Platform for Open Discovery

Anthony Helm, Brown University
Elisabeth Long, University of Chicago
Emily Morton-Owens, UPenn
Esmé Cowles, Princeton University
Hector Correa, Brown University

Joe Zucca, UPenn
Mark A. Matienzo, Stanford University
Simeon Warner, Cornell University
Tim McGearry, Duke University
Tom Cramer, Stanford University
PLATFORM FOR OPEN DISCOVERY

POD is working to create a platform that positions data reuse and service integration as strategic assets. Through open, iterative development and leveraging the investment in our libraries’ internal capacities, we will meet multiple library needs and enable innovation in ways that cannot be done through a series of one-off solutions or relying on vendors and external systems.
In Other Words...

1. **Gather** data from IPLC institutions
2. **Pool** the data for easy reuse
3. **Enrich** the aggregated data
4. **Deploy** to support varying needs

and do this in a way that...

5. **Enables innovation** by reducing friction
6. **Builds capacity** within IPLC
7. Recognizes **data as a reusable asset**
USE CASES

1. Resource sharing: discovery
2. Collections analysis
3. Linked data & data sharing
4. Data mining
5. AI
ONE DATA FEED, MANY POSSIBLE USES

**MARC Staged** from *n* Contributing Institutions

Upper Lake

MARC Normalized as needed

Middle Lake

MARC Optimized for Analysis

Lower Lake

Export for Analysis, etc.

Single feed per participating institution

Feeds for ILL discovery

Potential data feeds for collection analysis, et al. (e.g., Gold Rush)
DATA FLOW

1. Publish Data
2. Aggregate
3. Refine
4. Operationalize
5. Reuse
DATA FLOW: ILL DISCOVERY

1. Publish Data
   - Export records

2. Aggregate
   - Gather in data lake

3. Refine
   - Normalize (as needed)

4. Operationalize
   - Locally load & index

5. Reuse
   - Blacklight / Alma / VUFind
PROJECT STATUS

- Organization: 9 participating institutions with lightweight org and governance
- Data: 75M records amalgamated from 9 sources + Hathi Trust
- Proof-of-concept Blacklight environment operational (!), proves approach
- Consensus requirements for data feeds and interaction with ILL fulfillment
- Explore possible extensions beyond ILL into collections analysis
- Build data lake for aggregating and syndicating data feeds (**work in progress**)
A **data lake** is a repository for structured, unstructured, and semi-structured data, allowing data to be in its rawest form without needing to be converted and analyzed first.

Source: [https://learn.g2.com/what-is-a-data-lake](https://learn.g2.com/what-is-a-data-lake)
THE DATA LAKE MVP ELEVATOR PITCH

The POD Data Lake MVP will produce a openly developed **minimum viable product** to receive and transmit MARC bibliographic and holdings data from multiple institutions.

It will support **receiving and transmitting both full dumps and delta change sets** for external normalization, analysis, and discovery indexing, and basic reporting and data reconciliation.

It will also serve as a project to **connect an engineering team with data providers and consumers**, and will help POD **refine and develop operational models** for a service based on the MVP, situating data transfer and aggregation within the POD ecosystem.
DATA LAKE MVP PROJECT TEAM

- Product Owner: Mark Matienzo (Stanford)
- Tech Lead: Chris Beer (Stanford)
- Scrum Master: Jessie Keck (Stanford)
- Developers:
  - Stanford: Chris Beer, Jessie Keck, Jack Reed, Camille Villa
  - Brown: Adam Bradley, Hector Correa, Birkin Diana, Justin Uhr
- UX Design Consultant: Gary Geisler (Stanford)
- Data Liaison/Wrangler: Bob Persing (Penn)
UPLOADER

- Supports uploads from dashboard, API, or remote URL (on a web server)

New Upload

You can also upload by submitting a POST with an http client (e.g. curl command). Note that you should only provide either upload[files]

```
curl -F 'upload[name]=[NAME_OF_YOUR_FILE]' \
     -F 'upload[files][]=[LOCATION_TO_YOUR_FILE]' \
     -H 'Authorization: Bearer ey 
     https://pod.stanford.edu/organizations/stanford/uploads
```

Name

Upload file

OR

Upload via URL

https://example.com/12345.marc

Create Upload
DATA PROFILING TOOLS

- Summary information (inclusion of 001s, multilingual data, holdings)

<table>
<thead>
<tr>
<th>Description</th>
<th>Fields</th>
<th>Info</th>
</tr>
</thead>
<tbody>
<tr>
<td>001</td>
<td>001</td>
<td>8243438 of 8941463</td>
</tr>
<tr>
<td>MHLD information</td>
<td>85x, 86x, 87x</td>
<td>85x: true 86x: true 87x: true</td>
</tr>
<tr>
<td>Mutilingual data</td>
<td>88x</td>
<td>true</td>
</tr>
</tbody>
</table>
**DATA PROFILING TOOLS**

- Summary information (inclusion of 001s, multilingual data, holdings)
- Histogram of MARC field and subfield occurrence

