DPLA

DIGITAL PUBLIC LIBRARY OF AMERICA

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/

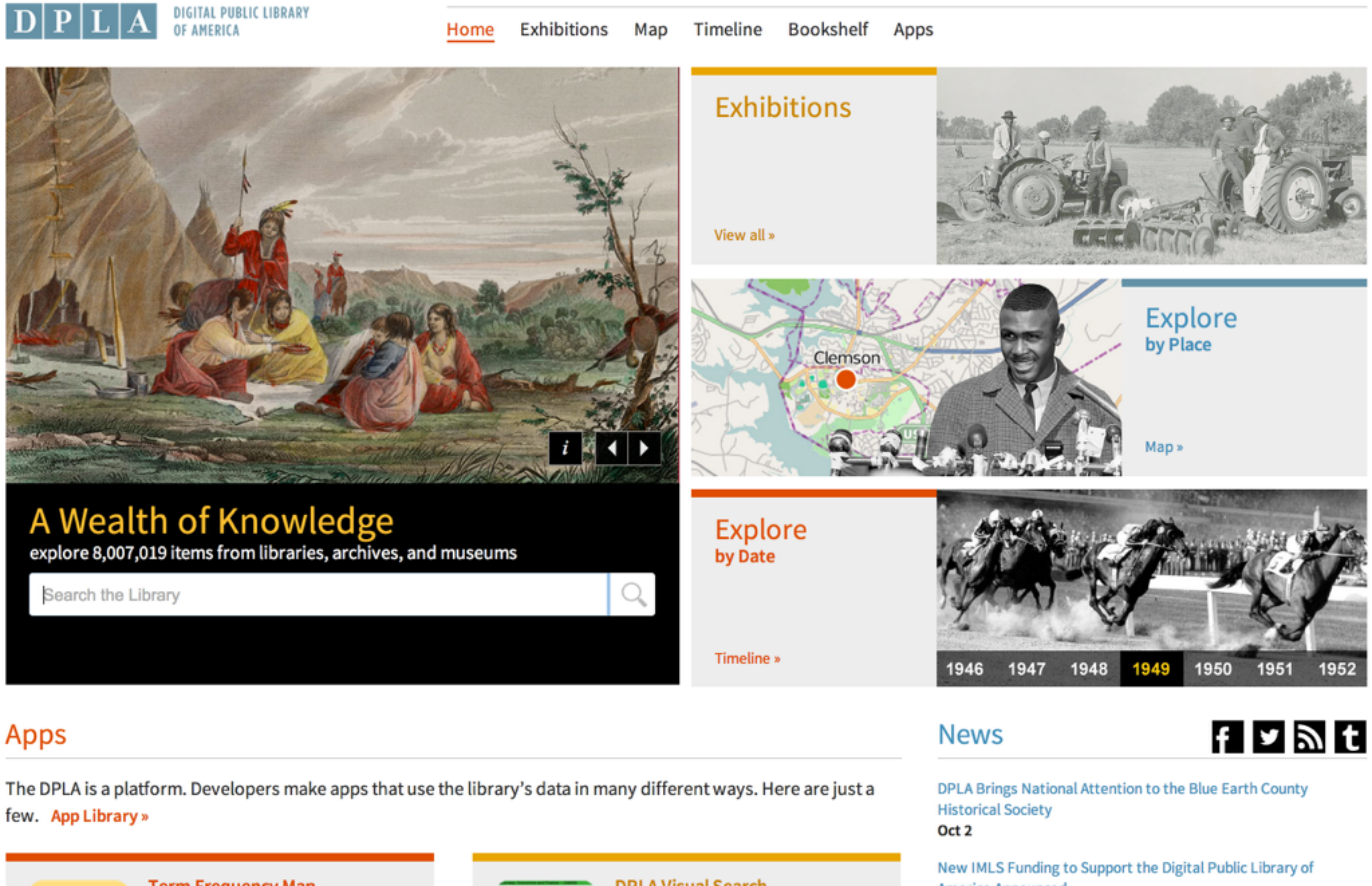
Amy Rudersdorf
amy@dp.la

International Conference on Dublin Core & Metadata Applications
October 9, 2014

Outline

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



DIGITAL PUBLIC LIBRARY OF AMERICA

A Wealth of Knowledge

explore 8,007,019 items from libraries, archives, and museums

Search the Library

Exhibitions

View all »

