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?

BUILDING THE 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

http://dp.la/
Searching DPLA

Your search for kittens returned 143 results.

Refine

By Format

- image
  - 97
- text
  - 41
- sound
  - 1

Contributing Institution

- Boston Public Library
  - 14

http://dp.la/search/
Your search for philadelphia returned 99,164 results. Only results with time data are shown below.

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/
Your search for philadelphia returned 46,358 results. Only book and periodical results are shown below.

About the DPLA Bookshelf

The bookshelf is an easy way to search DPLA’s books, manuscripts, digital newspapers, maps, 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.
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.

http://dp.la/exhibitions/
A Platform to Build On

http://dp.la/apps/
A Strong Public Option

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

**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
- David Rumsey
- HathiTrust
- ArtStor
Soon to be 19 Service Hubs and 16 Content Hubs
Partner diversity...

... leads to **diverse metadata**

- 11.4 million records
- 9 metadata schemas
- 27 crosswalks representing 1,600+ institutions’ data
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
<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**
<table>
<thead>
<tr>
<th>Value</th>
<th>DPLA MAP equivalent</th>
<th>Obligation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contributing Institution’s name</td>
<td>Aggregation.dataProvider</td>
<td>Required when available</td>
</tr>
<tr>
<td>Collection title</td>
<td>Collection.title</td>
<td>Required when available</td>
</tr>
<tr>
<td>Date of creation</td>
<td>SourceResource.date</td>
<td>Strongly recommended</td>
</tr>
<tr>
<td>Hub’s name</td>
<td>Aggregation.provider</td>
<td>Required</td>
</tr>
<tr>
<td>Link to original record</td>
<td>Aggregation.isShownAt</td>
<td>Required</td>
</tr>
<tr>
<td>Link to thumbnail</td>
<td>Aggregation.object</td>
<td>Required when available</td>
</tr>
<tr>
<td>“Place” of resource</td>
<td>SourceResource.spatial</td>
<td>Strongly recommended</td>
</tr>
<tr>
<td>Rights of resource</td>
<td>SourceResource.rights</td>
<td>Required</td>
</tr>
<tr>
<td>Subject of resource</td>
<td>SourceResource.subject</td>
<td>Strongly recommended</td>
</tr>
<tr>
<td>Title of resource</td>
<td>SourceResource.title</td>
<td>Required</td>
</tr>
<tr>
<td>Type of resource (from Dublin Core Type Vocabulary)</td>
<td>SourceResource.type</td>
<td>Required when available</td>
</tr>
</tbody>
</table>
"id": "http://dc.lib.unc.edu/u?/dig_nccpa,7197",
"provider": {
  "@id": "http://dp.la/api/contributor/digitalnc",
  "name": "North Carolina Digital Heritage Center"
},
"isShownAt": "http://dc.lib.unc.edu/u?/dig_nccpa,7197",
"ingestDate": "2014-07-11T01:02:29.518059",
"@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 Photographs, University of North Carolina at Chapel Hill"
  ],
  "rights": "The current policies of the North Carolina Collection define the conditions of use. See http for more information."
},
"description": [
],
"format": "Images",
"@id": "http://dp.la/api/items/f301b671789898d8b7905db119f53db4#sourceResource",
"title": [
],
"collection": {
  "description": "",
  "@id": "http://dp.la/api/collections/16b3df7b857ee0b2a12b733dd8096d2",
  "id": "16b3df7b857ee0b2a12b733dd8096d2",
  "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

