Curating Big Data for Reuse

Harvard Dataverse

Open Repositories 2023

Katie Mika
Data Services Librarian
Harvard Library & IQSS

Gustavo Durand
Tech Lead / Architect
IQSS

Leonid Andreev
Senior Software Developer
IQSS

Challenges

Volume

- Large files
- Large number of files

Variety

- Diverse data types

Velocity

- Instrument data
- Streaming data

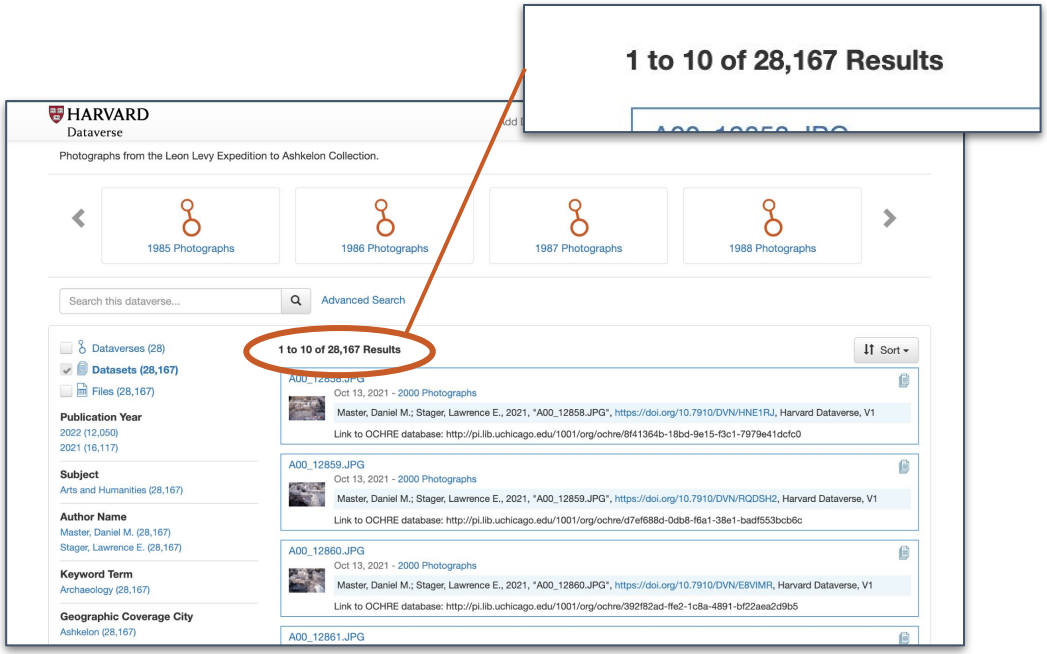
Veracity

- Data quality

