The Digital Public Library of America
Ingestion Ecosystem

Lessons Learned After One Year of
Large-Scale Collaborative Metadata Aggregation

Mark A. Matienzo
mark@dp.la

Digital Public Library of America
http://dp.la/

Amy Rudersdorf
amy@dp.la

International Conference on Dublin Core & Metadata Applications
October 9, 2014
1. Introduction to DPLA
2. DPLA Metadata Application Profile
3. DPLA ingestion system
4. Challenges with the ingestion system and process
5. Challenges with partner metadata
6. Responses and requests from DPLA Hubs (partners)
7. Planning for needed improvements
8. Conclusion
Introduction
DPLA Hubs

MINNESOTA DIGITAL LIBRARY

Kentucky Digital Library

David Rumsey Map Collection

SCDL SOUTH CAROLINA DIGITAL LIBRARY

Mountain West Digital Library

Digital Commonwealth

UF UNIVERSITY OF FLORIDA

EMLR EP FINE STATE DIGITAL NETWORK

The Portal to Texas History

EMBARK on a Voyage of Discovery

GPO

GPO

North Carolina Digital Heritage Center

NATIONAL ARCHIVES

Biodiversity Heritage Library

INTERNET ARCHIVE

HATHI TRUST

Seal of the State of Indiana

University of Southern California

Smithsonian

VII E I

TAS

Harvard

University of Virginia

University of Illinois

1816

1880

1889

1901

1867
Infrastructure

- Frontend (Ruby on Rails)
- API (Ruby on Rails)
- CouchDB
- Elasticsearch
- PostgreSQL
- Ingestion system (Python)
Metadata Application Profile

http://dp.la/info/developers/map/
DPLA Ingestion System

- Python application written using Akara framework
- CouchDB (BigCouch) as primary persistence layer
- Elasticsearch as indexing and search layer
- Code released as open source (Affero GPL 3.0)
- [https://github.com/dpla/ingestion/](https://github.com/dpla/ingestion/)
Transformation & enrichment

Sample pipeline for Portal to Texas History

Challenges: ingestion

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

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

• Mapping and validation as implemented is inadequate

• System has lack of awareness of MAP data as RDF

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

• Unqualified Dublin Core requires the most work in terms of mapping and transformation

• DCMES elements used very differently across partners

• OAI-PMH providers do not always have documented mappings from origin schemas (??? → oai_dc)

• Usage of controlled vocabularies not always clear
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
Tools developed by Hubs

- Bplgeo (Digital Commonwealth): [https://github.com/projecthydra-labs/Bplgeo](https://github.com/projecthydra-labs/Bplgeo)

- NCDHC Aggregation Tools: [https://github.com/ncdhc/dpla-aggregation-tools](https://github.com/ncdhc/dpla-aggregation-tools)  
  [https://github.com/ncdhc/dpla-submission-precheck](https://github.com/ncdhc/dpla-submission-precheck)

- Minnesota Digital Library: [https://github.com/umnlibraries?query=dpla](https://github.com/umnlibraries?query=dpla)
Developing a new system

- DPLA starting development on new ingestion system and metadata repository in October 2014
- Collaborative project across both DPLA Content and Technology teams
- Work will serve as a basis for an “aggregation system in a box,” intended for use by DPLA Hubs and others
Conclusion

• DPLA successfully aggregated 8 million records from 24 Hubs using lightweight infrastructure

• Limitations of existing system allowed DPLA and its Hubs to identify shared needs and opportunities for collaboration

• DPLA uniquely situated to develop resources and community of practice for national-level aggregation, remediation, and enhancement of metadata
Thank You!

Mark A. Matienzo
mark@dp.la

Digital Public Library of America
http://dp.la/

Amy Rudersdorf
amy@dp.la
References