What’s New in DPLA Technology?

Mark A. Matienzo
Director of Technology, DPLA
DPLAFest – Indianapolis, IN – April 17, 2015
Overview

- Heiðrún: DPLA’s new ingestion system
  - Motivation for development/overall goals
  - Development progress/current features
  - Differences in workflow
  - Future work

- New IMLS grant on Hydra development
  - Overall goals
  - Opportunities for feedback
Ingestion Workflow

Heiðrún: DPLA’s New Ingestion System

- Development began October 2014
- Project team: Audrey Altman, Mark Breedlove, Gretchen Gueguen, Tom Johnson, Mark Matienzo, Amy Rudersdorf
- Initial objective: implementation of a new ingestion and partner management system for DPLA’s metadata aggregation activities
- Long-term objective: develop generalized “aggregation system in a box” for DPLA Hubs and others
Motivation for development

- Architectural & scale issues

- DPLA MAP is an RDF model, but previous system did not deal with data as such

- Workflow was opaque to non-developer and partner staff and was an unbalanced workload

- Workflow too tightly coupled

- Inadequate tools for process/partner management
Overall Goals

- New system for scalable harvesting, mapping, enrichment, storage, and indexing of metadata
- Native support for DPLA Metadata Application Profile v4 as RDF
- Web-based “dashboards” for DPLA and Hub staff, to support ingest scheduling, partner management, QA, and metadata mapping
Development Progress and Current Features

- Brand new modular architecture
- Metadata mapping language
- Better decoupled workflow
- QA improvements
Improved Architecture

- Real RDF: ActiveTriples & DPLA MAP modules
- Marmotta LDP server & triple store
- Easier to develop QA system using Blacklight, Solr, SPARQL queries
- Queueing and scheduling system
Krikri is:
- The core component of Heidrun.
- A Rails engine.

Krikri provides:
- Basic harvesters, mappers, and enrichments.
- User interfaces for managing workflows.
- CRUD interfaces for Marmotta, Solr/ElasticSearch, and a relational database (PostgreSQL or MySQL).
Ingestion Workflow

Harvest process → Original Record → Mapping process → DPLA MAP Resource → Enrichment processes → DPLA MAP Resource

Marmotta
• Used to define mappings from provider metadata to DPLA MAP properties and classes

• Allows for matching elements based on values, attributes, etc.

• Under continual improvement as complexity of mappings is addressed
QA improvements

• Better reporting, using search index

• Introduce Provenance Ontology (PROV-O) to track info about ingestion activities
1. http://ldp.dp.la/ldp/items/8eaa4c2a8a26819aa02426fddf320db0
   
   **Title:** View of Geneva to the West taken from the Nestor Hotel
   
   **Description:** View of Geneva to the West taken from the Nestor Hotel
   
   **Rights:** This digital image may be used for educational purposes, as long as it is not altered in any way. Prior written permission is required from the Geneva Historical Society for any other use of the image.

2. http://ldp.dp.la/ldp/items/0fd6a7f6e06111c59e8c303227982d7b
   
   **Title:** Empire state tax news. - WIN-2001
   
   **Rights:** This document or image is provided for education and research purposes. Rights may be reserved. Responsibility for securing permissions to distribute, publish, reproduce or use it in any way rests with the user. For additional information, see the New York State Library’s Copyright and Use Statement, available at http://www.nysl.nysed.gov/scandocs/rights.htm.
<table>
<thead>
<tr>
<th>ID</th>
<th>Title</th>
<th>Is Shown At</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://dp.dp.la/ldp/items/1d3b8694a8699d956476c40be62481d8">http://dp.dp.la/ldp/items/1d3b8694a8699d956476c40be62481d8</a></td>
<td>Street scene during a May Day celebration in Moscow</td>
<td><a href="http://cdm16786.contentdm.oclc.org/cdm/ref/collection/eurasia/id/5620">http://cdm16786.contentdm.oclc.org/cdm/ref/collection/eurasia/id/5620</a></td>
</tr>
<tr>
<td><a href="http://dp.dp.la/ldp/items/f213e252248be35437b6cf6c4ab1f6282">http://dp.dp.la/ldp/items/f213e252248be35437b6cf6c4ab1f6282</a></td>
<td>Santa Barbara County Courthouse</td>
<td><a href="http://cdm16786.contentdm.oclc.org/cdm/ref/collection/buildings/id/11092">http://cdm16786.contentdm.oclc.org/cdm/ref/collection/buildings/id/11092</a></td>
</tr>
<tr>
<td><a href="http://dp.dp.la/ldp/items/b0e06c602b8d90035eda45f7e45ed2f0">http://dp.dp.la/ldp/items/b0e06c602b8d90035eda45f7e45ed2f0</a></td>
<td>opp2-4</td>
<td><a href="http://cdm16786.contentdm.oclc.org/cdm/ref/collection/pltec/id/1183">http://cdm16786.contentdm.oclc.org/cdm/ref/collection/pltec/id/1183</a></td>
</tr>
</tbody>
</table>
Record
http://ldp.dp.la/ldp/items/1d3b8694a8699d956476c40be62481d8

