



Heiðrun

Building DPLA's New Metadata Ingestion System

Mark A. Matienzo <mark@dp.la>

Digital Public Library of America

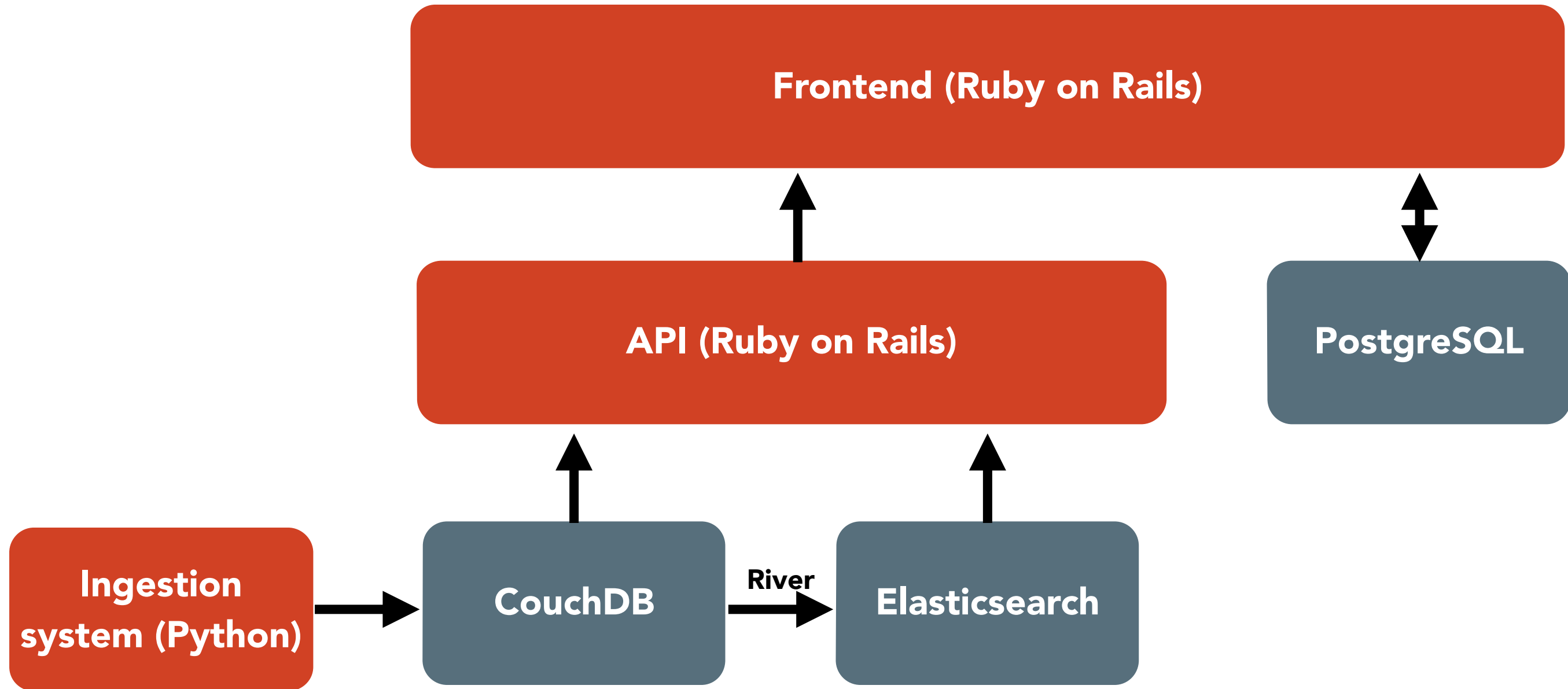
Metropolitan New York Library Council Annual Conference

