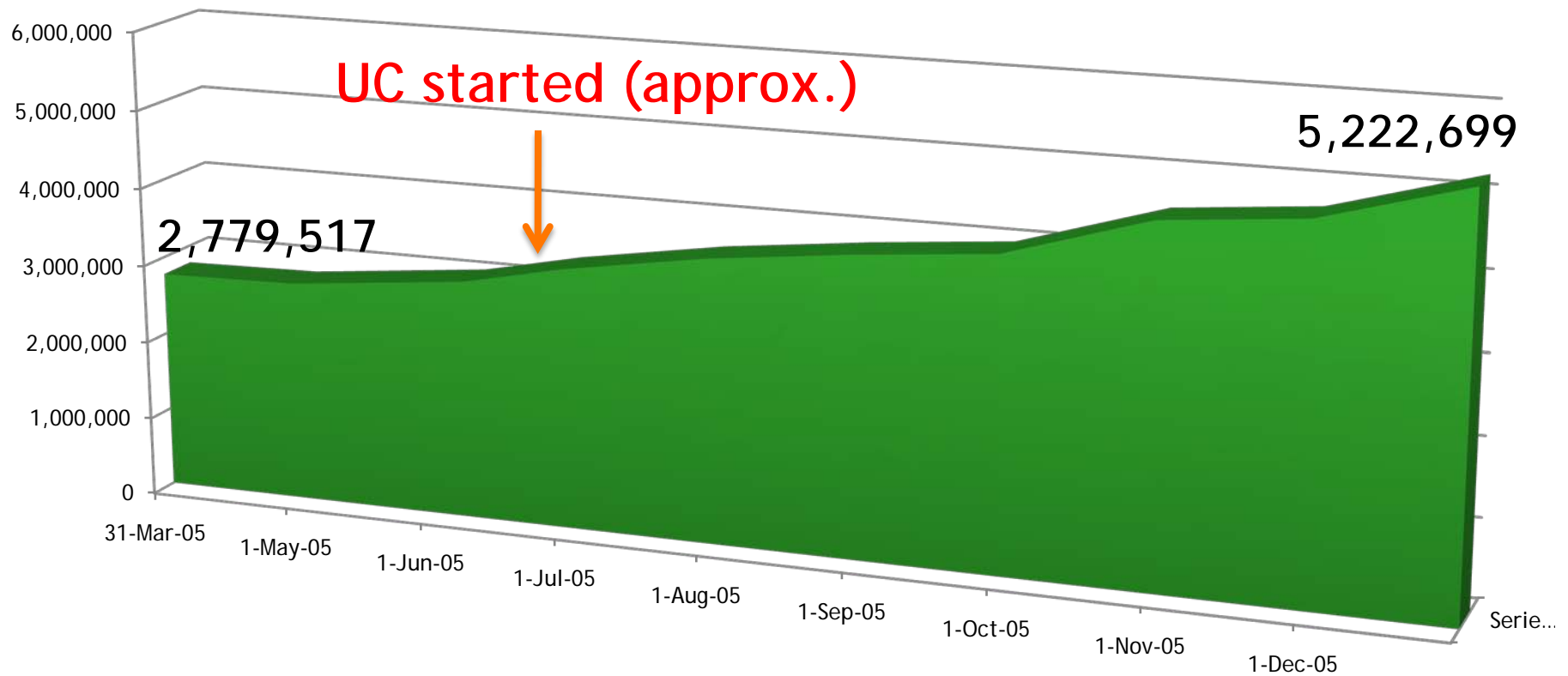
Make Up Your Own Mind



