



Improving Metadata and Reuse Across the Network

Mark A. Matienzo <mark@dp.la>

Digital Public Library of America

New York University — Advanced Archival Description

November 23, 2015

What is DPLA?



DPLA is...

A ***portal*** for discovery

A ***platform*** to build on

A strong advocate for a ***public option***

A network of ***partnerships***

A Portal for Discovery



A Wealth of Knowledge

explore 8,007,019 items from libraries, archives, and museums



Exhibitions

[View all »](#)



Explore
by Place

[Map »](#)

Explore by Date

[Timeline »](#)



1946 1947 1948 **1949** 1950 1951 1952

Apps

The DPLA is a platform. Developers make apps that use the library's data in many different ways. Here are just a few. [App Library »](#)

News



[DPLA Brings National Attention to the Blue Earth County Historical Society](#)
Oct 2

Searching DPLA

[About](#)[Get Involved](#)[For Developers](#)[Help](#)[Follow](#)[Contact](#)[Donate](#)[Login](#)[Sign Up](#)

DIGITAL PUBLIC LIBRARY
OF AMERICA

[Home](#)[Exhibitions](#)[Map](#)[Timeline](#)[Bookshelf](#)[Apps](#)

View:



kittens



Search Results

[Save](#)[Share](#)

Your search for **kittens** returned 143 results.

Items per page: 10

Sort by: Relevance

[1](#) [2](#) [3](#) ... [14](#) [15](#)

Refine

By Format

[image](#) 97[text](#) 41[sound](#) 1

Contributing Institution

[Boston Public Library](#) 14

IMAGE

Kittens

Jones, Leslie, 1886-1967

Title from information provided by Leslie Jones or the Boston Public Library on the negative or negative sleeve.. Date supplied by cataloger.

[View Object](#)

IMAGE

<http://dp.la/search/>

Timeline

Your search for **philadelphia** returned 99,164 results.

Only results with time data are shown below.

Show >

1000 1100 1200 1300 1400 1500 1600 1700 1800 1900 2000

View: Decades Years

1994

41 items

IMAGE

Actors Kerry O'Malley and Mal White in a scene from the Philadelphia Drama Guild's production of the play "The Plough And the Stars." (Philadelphia)

Swope, Martha

SW-274

View Object 



IMAGE

Actors Madeleine Potter and Des Keogh in a scene from the Philadelphia Drama Guild's production of the play "The Plough And the Stars." (Philadelphia)

Swope, Martha



<http://dp.la/timeline/>

Bookshelf

Your search for **philadelphia** returned 46,358 results.

Only book and periodical results are shown below.

Refine

Contributing Institution

University of Michigan	7944
University of California	7621
New York Public Library	6439
Library of Congress	6016
Harvard University	4929

[More »](#)

Partner

HathiTrust	46210
United States Government Printing Office (GPO)	108
Digital Library of Georgia	20
The Portal to Texas History	15
South Carolina Digital Library	5

Sort by: Relevance



About the DPLA Bookshelf

The bookshelf is an easy way to search DPLA's books, serials, and journals. The darker the shade of blue, the more relevant the results. Click on a spine for details and related images. Book thickness indicates the page count, and the horizontal length reflects the book's actual height.

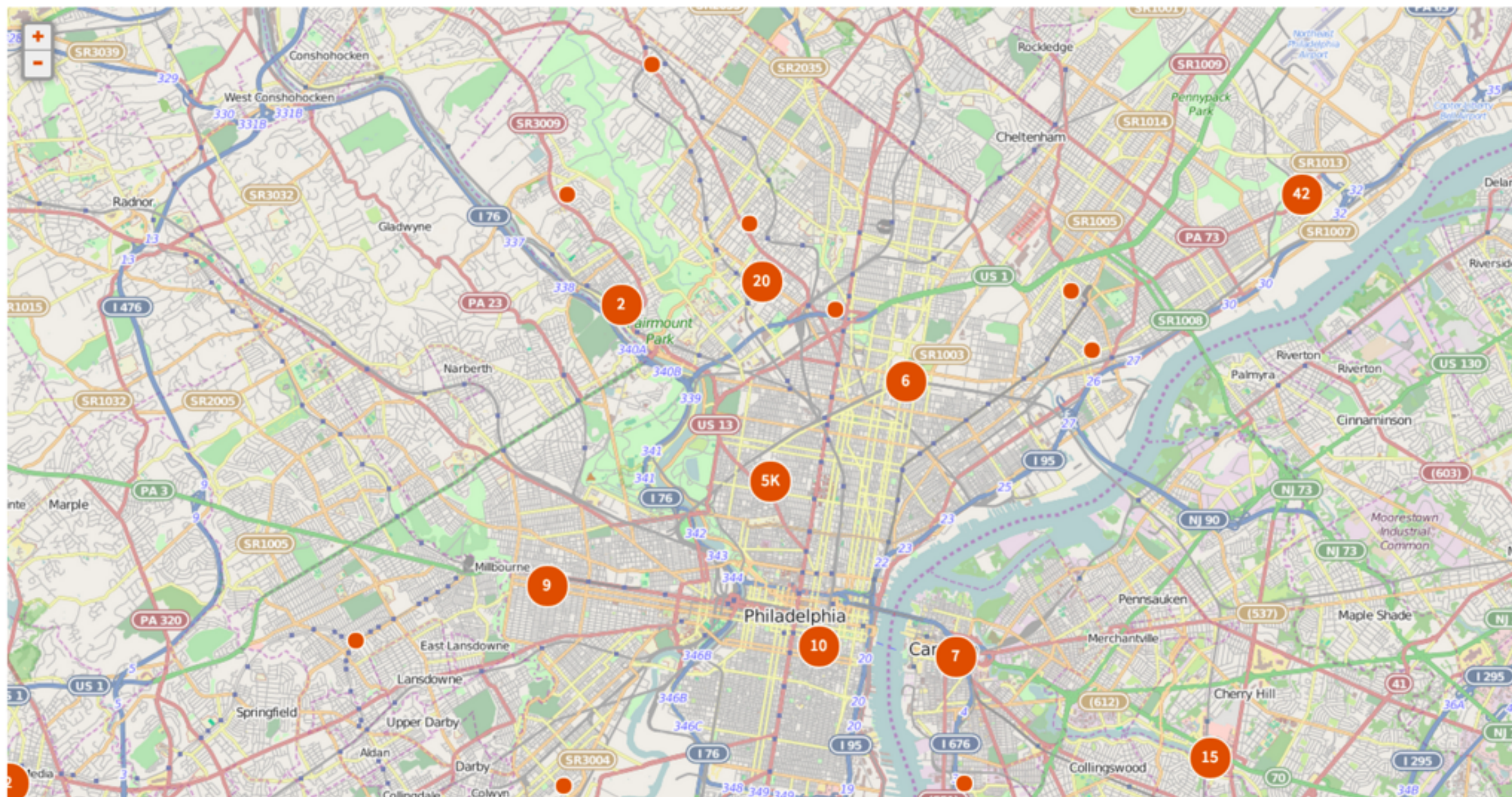
<http://dp.la/bookshelf/>

Map view

Your search for **philadelphia** returned 99,164 results.

Only results with location data are shown below.

Show >



<http://dp.la/map/>

Exhibitions



Staking Claims: The Gold Rush in Nineteenth-Century America



The Show Must Go On! American Theater in the Great Depression



Leaving Europe: A new life in America



Boston Sports Temples



History of Survivance: Upper Midwest 19th Century Native American Narratives



America's Great Depression and Roosevelt's New Deal



Indomitable Spirits: Prohibition in the United States



Activism in the U.S.

