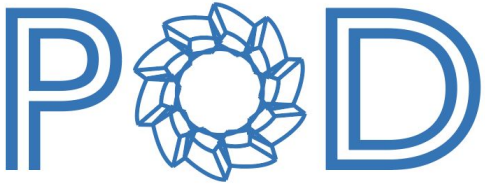
Platform for Open Discovery

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 McGeary, 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

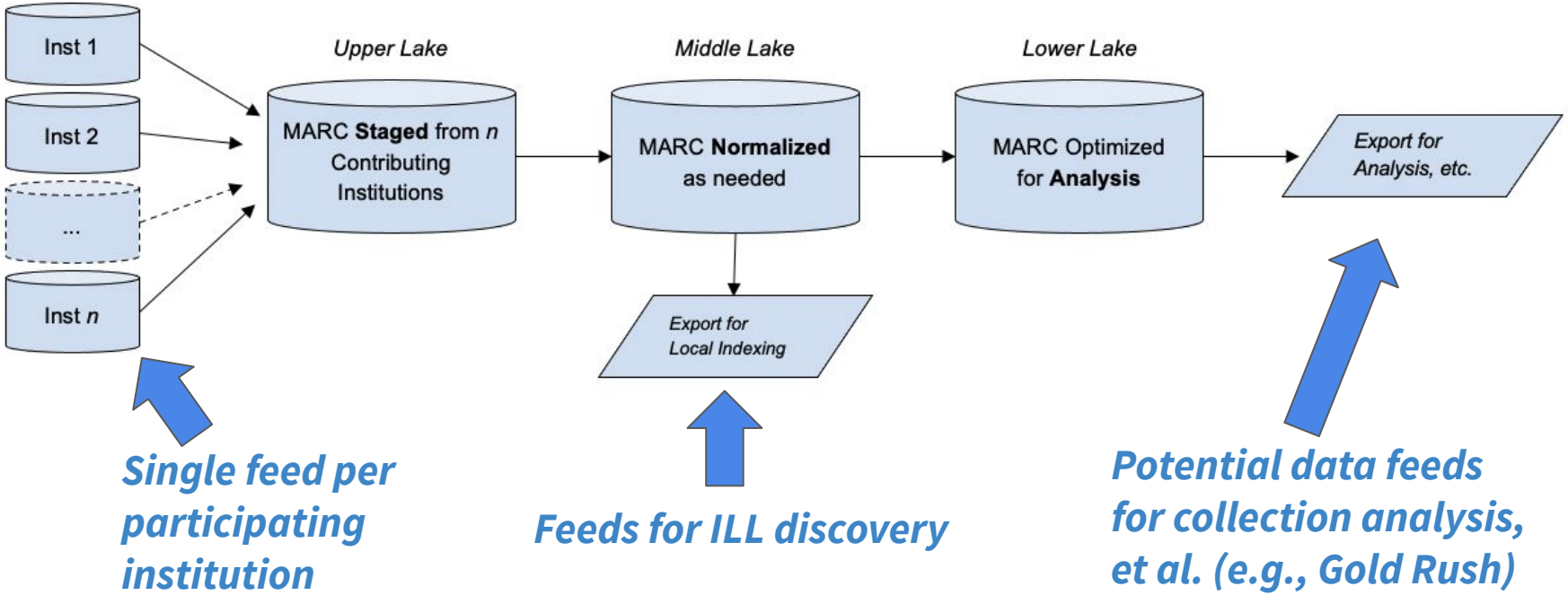
and do this in a way that...

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



DATA FLOW

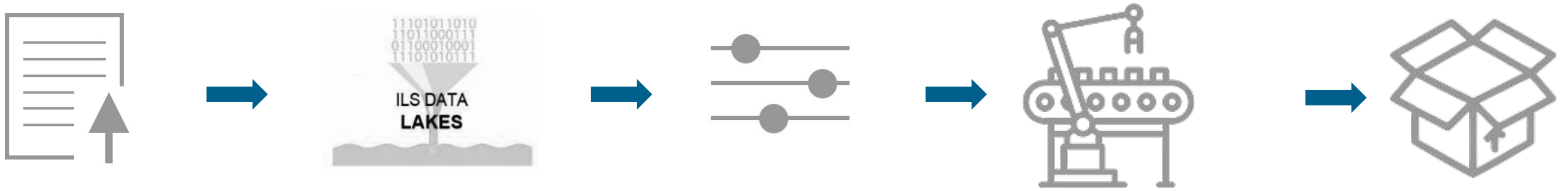
1 Publish Data

2 Aggregate

3 Refine

4 Operationalize

5 Reuse



DATA FLOW: ILL DISCOVERY

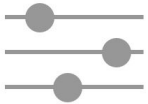
1 Publish Data

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***)

DATA LAKE MVP DEVELOPMENT

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

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)

UPLOADS

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

POD Aggregator Organizations Activity Users Mark Matienzo (matienzo@stanford.edu) Logout

Stanford

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][]` **OR** `upload[url]`.