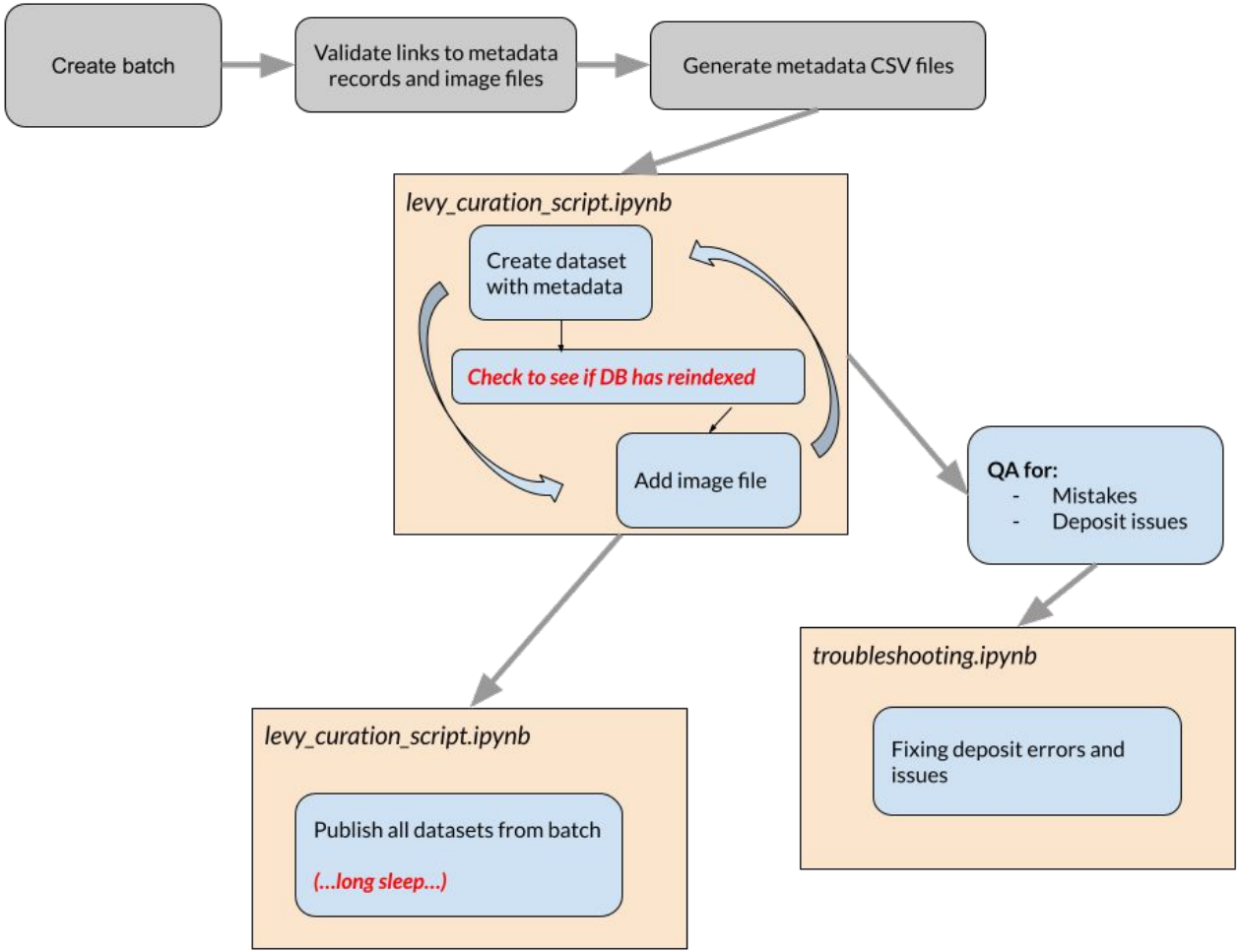
Repositories & Big Data

1. Review several projects from Harvard Dataverse that addressed “big data” in some fashion. What challenges did they present & how did we address them?
2. A few updates from Dataverse on how we’re thinking about addressing several of these
3. Lingering questions for the community about how repositories and technical infrastructure could meet these challenges?