A Platform to Build On

DPLA
hack!



<http://dp.la/apps/>

A Strong Public Option



Andrew Carnegie and wife. Courtesy Boston Public Library via Digital Commonwealth

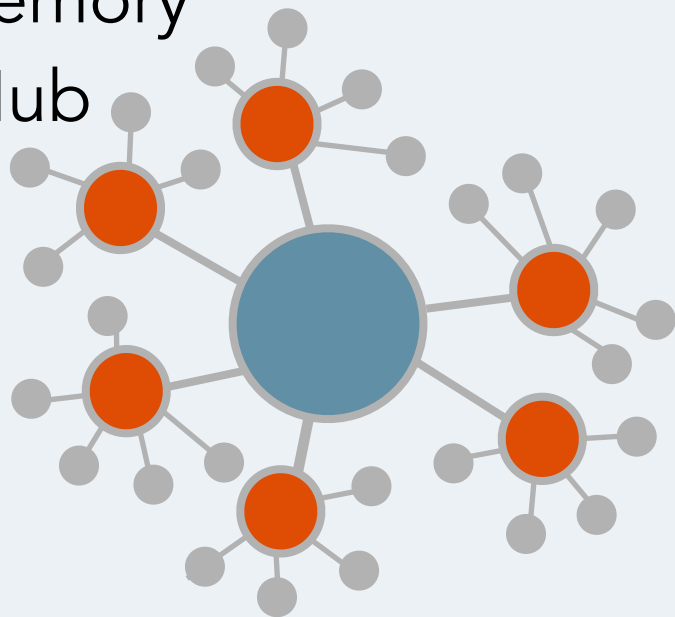
Partnerships



Ted Shawn and Hazel Walleck. Courtesy The New York Public Library.

Service Hubs

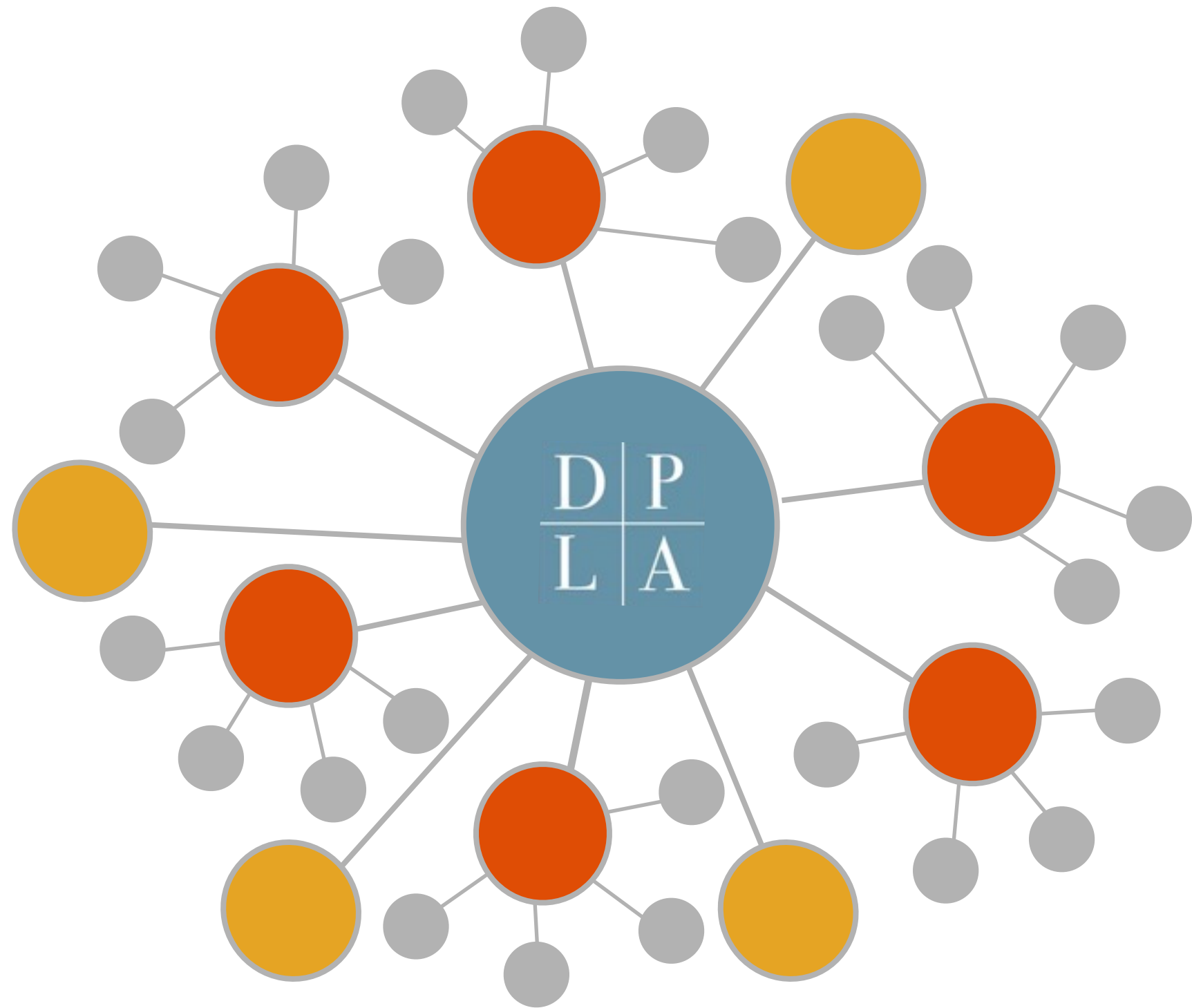
Mountain West Digital Library
South Carolina Digital Library
Empire State Digital Network
Digital Commonwealth (MA)
NC Digital Heritage Center
The Portal to Texas History
Digital Library of Georgia
Minnesota Digital Library
Kentucky Digital Library
Indiana Memory
Missouri Hub



