

# Harvesting and normalization at the Digital Public Library of America

## Lessons from a diverse aggregation

DIGITAL LIBRARY of GEORGIA

Kristy Dixon, DPLA coordinator



Sandra McIntyre, director



Amy Rudersdorf, assistant director for content

# Overview

Brief intro to DPLA

Challenges of aggregation

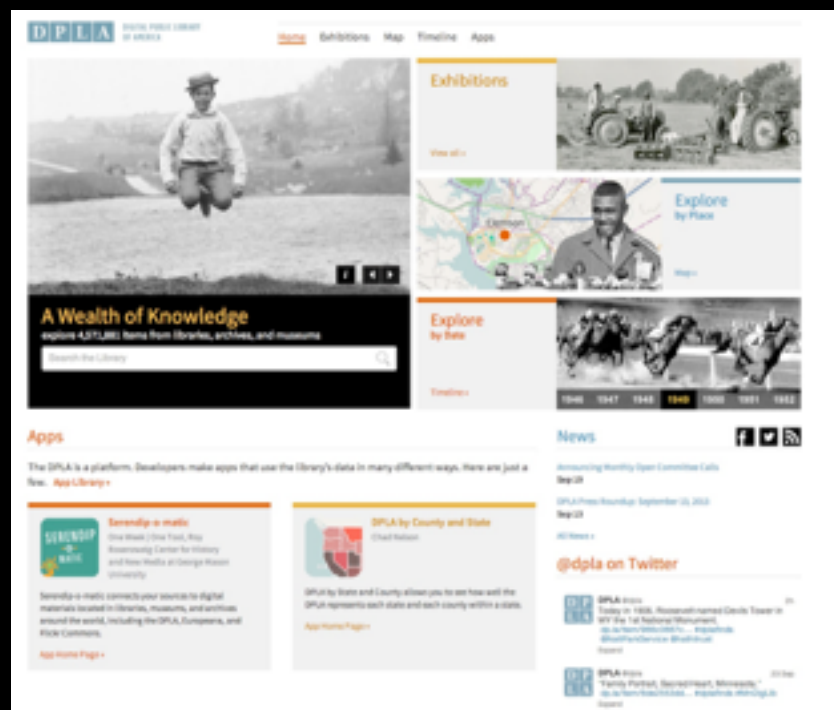
Addressing challenges

Getting started

# Digital Public Library of America (DPLA)

<http://dp.la>

# DPLA is a . . .



**Portal** for discovery

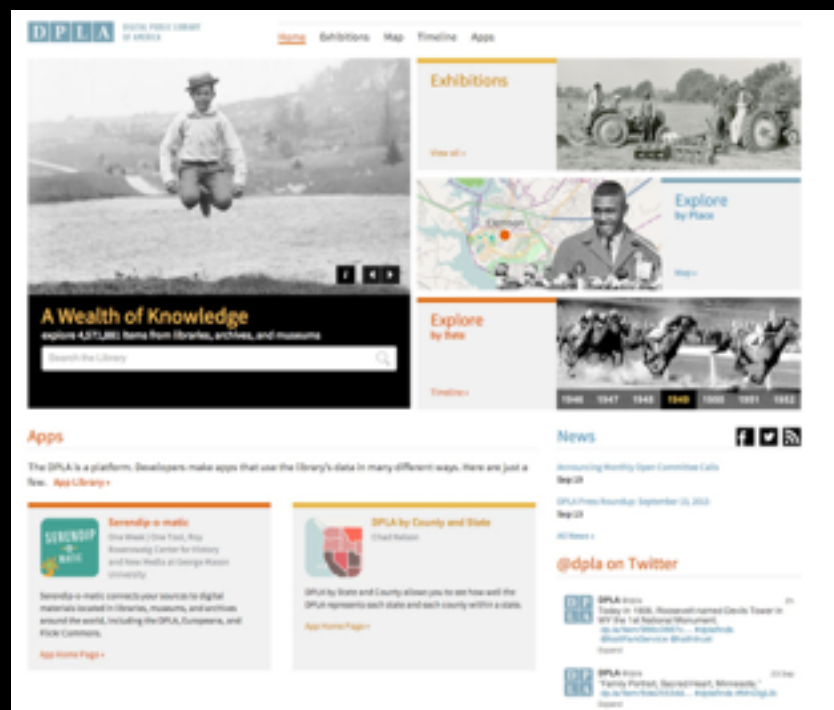


**Public** option



**Platform** to  
build upon

# DPLA is a . . .



**Portal** for discovery



**Public** option



**Platform** to  
build upon



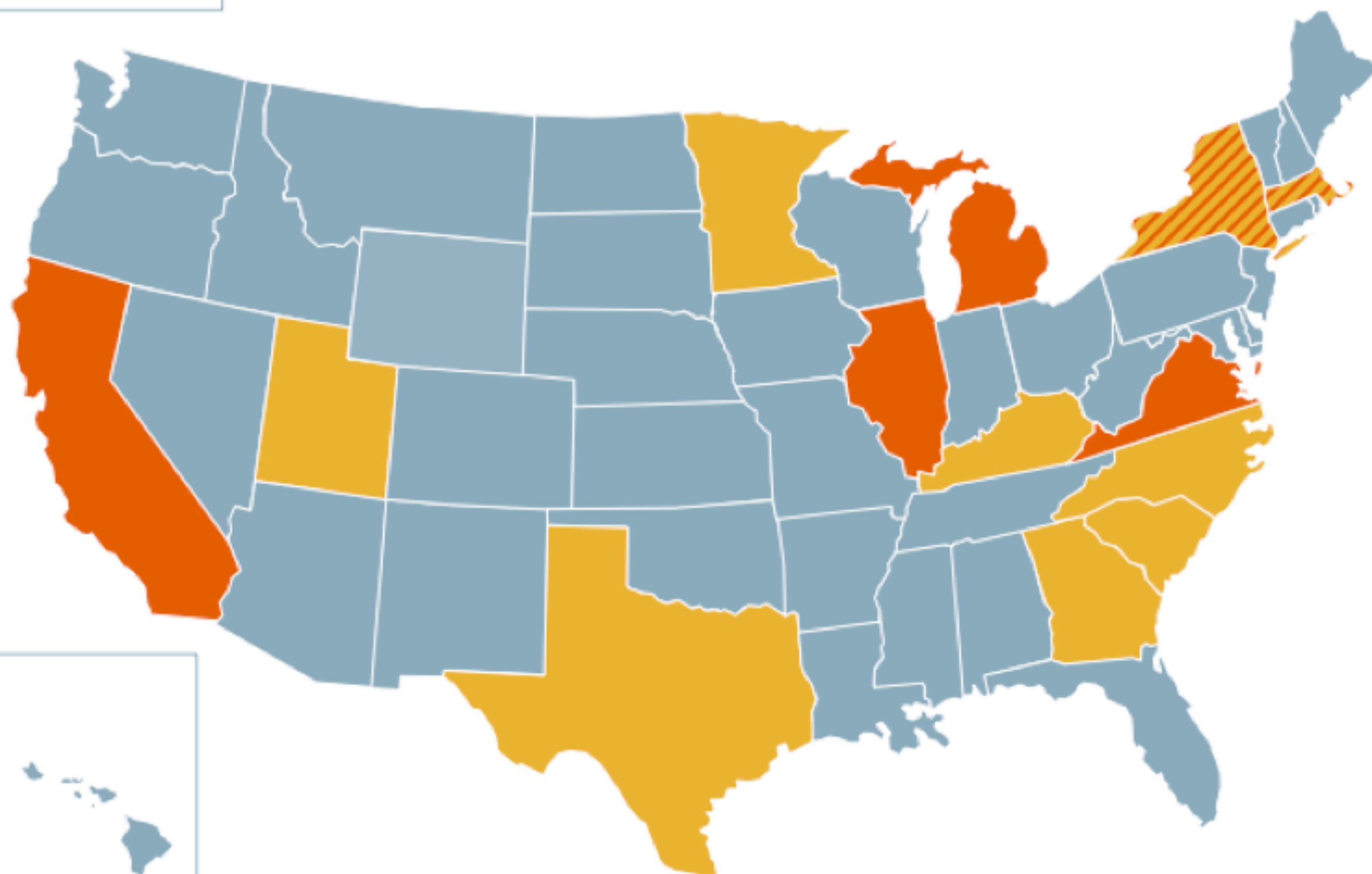


*David  
Rumsey  
Map  
Collection*

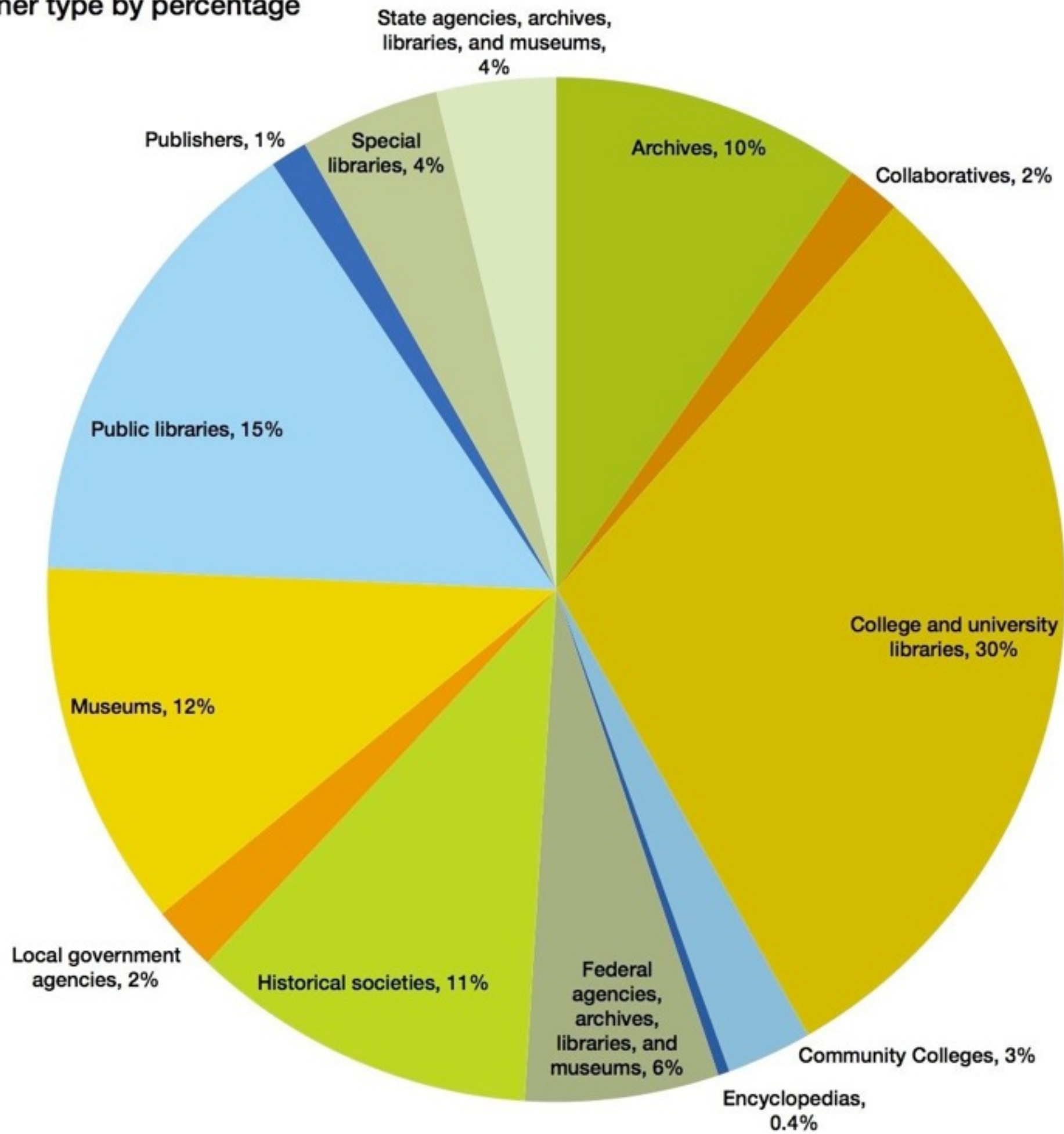


DIGITAL  
LIBRARY of GEORGIA  
SHARING GEORGIA'S HISTORY AND CULTURE ONLINE





Partner type by percentage





# ... leads to diverse metadata

5.6+ million records

9 metadata schemas

25 crosswalks representing 1,100

institutions' data

<http://bit.ly/dpla-crosswalks>

# Challenges of Aggregation

# Contributing institution

5 Dublin Core definitions

source, publisher, contributor,  
combo, external