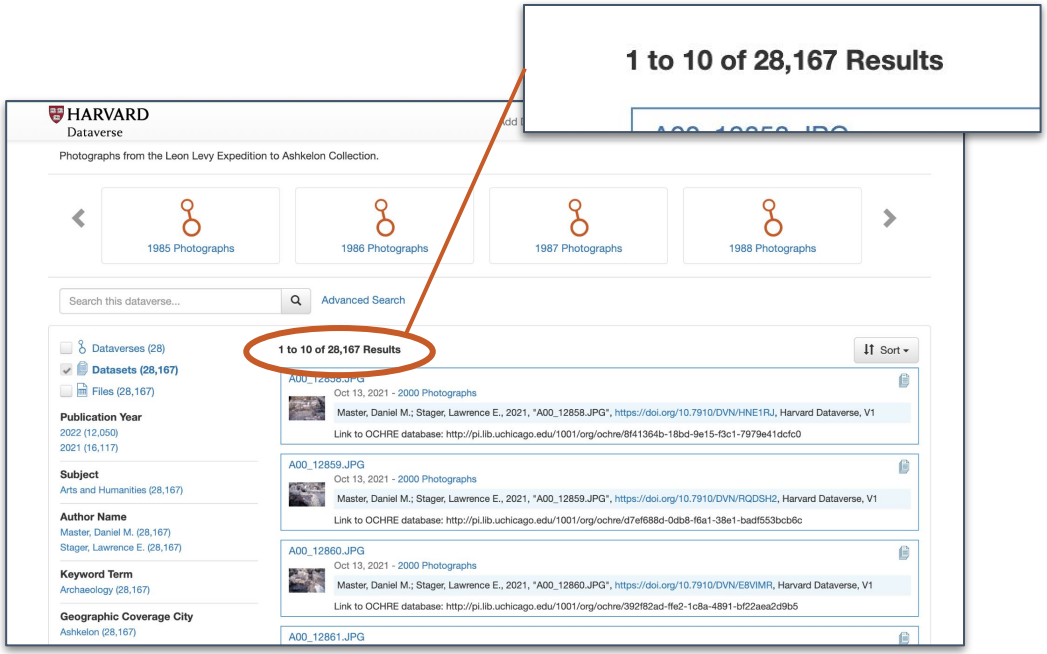
Levy Collection



- 60k files
- Researcher wanted unique DOIs
- Each file would be accessed individually by external image viewer
- Emphasized importance of communicating best practices to users

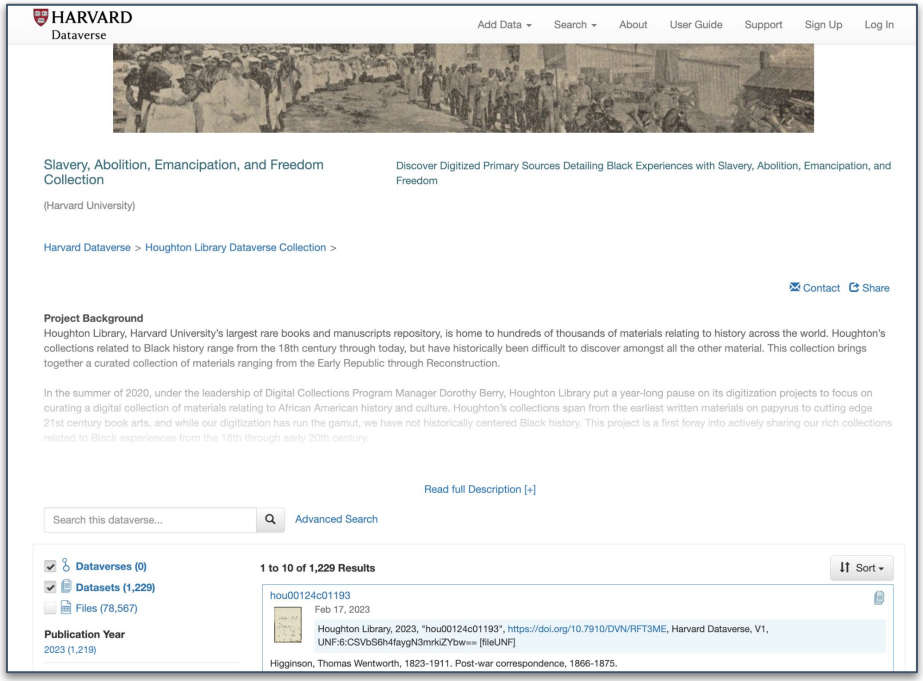

Create batch
Validate links to metadata records and image files
Generate metadata CSV files
levy_curation_script.ipynb
Create dataset with metadata
Check to see if DB has reindexed
Add image file
QA for:
- Mistakes
- Deposit issues
troubleshooting.ipynb
Fixing deposit errors and issues
levy_curation_script.ipynb
Publish all datasets from batch
(...long sleep...)

Levy Collection



- 60k files
- Researcher wanted unique DOIs
- Each file would be accessed individually by external image viewer
- **Emphasized importance of communicating best practices to users**

SAEF & Historic Datasets



- Large number of files
- Diversity of content: images, XML, json, text, csv, documentation, code
- How to curate data and aux files for diverse reuse cases?
- How do you describe and make files findable? Browseable?

Dataset Metadata

Citation Metadata

Dataset Persistent ID	doi:10.70122/FK2/9XXPIX
Title	hou00201c00195
Author	Houghton Library (Harvard University)
Contact	Use email button above to contact. Jacobson, Christine (Houghton Library)
Description	[Unidentified author]. Rev. Mr. Broadwater (says) Every child is allowed to go to the government school, but the colored people can have separate schools [first line] : Ms (in unidentified hand), [no place], 1863?
Subject	Arts and Humanities
Kind of Data	Institutional papers
Origin of Sources	URN-3:FHCL.HOUGH:100528159

Geospatial Metadata

Geographic Coverage	Tuskegee Alabama Canada

SAEF Metadata

MMD ID	990091469160203941
Created	1863
Person/Org Tags	Broadwater, Rev. [?]
Theme	Education; Freedmen
Genre	Government documents