Content Hubs

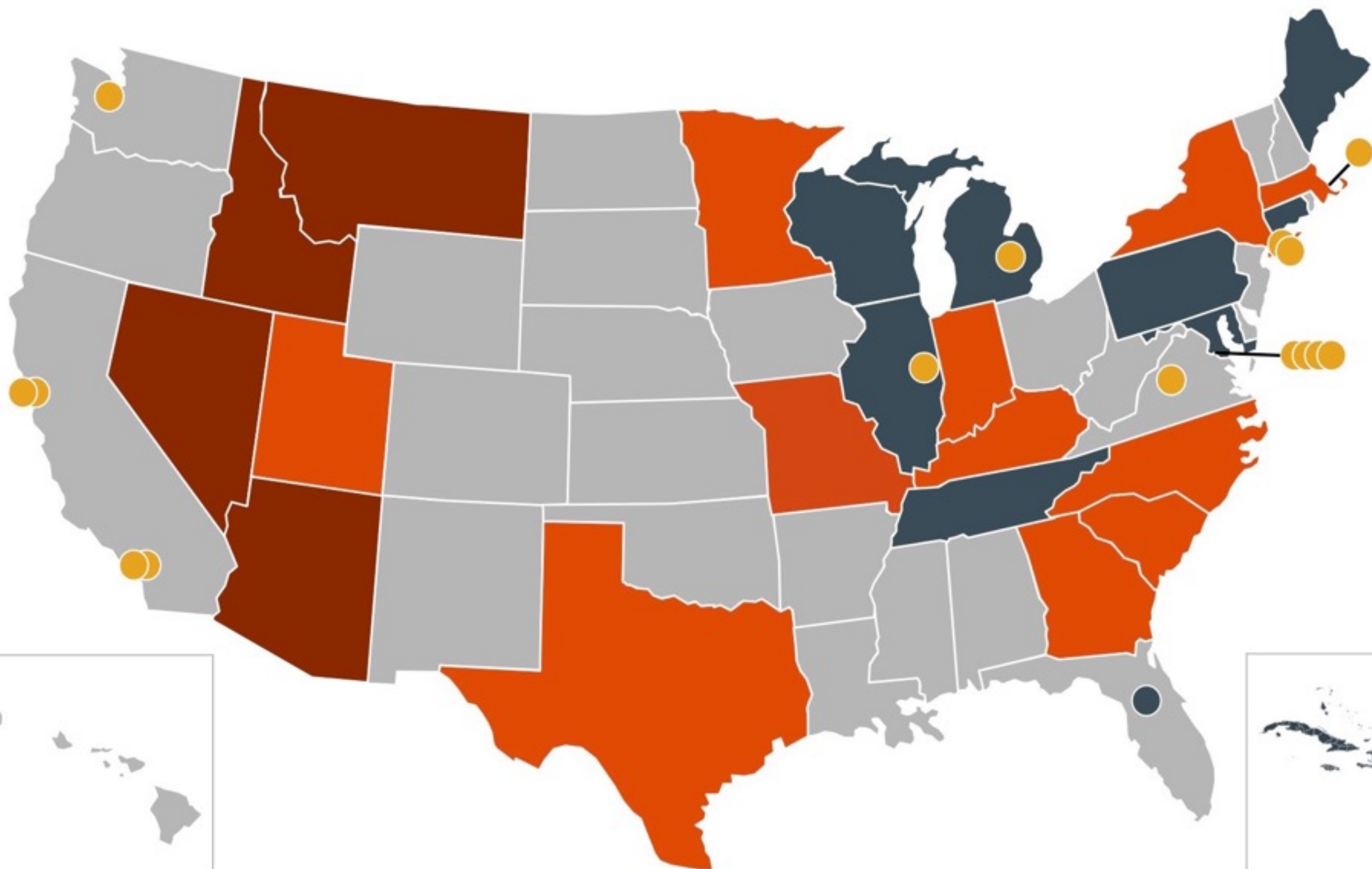
Government Publishing Office (GPO)
National Archives & Records Admin.
Univ. of Illinois Urbana-Champaign
University of Southern California
Biodiversity Heritage Library
The New York Public Library
Harvard University Library
California Digital Library
Smithsonian Institution
University of Virginia
J. Paul Getty Trust





DPLA Hubs locations

- Service Hubs
- Service Hub partner states
- Content Hubs
- Hubs in active development



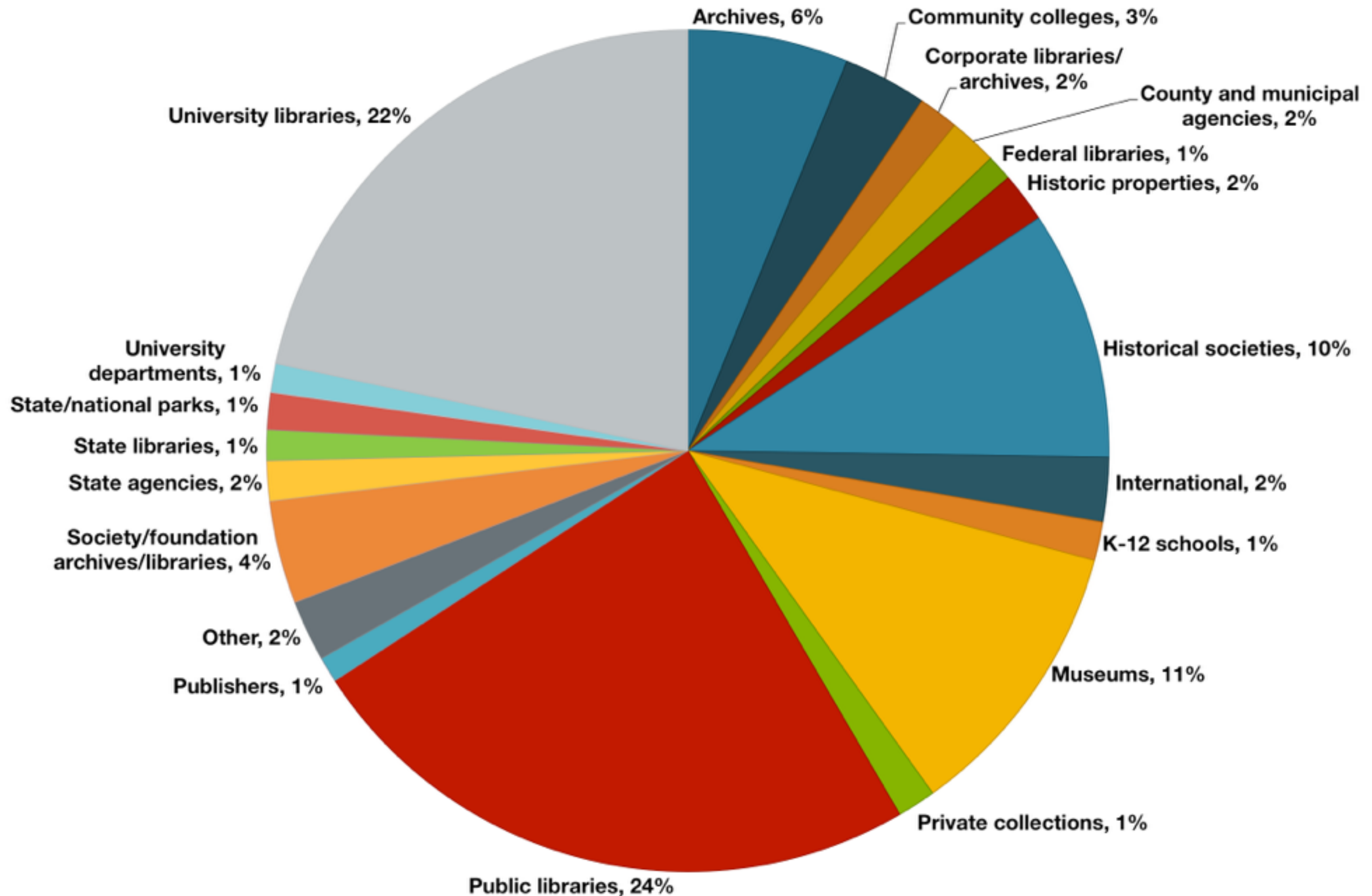
Hub Math

15 Content Hubs
+ 11 Service Hubs

1,600 Partners

Soon to be 19 Service Hubs and 16 Content Hubs

DPLA Partners



Partner diversity...

