Heiðrun

Building DPLA’s New Metadata Ingestion System

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

Metropolitan New York Library Council Annual Conference
January 15, 2015
1. Original DPLA Infrastructure
2. The DPLA ingestion process
3. Challenges with ingestion
4. Feedback from DPLA Hubs
5. Planning for needed improvements
6. Building Heiðrun
Original Infrastructure

- Frontend (Ruby on Rails)
- API (Ruby on Rails)
- CouchDB
- Elasticsearch
- PostgreSQL
- Ingestion system (Python)
Ingestion workflow

1. Harvest hub metadata
2. Transform to DPLA MAP
3. Enrich metadata to clean/add value
4. Remove deleted records
5. Perform QA if needed
6. Save/index enriched metadata
Sample pipeline for Portal to Texas History

Challenges with ingestion

• Ingestion process very hands-on; requires significant staff time despite use of common standards

• Ingestion process not modular and flexible enough to support partial reharvesting or enrichment

• System has lack of awareness of MAP data as RDF

• Some enrichment processes (e.g. geocoding) introduce and expose metadata inconsistencies

• Unqualified Dublin Core requires the most work in terms of mapping and transformation
Feedback from DPLA Hubs

• Greater control over and feedback during the ingestion process

• Access to data quality reports

• Provide mechanism to receive enrichments applied by DPLA ingestion process

• Collaborate on further tool and infrastructure development
Planning for improvements

• Improvement of documentation for metadata model and ingestion process

• Revision of the DPLA Metadata Application Profile

• Reassessment of “data quality” and “validation” in the context of DPLA

• Encouraging Hubs to undertake metadata transformation and enrichment locally and to develop appropriate tools

• Replacement of the DPLA ingestion system
• DPLA started development on new ingestion system and metadata repository in October 2014

• Collaborative project across both DPLA Content and Technology teams
Development goals

• Make it easier to harvest and map metadata from various sources/schemas into DPLA MAP

• Improve enrichment using external sources

• Actively involve partners in ingestion process through better tools

• Native support for DPLA MAP as RDF data model
Current features

• Improved harvesting, including support for partial harvests

• Domain-specific language for metadata mapping

• Improved scoping of enrichments as field- or record-based

• Basic QA environment
Future plans

• Ingest dashboard for DPLA and hub staff
• Improved QA tools and reports
• Browser-based GUI metadata mapping tool
• Building an “aggregation system in a box” for use by DPLA hubs and others
• More control for both DPLA Content Team and Hubs staff
Thank You!

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

This work is licensed under a Creative Commons Attribution 4.0 International License.
http://creativecommons.org/licenses/by/4.0/
Heiðrun Architecture

- QA and management interface
- Generic (shared) enrichments
- Hub-specific enrichments and enrichment profiles
- Generic (shared) metadata mappings
- Hub-specific metadata mappings
- Harvesters
- Harvester settings for specific hubs
- DPLA MAP models
- ActiveTriples
- RDF.rb
- Apache Marmotta
- RDBMS (PostgreSQL or MySQL)
- Solr/Elasticsearch
• DPLA new ingestion system code bases.
  • https://github.com/dpla/heidrun
  • https://github.com/dpla/KriKri