<table>
<thead>
<tr>
<th>Class</th>
<th>Label</th>
<th>Property</th>
<th>QDC</th>
<th>MARC</th>
<th>MODS</th>
</tr>
</thead>
<tbody>
<tr>
<td>dpla:SourceResource</td>
<td>Alternative Title</td>
<td>dc:terms:alternative</td>
<td>none</td>
<td>none</td>
<td>&lt;titleInfo&gt;&lt;titletype=&quot;alternative&quot;/&gt;</td>
</tr>
<tr>
<td>dpla:SourceResource</td>
<td>Collection</td>
<td>dc:terms:isPartOf</td>
<td>oai set info</td>
<td>see below</td>
<td>&lt;relatedItem type=&quot;host&quot; displayLabel=&quot;Project&quot;&gt;&lt;titleInfo/&gt;</td>
</tr>
<tr>
<td>dpla:SourceResource</td>
<td>Creator</td>
<td>dc:terms:contributor</td>
<td>none</td>
<td>700/710/711/720$ when there is not a $e that says &quot;joint author,&quot; &quot;jt author,&quot; &quot;aut&quot; or &quot;cre&quot; – i.e. sometimes there will only be $a, only exclude is those $e values exist</td>
<td>&lt;name&gt;&lt;namePart&gt;[value]&lt;namePart&gt;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>dc:creator</td>
<td>dc:creator</td>
<td>700/710/711/720$ when there is not a $e that says &quot;joint author,&quot; &quot;jt author,&quot; &quot;aut&quot; or &quot;cre&quot; – i.e. sometimes there will only be $a, only exclude is those $e values exist</td>
<td>&lt;role&gt;&lt;roleTerm&gt;Contributor&lt;/roleTerm&gt;&lt;role/&gt;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>dc:date</td>
<td>first dc:date</td>
<td>Take the value of $a for 100, 110, 111 and the value of $a for 700 if the value of $e is &quot;joint author&quot; or &quot;jt author&quot; ... so $a only is actually mapped</td>
<td>&lt;name&gt;&lt;namePart&gt;[value]&lt;namePart&gt;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>dc:description</td>
<td>dc:creator</td>
<td>Take the value of $a for 100, 110, 111 and the value of $a for 700 if the value of $e is &quot;joint author&quot; or &quot;jt author&quot; ... so $a only is actually mapped</td>
<td>&lt;role&gt;&lt;roleTerm&gt;Creator&lt;/roleTerm&gt;&lt;role/&gt;</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(use all instances. Some have qualifier indicating EDTF format and keyDate, others are qualified as &quot;approximate&quot;)</td>
<td>&lt;originInfo&gt;&lt;dateCreated&gt; (use all instances. Some have qualifier indicating EDTF format and keyDate, others are qualified as &quot;approximate&quot;)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>dcterms:description</td>
<td>dc:creator</td>
<td>Take the value of $a for 100, 110, 111 and the value of $a for 700 if the value of $e is &quot;joint author&quot; or &quot;jt author&quot; ... so $a only is actually mapped</td>
<td>&lt;abstract/&gt;</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(use all instances. Some have qualifier indicating EDTF format and keyDate, others are qualified as &quot;approximate&quot;)</td>
<td>&lt;physicalDescription&gt;extent</td>
</tr>
<tr>
<td></td>
<td></td>
<td>dcterms:extent</td>
<td>dc:creator</td>
<td>Take the value of $a for 100, 110, 111 and the value of $a for 700 if the value of $e is &quot;joint author&quot; or &quot;jt author&quot; ... so $a only is actually mapped</td>
<td>&lt;abstract/&gt;</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(use all instances. Some have qualifier indicating EDTF format and keyDate, others are qualified as &quot;approximate&quot;)</td>
<td>&lt;physicalDescription&gt;extent</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(use all instances. Some have qualifier indicating EDTF format and keyDate, others are qualified as &quot;approximate&quot;)</td>
<td>&lt;physicalDescription&gt;extent</td>
</tr>
</tbody>
</table>

**Notes:**
- QDC: Qualitative Data Code
- MARC: Machine-Readable Authority Control
- MODS: Metadata Object Description Syntax
Sample pipeline for Portal to Texas History

<table>
<thead>
<tr>
<th>Created Date</th>
<th>ca. 1947</th>
</tr>
</thead>
<tbody>
<tr>
<td>Partner</td>
<td>North Carolina Digital Heritage Center</td>
</tr>
<tr>
<td>Contributing Institution</td>
<td>University of North Carolina at Chapel Hill</td>
</tr>
</tbody>
</table>
| 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": [
  {
    "name": "Postcards--North Carolina"
  }
]