January 15, 2015

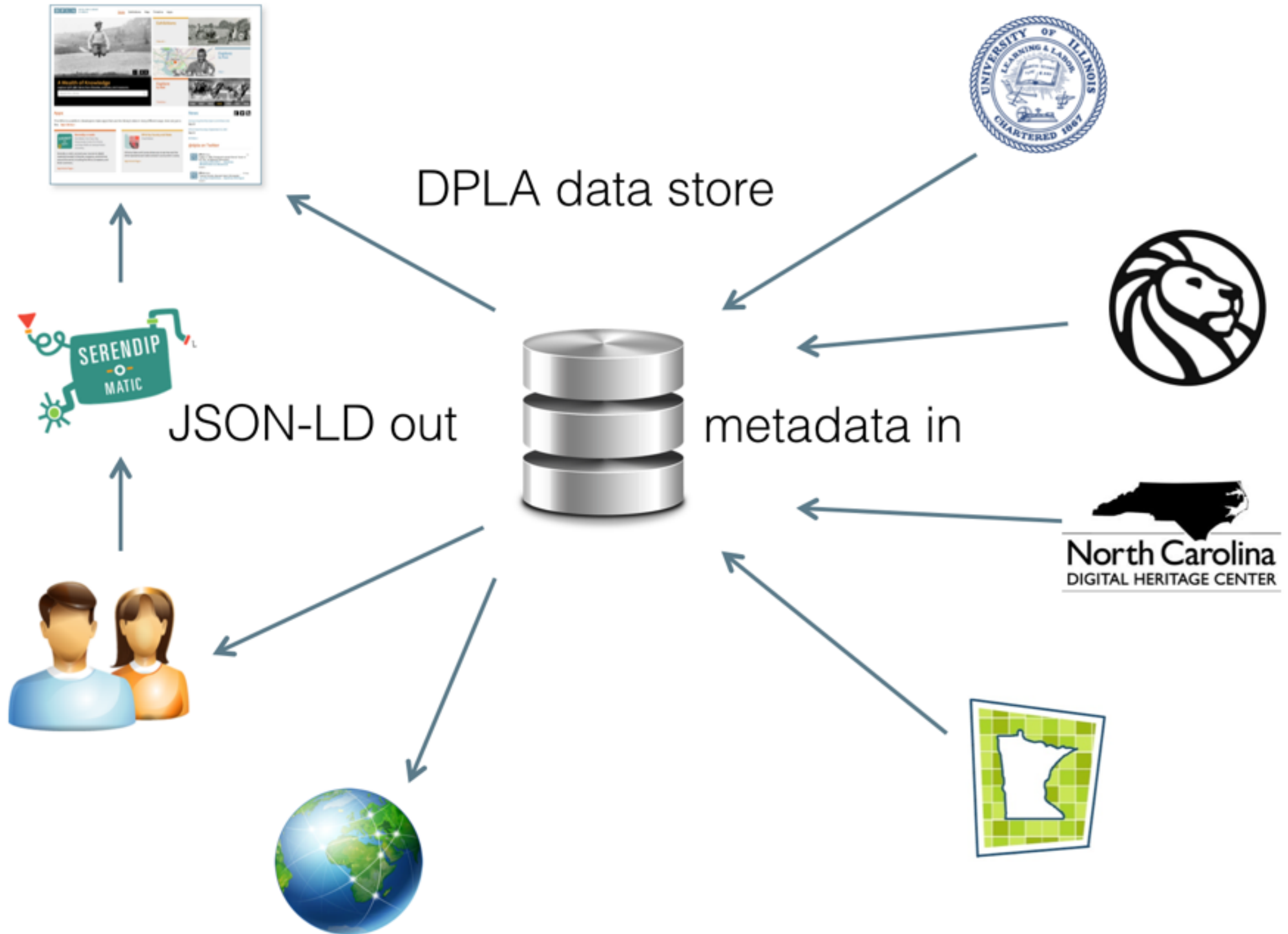
Outline

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

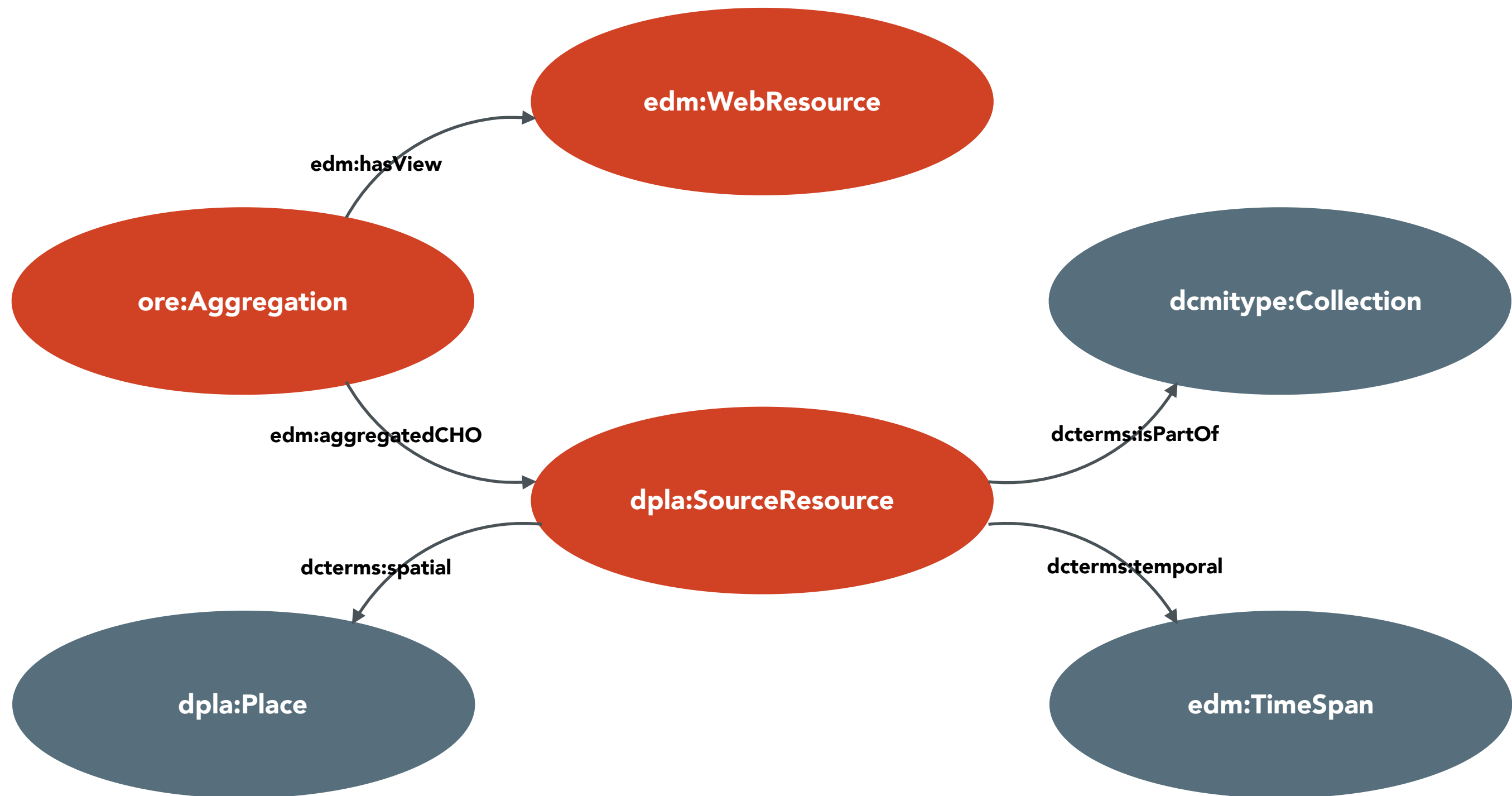
Original Infrastructure



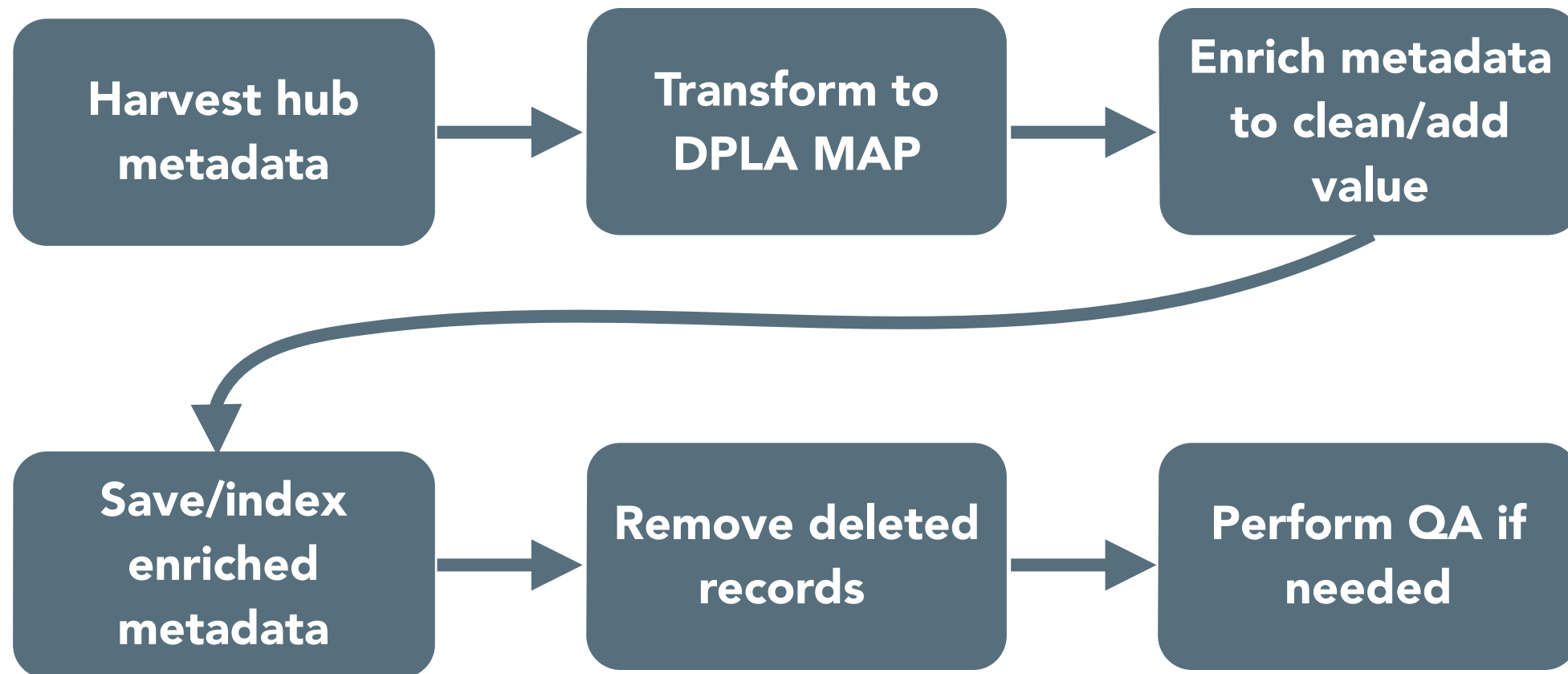
The DPLA Ingestion Process



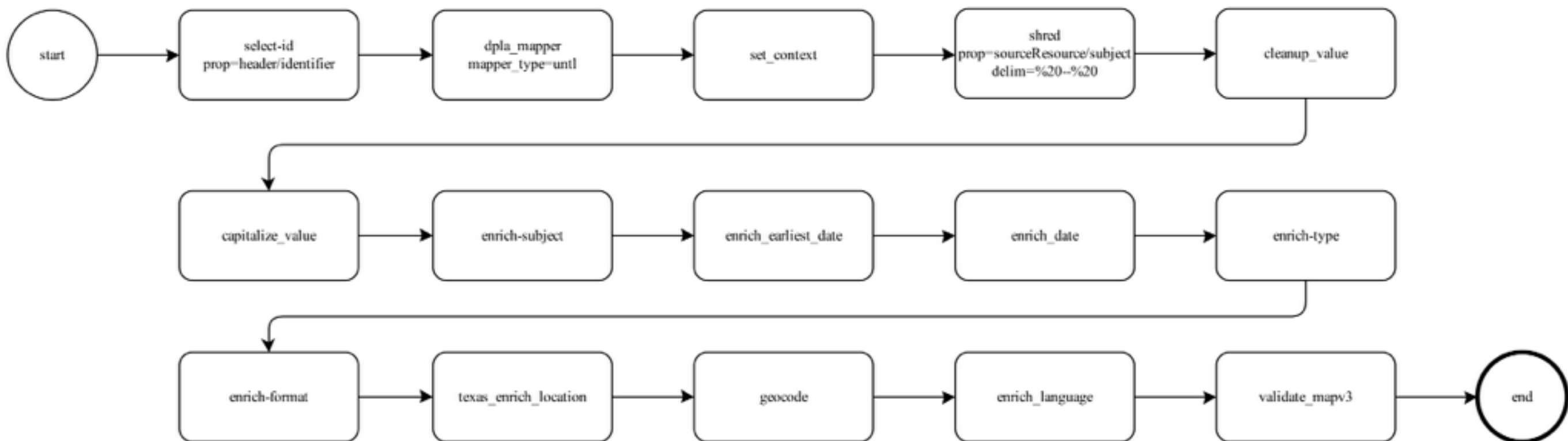