... leads to **diverse metadata**

11.4 million records

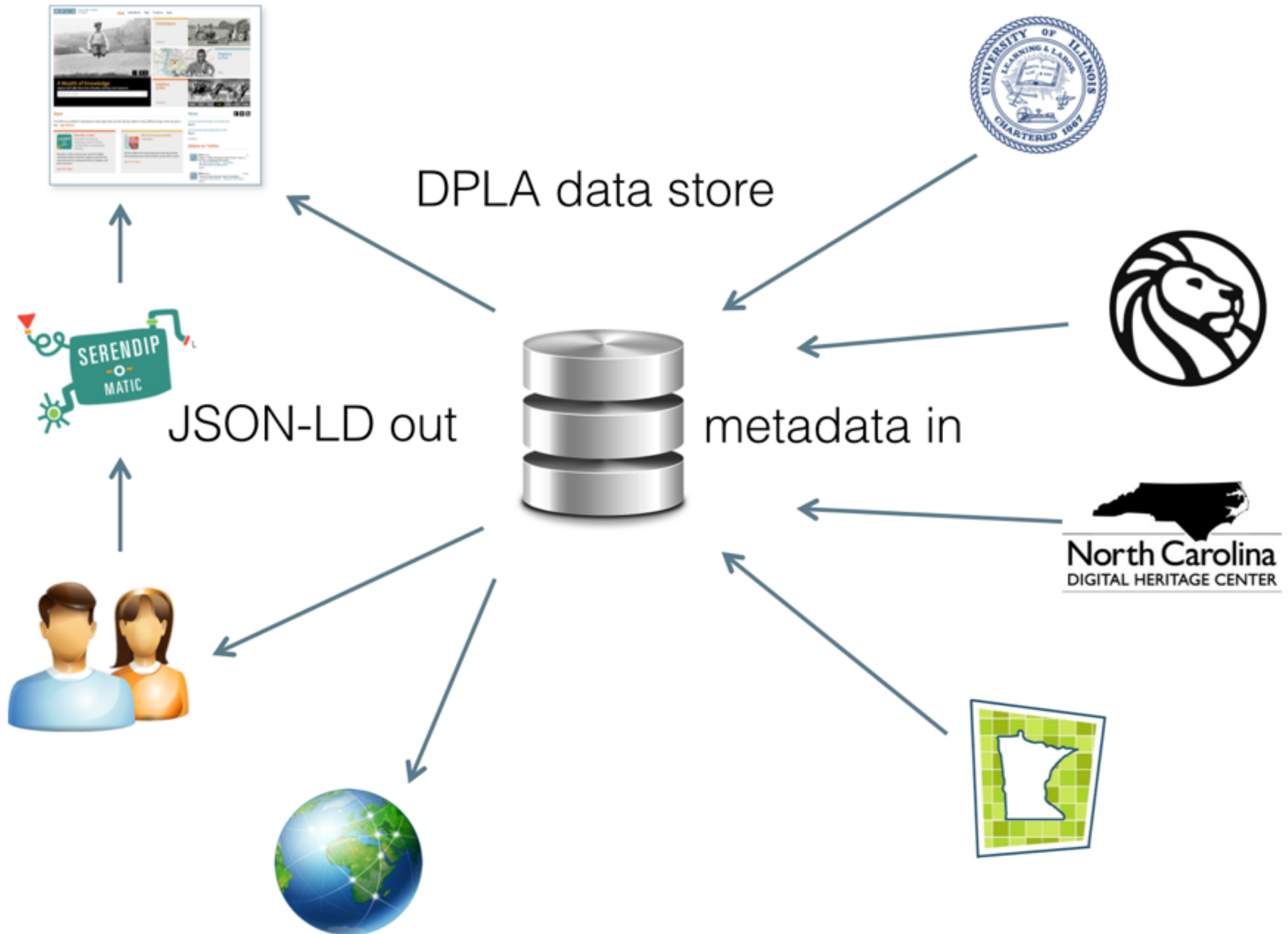
9 metadata schemas

27 crosswalks representing 1,600+ institutions' data

... leads to challenges

We need to understand what the metadata is trying to tell us before we can transform, normalize, or enrich.

The DPLA Ingestion Process



Harvesting Metadata



Leslie Jones, "Farmer with harvester." Boston Public Library

<subject>Dicotyledonae</subject>

<dc:subject>figures (representations); trees;
vines</dc:subject>

<dc:subject>Korea—History-- Japanese occupations,
1910-1945</dc:subject>

<dc:identifier>

KADA-shyun03-002~1; KADA-shyun03-002~2; KADA-
shyun03-002~3; KADA-shyun03-002~4; KADA-
shyun03-002~5; KADA-shyun03-002~6; KADA-
shyun03-002~7; KADA-shyun03-002~8

</dc:identifier>

<place>

<placeTerm>New York</placeTerm>

</place>

DPLA is **more than a data harvester**

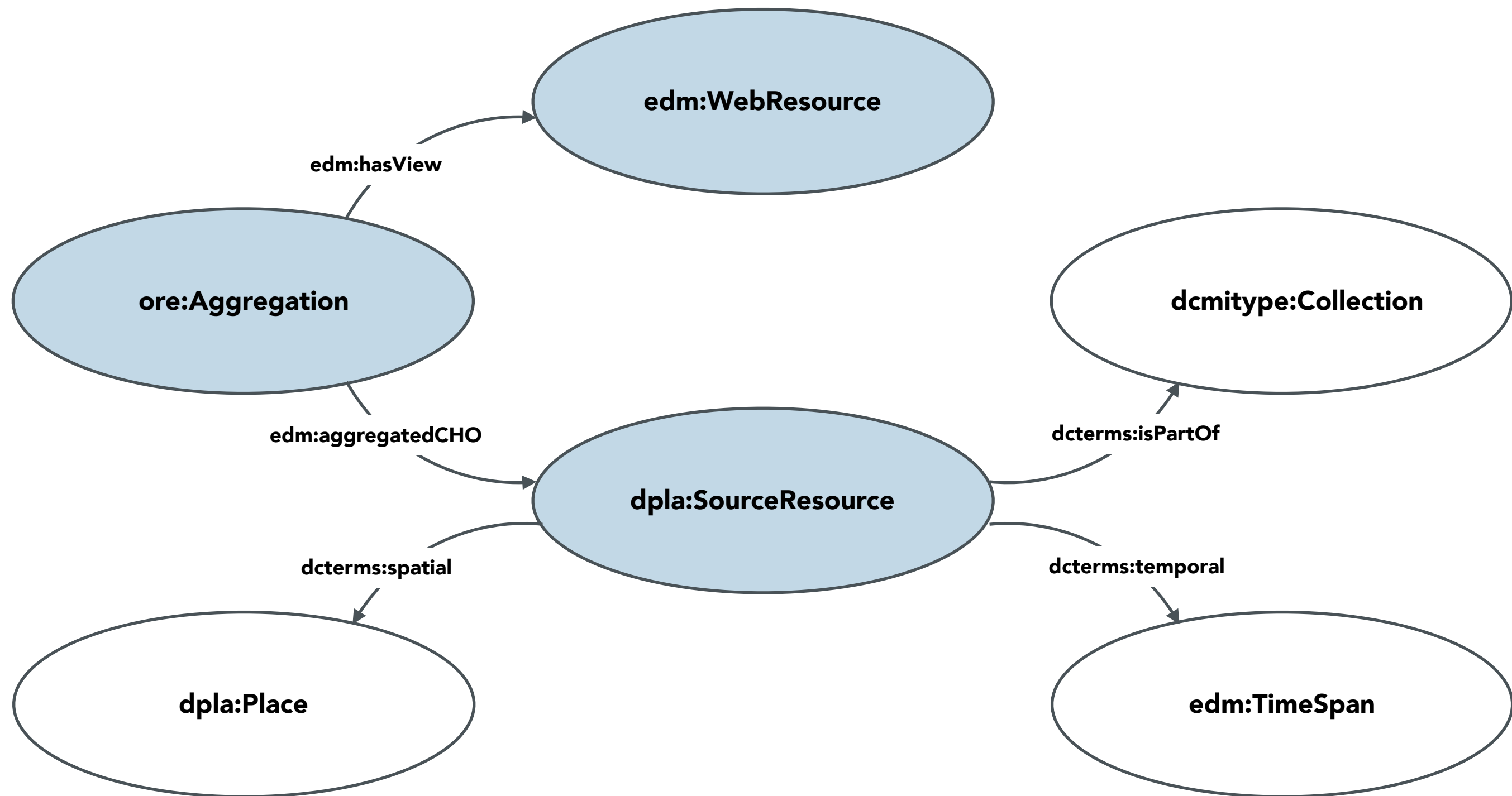
The **better** your data is,
the more **useful** it becomes,
and the more it is **used**