4 MODS

2 MARC XML

5 local metadata application profiles

# Subject

- Commando automobile Neighborhoods Riot control United States. Army--African American troops King, Martin Luther, Jr., 1929-1968 United States. Army
- King, Martin Luther Jr., 1929-1968
- King, Martin Luther, 1899-1984
- King, Martin Luther, 1899-1984--Homes and haunts--Georgia--Atlanta
- King, Martin Luther, 1957-
- King, Martin Luther, III
- King, Martin Luther, Jr
- King, Martin Luther, Jr. 1929-1968
- King, Martin Luther, Jr., 1929-1968
- King, Martin Luther, Jr., 1929-1968--Arrest
- King, Martin Luther, Jr., 1929-1968--Assassination
- King, Martin Luther, Jr., 1929-1968--Birthplace
- King, Martin Luther, Jr., 1929-1968--Correspondence
- King, Martin Luther, Jr., 1929-1968--Death and burial
- King, Martin Luther, Jr., 1929-1968--Homes and haunts
- King, Martin Luther, Jr., 1929-1968--Homes and haunts--Georgia--Atlanta
- King, Martin Luther, Jr., 1929-1968--Imprisonment
- King, Martin Luther, Jr., 1929-1968--Influence
- King, Martin Luther, Jr., 1929-1968--Portrait
- King, Martin Luther, Jr., 1929-1968--Tomb
- King, Martin Luther, Jr., 1929-1968--Trials, litigation, etc
- King, Martin Luther, Jr., 1929-1968
- King, Martin Luther, Jr., 1929-1968--Anniversaries, etc
