DPLA/ESDN/Metadata

John Mignault

November 15, 2016

What is DPLA?

- The Digital Public Library of America
  - A national digital collections project
  - Aggregates metadata from digital collections presented in their portal at `http://dp.la`

The Hub model

- Hubs supply DPLA with metadata
  - 2 types of hub
    - Content hubs
      - large digital collections that provide metadata for their own content, such as NYPL or BHL
    - Service hubs
      - Hubs that act as a conduit for DPLA ingestion, such as ESDN

What is ESDN?

- Empire State Digital Network
  - NY state service hub
  - formed in 2013
- ESDN coordinates metadata harvest, transformation, and ingest for NY state

- ESDN provides services to state regional council (ESLN) members seeking to get their metadata into DPLA
- Currently funded by and based at METRO

Second generation hub

- Most of the first generation of DPLA hubs were digital collections websites that provided OAI feeds from their content management systems
  - Biodiversity Heritage Library, NYPL, Digital Commonwealth (BPL)
- ESDN part of a "second wave" of hubs

The "Second Wave"

- Pure aggregators
  - we don't host any content ourselves
  - we have no public interfaces
- Currently we harvest partner metadata and transform it into a DPLA-approved format which they then harvest

What is our workflow?

Prep - Harvest - Transformation - Ingest

Prep

- Regional council members contact their ESDN liaison
- The member provides a letter granting permission to use their metadata
- The provider's contact person contacts Chris Stanton, ESDN's Metadata Specialist
- Chris works with the provider to normalize and standardize their data
  - Will often work on normalizations and tweaks using OpenRefine

Harvest

- We then pull their data into Repox
  - By various methods
    * But mostly and ultimately via OAI-PMH
    * There are a few others, like CSV import into Omeka

Transformation

- Their data is mapped to the ESDN Metadata Application Profile (MAP)
  - a variant of MODS
  - which is itself mapped to the DPLA MAP
  - it's an iterative process
    * often there will be multiple rounds of writing and rewriting the transforms
  - it's a very manual process
    * tweaking the output

Ingest

- Once the data has been sufficiently massaged it is ingested by DPLA
- They perform additional transformations and enrichments
- The result is then QA'd
  - sometimes we need to go back and make additional adjustments
- The records then finally go live on DPLA

Issues

what are we trying to achieve?

- We try to strike a balance between asking the provider to normalize data and writing horrendous special cases
- Problems are mostly dependent on how consistently the data has been entered into the CMS

  - minor inconsistencies can be written around
  - major inconsistencies require the provider to go back and edit the data

Sidebar: the date field, aka "dumpster fire"

- roman numerals: MCMXVIII
- natural language dates: June 16, 1904
- CONTENTdm timespans: 1932;1933;1934;1935;1936 to represent 1932-1936
- People will put anything and their dog in the date field

What is Repox?

OAI aggregator software

Past

- originally written for the Europeana project
- appeared to have been abandoned in 2014
  - as such, documentation is spotty
- used by a number of more recent DPLA hubs

Present

- DPLA hub user community
  - mailing list
- Has a number of undocumented quirks
  - the "records per page" quirk
- support folklore has sprung up around it
  - everyone uses it by necessity
  - its relative limbo makes it less than ideal

Future

- the long-awaited "Repox killer"
  - DPLA has been rumored to be be working on an aggregator appliance
  - part of the Hydra in a Box project
  - rumors of its birth are greatly exaggerated

How does Repox work?

Overview

- We harvest partner feeds in various formats and protocols
- Our outgoing format is always ESDN MAP MODS
- we define "data sets" that specify the incoming format depending on the CMS in use
- we then write XSL stylesheets that transform the harvested data to ESDN MAP MODS
- we attach those stylesheets to the data set
- when DPLA ingests data they harvest the entire repository in ESDN MAP format

XSLT

- we built on the great work done at NCDHC
- we forked their Github repository of XSLT style-sheets for use with Repox
- NCDHC mainly works with CONTENTdm providers
- As the number of different CMSes we saw grew, we developed a CSS-like cascading model for stylesheets
- our repository is on Github at `http://github.com/esdnhub/dpla-custom-repox-xslt`

Lessons

Metadata Improvements

- DPLA as a national project is actually improving metadata at the state and local levels
  - It provides an impetus for folks to address issues in their data
  - it provides justification to administrators to lend staff time to clean-up
- it has led to the formation of the ESDN metadata group (`http://empirestate.digital/governance/metadata-working-group/`
  - creating best practices for creating shareable metadata.

Local history, global data

Data re-use

- No public front-end
- scope of available data constrained by MAP
  - repox cannot link to external resources
    - no LOD
    - limitation of XSLT
- our partners need reporting and statistics tools
- so, we attempted to build a basic tool

The "collstool"

- uses MODS data harvested from our Repox instance
  - XML -> JSON -> YAML using XSLT and json2yaml
  - Jekyll reads resulting YAML file
- ungainly, manual process that output inaccurate results
- Running on Github Pages at `http://esdnhub.github.io/collstool/`
- source available at `http://github.com/ESDNHub/collstool`

The ESDN portal

- Plans to work on a NY state wide search portal
- sort of a "tiny DPLA" for NY state
- Blacklight Rails application
  - Based on Ben Armintor's "DBLA" gem `https://github.com/barmintor/dbla`

How does it work?

- makes Blacklight think it's talking to a solr repository
  - when it's actually talking to the DPLA API
- extended to integrate additional "external vocabularies"
  - Sub-collection info
  - Council info
  - Possibly LCSH
- Build additional reporting and search capabilities
- Prototype is up at `https://afternoon-shelf-4328.herokuapp.com`

DPLA Exhibitions

- Avilable at `http://dp.la/exhibitions`
- Built on Omeka (`http://omeka.org`)
- ESDN will begin providing DPLA exhibition hosting for partners
  - Later in 2016

Questions?

- Thanks!
- John Mignault (jmignault@metro.org)
- METRO (`http://metro.org`)

- ESDN (`http://empirestate.digital`)
- Empire State Library Network (esln.org)
- DPLA (`http://dp.la`)