Transforming Metadata



Image Credit: [Sham Hardy](#)

Metadata Application Profile



Value	DPLA MAP equivalent	Obligation
Contributing Institution's name	Aggregation.dataProvider	Required when available
Collection title	Collection.title	Required when available
Date of creation	SourceResource.date	Strongly recommended
Hub's name	Aggregation.provider	Required
Link to original record	Aggregation.isShownAt	Required
Link to thumbnail	Aggregation.object	Required when available
"Place" of resource	SourceResource.spatial	Strongly recommended
Rights of resource	SourceResource.rights	Required
Subject of resource	SourceResource.subject	Strongly recommended
Title of resource	SourceResource.title	Required
Type of resource (from Dublin Core Type Vocabulary)	SourceResource.type	Required when available

JSON-LD

```
{
  "@id": "http://dp.la/api/items/f301b671789898dbc7905db119f53db4",
  "id": "f301b671789898dbc7905db119f53db4",
  "_id": "digitalnc--urn:brevard.lib.unc.eduunc_dig_nccpa:oai:dc.lib.unc.edu:dig_nccpa/7197",
  "@context": "http://dp.la/api/items/context",
  "object": "http://dc.lib.unc.edu/utis/getthumbnail/collection/dig_nccpa/id/7197",
  "aggregatedCHO": "#sourceResource",
  "ingestDate": "2014-07-11T01:02:29.518959",
  "@type": "ore:Aggregation",
  "ingestionSequence": 1,
  "provider": {
    "@id": "http://dp.la/api/contributor/digitalnc",
    "name": "North Carolina Digital Heritage Center"
  },
  "isShownAt": "http://dc.lib.unc.edu/u?/dig_nccpa,7197",
  "ingestType": "item",
  "dataProvider": "University of North Carolina at Chapel Hill",
  "sourceResource": {
    "publisher": [
      [
        "University of North Carolina at Chapel Hill, Wilson Library, North Carolina Collection Photographi",
        "University of North Carolina at Chapel Hill"
      ]
    ],
    "rights": "The current policies of the North Carolina Collection define the conditions of use. See http",
    "description": [
      "Series 2. North Carolina Scenes/Subjects,circa 1837-1908."
    ],
    "format": "Images",
    "@id": "http://dp.la/api/items/f301b671789898dbc7905db119f53db4#sourceResource",
    "title": [
      "Folder 0038: Civil War: New Hanover County: Fort Fisher: The War in America: Interior of Fort Fisher"
    ],
    "collection": {
      "description": "",
      "@id": "http://dp.la/api/collections/16b3df7b857ee0b2a12bd733dd8096d2",
      "id": "16b3df7b857ee0b2a12bd733dd8096d2",
      "title": "Digital North Carolina Collection Photographic Archives"
    },
    "isPartOf": [
      "P0005"
    ],
    "stateLocatedIn": [
      {
        "name": "North Carolina"
      }
    ],
    "identifier": [
      "P0005_0038.tif",
      "http://dc.lib.unc.edu/u?/dig_nccpa,7197"
    ],
    "type": "image"
  }
}
```

The DPLA Mapping Process

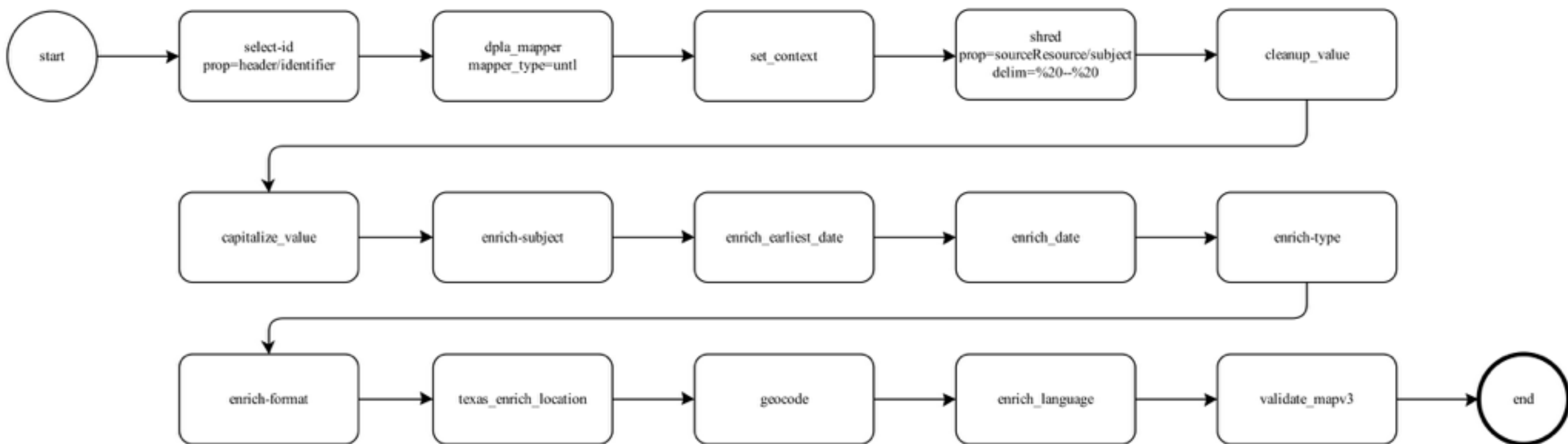
DPLA Partners Crosswalk, MAP v4.0

<http://bit.ly/dpla-MAP4-crosswalk>

DPLA metadata application profile

			QDC	MARC	MODS
Class	Label	Property	University of Washington	University of Florida	Tennessee
dpla:SourceResource	Alternative Title	dcterms:alternative	none	none	<titleInfo><titletype="alternative">
dpla:SourceResource	Collection	dcterms:isPartOf	oai set info	see below	<relatedItem type="host" displayLabel="Project"><titleInfo><title>
dpla:SourceResource	Contributor	dcterms:contributor	none	700/710/711/720\$a when there is not a \$e that says "joint author," "jt author," "aut" or "cre" – i.e. sometimes there will only be \$a, only exclude is those \$e values exist	<name><namePart>[value]</namePart><role><roleTerm>Contributor</roleTerm></name>
dpla:SourceResource	Creator	dcterms:creator	dc:creator	Take the value of \$a for 100, 110, 111 and the value of \$a for 700 if the value of \$e is "joint author" or "jt author" ... so \$a only is actually mapped	<name><namePart>[value]</namePart><role><roleTerm>Creator</roleTerm></name>
dpla:SourceResource	Date	dc:date	first dc:date	260\$c	<originInfo><dateCreated> (use all instances. Some have qualifier indicating EDTF format and keyDate, others are qualified as "approximate")
dpla:SourceResource	Description	dcterms:description	dc:description	5XX; not 506, 538, 536, 535, 533, 510, 540	<abstract>
dpla:SourceResource	Extent	dcterms:extent		300a; 300c; 340b	<physicalDescription><extent
dpla:SourceResource	Format	dc:format		007 position 00 [see http://www.loc.gov/marc/bibliographic/ position 06 in Leader [see "06 - Type	

Transformation & enrichment



Sample pipeline for Portal to Texas History

Fort Raleigh Grill, Elizabeth City, N.C.