Explore by Place

Map »

Explore by Date

Timeline »

Apps

The DPLA is a platform. Developers make apps that use the library's data in many different ways. Here are just a few. App Library »

Term Frequency Map

DPLA Visual Search

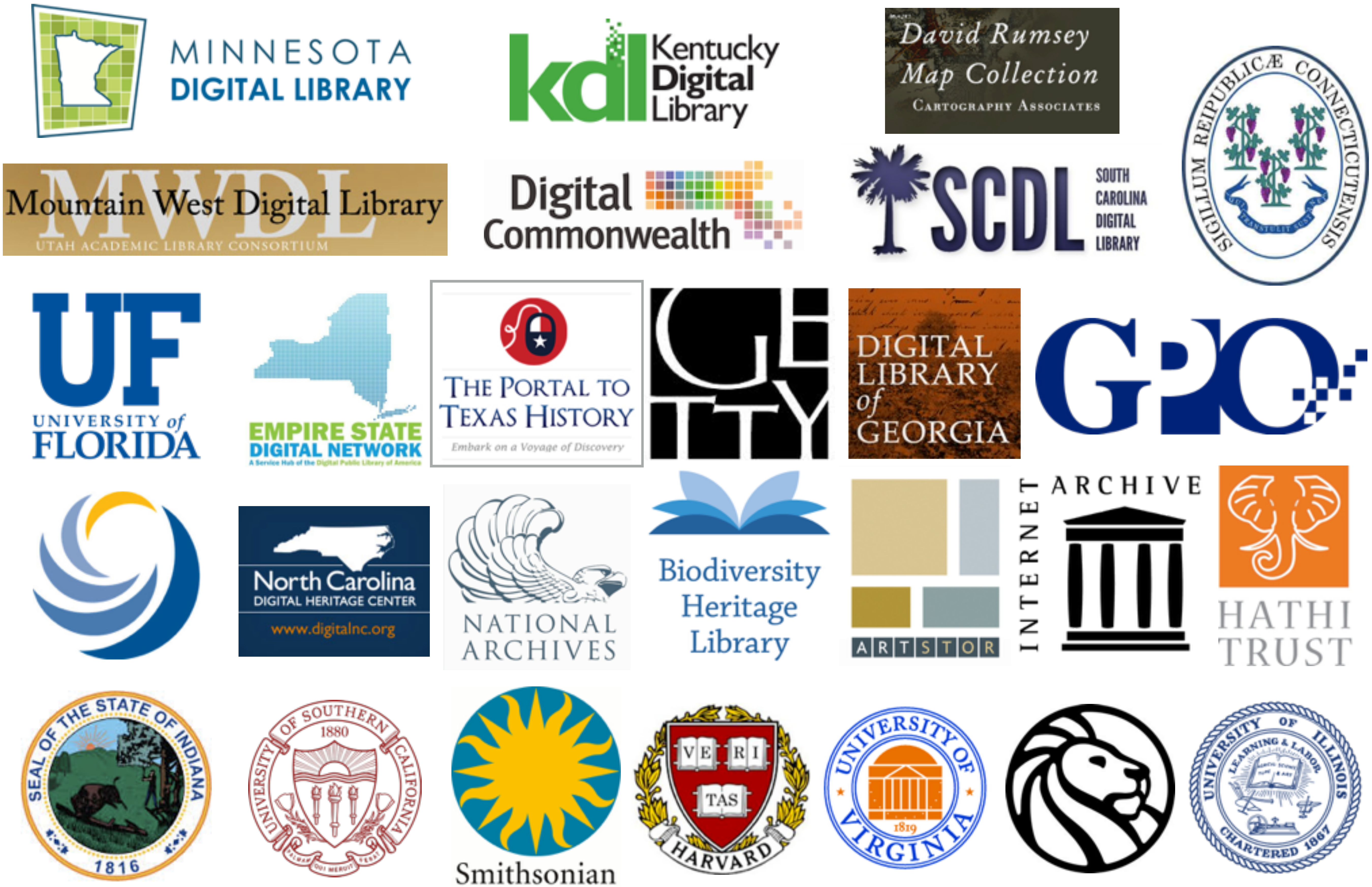
News

DPLA Brings National Attention to the Blue Earth County Historical Society

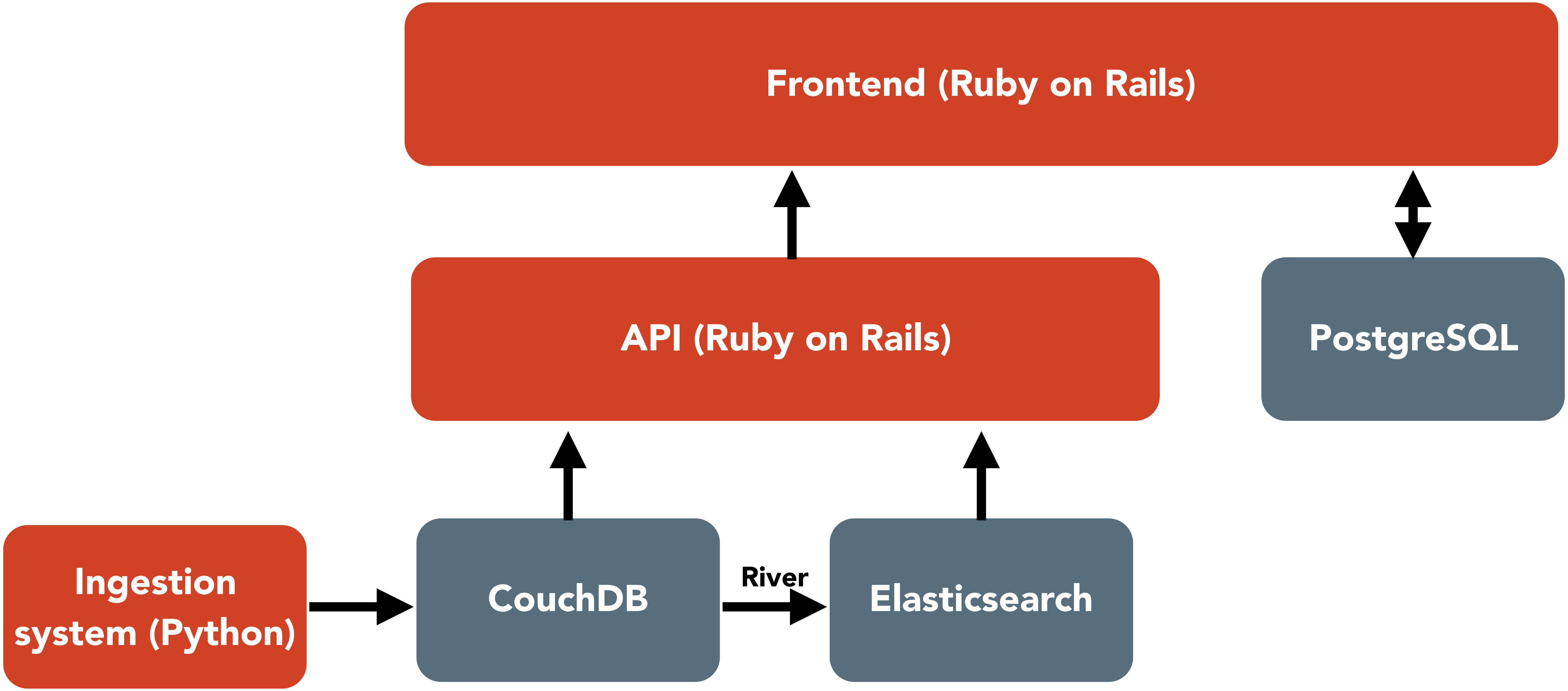
Oct 2

New IMLS Funding to Support the Digital Public Library of