Custom metadata block supports discovery by project metadata

File Metadata & Tags

hou00201c00009_0003.jpg

JPEG Image - 776.7 KB

Deposited Jun 9, 2022

MD5: a40...e00

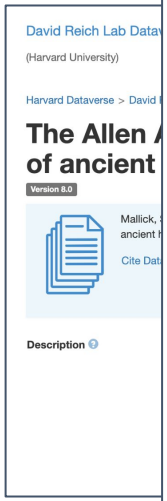
File associated with: Balch, F. V. (Francis Vergnies), 1839-1898. Letters to the Commission, Washington, D.C., 1864 Jan. 11 and 14. Origin of source: https://nrs.harvard.edu/URN-3:FHCL.HOUGH:100526440

City:Washington | Country:United States | Created:1864 | Data | Genre:Correspondence | Person/Org:Balch, F. V. (Francis Vergnies) | Physical Format:Institutional papers | State:District of Columbia | Theme:Freedmen | UID:hou00201c00009

Auto-Generated Documentation: Relationship Files

	filename_source	filename_target	relationshps	source_file_format	target_file_format	relationships
1	hou00201c00009_0001.jpg	hou00201c00009_mets.xml	belongs_to	JPEG 2000 JP2	Extensible Markup Language	
2	hou00201c00009_mets.xml	hou00201c00009_0001.jpg		Extensible Markup Language	JPEG 2000 JP2	contains
3	hou00201c00009_0002.jpg	hou00201c00009_mets.xml	belongs_to	JPEG 2000 JP2	Extensible Markup Language	
4	hou00201c00009_mets.xml	hou00201c00009_0002.jpg		Extensible Markup Language	JPEG 2000 JP2	contains
5	hou00201c00009_0003.jpg	hou00201c00009_mets.xml	belongs_to	JPEG 2000 JP2	Extensible Markup Language	
6	hou00201c00009_mets.xml	hou00201c00009_0003.jpg		Extensible Markup Language	JPEG 2000 JP2	contains
7	hou00201c00009_0004.jpg	hou00201c00009_mets.xml	belongs_to	JPEG 2000 JP2	Extensible Markup Language	
8	hou00201c00009_mets.xml	hou00201c00009_0004.jpg		Extensible Markup Language	JPEG 2000 JP2	contains

HMS Reich Lab



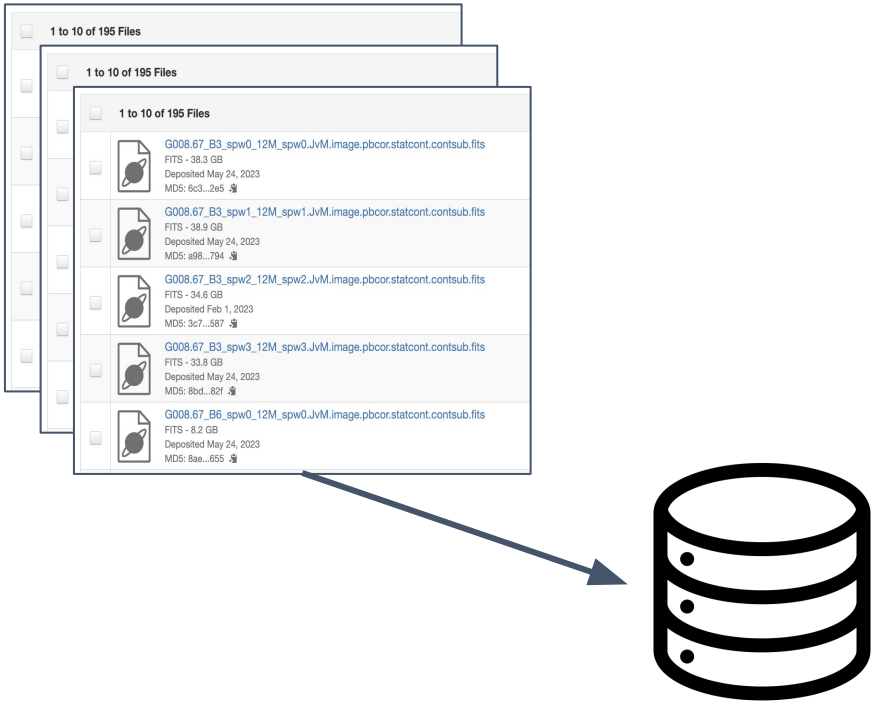
	Dataset Version	Summary	Contributors	Published on
☐	8.0	**Citation Metadata:** Description (1 Changed); **Files (Added: 13; Removed: 12);** View Details	Katherine Mika, Swapan Mallick	2023-04-05
☐	7.0	**Files (Added: 12; Removed: 12);** View Details	Swapan Mallick	2023-04-05
☐	6.0	**Citation Metadata:** Description (1 Changed); **Files (Added: 12; Removed: 12);** View Details	Swapan Mallick	2023-04-05
☐	5.0	**Files (Added: 12; Removed: 12);** View Details	Swapan Mallick	2023-04-05
☐	4.0	**Files (Added: 12; Removed: 13);** View Details	Swapan Mallick	2023-04-05
☐	3.0	**Citation Metadata:** Author (1 Added); Title (Changed); **Files (Added: 13; Removed: 9);** View Details	Katherine Mika, Swapan Mallick	2023-04-05
☐	2.0	**Files (Added: 9; Removed: 9);** View Details	Katherine Mika, Swapan Mallick	2023-04-05
☐	1.0	This is the first published version.	Swapan Mallick, David Reich, Katherine Mika	2023-04-01