[View Object](#) 

Created Date ca. 1947

Partner North Carolina Digital Heritage Center

Contributing Institution University of North Carolina at Chapel Hill

Publisher North Carolina Collection Photographic Archives,
Wilson Library, University of North Carolina at
Chapel Hill

Location Elizabeth City (N.C.)
Pasquotank County (N.C.)

Type image

Subject [Postcards--North Carolina](#)

Rights This item is presented courtesy of the North Carolina Collection, UNC-Chapel Hill, for research and educational purposes. Prior permission from the North Carolina Collection is required for any commercial use.

URL http://dc.lib.unc.edu/u/?/nc_post,4365

Subject Topical Postcards--North Carolina.

```
"subject": [  
  {  
    "name": "Postcards--North Carolina"  
  }  
]
```

Location	Elizabeth City (N.C.) Pasquotank County (N.C.)
----------	---

```
"spatial": [  
  {  
    "county": "Pasquotank County",  
    "name": "Elizabeth City (N.C.)",  
    "state": "North Carolina",  
    "coordinates": "36.3014984131, -76.2197570801",  
    "country": "United States"  
  },  
  {  
    "county": "Pasquotank County",  
    "name": "Pasquotank County (N.C.)",  
    "state": "North Carolina",  
    "coordinates": "36.2649002075, -76.2491378784",  
    "country": "United States"  
  }  
]
```


Simple things make
your metadata
better.

Be a **copy cat**.



Courtesy The New York Public Library

Search Results

Save Share

Your search for **sloth** returned 53 results.

Refined by: image

Items per page: 10

Sort by: Relevance

1 2 3 ... 5 6

Refine

By Format

image

Contributing Institution

IMAGE

The Sloth

[View Object](#)



Populate required and recommended fields

George Arents Collection. The New York Public Library	1
NMNH - Anthropology Dept.	6
General Research Division. The New York Public Library	4
More »	

Sloth

Written on border: "ca. 1840s"

[View Object](#)



Partner	
The New York Public Library	29
University of Southern California. Libraries	10
Smithsonian Institution	10
North Carolina Digital Heritage Center	2
The Portal to Texas History	1

IMAGE

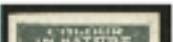
[Sloth.]

[View Object](#)



IMAGE

The sloth



Search Results

Your search for **sloth** returned 53 results.

Refined by: image

Items per page: 10

Sort by: Relevance

123...56

Refine

By Format

image

Contributing Institution

Art and Picture Collection. The New York Public Library17

California Historical Society8

George Arents Collection. The New York Public Library7

NMNH - Anthropology Dept.6

General Research Division. The New York Public Library4

More »

Partner

The New York Public Library29

University of Southern California. Libraries10

Smithsonian Institution10

North Carolina Digital Heritage Center2

The Portal to Texas History1

IMAGE

The Sloth

[View Object](#)



IMAGE

Sloth

Written on border: "ca. 1840s"

[View Object](#)



IMAGE

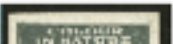
[Sloth.]

[View Object](#)



IMAGE

The sloth



Search Results

Your search for **sloth** returned 53 results.

Refined by: image

Items per page: 10

Sort by: Relevance

1 2 3 ... 5 6

Refine

By Format

image

Contributing Institution

Art and Picture Collection. The New York Public Library17

California Historical Society8

George Arents Collection. The New York Public Library7

NMNH - Anthropology Dept.6

General Research Division. The New York Public Library4

More »

Partner

The New York Public Library29

University of Southern California. Libraries10

Smithsonian Institution10

North Carolina Digital Heritage Center2

The Portal to Texas History1

IMAGE

The Sloth

[View Object](#)

IMAGE

Sloth

Written on border: "ca. 1840s

[View Object](#)

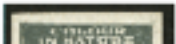
IMAGE

[Sloth.]

[View Object](#)

IMAGE

The sloth

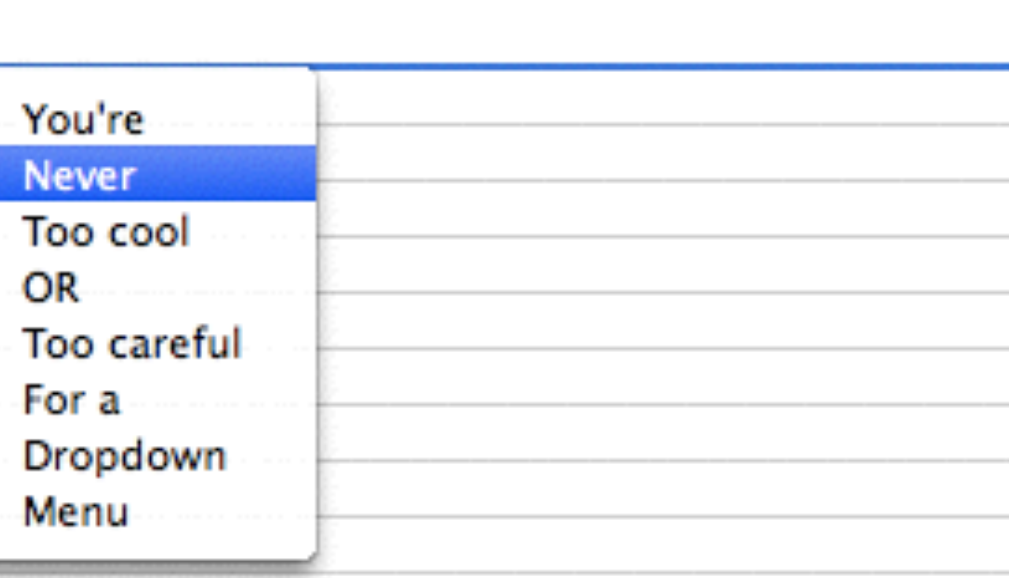


Be **consistent**, consistently.

edm:dataProvider	dc:contributor	dc:publisher	"David Rumsey" (hard coded)	"University of Southern California. Libraries" (hard coded). *Exception for set "chs," which should pull from the first dc:source value in each record.
-------------------------	----------------	--------------	-----------------------------	---

dc:publisher	dc:publisher	dc:publisher (when value is NOT "University of Southern California. Libraries"	dc:source	n/a
---------------------	--------------	--	-----------	-----