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

Name

Upload file Browse

OR

Upload via URL

https://example.com/12345.marc

Create Upload

DATA PROFILING TOOLS

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

Stanford

Streams

full_20201124

Summary

Description	Fields	Info
	001	8243438 of 8941463
MHLD information	85x, 86x, 87x	85x: true 86x: true 87x: true
Mutilingual data	88x	true

DATA PROFILING TOOLS

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

Histogram

Field	%	Occurences	Subfields
001	92.193%	1x: 8243438	
002	0.000%	1x: 4	
003	92.193%	1x: 8243438	
005	92.193%	1x: 8243433	
006	15.272%	1x: 1362260, 2x: 3218, 3x: 68, 4x: 15, 5x: 3	
007	20.426%	1x: 1579781, 2x: 233756, 3x: 12749, 4x: 53, 5x: 6, 6x: 1	
008	92.163%	1x: 8240730, 2x: 7	
009	0.000%	1x: 9	
010	40.934%	1x: 3634834, 2x: 24298, 3x: 949, 4x: 14, 5x: 1, 7x: 1	>
260	77.410%	1x: 6917548, 2x: 3326, 3x: 519, 4x: 105, 5x: 36, 6x: 12, 7x: 3, 8x: 1, 12x: 1	>

Subfield	%	Occurences
	0.000%	1x: 26
	0.000%	1x: 1
	0.000%	1x: 2
1	0.001%	1x: 81
2	0.000%	1x: 28
3	0.050%	1x: 1462, 2x: 1594, 3x: 330, 4x: 69, 5x: 25, 6x: 11, 7x: 1, 11x: 1
6	8.915%	1x: 616913, 2x: 148, 3x: 15, 5x: 1

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

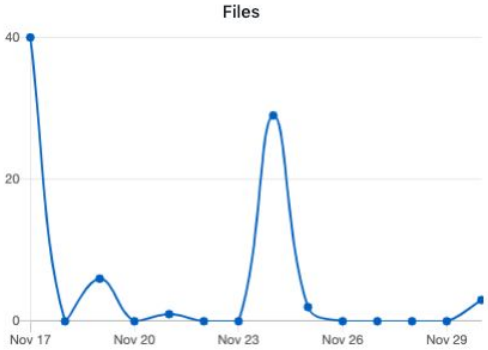
Field	Subfield	Sample values
010		
		2014432671, /agb985468
	1	0912300
	c	CSt, CSt-H
	d	CSt, RCJ, W1215900 ›
	e	editor2017944106, dacs, W1215900
	f	N0371300
	g	N0371300
	o	ocm11043999
	q	(electronic), (USB), (print-on-demand) ›

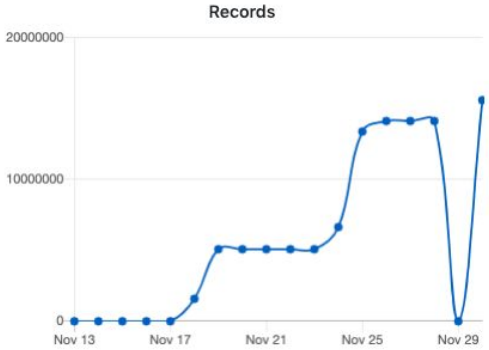
STATISTICS

Organizations

Name	Files	Size	Unique records	Total records	Last updated
Brown	40	2.06 GB	4,326,101	5,616,145	2020-11-17 16:32:27 UTC
Chicago	0	0 Bytes	0	0	
Columbia	0	0 Bytes	0	0	
Cornell	0	0 Bytes	0	0	
Dartmouth	0	0 Bytes	0	0	
Duke	0	0 Bytes	0	0	
Harvard	60	4.63 GB	0	0	2020-11-19 23:25:29 UTC
Ivy University	6	210 MB	508	551	2020-11-21 00:24:54 UTC
Johns Hopkins	0	0 Bytes	0	0	
Library of Congress	3	193 MB	750,000	750,000	2020-11-19 02:19:14 UTC
markm test	3	2.48 GB	2,257,621	6,876,489	2020-11-25 01:51:28 UTC
MIT	0	0 Bytes	0	0	
Penn	0	0 Bytes	0	0	
Princeton	0	0 Bytes	0	0	
Stanford	28	13.5 GB	8,243,438	8,243,438	2020-11-24 23:32:21 UTC
test	0	0 Bytes	0	0	
Yale	0	0 Bytes	0	0	
	140	23 GB	15,577,668	21,486,623	