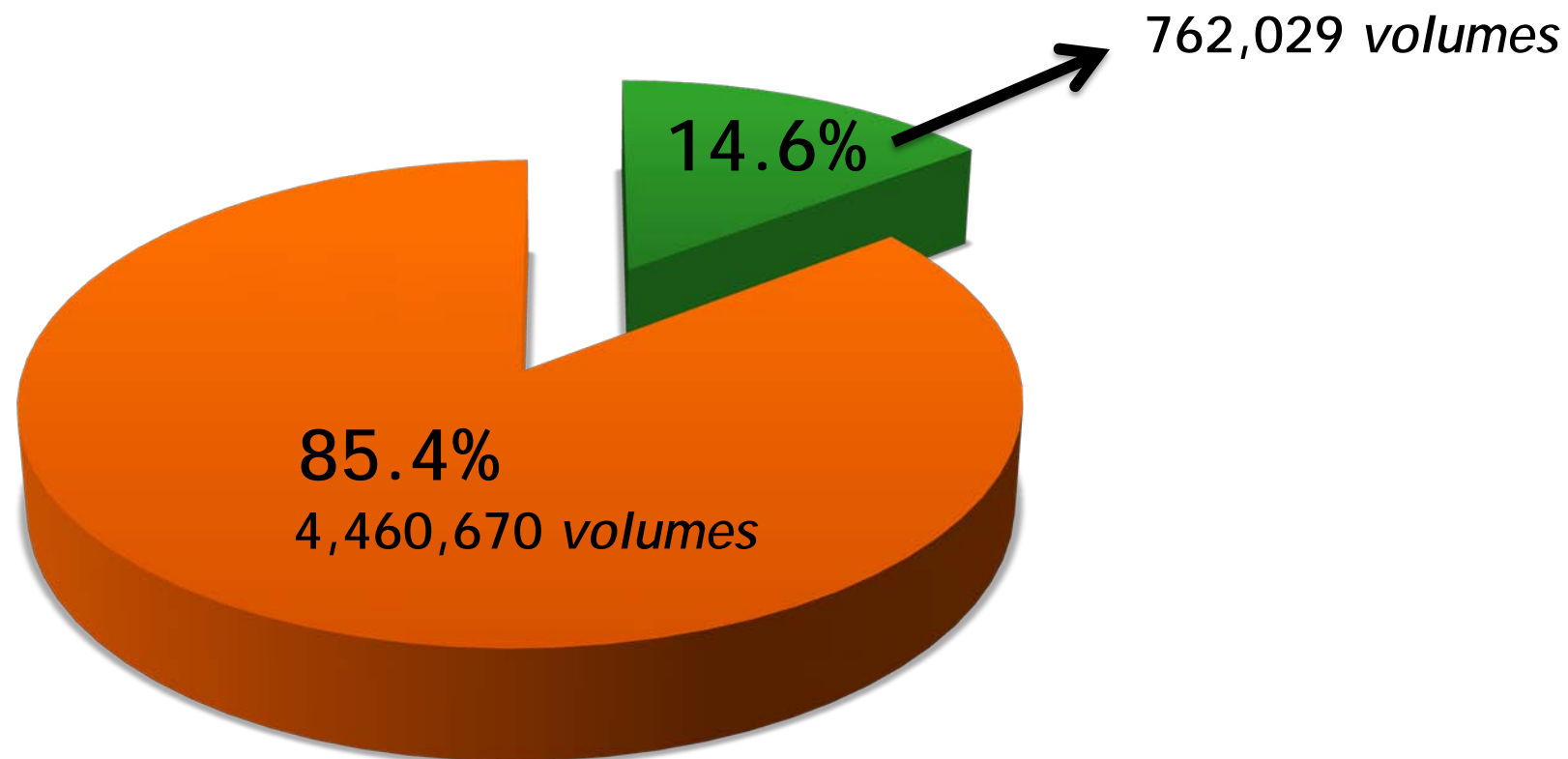
OCLC™

The world's libraries.  
Connected.