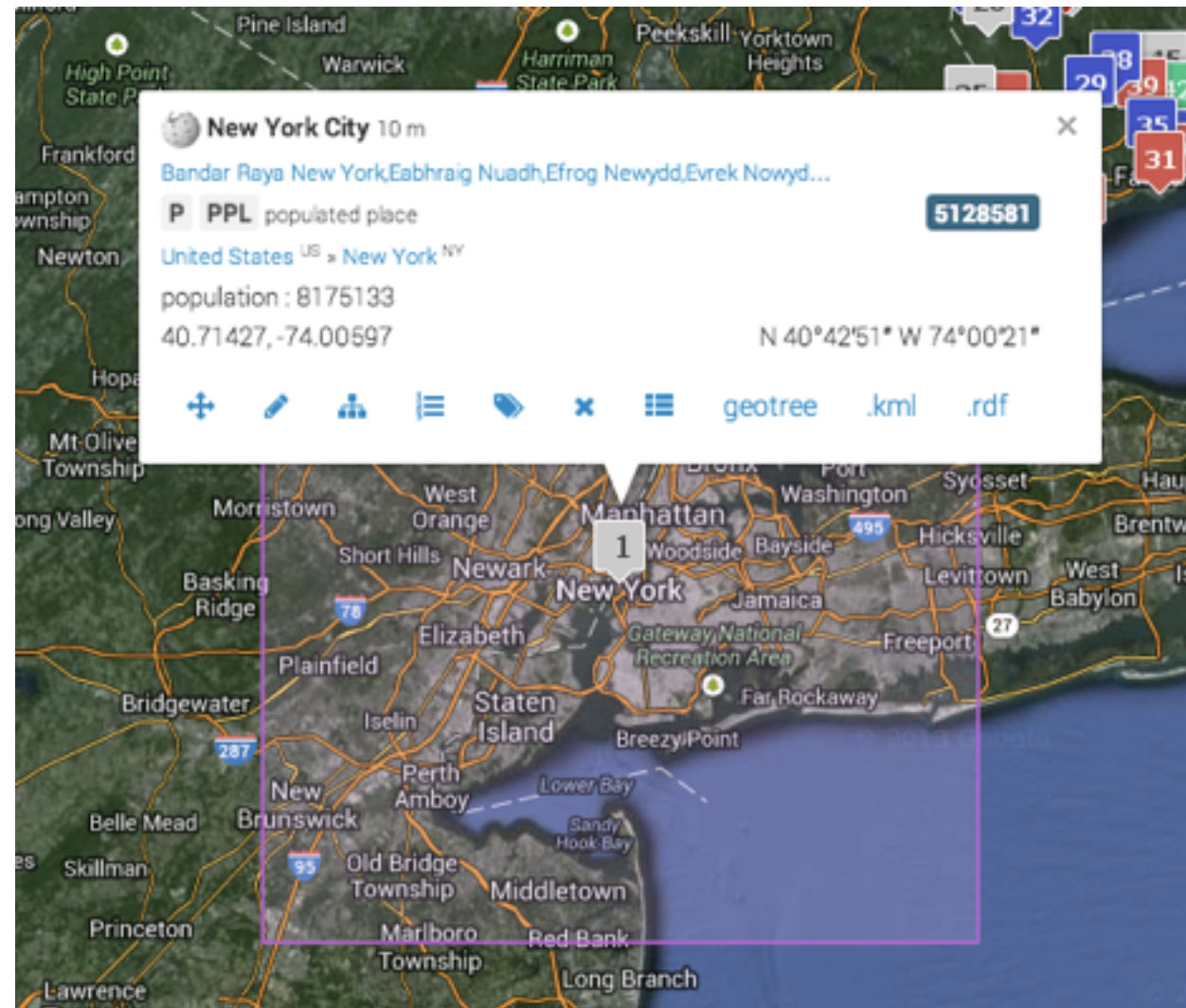
Control and **explain** yourself.



The screenshot shows a Google Sheet with a dropdown menu open. The menu is positioned over a cell in the first row. The menu options are: "You're", "Never", "Too cool", "OR", "Too careful", "For a", "Dropdown", and "Menu". The "Never" option is highlighted in blue, indicating it is the selected item. The background of the sheet is a light gray grid.


```
<place>  
  <placeTerm>New York</placeTerm>  
</place>
```

```
<place>
  <placeTerm valueURI="http://sws.geonames.org/5128581">
    New York</placeTerm>
</place>
```



Check your wok.

Standardizing Rights Statements

A photograph of a wooden wall with a sign that reads "REFUSE THINGS". The sign is made of white paper with blue letters, and the letters are cut out and pinned to the wall. The wall is made of vertical wooden planks. The sign is slightly tilted to the right. The word "REFUSE" is on the left and "THINGS" is on the right, with a gap between them. The letters are in a bold, sans-serif font. The background is a warm, reddish-brown color. There are some dark spots on the wall, possibly from nails or paint. The overall tone is serious and defiant.

REFUSE THINGS

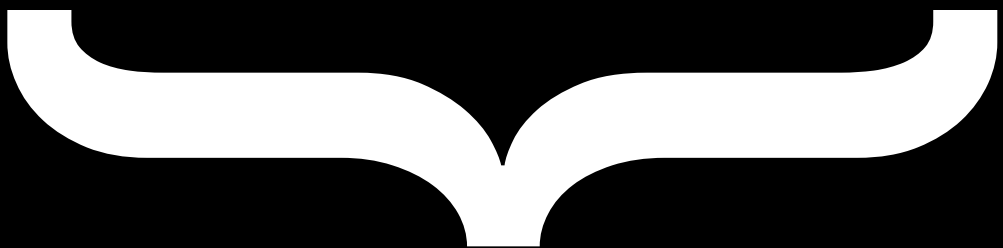
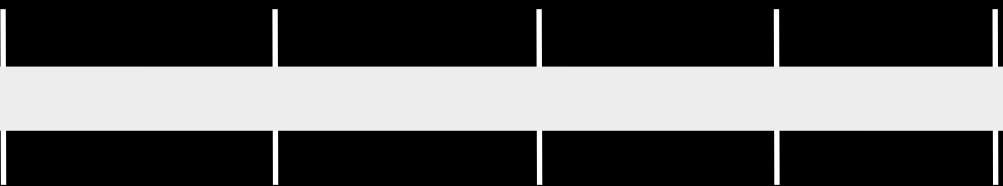
Image Credit: Orin Zebest (<https://www.flickr.com/photos/orinrobertjohn/5673450686>)

Statement of Problem

- Over 87,000 unique rights statements in DPLA
- Lack of standardized mechanism to express rights statements for digitized cultural heritage material
- Wide variety of *textual* rights statements make potential for reuse of digital objects unclear
- Metadata requirements for aggregators often lead to implementation of “boilerplate” statements

The problem ...

1989 2015



the web

The problem ...

1600

1700

1800

1900



other human expression

Rights statements in DPLA

- Analysis undertaken in October 2014 on 8.1 million records in the DPLA Metadata Application Profile
- Reports (CouchDB view) generated from DPLA's primary metadata repository to return JSON-encoded information
- Reports contained rights statements, Hub, and contributing institution, and count

Analysis process

- Transformed into CSV and imported into OpenRefine
- Associated institutions removed
- Normalized (whitespace and multipass clustering process) to reduce and aggregate near-duplicates
- Aggregated counts created based on method described by Morris 2013 (<http://bit.ly/morris2013>)

Findings

- Incredible diversity and absence of rights statements in 8.1 million DPLA records
- 87,610 “unique” rights statements after normalization
- ~1.01M (~12.5%) records missing rights statements
- ~2.4M (~29.9%) “in copyright/(c)/all rights reserved”
- ~1.6% under a Creative Commons license
- Rights statements matching multiple “categories” that could be confusing to end users

A shared rights framework

KF Knight News Challenge: Strengthening the Internet

'Getting It Right on Rights'



