DPLAH: A Hydra-Based DPLA Service Hub Model

Aggregator Prototype Team

This project was a statewide effort in Pennsylvania. Developers, metadata specialists, archivists, and others from Penn State, Temple University, the State Library, and the University of Scranton collaborated from their respective institutions to make this project come together.

- Linda Ballinger, Penn State
- Doreva Belfiore, Temple University
- Mohamed Berray, Penn State
- William Fee, State Library
- Andrew Gearhart, Penn State
- Ben Goldman, Penn State
- Patricia Hswe, Penn State
- Delphine Khanna, Temple University
- Katherine Lynch, Temple University
- Steven Ng, Temple University
- Kristen Yarmey, University of Scranton

Links

- DPLAH on Github: https://github.com/tulibraries/dplah
- Information about PDCP: http://www.powerlibrary.org/librarians/special-projects-office-of-commonwealth-libraries/project-for-dpla

What is DPLAH?

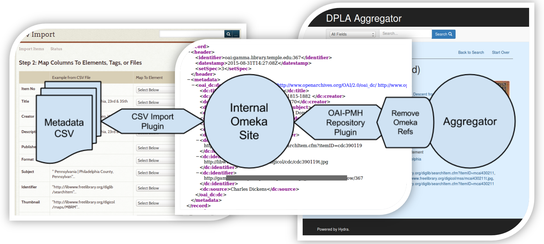
DPLAH is a Hydra-powered aggregator of OAI-PMH metadata with features specific to the DPLA's metadata and discovery layer needs. It was developed as part of the Pennsylvania Digital Collections Project for DPLA (PDCP), and will power the official Service Hub to the DPLA for the state of Pennsylvania.

Using the ruby-oai gem, the head harvests metadata from OAI seeds as XML. It normalizes the data with XSLT and Ruby. Normalized objects are added to Fedora and are made harvestable with OAI and are indexed in Solr to populate the Blacklight front end.

Metadata enhancements for objects in this head include: sanitizing for improved crosswalks, tokenized thumbnail and identifier linkback patterns, thumbnail detection drivers for common repository types, withholding items from harvest on a per-object basis, advanced mapping to encodings such as DCMI Types and ISO 639-1, and more. Mapping enhancements are optional and configured on a per-seed basis. Seeds are added to the head through an admin interface, from which seed jobs are also triggered. Jobs include harvesting and purging per-seed, per-institution, and for the entire repository. Certain dashboard jobs are also available as rake tasks to allow metadata specialists to do in-depth analysis of problematic records at each step of the harvesting process.

Passthrough Workflow

To support content providers whose repositories do not support any form of harvestable metadata, the DPLAH head offers the Passthrough Workflow. Metadata is delivered as a CSV document and includes the URI to the resource, and to its thumbnail. This is uploaded to an Omeka site using the CSV Import plugin and mapping the elements to the appropriate DC fields. OAI-PMH is exposed for harvesting from within the Omeka site using the OAI-PMH Repository plugin. Passthrough data is added to a seed in the aggregator, which harvests and then performs additional transformations on passthrough seeds to remove all passthrough references and properly detect and map the thumbnail.



Front End Preview

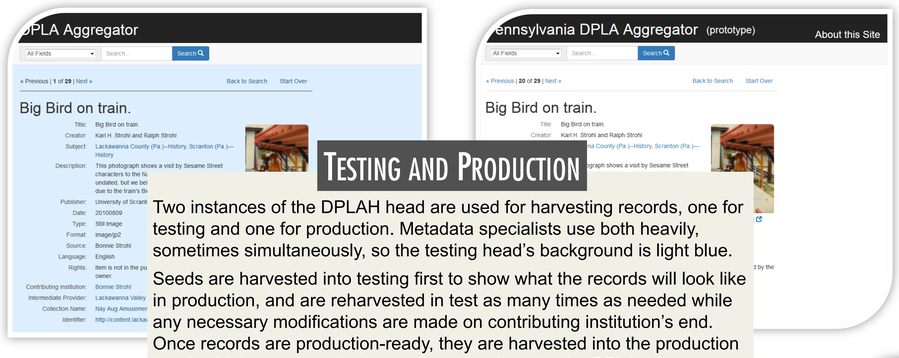
The Blacklight front end of the DPLAH head functions as a preview for metadata that has been transformed by humans and by the aggregator's automated and opt-in transformations, show how it will look for the DPLA. In the testing aggregator, this allows metadata specialists to clearly identify and communicate to correct issues before including in the production aggregator, and gives an idea to stakeholders what their data will look like in the DPLA before it is included.

Supported Metadata Formats

DPLAH currently supports harvesting the following OAI-PMH metadata formats.

OAI-DC: an XML scheme that supports simple Dublin Core under the Dublin Core Metadata Element Set base scheme for resource description. Passthrough Workflow metadata is outputted in this format.

OAI-QDC: like OAI-DC, except its scheme includes additional qualifier elements that refine the meaning and precision of metadata description of a resource (for instance, accessRights and rightsHolder in addition to rights).