Subject Topical Postcards--North Carolina.
"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**.
Populate required and recommended fields
Be consistent, consistently.
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>dc:publisher</td>
<td>dc:publisher</td>
<td>dc:source</td>
<td>n/a</td>
<td></td>
</tr>
</tbody>
</table>
Control and explain yourself.
<place>
    <placeTerm>New York</placeTerm>
</place>
<place>
  <placeTerm valueURI="http://sws.geonames.org/5128581">New York</placeTerm>
</place>
Check your wok.
Standardizing Rights Statements

REUSE 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 ...

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
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

Knight News Challenge: Strengthening the Internet

‘Getting It Right on Rights’

WINNER
Digital Public Library of America
@DPLA

Dan Cohen
@DANCOHEN

Emily Gore
@NCSCHISTORY

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 Pekel (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)


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
<table>
<thead>
<tr>
<th>Source</th>
<th>Prefix abbreviation</th>
<th>Namespace</th>
</tr>
</thead>
<tbody>
<tr>
<td>Creative Commons Rights Expression Language (ccREL)</td>
<td>cc:</td>
<td><a href="http://creativecommons.org/ns#">http://creativecommons.org/ns#</a></td>
</tr>
<tr>
<td>Dublin Core Elements 1.1</td>
<td>dc:</td>
<td><a href="http://purl.org/dc/elements/1.1/">http://purl.org/dc/elements/1.1/</a></td>
</tr>
<tr>
<td>DCMI Type Vocabulary</td>
<td>dcmitype:</td>
<td><a href="http://purl.org/dc/dcmitype/">http://purl.org/dc/dcmitype/</a></td>
</tr>
<tr>
<td>DCMI Metadata Terms</td>
<td>dcterms:</td>
<td><a href="http://purl.org/dc/terms/">http://purl.org/dc/terms/</a></td>
</tr>
<tr>
<td>Europeana Data Model</td>
<td>edm:</td>
<td><a href="http://www.europeana.eu/schemas/edm/">http://www.europeana.eu/schemas/edm/</a></td>
</tr>
<tr>
<td>ODRL</td>
<td>odrl:</td>
<td><a href="http://www.w3c.org/ns/odrl/2/">http://www.w3c.org/ns/odrl/2/</a></td>
</tr>
<tr>
<td>PREMIS Copyright Status</td>
<td>premiscopy:</td>
<td><a href="http://id.loc.gov/vocabulary/preservation/copyrightStatus/">http://id.loc.gov/vocabulary/preservation/copyrightStatus/</a></td>
</tr>
<tr>
<td>SKOS</td>
<td>skos:</td>
<td><a href="http://www.w3.org/2004/02/skos/core#">http://www.w3.org/2004/02/skos/core#</a></td>
</tr>
<tr>
<td>OWL</td>
<td>owl:</td>
<td><a href="http://www.w3.org/2002/07/owl#">http://www.w3.org/2002/07/owl#</a></td>
</tr>
<tr>
<td>ODRS</td>
<td>odrs:</td>
<td><a href="http://schema.theodi.org/odrs#">http://schema.theodi.org/odrs#</a></td>
</tr>
</tbody>
</table>
Vocabulary Alignment
<table>
<thead>
<tr>
<th>Short name of RS</th>
<th>Name of Rights Statement</th>
</tr>
</thead>
<tbody>
<tr>
<td>dc:identifier</td>
<td>skos:prefLabel</td>
</tr>
</tbody>
</table>

**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.

*not mapped*

**Text of the Rights Statement**

skos:definition

**Notices: One or more notices related to the Rights Statement**

skos:note

**Disclaimer regarding this being a Rights Statement and not a legally operative License summary.**

*not mapped*

**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.

skos:scopeNote

**Extra metadata**

*not mapped*

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
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