Dan Cohen
@DANCOHEN



Emily Gore
@NCSCHISTORY

WINNER
Digital Public Library of America
@DPLA

Knight News Challenge @knightfdn | #Newschallenge

Image credit: The John S. and James L. Knight Foundation

IRSWG Deliverables

- Shared framework/data model for rights statements under common namespace external to partners
- Best practice guidelines for aggregators and cultural heritage institutions to adopt rights statements
- Governance model to maintain framework

Working Group Contributors

- Paul Keller (co-chair, Kennisland)
- Marie-Claire Dangerfield (Europeana)
- Julia Fallon (Europeana)
- Ranu Gayadin (Europeana)
- Lucie Guibault (Inst. for Information Law)
- Antoine Isaac (Europeana)
- Lyubomir Kamenov (Europeana)
- Patrick Peiffer (B.N. Luxembourg)
- Joris Pikel (Europeana)
- Henning Scholz (Europeana)
- Maarten Zeinstra (Kennisland)
- Emily Gore (co-chair, DPLA)
- Greg Cram (New York Public Library)
- Karen Estlund (University of Oregon)
- Dave Hansen (University of North Carolina)
- Matt Lee (Creative Commons)
- Melissa Levine (University of Michigan)
- Mark Matienzo (DPLA)
- Diane Peters (Creative Commons)
- Amy Rudersdorf (DPLA)
- Richard Urban (Florida State University)

IRSWG Implementation

- Vocabulary modeled using RDF/Simple Knowledge Organization System
- Aligned with other rights vocabularies (e.g. PREMIS Copyright Status and Europeana rights framework)
- White paper on framework: <http://bit.ly/rsorg-whitepaper>
- Technical white paper: <http://bit.ly/rstech-whitepaper>

The Statements

- In Copyright
- In Copyright - EU Orphan Work
- In Copyright - Rightsholder(s) Unlocatable or Unidentifiable
- In-Copyright - Educational Use Permitted
- In Copyright - Non-Commercial Use Permitted
- Out Of Copyright - Non-Commercial Use Only
- No Copyright - Contractual Restrictions
- No Copyright - Jurisdiction-Specific
- No Copyright - Other Known Legal Restrictions
- No Known Copyright
- Copyright Not Evaluated

Design Considerations

- Rights statements under development are not licenses; try to reflect that in implementation
- Develop as an RDF vocabulary for broad reuse, following best practices on publication
- Considerations of extensibility
- Provide implementation guidelines for aggregators (e.g. DPLA and Europeana) and their partners

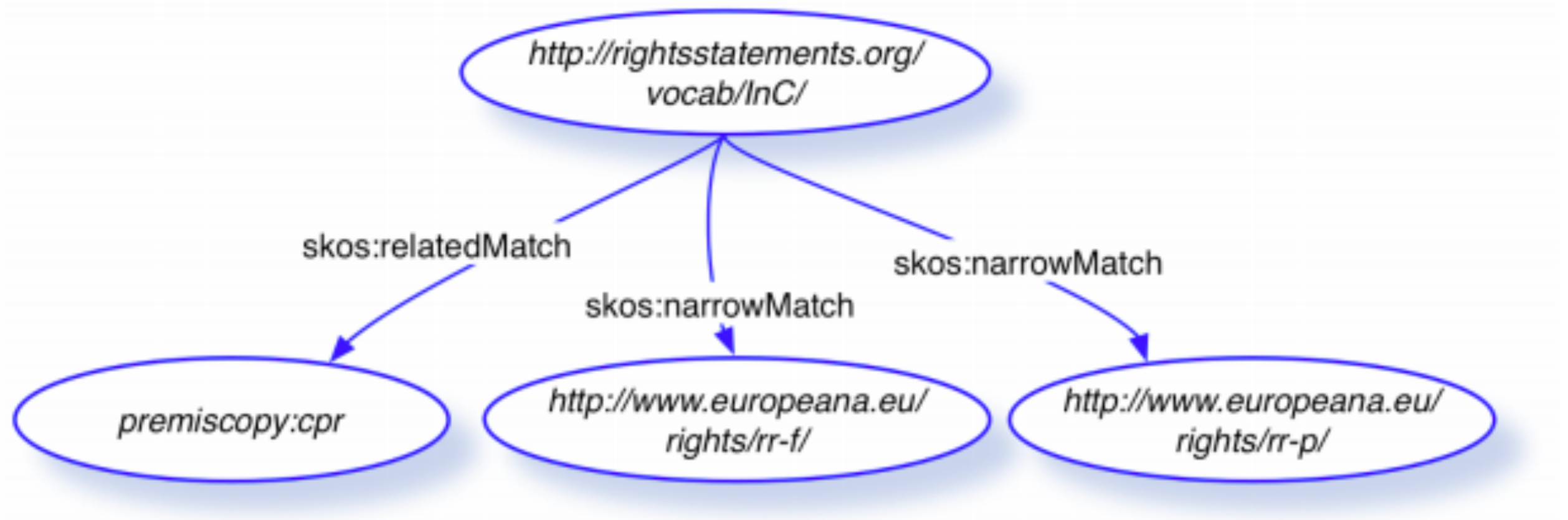
Data Model

- Published as a SKOS vocabulary:
<http://bit.ly/rights-data-model>
- All statements are modeled as both `skos:Concepts` and `dcterms:RightsStatements`
- Allows for interoperability with other frameworks for rights

Vocabulary Reuse

Source	Prefix abbreviation	Namespace
Creative Commons Rights Expression Language (ccREL)	cc:	http://creativecommons.org/ns#
Dublin Core Elements 1.1	dc:	http://purl.org/dc/elements/1.1/
DCMI Type Vocabulary	dcmitype:	http://purl.org/dc/dcmitype/
DCMI Metadata Terms	dcterms:	http://purl.org/dc/terms/
Europeana Data Model	edm:	http://www.europeana.eu/schemas/edm/
ODRL	odrl:	http://www.w3.org/ns/odrl/2/
PREMIS Copyright Status	premiscopy:	http://id.loc.gov/vocabulary/preservation/copyrightStatus/
SKOS	skos:	http://www.w3.org/2004/02/skos/core#
OWL	owl:	http://www.w3.org/2002/07/owl#
ODRS	odrs:	http://schema.theodi.org/odrs#

Vocabulary Alignment



Mapping to the data model

Short name of RS <code>dc:identifier</code>	Name of Rights Statement <code>skos:prefLabel</code>
URI for Rights Statement	
One sentence description of the Rights Statement. This will not be displayed as part of the Rights Statement. Intended for use in documents or on websites describing the Rights Statements. <code>not mapped</code>	
Text of the Rights Statement <code>skos:definition</code>	
Notices: One or more notices related to the Rights Statement <code>skos:note</code>	
Disclaimer regarding this being a Rights Statement and not a legally operative License summary. <code>not mapped</code>	
Generic selection criteria for the Rights Statement. Short text that describes when this Rights Statement should be used, aimed primarily at data providers. This text will not be displayed as parts of the Rights Statement. <code>skos:scopeNote</code>	
Extra metadata <code>not mapped</code>	<div><input type="checkbox"/></div> <p>For some statements it is possible to provide additional metadata which triggers the display of optional information at the text of the Rights Statement (and above the notices). If this is the case this will be noted here. Specific behavior is specified by keywords in bold as described by RFC2119.³⁴</p>

What You Can Do

- Once the framework is established, work with your digital collections and/or those of your partners to implement
- DPLA's plan for implementation will utilize Hubs Network to train the current 1,600+ DPLA contributing institutions

Thank You!

Mark A. Matienzo <mark@dp.la>
Digital Public Library of America