# Collection Growth

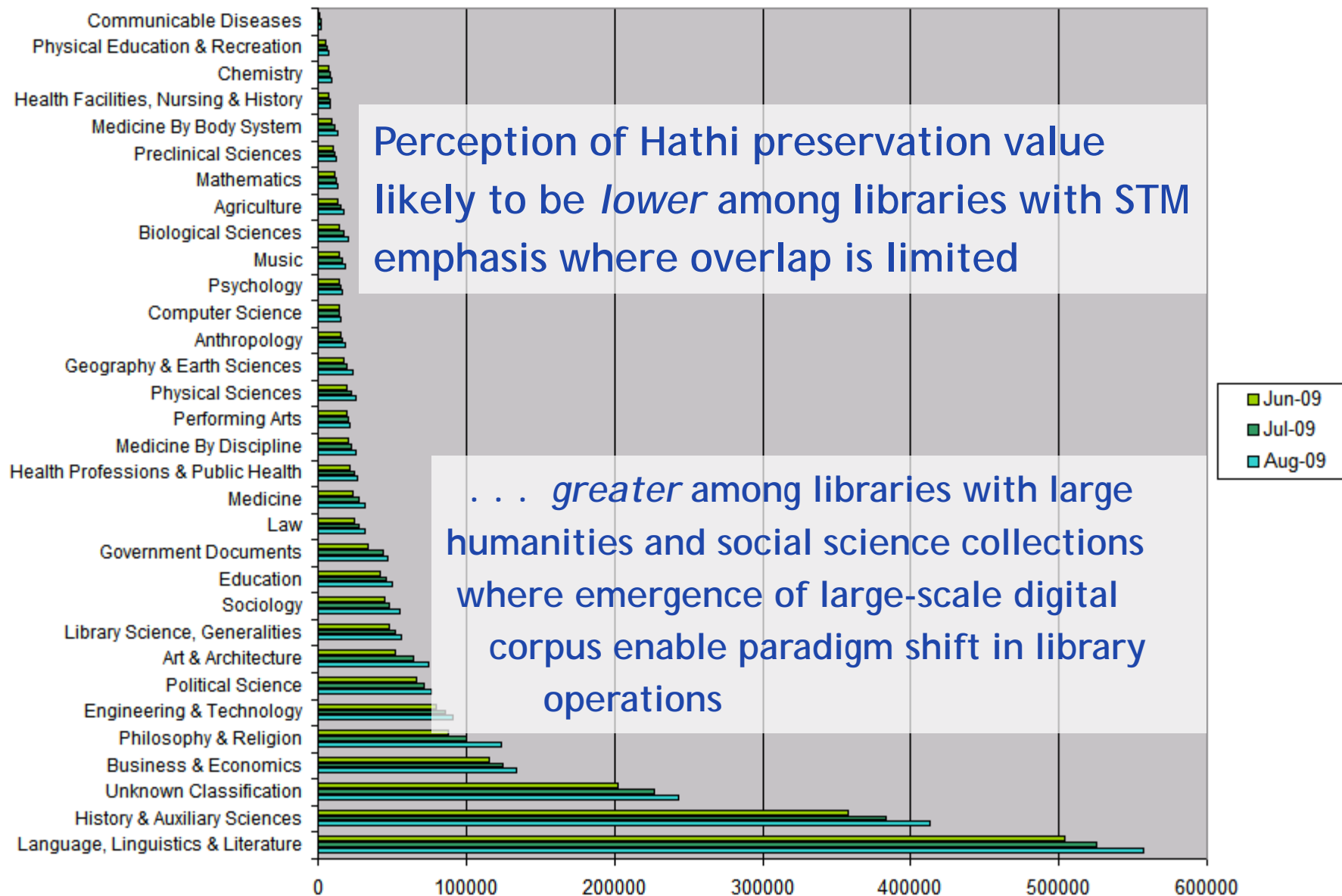


# Public Domain vs. In Copyright



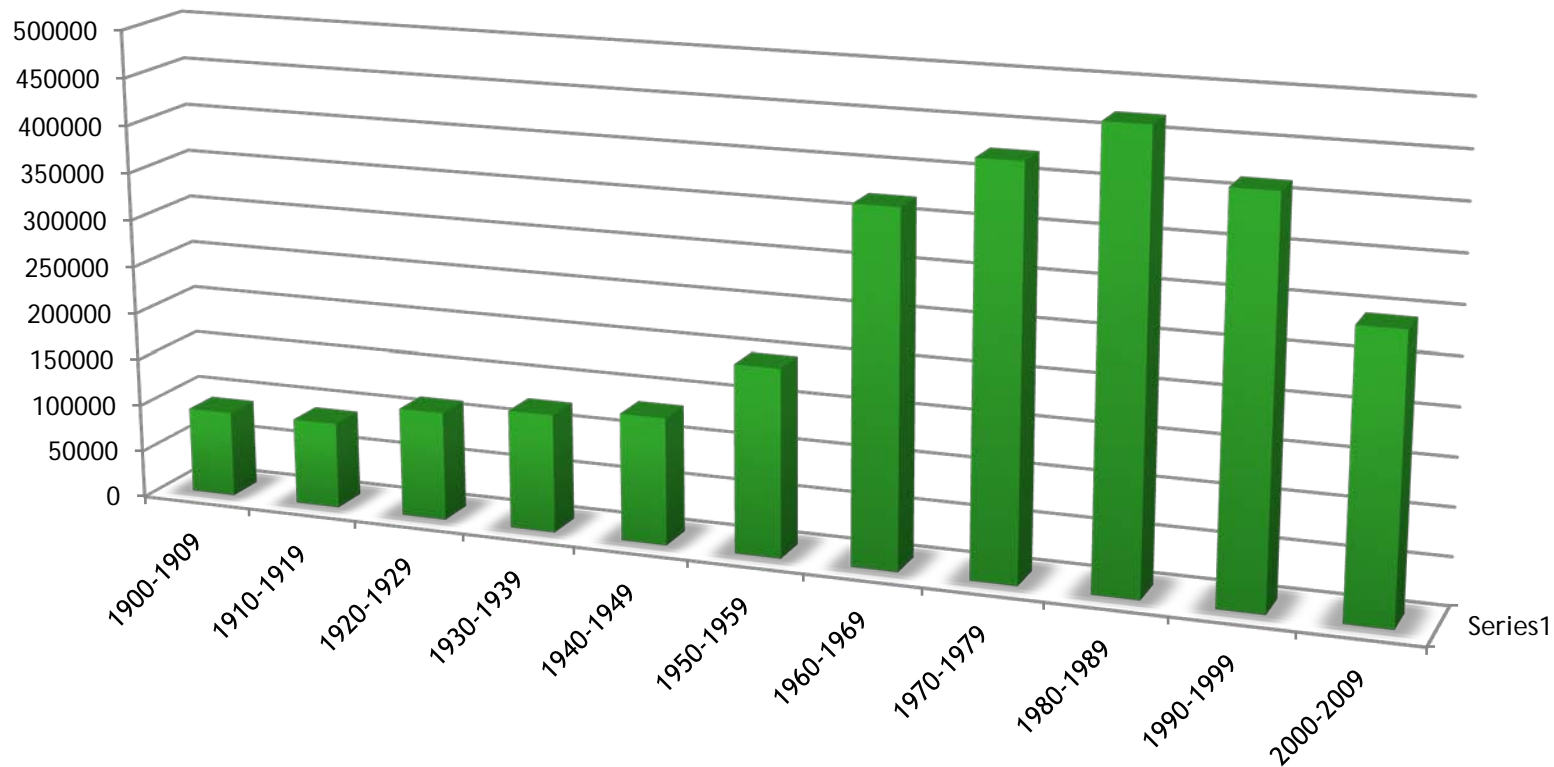
Hathi Trust data for December 2009

# Subject Distribution





# Date Range Distribution, 1900-Present (December 2009 data)



Note: WorldCat subset



Wow

# Where We're Going With This

So this hassle may be worth it after  
all?



OCLC™

The world's libraries.  
Connected.