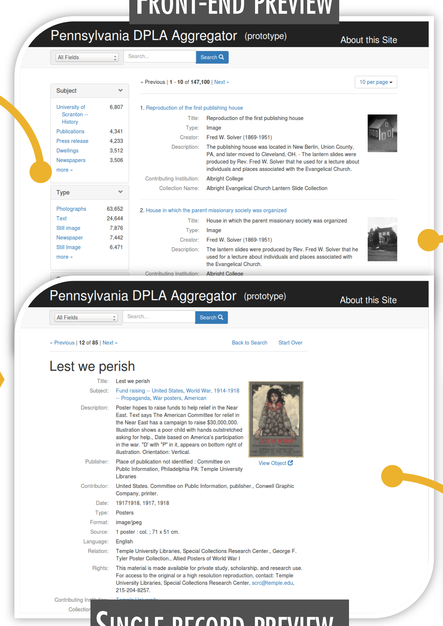


Testing and Production

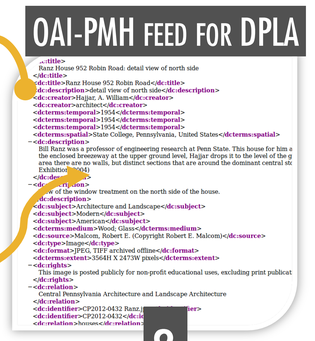
Two instances of the DPLAH head are used for harvesting records, one for testing and one for production. Metadata specialists use both heavily, sometimes simultaneously, so the testing head's background is light blue.

Seeds are harvested into testing first to show what the records will look like in production, and are reharvested in test as many times as needed while any necessary modifications are made on contributing institution's end. Once records are production-ready, they are harvested into the production head, which serves as the harvesting repository for the DPLA.

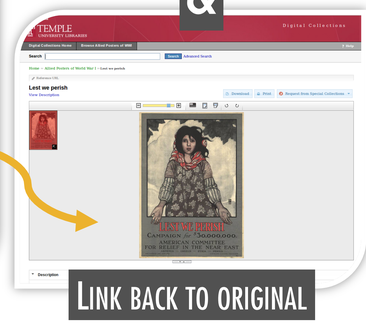
Front-end preview



OAI-PMH feed for DPLA



&

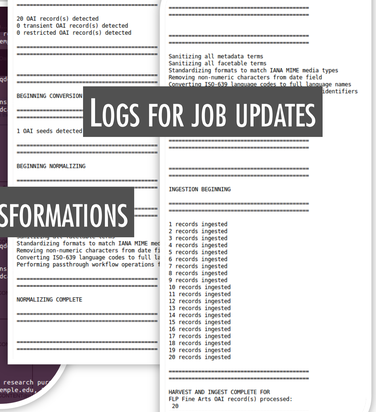


Create OAI Seed for each collection

Trigger jobs from Dashboard

Harvest metadata

Logs for job updates



Metadata Transformations

Single record preview

Link back to original

Harvesting Workflow

View: PDCP Project Timeline

August 2014

A group representing various Pennsylvania libraries and cultural heritage institutions met to discuss enhancing support for PA digital collections by exposing them to the DPLA through an official Service Hub in the state of Pennsylvania. This meeting sparked the Pennsylvania Digital Collections Project for the DPLA (PDCP).

October 2014

Temple University Libraries began working on a simple proof-of-concept metadata harvesting aggregator Hydra head, to investigate Hydra's viability for powering a DPLA Service Hub that could be made generic enough to share across Service Hubs and supporting institutions. The successful prototype included rudimentary functionality for harvesting metadata from, simple sanitizing it, and exposing it through OAI-PMH and Blacklight/Solr.

December 2014

Based on this early work, the PDCP Developers Team consisting of developers from Penn State and Temple developed DPLAH, a Hydra head based on the October prototype.

April 2015

The refined and improved prototype was completed and shared with the PDCP Teams, PA stakeholders, and the DPLA.

June & July 2015

An additional Developers' code sprint was held to add more features and enhancements to the head based on feedback and work by the PDCP Metadata Team and make it scalable / production-ready.

August 28, 2015

Just about one year to the day from the initial meeting, with 147,100 records from 165 collections at 29 PA institutions, the PDCP Service Hub was approved as the DPLA Service Hub for the state of Pennsylvania, with the Hydra-powered DPLAH as the software layer. Members of the PDCP are working with the DPLA now to prepare for our content's inclusion in the growing collections of the Digital Public Library of America.

The Pennsylvania Digital Collections Project (PDCP) is made possible in part by a grant from the Institute of Museum and Library Services (IMLS) as administered by the Pennsylvania Department of Education through the Office of Commonwealth Libraries, and the Commonwealth of Pennsylvania, Tom Wolf, Governor. Other sponsors include the Free Library of Philadelphia, The Pennsylvania State University, Temple University, and the University of Pennsylvania.