What We Learned By Aggregating Metadata for 7 Million Items

Mark Matienzo
Director for Technology
mark@dp.la / @anarchivist
DPLA is partnerships
Service Hubs

- The Portal to Texas History
- Digital Commonwealth
- Mountain West Digital Library
- North Carolina Digital Heritage Center
- Kentucky Digital Library
- South Carolina Digital Library
- Empire State Digital Network
- Minnesota Digital Library

Digital Library of Georgia
12 Content Hubs
+ 9 Service Hubs

1,300+ Partners
One-to-one partnerships

Service/Content Hub Model
DPLA is a portal for discovery

A Wealth of Knowledge
explore 7,154,010 items from libraries, archives, and museums

Search the Library
Search Results

Your search for Portland returned 1,181 results.

Refined by: Oregon

Items per page: 10
Sort by: Relevance

By Format
- text: 904
- image: 265
- sound: 11

Contributing Institution
- University of Minnesota: 466
- University of Michigan: 196
- University of California: 71
- Boston Public Library: 56
- California Historical Society: 37

Partner
- HathiTrust: 886
- Smithsonian Institution: 76
- Digital Commonwealth: 55
- Mountain West Digital Library: 54
- University of Southern California. Libraries: 42

Image
- Portland, Oregon
  - Ken Vermillion
  - Postcard of a view of Downtown Portland, Oregon. As seen from the new Marquam Bridge. Photographic postcards

Image
- Hotel Portland, Portland, Oregon (2)
  - Postcard is of the old Portland Hotel in Portland, Oregon. Picture postcards

Image
- Hotel Portland, Portland, Oregon (1)
  - Postcard is of the old Portland Hotel, which is no longer standing, in Portland, Oregon. Picture postcards

Text
- The Portland Woman's Exchange cook book
**Oregon och Washington**
Skarstedt, Ernst Teofil, 1857- [from old catalog]

**View Object**

**IMAGE**

How forest giants are hauled up the slip into a saw-mill, harborside at Portland. 9855 interpositive
Currently stored in box 3.2.32 [150]

**TEXT**
The Oregon Humane Society: containing an abstract of the state laws, the constitution and by-laws of the society, what the Oregon Humane Society is doing, ways of cruelty and ways of mercy, humane work throughout the world, general intelligence

**View Object**
... leads to diverse metadata

~7.4 million records
9 metadata schemas
27 crosswalks representing 1,300+ institutions’ data

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

. . . which leads to *challenges*
The DPLA Ingestion System

https://github.com/dpla/ingestion

Courtesy National Archives and Records Administration
Ingestion Profiles
Ingestion Workflow
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

Sham Hardy, “Bumblebee, Transformers Toys.”
Digital Public Library of America Domain Model, v3.1

Core classes highlighted in blue.

- edm:WebResource
- ore:Aggregation
- edm:aggregatedCHO
- dpla:SourceResource
- dcterms:temporal
- dcterms:spatial
- dcmitype:Collection
- edm:TimeSpan
- dcterms:isPartOf
<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>
The current policies of the North Carolina Collection define the conditions of use. See http://dc.lib.unc.edu/u/?dig_nccpa,7197

## Mapping to the DPLA MAP

### DPLA Partners Crosswalk

<table>
<thead>
<tr>
<th>DPLA metadata application profile</th>
<th>Simple Dublin Core</th>
<th>ArtStor</th>
<th>BHL</th>
<th>DC</th>
<th>DLG</th>
</tr>
</thead>
<tbody>
<tr>
<td>dpla:SourceResource</td>
<td>Collection</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>dpla:SourceResource</td>
<td>Contributor</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>dpla:SourceResource</td>
<td>Date</td>
<td>dc:date</td>
<td>dc:date</td>
<td>dc:date</td>
<td>dc:coverage (ignore dc:date)</td>
</tr>
<tr>
<td>dpla:SourceResource</td>
<td>Extent</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>dpla:SourceResource</td>
<td>Format</td>
<td>dc:format; dc:type</td>
<td>dc:type</td>
<td>dc:format</td>
<td>dc:format</td>
</tr>
<tr>
<td>dpla:SourceResource</td>
<td>Identifier</td>
<td>header_dc:identifier</td>
<td>n/a</td>
<td>dc:identifier</td>
<td>n/a</td>
</tr>
<tr>
<td>dpla:SourceResource</td>
<td>Language</td>
<td>n/a</td>
<td>dc:language</td>
<td>dc:language</td>
<td>dc:language</td>
</tr>
<tr>
<td>dpla:SourceResource</td>
<td>Place</td>
<td>n/a</td>
<td>n/a</td>
<td>dc:coverage</td>
<td>dc:coverage</td>
</tr>
<tr>
<td>dpla:SourceResource</td>
<td>Publisher</td>
<td>n/a</td>
<td>dc:creator</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>dpla:SourceResource</td>
<td>Relation</td>
<td>n/a</td>
<td>n/a</td>
<td>dc:relation</td>
<td>dc:relation</td>
</tr>
<tr>
<td>dpla:SourceResource</td>
<td>Rights</td>
<td>dc:_rights</td>
<td>n/a</td>
<td>dc:rights</td>
<td>dc:rights</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Fort Raleigh Grill, Elizabeth City, N.C.</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Created Date</strong></td>
</tr>
<tr>
<td><strong>Partner</strong></td>
</tr>
<tr>
<td><strong>Contributing Institution</strong></td>
</tr>
<tr>
<td><strong>Publisher</strong></td>
</tr>
<tr>
<td><strong>Location</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td><strong>Type</strong></td>
</tr>
<tr>
<td><strong>Subject</strong></td>
</tr>
<tr>
<td><strong>Rights</strong></td>
</tr>
<tr>
<td><strong>URL</strong></td>
</tr>
</tbody>
</table>
The “Pipeline”

http://bl.ocks.org/anarchivist/a436585ee39fb5ea1c3a

http://bl.ocks.org/anarchivist/c8452569f65531d86669
"subject": [
{
"name": "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"
  }
]
Ingestion challenges
Working together with Hubs
Thanks!

Questions to:
mark@dp.la / @anarchivist