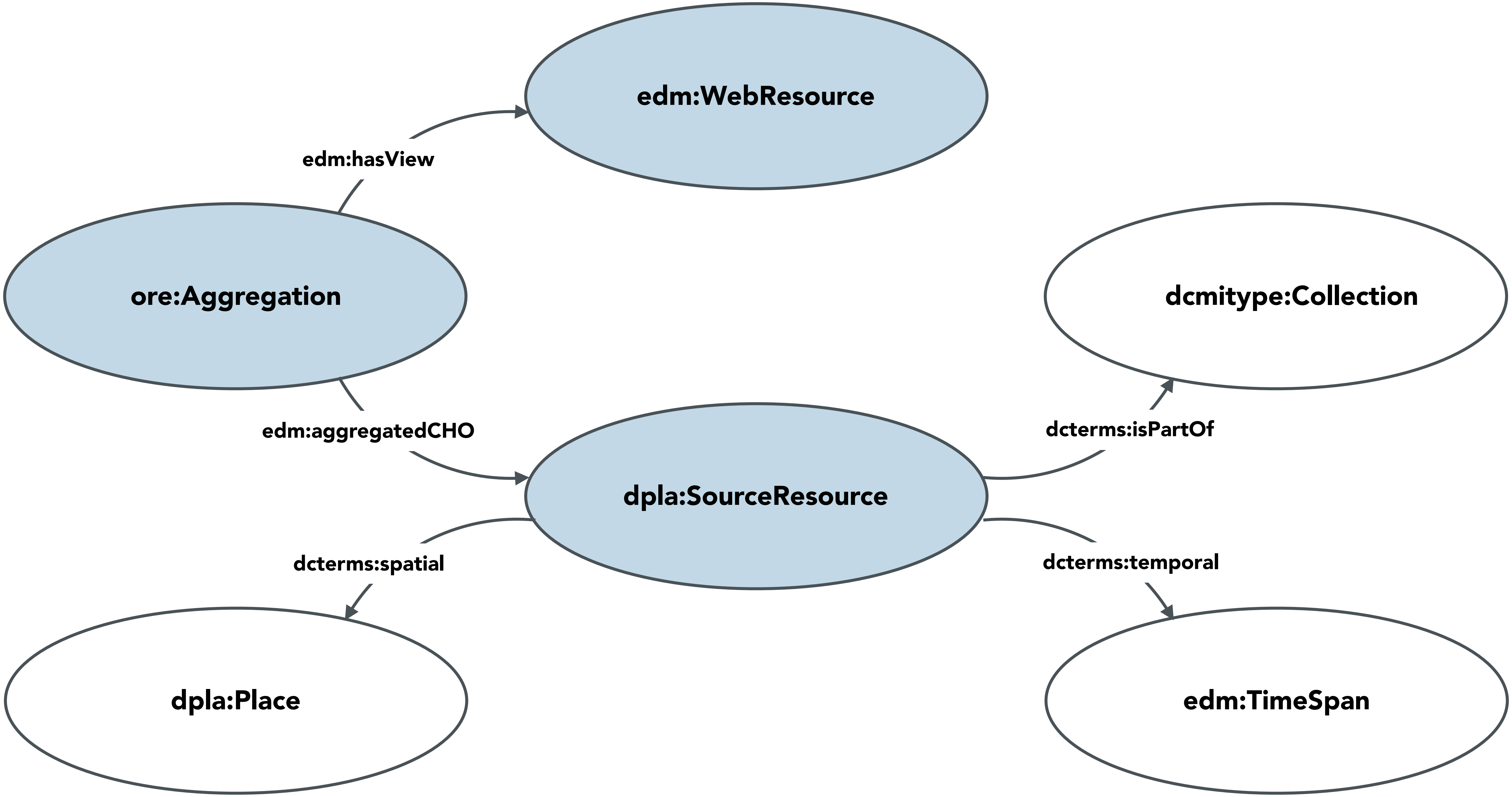
DPLA Hubs


MINNESOTA DIGITAL LIBRARY
kdl Kentucky Digital Library
David Rumsey Map Collection
Cartography Associates
SIGILLUM REIPUBLICÆ CONNECTICUTENSIS
QUI TRANSTULIT SUSTINET
MWDL Mountain West Digital Library
UTAH ACADEMIC LIBRARY CONSORTIUM
Digital Commonwealth
SCDL SOUTH CAROLINA DIGITAL LIBRARY
UF UNIVERSITY of FLORIDA
EMPIRE STATE DIGITAL NETWORK
A Service Hub of the Digital Public Library of America
THE PORTAL TO TEXAS HISTORY
Embark on a Voyage of Discovery
GETTY
DIGITAL LIBRARY of GEORGIA
GPO
North Carolina DIGITAL HERITAGE CENTER
www.digitalnc.org
NATIONAL ARCHIVES
Biodiversity Heritage Library
ARTSTOR
INTERNET ARCHIVE
HATHI TRUST
SEAL OF THE STATE OF INDIANA
1816
UNIVERSITY OF SOUTHERN CALIFORNIA
1880
PALMAM QUI MERUIT FERAT
Smithsonian
VE RI TAS
HARVARD
UNIVERSITY OF VIRGINIA
1819
UNIVERSITY OF ILLINOIS
LEARNING & LABOR
CHARTERED 1867