- King, Martin Luther, Jr., 1929-1968--Assassination
- Los Angeles--Streets--Santa Barbara Avenue (Martin Luther King)
- Martin Luther King, Jr. Day
- Martin Luther King, Jr. International Chapel Morehouse College (Atlanta, Ga.), Carter, Jimmy, 1924- , Southview Cemetery, Dorn, William Jennings Bryan (1916-2005), King, Coretta Scott, 1927-2006 , Omega Psi Phi Fraternity
- Martin Luther King, Jr., Day
- Martin Luther King, Jr., Federal Building (Atlanta, Ga.)
- Martin Luther King, Jr., Memorial (Washington, D.C.)
- Martin Luther King, Jr., National Historic Site (Atlanta, Ga.)
- Martin Luther King, Jr., Shaw AFB, S.C., National Negro Hymn, Newman, I. DeQuincey (Isaiah DeQuincey), 1911-1985
- Newman, I. DeQuincey (Isaiah DeQuincey), 1911-1985, King, Martin Luther, Jr., 1929-1968
- Orangeburg, S.C. Martin Luther King, Jr. Auditorium, Newman, I. DeQuincey (Isaiah DeQuincey), 1911-1985, Claflin College (Orangeburg, S.C.)
- President Crowley, Joy Crowley, Michael Coray, Ashok Dhingra, Patricia Miltenberger at 6th annual Dr. Martin Luther King, Jr., dinner