Uploads

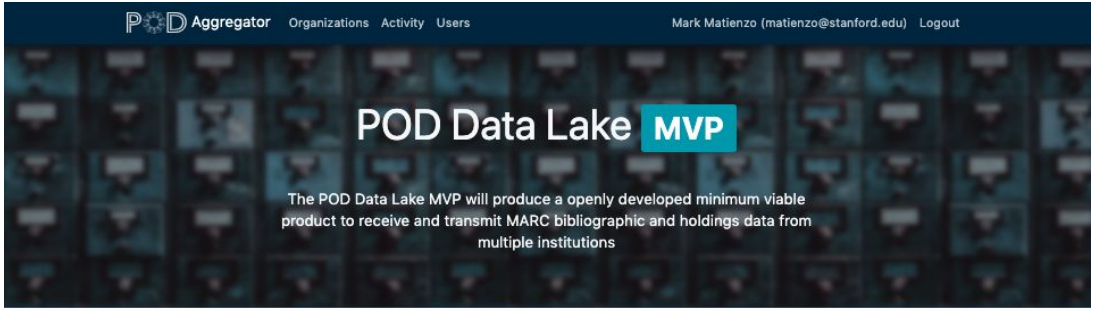




Organization	Name	Updated	File count	File size	Records	
Ivy University	2020-11-30T21:09:25Z	about 2 hours ago	28	120 KB	490	
	vernacularSearchTests.mrc	application/marc		12.4 KB	24	✅
	vernacularNonSearchTests.mrc	application/marc		6.48 KB	14	✅

FOR MORE DATA LAKE INFO…

- Data Lake MVP: pod.stanford.edu
- GitHub Repository: github.com/ivplus/aggregator



MVP Goal
Support receiving and transmitting both full dumps and delta change sets

MVP Goal
Enable external normalization, analysis, and discovery indexing, and basic reporting and data reconciliation

MVP Goal
Help POD refine and develop operational models for a service based on the MVP

FRANKLIN DEMO

- 7M Penn records
- 43M titles total
- Default limit to Penn
- Expand to POD
- Direct link to partner libraries

Penn Libraries Franklin

Known Issues | Franklin Help | Contact us | Bookmarks (0) | Library Home | Login

Keyword | Find books, journals, videos, & more | Search | Advanced Search

Everything | Catalog | Articles+ | Databases | Website

FEEDBACK

The Penn Libraries buildings are closed due to COVID-19. To perform a catalog search with results limited to online materials, start here. You can also use the bookmarks feature to save a list of materials to use later, but please log in first.

Limit your search

Search domain

Include Partner Libraries	42,592,409

Access

Online	3,117,950
At the library	4,287,725

Format

Book	6,157,525
Government document	1,056,464
Journal/Periodical	317,752

« Previous | 1 - 25 of 7,281,538 | Next »

Sort by Relevance | 25 per page

1. National geographic. Bookmark

Publication: [Washington, D.C.] : National Geographic Society, ©1959-

Format/Description: Journal/Periodical

Online resource: http://magma.nationalgeographic.com/ngm/data/html/home_refresh.html
HathiTrust Digital Library Connect to full text

Available. Kislak Center for Special Collections - Tehon Collection. Art and Design G1 .N27. Request to view
See options. Penn Museum Library. G1 .N27
See options. Van Pelt Library. G1 .N27
Available. LIBRA. G1 .N27
Available. LIBRA Special. G1 .N27. Request to view

Options

2. Guide to Indian periodical literature : social sciences and humanities. Bookmark

Publication: Gurgaon : Indian Documentation Service, 1964-

REQUIREMENTS FOR DATA FEEDS & USE

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

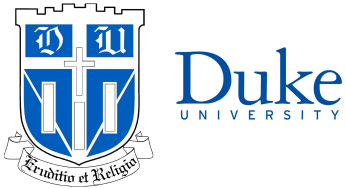
DISCOVERY & FULFILLMENT

- 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?


BROWN
UNIVERSITY


CORNELL UNIVERSITY
FOUNDED A.D. 1865
Cornell University


Duke
UNIVERSITY
Eruditio et Religio


Stanford
University


THE UNIVERSITY OF
CHICAGO


Penn
UNIVERSITY of PENNSYLVANIA


Dartmouth
1769


PRINCETON
UNIVERSITY


VE RI TAS
HARVARD
LIBRARY

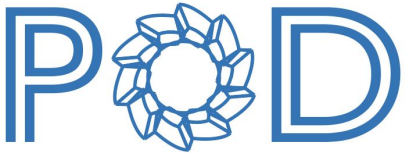
COMMUNICATIONS

- 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*
- Review the Data Usage Framework: https://bit.ly/pod-data-framework
- 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