Metadata Application Profile



Ingestion workflow



Transformation & enrichment



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

Building Heiðrun

- 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

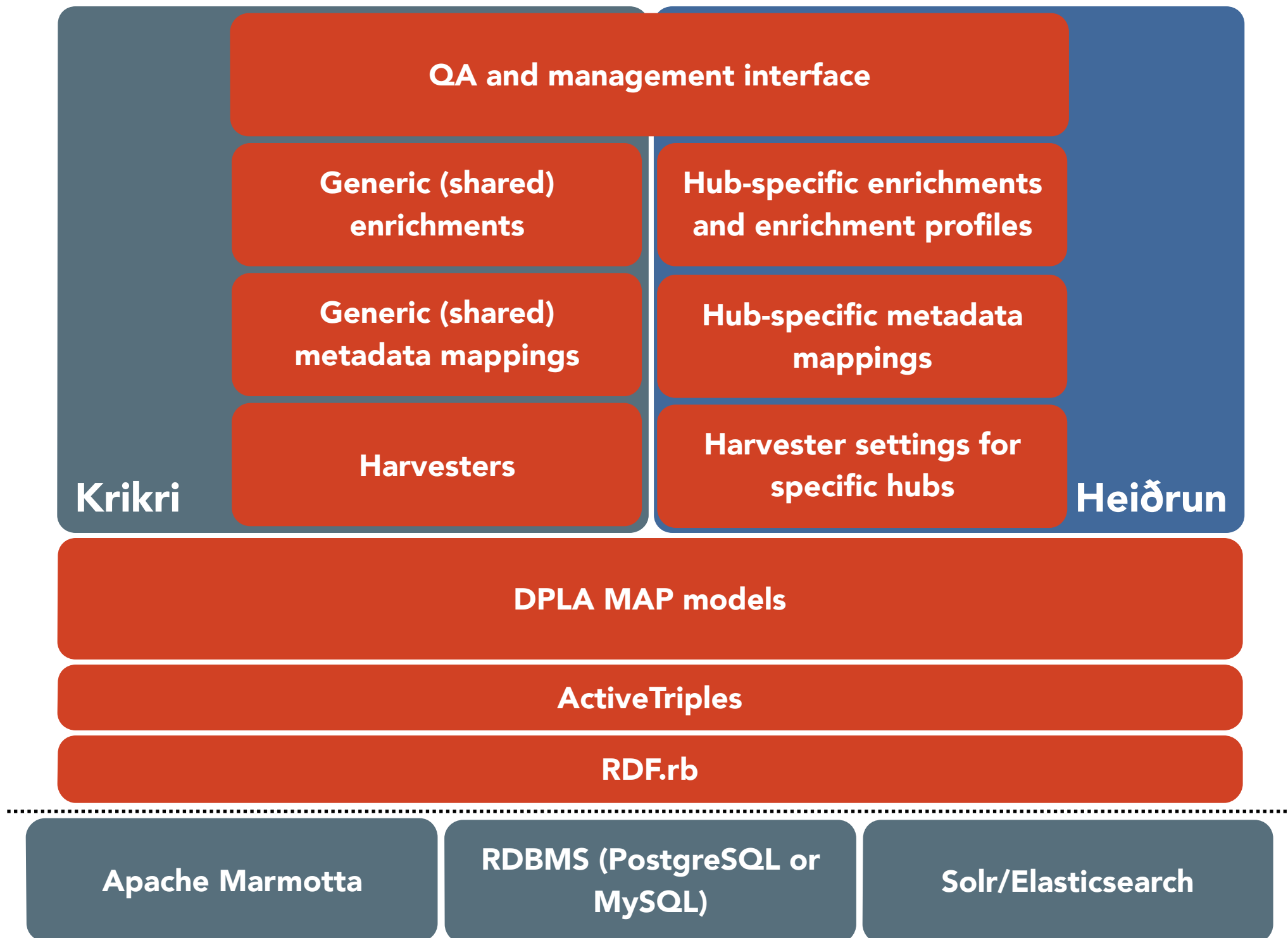
Thank You!

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



Resources

- DPLA ingestion system ("legacy" system). <https://github.com/dpla/ingestion>.
- DPLA new ingestion system code bases.
 - <https://github.com/dpla/heidrun>
 - <https://github.com/dpla/KriKri>
- Matienzo, Mark A. and Rudersdorf, Amy. The Digital Public Library of America Ingestion Ecosystem: Lessons Learned After One Year of Large-Scale Collaborative Metadata Aggregation. *Proc. Int. Conf. on Dublin Core and Metadata Applications, 2014*. <http://dcpapers.dublincore.org/pubs/article/view/3700>.