- United States--History--20th century Civil rights movements--United States Vietnam War, 1961-1975 Vietnam War, 1961-1975--Protest movements Vietnam War, 1961-1975--Protest movements--United States Vietnam War, 1961-1975--Public opinion Cold War Religion and politics Draft United States--Foreign relations--1945-1989 United States--Foreign relations--Asia Vietnamese reunification question (1954-1976) Vietnam War, 1961-1975--Personal narratives, American Vietnam--Politics and government War and society Youth and war Imperialism Vietnam War, 1961-1975--United States United States--History--1961-1969 King, Martin Luther, Jr., 1929-1968
- United States. Martin Luther King, Jr., Federal Holiday Commission
- United States. Task Force to Review the FBI Martin Luther King, Jr., Security and Assassination Investigations

# Spatial

Lack of context



Washington County

Boulder [UT or CO?]

Upstate [SC or NY?]

Colorado City [AZ or TX?]

Relying on local field labels for meaning

<i>County (Getty TGN)</i>	• <a href="#">Clark</a> 
<i>State (Getty TGN)</i>	• <a href="#">Nevada</a> 

<b>Latitude</b>	42.536667
<b>Longitude</b>	-111.794139



# Data ambiguity

What works locally may not work globally

“Paris”

Over 150 hits in geonames

<http://www.geonames.org/2988507>

Martha Paris with jar of honey. Georgia State University Libraries. Digital Library of Georgia.