# Value of partnership increases as number of participants grows

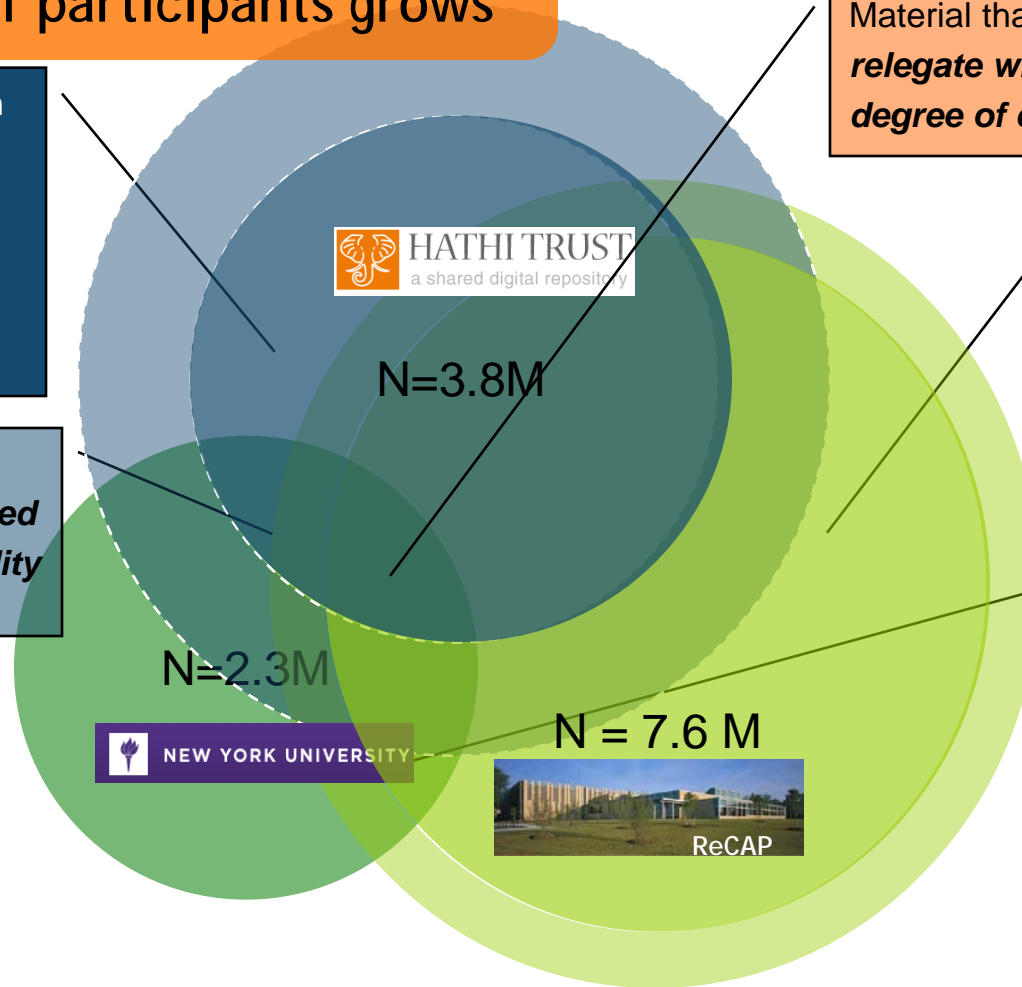
Material that NYU can  
obtain through HT  
dependent on  
copyright status –  
means of **enhancing**  
**'local' collection**

Material that NYU may  
choose to **relegate based**  
**on copyright/ availability**

Material that NYU can  
**relegate with a high**  
**degree of confidence**

Material that NYU  
can already  
source through  
existing ILL –  
**enhance local**  
**collection**

Material that NYU may  
choose to **relegate**  
**with appropriate**  
**service level**  
**agreement**