Random record from University of Washington

Enriched Record
"identifier": [  
  "5287",  
  "http://cdm16786.contentdm.oclc.org/cdm/ref/collection/eurasia/id/5620"
],
"rights": "Augerot, James",
"spatial": [  
  {  
    "@id": ".:_b6",
    "@type": "edm:Place",
    "providedLabel": "Russia"
  },  
  {  
    "@id": ".:_b5",
    "@type": "edm:Place",
    "providedLabel": "Tsentralnyy Ekonomicheskiy Rayon"
  }
],

Original Record
<dc:title>Street scene during a May Day celebration in Moscow</dc:title>
<dc:creator>Augerot, James</dc:creator>
<dc:date>1962</dc:date>
<dc:subject>City and town life; Celebrations; Parades and processions; Festive decorations; Flags; Streets</dc:subject>
<dc:coverage>Russia</dc:coverage>
<dc:coverage>Tsentralnyy Ekonomicheskiy Rayon</dc:coverage>
<dc:coverage>Moskovskaya Oblast</dc:coverage>
<dc:coverage>Moskva</dc:coverage>
<dc:contributor>Augerot, James</dc:contributor>
Future Plans

Tools to automate and ease workflow

- Responsive metadata mapping interface.
- QA reports that target common ingestion errors.
- Ingestion dashboard for scheduling and monitoring harvests.

A metadata aggregation toolkit for the community

- Define an RDF model for your data using ActiveTriples.
- Harvest metadata (RDF or non-RDF) from a various sources.
- Map harvested metadata to your RDF model.
- Enrich metadata by linking to other RDF sources.
# Hub Dashboard

## Past Ingest Events

<table>
<thead>
<tr>
<th>Ingest Date</th>
<th>Start Time</th>
<th>End Time</th>
<th>Records Retrieved</th>
<th>Errors</th>
<th>Live Records</th>
<th>Live Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>8/5/2014</td>
<td>8:15am</td>
<td>3:25pm</td>
<td>55,774</td>
<td>347</td>
<td>55,400</td>
<td>8/7/2014</td>
</tr>
<tr>
<td>12/7/2014</td>
<td>10:45am</td>
<td>3:32pm</td>
<td>25,117</td>
<td>17</td>
<td>-cancelled</td>
<td>-cancelled</td>
</tr>
</tbody>
</table>

## Current Ingest

<table>
<thead>
<tr>
<th>Ingest Date</th>
<th>Start Time</th>
<th>End Time</th>
<th>Records Retrieved</th>
<th>Errors</th>
<th>QA</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>1/9/2015</td>
<td>8:15am</td>
<td>3:25pm</td>
<td>55,774</td>
<td>347</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Start New Ingest
More Information on Krikri/Heiðrún

● Project page:

● GitHub projects:
  ○ http://github.com/dpla/krikri
  ○ http://github.com/dpla/heidrun
“Hydra-in-a-Box” Project

- Project partners: DPLA, Stanford University, DuraSpace
- $2M, 30 month grant awarded by Institute of Museum and Library Services, in their National Leadership Grants program
- Focuses on fostering a new, national, library network through a community-based repository
- Leverages successes and infrastructure from the Hydra project and community, while contributing back to it
“Hydra-in-a-Box” Goals

- Development of turnkey, Hydra-based application
- Improve and DPLA’s metadata ingestion system into an “aggregator system in a box”
- Connect components with pieces with DPLA hubs, current Hydra partners, and prospective Hydra adopters
- Work toward a hosted service
Coming soon!
Contact mark@dp.la

More information on “Hydra-in-a-Box”
Contributor Agreements & Licensing for DPLA Software Projects

- Addressing need to define terms under which code has been contributed to DPLA projects
- Development of draft guidelines for DPLA staff to select FLOSS licenses
- DPLA in the process of determining best options for contributor agreements, e.g.:
  - What is the state of the art?
  - Should we use contributor license agreements or copyright assignment agreements?
Planning for the future: DPLA and the Technical Advisory Committee

- Survey for feedback:
  - Feel free to submit after DPLAFest
  - Open to all

- Small group discussion:
  - What should we be thinking about over the next year or more?
  - Who should be in the conversation?
  - How can the Technical Advisory Committee help DPLA?

- Report back and next steps