(Wrong Paris. Sorry, Martha.)



# Addressing Challenges (today):

Standardizing Data

# Controlled vocabularies

Names/agents

ULAN, LCNAF, VIAF, DBpedia, EAC-CPF

Spatial

Geonames.org, DBpedia, TGN

Subject

LCSH, MeSH, DBpedia

Format

AAT, LCSH, TGM

Extended Date/time Format - LC



Leo J. "Scoop"  
Leeburn, 1950. Idaho  
State Historical Society.  
Mountain West Digital  
Library

# Extending local/regional vocabularies

Examples:

Gazeteers for a region's place names

Parishes for Catholic Church collections

Wards for Mormon Church collections

# Data Entry Workflows

Document and apply them.



Set of weights, standards for weight presented to UNC in 1883. Wilson Library, University of North Carolina, Chapel Hill. North Carolina Digital Heritage Center.



# Challenge ourselves

Free your metadata  
CC0\*/Public Domain

Share your resources  
CC licenses\*\*



Silo and barns Virginia Normal  
and Industrial Institute, 1916.  
University of Virginia Library

\*<http://creativecommons.org/publicdomain/zero/1.0/>

\*\*<http://creativecommons.org/choose/>

# Addressing Challenges (today? tomorrow? soon?)

With Linked Open Data

# Challenge ourselves

Use machine-readable values (too)

- Include IDs and URIs; coordinates

1 2 3 4 W M 0 1 5 0 3 Un 0 6 12 0 6 12 Me NH VT OH MCH IA SD  
5 6 7 8 B F 10 15 18 1 4 S 1 7 13 1 7 13+ MAS RI CT IND WIS MO NBR  
1 2 3 4 Ch 20 21 25 30 3 MO 2 8 14 2 8 N NY NJ PA ILL MIN ND KAN  
5 6 7 8 Jp 35 40 45 50 2 MI 3 9 15 3 9 F ND NW WVA NY TEN ALA CLF  
1 2 3 4 In 55 60 65 70 3 Wd 4 10 16 4 10 DEL AK NC SC MS LA TEX OK OK LX WSH  
5 6 7 8 75 80 85 90 95+ Un D 5 11 17 5 11 DC SA FLA DEL IT CA ARK OK IDA NEV  
1 2 3 4 En OK 0 a 4 17 11 5 Un 15 2 0 US Un En US Un En UTA AZ  
5 6 7 8 Ot NR 1 b 5 Ot 12 6 NG 20+ 3 1 Gr Ir Sc Gr Ir Sc NM PMP COL  
1 2 3 4 2 NW 4 c 6 0 13 7 1 Va 4 Au Sw CE Wa Sw CE Wa WYO MNT  
5 6 7 8 4 0 7 d 7 1 14 8 2 Po 5 Sz Nw CF Hu Nw CF Hu ALA AR  
1 2 3 4 6 12 10 e 8 2 15 9 3 Al 6 Po Ok Fr It Ok Fr It Au SEA  
5 6 7 8 8+ Un g f 9 3 16 10 4 Jn 10 Ot Ru Bo Ot Ru Bo Sz Po NS

Hollerith Punch Card for the 1900 Census of Population, Replica, ca1950.  
National Museum of American History. Smithsonian Institution

# DPLA Data structure

Stored in JSON-LD

Based on DPLA MAP specifications  
(<http://dp.la/info/map>)

JSON-LD is a method of transporting  
LOD in a human readable, structured  
data interchange format

JSON-LD needs URIs in addition to or instead of string values to make the next jump truly "linkable" data.

OR Hubs/DPLA need identifiable (standardized and qualified) string values to build systems to "pair" them to URIs.



```
"subject":[
  {
    "name":"King, Martin Luther,. Jr., 1929-1968"
  },
  ],
```

```
"mods:subject":[
  {
    "authority":"lcnaf",
    "mods:topic":"King, Martin Luther, Jr.,
1929-1968"
  },
  ],
```

```
"mods:subject":[
  {
    "authority":"lcnaf",
    "mods:topic":"King, Martin Luther,. Jr., 1929–1968"
    "@id":"http://id.loc.gov/authorities/names/n79084324"
  },
  ],
```

@prefix dcterms <<http://purl.org/dc/terms/>> .