### Histogram

<table>
<thead>
<tr>
<th>Field</th>
<th>%</th>
<th>Occurrences</th>
</tr>
</thead>
<tbody>
<tr>
<td>001</td>
<td>92.19%</td>
<td>1x: 8243436</td>
</tr>
<tr>
<td>002</td>
<td>0.00%</td>
<td>1x: 4</td>
</tr>
<tr>
<td>003</td>
<td>92.19%</td>
<td>1x: 8243438</td>
</tr>
<tr>
<td>005</td>
<td>92.19%</td>
<td>1x: 8243433</td>
</tr>
<tr>
<td>006</td>
<td>15.27%</td>
<td>1x: 1362260, 2x: 3218, 3x: 68, 4x: 15, 5x: 3</td>
</tr>
<tr>
<td>007</td>
<td>20.42%</td>
<td>1x: 1579781, 2x: 233756, 3x: 12749, 4x: 53, 5x: 6, 6x: 1</td>
</tr>
<tr>
<td>008</td>
<td>92.16%</td>
<td>1x: 8240730, 2x: 7</td>
</tr>
<tr>
<td>009</td>
<td>0.00%</td>
<td>1x: 9</td>
</tr>
<tr>
<td>010</td>
<td>40.93%</td>
<td>1x: 3634834, 2x: 24298, 3x: 949, 4x: 14, 5x: 1, 7x: 1</td>
</tr>
<tr>
<td>260</td>
<td>77.41%</td>
<td>1x: 6917548, 2x: 3326, 3x: 519, 4x: 105, 5x: 36, 6x: 12, 7x: 3, 8x: 1, 12x: 1</td>
</tr>
</tbody>
</table>

### Subfield Occurrences

<table>
<thead>
<tr>
<th>Subfield</th>
<th>%</th>
<th>Occurrences</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.000%</td>
<td>1x: 28</td>
</tr>
<tr>
<td></td>
<td>0.000%</td>
<td>1x: 1</td>
</tr>
<tr>
<td></td>
<td>0.000%</td>
<td>1x: 2</td>
</tr>
<tr>
<td>1</td>
<td>0.001%</td>
<td>1x: 81</td>
</tr>
<tr>
<td>2</td>
<td>0.000%</td>
<td>1x: 28</td>
</tr>
<tr>
<td>3</td>
<td>0.050%</td>
<td>1x: 1482, 2x: 1594, 3x: 330, 4x: 69, 5x: 25, 6x: 11, 7x: 1, 11x: 1</td>
</tr>
<tr>
<td>6</td>
<td>8.915%</td>
<td>1x: 616913, 2x: 148, 3x: 15, 5x: 1</td>
</tr>
</tbody>
</table>
DATA PROFILING TOOLS

- Summary information (inclusion of 001s, multilingual data, holdings)
- Histogram of MARC field and subfield occurrence
- Listing of non-standard field usage

Non-standard subfields

<table>
<thead>
<tr>
<th>Field</th>
<th>Subfield</th>
<th>Sample values</th>
</tr>
</thead>
<tbody>
<tr>
<td>010</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2814432671, /agb985468</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>0912300</td>
</tr>
<tr>
<td>c</td>
<td></td>
<td>CST, CST-H</td>
</tr>
<tr>
<td>d</td>
<td></td>
<td>CST, RCJ, W1215900</td>
</tr>
<tr>
<td>e</td>
<td></td>
<td>editor2017944106, dacs, W1215900</td>
</tr>
<tr>
<td>f</td>
<td></td>
<td>N0371300</td>
</tr>
<tr>
<td>g</td>
<td></td>
<td>N0371300</td>
</tr>
<tr>
<td>o</td>
<td></td>
<td>ocm11043999</td>
</tr>
<tr>
<td>q</td>
<td></td>
<td>(electronic), (USB), (print-on-demand)</td>
</tr>
</tbody>
</table>
FOR MORE DATA LAKE INFO...

- Data Lake MVP: pod.stanford.edu
- GitHub Repository: github.com/ivplus/aggregator
- 7M Penn records
- 43M titles total
- Default limit to Penn
- Expand to POD
- Direct link to partner libraries
Enable innovation and reduce bureaucratic drag
Appropriately apportion risk, responsibility and accountability to actors

**Data Provider Minimum Standards**
- Contribute only data that may be shared to facilitate public discovery
- Agree that data may be used by other POD libraries for purposes such as research, testing, or development
- Encouraged to provide data under permissive terms, with license (e.g., a CC0 waiver).

**Data Consumers**
- Data in the lake is available to any POD participant to extract and reuse.
- Participant must consume and reuse the data responsibly.
- Contact originating library for uses outside minimum standards

[Full POD Data Provider & Usage Framework](#)
DISCOVERY & FULFILLMENT

- Initial vision had a cloud hosted shared index of POD data (à la BDSI) or even a shared Blacklight instance

- Current focus on **indexing POD data locally**
  - Bibliographic and holdings records
  - Does not include availability

- Fulfillment via existing Relais (or successor)
WHO IS POD?
COMMUNICATIONS

● Mailing Lists
  ○ Slack (Open): pod4lib.slack.com
  ○ Announcements (Open): POD-announce@googlegroups.com
  ○ Technical Coordination (Open): POD-tech@googlegroups.com
  ○ Contact: LIBpod@o365lists.upenn.edu

● Monthly Reports

● Biweekly meetings of the POD Tech group
CURRENT POD DATA SOURCES

- 9 participating institutions
- CLIO Open Data (Columbia) - https://library.columbia.edu/bts/clio-data.html
- HathiTrust
- We welcome data contributions from additional sources
CALL TO ACTION

● Explore the Demo site
  ○ Penn: https://blacklight-test.library.upenn.edu/catalog
  ○ Username = pod, Password = dolphin


● Provide feedback via email: LIBpod@o365lists.upenn.edu

● Join our communication channels/listservs

● Contribute your data

● Join the project
LOOKING FORWARD TO 2021

- More work on the MVP
- Sustaining governance while increasing the number of partners
- Engaging local solutions using POD data for BorrowDirect fulfillment
- Explore solutions using POD data for Collections analytics
- Review and assess initial infrastructure choices
- Exploring intersections with other Ivy Plus priorities