Infrastructure



Metadata Application Profile

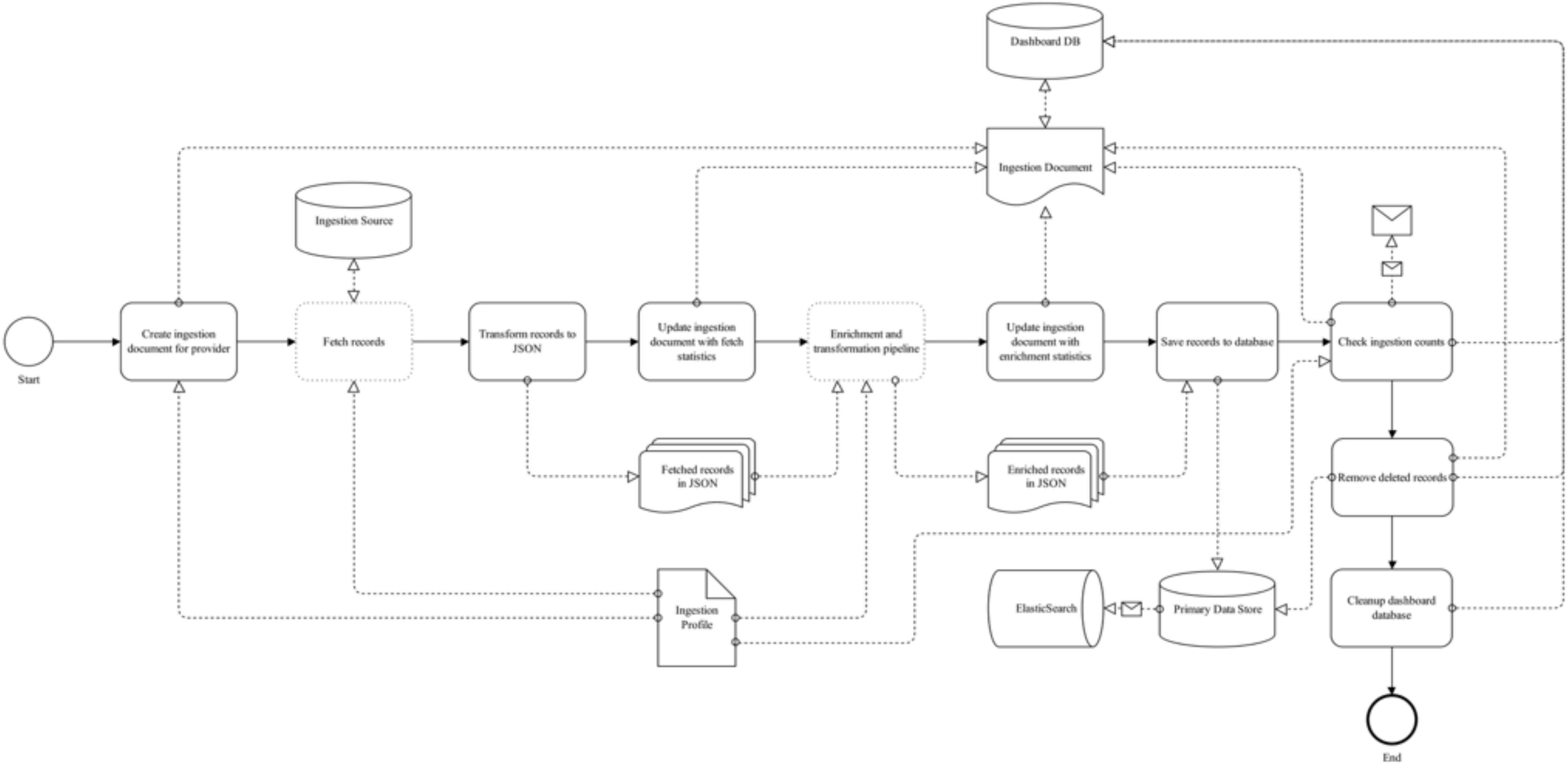


http://dp.la/info/developers/map/

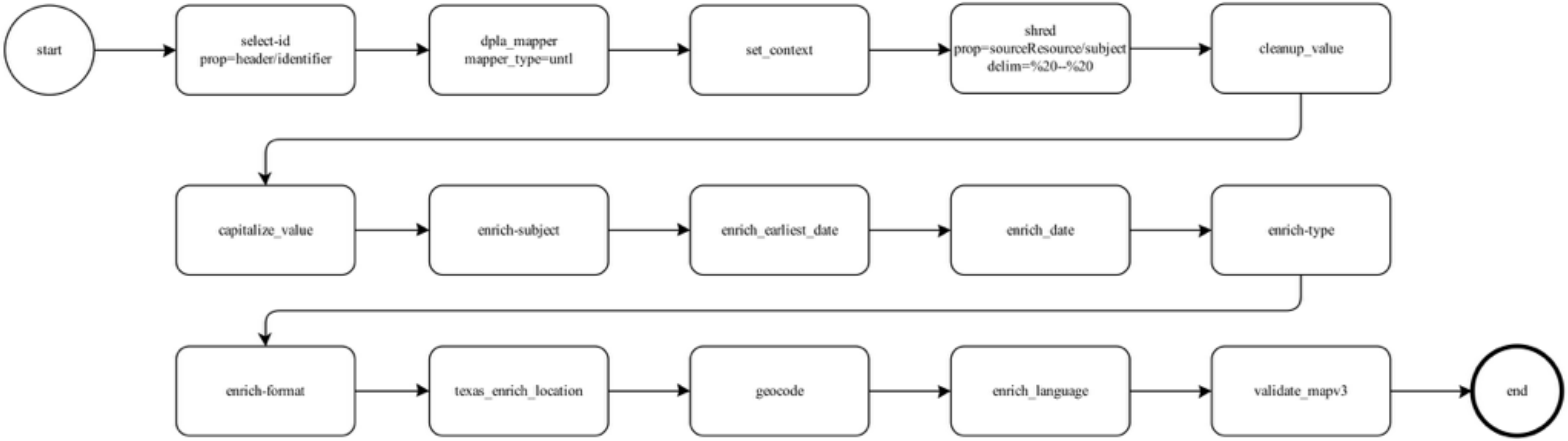
DPLA Ingestion System

- 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/

Ingestion workflow



Transformation & enrichment



Sample pipeline for Portal to Texas History

http://bit.ly/dpla-ingest-workflows

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

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

- Minnesota Digital Library: 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