@prefix i <<http://dp.la/item/>> .

@prefix lcnaf <<http://id.loc.gov/authorities/names/>> .

@prefix geonames <<http://www.geonames.org>> .

i:9f695d8c16a4978061a25076 dcterms:subject **lcnaf:n79084324** .

i:9f695d8c16a4978061a25076 dcterms:spatial geonames:4183143 .

# LOD-interpretive interfaces

Sooner rather than later

Visible results of work on LOD entries

Yield the true power of LOD



# Implications for Users

Great precision

Thorough discoverability

Localization into multiple languages

Authority for topic

# Implications for DPLA Hubs (and other Digital Repositories)

Less monthly cleanup - need to touch  
data fewer times

Preparing data for aggregation means  
better data at home

Precision and ease:

e.g., don't have to specify the entire  
hierarchy/coordinates of a place

[Web](#)[Images](#)[Videos](#)[Books](#)[News](#)[More ▾](#)[Search tools](#)

About 253,000,000 results (0.43 seconds)

[Martin Luther King, Jr. - Wikipedia, the free encyclopedia](#)[en.wikipedia.org/wiki/Martin\\_Luther\\_King,\\_Jr.](#) ▾ Wikipedia ▾

**Martin Luther King, Jr.** (January 15, 1929 – April 4, 1968) was an American pastor, activist, humanitarian, and leader in the African-American Civil Rights ...

[Assassination of Martin Luther ...](#) - [I Have a Dream](#) - [James Earl Ray](#) - [Bernice King](#)

[Martin Luther King Jr. - Biographical - Nobelprize.org](#)[www.nobelprize.org/nobel\\_prizes/peace/.../king-bio.html](#) ▾ Nobel Prize ▾

**Martin Luther King, Jr.**, (January 15, 1929-April 4, 1968) was born Michael Luther King, Jr., but later had his name changed to Martin. His grandfather began the ...

[Martin Luther King Jr. - The Biography Channel](#)[www.biography.com/.../...](#) ▾ The Biography Channel ▾

Sep 28, 2011

**Martin Luther King Jr.** led the U.S. Civil Rights Movement from the mid-1950s until his assassination in 1968 ...

[News for martin luther king jr](#)[Did the Elites Have Rev. Martin Luther King, Jr. Killed?](#)

CounterPunch - 1 day ago

But there are credible theories of a conspiracy, possibly involving US Army intelligence, whose role in the life and death of **Martin Luther King** ...

[Rediscovering a Martin Luther King Jr. speech inspired by a Lincoln procla...](#)

PBS NewsHour - 4 days ago

[Martin Luther King Jr. at UC Berkeley ... are you in these photos?](#)

## Martin Luther King, Jr.

Civil rights activist

Martin Luther King, Jr. was an American pastor, activist, humanitarian, and leader in the African-American Civil Rights Movement. [Wikipedia](#)

**Born:** January 15, 1929, Atlanta, GA**Assassinated:** April 4, 1968, Memphis, TN**Spouse:** [Coretta Scott King](#) (m. 1953–1968)**Awards:** [Nobel Peace Prize](#), [Time's Person of the Year](#), [More](#)**Children:** [Dexter Scott King](#), [Yolanda King](#), [Martin Luther King III](#), [Bernice King](#)**Education:** [Boston University](#) (1954–1955), [More](#)

People also search for

# Getting Started

(Today)

# Where to Start

Clean your data - prepare for reconciliation

Tools for the job

- OpenRefine

- (transitioning from Google Refine)

- Data Wrangler

- GREL/regex

# Where to Start

*For potential Service Hubs:*

Crosswalking many  
institutions

Single feed to DPLA

Sustainable model for digital  
libraries



People on a crosswalk in front of the arches of the Los Angeles International Airport, ca. 1974. California Historical Society. University of Southern California Libraries.

DPLA (all Hub data)

Spatial data ID/coordinates assigned

Rights statements namespace

Minnesota Digital Library

Machine-pairing subject strings to LC URIs

Digital Commonwealth

Subject, format, spatial data with authority IDs/URIs

Mountain West Digital Library

Regional Gazetteer for six states