In 2019, the Ivy Plus Libraries Confederation (IPLC) launched a concerted effort to enhance discovery for its inter-library loan program, BorrowDirect.

**The challenge:** what is the best way to aggregate data from 13 institutions to meet IPLC’s consortial needs?
PLATFORM FOR OPEN DATA

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**
A DATA LAKE FOR IPLC

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.

ONE FEED, MANY USES

Source: https://learn.g2.com/what-is-a-data-lake
ONE DATA FEED, MANY POSSIBLE USES

Inst 1

Inst 2

... Inst n

**Upper Lake**

MARC Staged from n Contributing Institutions

**Middle Lake**

MARC Normalized as needed

**Lower Lake**

MARC Optimized for Analysis

Single feed per participating institution

Feeds for ILL discovery

Potential data feeds for collection analysis, et al. (e.g., Gold Rush)
THREE DISTINCT CLUSTERS OF USE CASES

Resource Discovery, Access & Sharing
- Physical and digital access across IPLC

Collections Analysis & Decision Support
- View on collective holdings to inform local action

Data Innovation & Enrichment
- Data mining
- Linked data
- AI
USE CASES -- RESOURCE DISCOVERY, SHARING & ACCESS

1. BorrowDirect Discovery & Fulfillment Facilitation
2. Digital Resource Sharing for public domain, OA, repository content
3. Special Collections: IPLC-wide Virtual Reading Room
4. Serials Analysis for Copyright Status & Enhanced Access
5. Controlled Digital Lending for BorrowDirect / IPLC
6. Catalog Enrichment & Remediation
7. Catalog Matching, Deduplication, Linked Data Transformation
8. Potential ERM Analysis, Pooling
USE CASES -- COLLECTIONS ANALYSIS

1. Collections Analysis (POD feed to Gold Rush)
2. Shared Print Retention Commitment Management & Analysis
   a. Deselection, Gap filling, understand rare/unique
3. Data Mining to Support Research
   a. …autosuggest URIs for publishers in cataloging UI
   b. …art & architecture acquisitions over 20 years
4. DEI Analysis of IPLC Collections
5. IPLC Collections Intelligence:
   a. Inform Digitization, Preservation, Replace Lost Item Decisions
6. E-resource Holdings Analysis to HathiTrust, Open Library, etc.
GETTING DATA INTO POD

- pod.stanford.edu
- Bib, item and holdings data
- Upload data by
  - Dashboard
  - API
  - Remote URL
- Support for
  - Full data dump
  - Incremental changes

DATA PROFILING TOOLS

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

<table>
<thead>
<tr>
<th>Name</th>
<th>Files</th>
<th>Size</th>
<th>Unique records</th>
<th>Total records</th>
<th>Last updated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brain</td>
<td>40</td>
<td>2.08 GB</td>
<td>4,326,101</td>
<td>5,610,145</td>
<td>2020-11-17 16:32:27 UTC</td>
</tr>
<tr>
<td>Chicago</td>
<td>0</td>
<td>0 bytes</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Columbia</td>
<td>0</td>
<td>0 bytes</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Cornell</td>
<td>0</td>
<td>0 bytes</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Dartmouth</td>
<td>0</td>
<td>0 bytes</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Duke</td>
<td>0</td>
<td>0 bytes</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Harvard</td>
<td>60</td>
<td>4.63 GB</td>
<td>0</td>
<td>0</td>
<td>2020-11-19 23:25:29 UTC</td>
</tr>
<tr>
<td>Ivy University</td>
<td>6</td>
<td>216 MB</td>
<td>558</td>
<td>551</td>
<td>2020-11-21 00:24:54 UTC</td>
</tr>
<tr>
<td>Johns Hopkins</td>
<td>0</td>
<td>0 bytes</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Library of Congress</td>
<td>3</td>
<td>193 MB</td>
<td>750,000</td>
<td>750,000</td>
<td>2020-11-19 02:19:14 UTC</td>
</tr>
<tr>
<td>markm test</td>
<td>3</td>
<td>2.48 GB</td>
<td>2,257,021</td>
<td>6,878,488</td>
<td>2020-11-25 01:51:20 UTC</td>
</tr>
<tr>
<td>MIT</td>
<td>0</td>
<td>0 bytes</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Penn</td>
<td>0</td>
<td>0 bytes</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Princeton</td>
<td>0</td>
<td>0 bytes</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>test</td>
<td>0</td>
<td>0 bytes</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Yale</td>
<td>140</td>
<td>23 GB</td>
<td>15,577,658</td>
<td>21,486,623</td>
<td></td>
</tr>
</tbody>
</table>

## Uploads

### Files

- Ivy University 2020-11-30T21:09:25Z
- application/marc
- 12.4 KB
- 24

### Records

- Ivy University 2020-11-30T21:09:25Z
- application/marc
- 6.48 KB
- 14

- vernacularSearchTests.mrc
- application/marc
- 12.4 KB
- 24

- vernacularNonSearchTests.mrc
- application/marc
- 6.48 KB
- 14
FRANKLIN DEMO

- 2020 proof of concept
- 7M Penn records
- +43M titles from POD
- Default search Penn
- Expand to POD
- Direct link to partner libraries
PROJECT STATUS

- Broad participation across IPLC; 2+ years of continuous work
- Established ongoing team from 10+ institutions and 25+ individuals
- Operational data lake running since Dec 2020
- Data: 60M unique records amalgamated from 13 IPLC sites
- Two proof-of-concept local discovery environments operational
- 2022 development workcycle planned to put POD into production
- Full population of data lake with 100M records (ETA May 2022)
- Production feeds to support ILL discovery & fulfillment (ETA June 2022)
- Extension to additional use cases (2022-23)
● pod.stanford.edu
● GitHub Repository: github.com/ivplus/aggregator
● Contact: LIBpod@o365lists.upenn.edu
Possible Discussion Questions

- How does the concept of a data lake fit with shared print (and other) needs?
- What data would be most interesting to have readily accessible?
- What tools would be most interesting to have connected to lakes?