- Akara. (2010). Retrieved August 7, 2014, from http://akara.info/.
- DigitalNZ. (2014). Supplejack documentation, version 0.1. Retrieved August 7, 2014, from http://digitalnz.github.io/supplejack/.
- Boston Public Library. (2014). Bplgeo. Retrieved October 7, 2014, from https://github.com/projecthydra-labs/Bplgeo.
- Digital Public Library of America. (2014a). Digital Public Library of America Metadata Application Profile, Version 3.1. Retrieved August 7, 2014, from http://dp.la/about/map.
- Digital Public Library of America. (2014b). The DPLA ingestion system, version 31.1. http://dx.doi.org/10.5281/zenodo.11226. Retrieved August 7, 2014, from https://github.com/dpla/ingestion.
- Digital Public Library of America. (2014c). An introduction to the DPLA metadata model. Retrieved August 7, 2014, from http://dp.la/info/2014/03/25/intro-dpla-metadata-model/.
- Digital Public Library of America (2014d). Content wiki. Retrieved August 7, 2014, from https://digitalpubliclibraryofamerica.atlassian.net/wiki/display/CT/Content.
- DPLA RDF application profile use cases. (2014). Retrieved August 7, 2014, from http://wiki.dublincore.org/index.php/DPLA_RDF_application_profile_use_cases.
- Europeana. (2013). Europeana Data Model primer. 14 July 2013. Retrieved August 7, 2014, from http://pro.europeana.eu/documents/900548/770bdb58-c60e-4beb-a687-874639312ba5.
- Europeana. (2014). Definition of the Europeana Data Model v5.2.5. 22 May 2014. Retrieved August 7, 2014, from http://pro.europeana.eu/documents/900548/0d0f6ec3-1905-4c4f-96c8-1d817c03123c.
- Galiegue, Francis, Kris Zyp, and Gary Court. (2013). JSON Schema: interactive and non interactive validation. IETF Internet-Draft, January 30, 2013. Retrieved August 7, 2014 from http://json-schema.org/latest/json-schema-validation.html.
- Gregory, Lisa, and Stephanie Williams. (2014). On being a hub: some details behind providing metadata for the Digital Public Library of America. *D-Lib Magazine, 20*(7/8). http://dx.doi.org/10.1045/july2014-gregory.
- Hillmann, Diane I., Naomi Dushay, and Jon Phipps. (2004). Improving metadata quality: augmentation and recombination. Proceedings of the International Conference on Dublin Core and Metadata Applications, 2004. Retrieved May 15, 2014 from http://hdl.handle.net/1813/7897.
- Lagoze, Carl, Dean Krafft, Tim Cornwell, Naomi Dushay, Dean Eckstrom, and John Saylor. (2006). Metadata aggregation and “automated digital libraries”: A retrospective on the NSDL experience. In G. Marchionini, M. L. Nelson, and C. Marshall (Eds.): *JCDL '06: Proceedings of the 6th ACM/IEEE-CS joint conference on digital libraries* (pp. 230-239). New York: Association for Computing Machinery.
- NCDHC. (2014a). dpla-aggregation-tools. Retrieved August 7, 2014, from https://github.com/ncdhc/dpla-aggregation-tools.
- NCDHC. (2014b). dpla-submission-precheck. Retrieved August 7, 2014, from https://github.com/ncdhc/dpla-submission-precheck.
- Phillips, Mark, Hannah Tarver, and Stacy Frakes. (2014). Implementing a collaborative workflow for metadata analysis, quality improvement, and mapping. *Code4lib Journal, 23*. Retrieved August 7, 2014, from http://journal.code4lib.org/articles/9199.
- Riley, Jenn, John Chapman, Sarah Shreeves, Laura Akerman, and William Landis. (2008). Promoting shareability: metadata activities of the DLF Aquifer initiative. *Journal of Library Metadata*, *8*(3).
- Sporny, Manu, Gregg Kellogg, and Markus Lanthaler (Eds.). (2014). JSON-LD 1.0: A JSON-Based Serialization of Linked Data. W3C Recommendation 16 January 2014. Retrieved August 7, 2014, from http://www.w3.org/TR/json-ld/.
- University of Minnesota Libraries. (2014a). dpla.client. Retrieved August 7, 2014, from https://github.com/UMNLibraries/dpla.client.
- University of Minnesota Libraries. (2014b). dpla.docs. Retrieved August 7, 2014, from https://github.com/UMNLibraries/dpla.docs.
- University of Minnesota Libraries. (2014c). dpla.services. Retrieved August 7, 2014, from https://github.com/UMNLibraries/dpla.services.