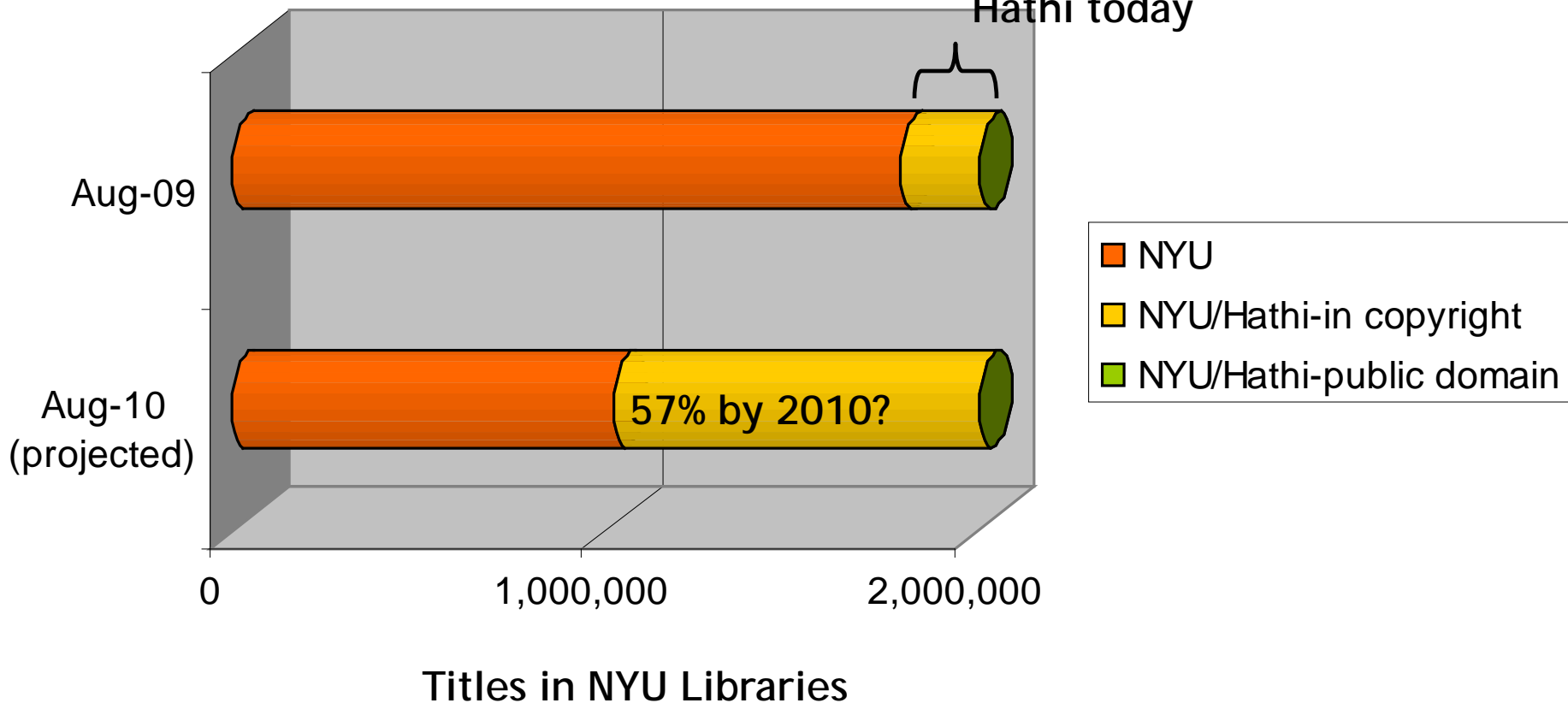
## Intersections

*Opportunities for Institutional Cooperation*  
*Shared Policy Frameworks*  
*Joint Service Agreements*  
*Increased Operational Efficiencies*

*What does a robust digital preservation guarantee for these holdings mean for NYU's long-term collections strategy?*

*How does it affect print preservation / access requirements?*

22% of titles in NYU Libraries are duplicated in Hathi today



Medium	Discounted Life Cycle Cost (per unit)	Total Life Cycle Cost (per unit)	Purchase Cost (per unit)	Total Cost / Purchase Cost (per unit)
Monographs	\$ 119.56	\$ 343.03	\$ 47.78	718%
Current serials	\$ 634.91	801.87	590.97	134
Microfilm				
CD-ROMs				
Internet resources				
Audio recordings				
Video recordings				
Video & Film	\$ 128.95	107.50	15.70	307
Computer files	\$ 0.17	0.07	0.01	331

**"monographs are overwhelmingly the largest source or driver of library costs . . .**

***If research libraries want to control their costs, they must work to control and reduce the life cycle costs of maintaining their monograph collections"*** S. Lawrence et al (2001)



Duh

# Lessons From the Experience

You call these lessons?



OCLC™

The world's libraries.  
Connected.

*Identifiers are  
essential*



*Standards are great  
but don't let them  
get in your way*

*When processing large amounts of data, always check your work (small factors can have large consequences)*

*When processing large  
amounts of data, carefully  
consider each decision  
(small factors can have  
large consequences)*

*Never  
underestimate  
the power of a  
prototype!*