g

server

- Hack versioning?

ALMA-IMF Large Program dataset



- Secured approval for very large files, but will require post-publication updates
- Versioning retains old files & explodes storage
- Do we keep files from old versions?

Solutions!

DVUploader – batch file uploads

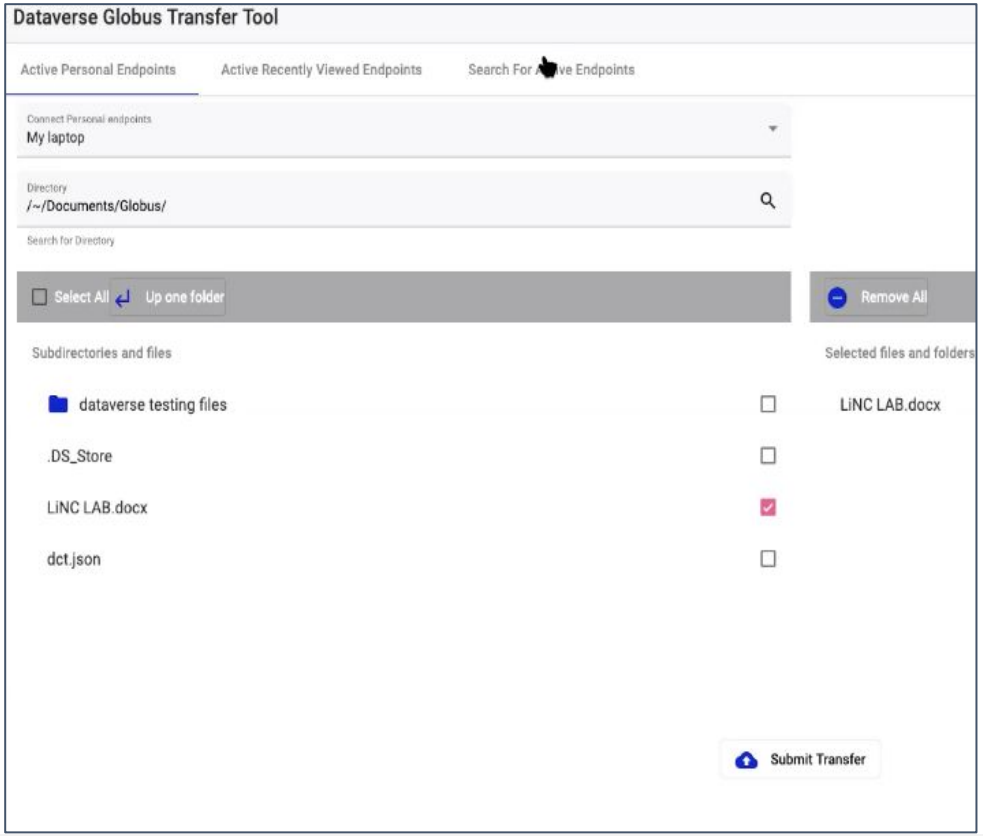
A Java program that uses the Dataverse API to upload a file(s)/a directory tree of files to Dataverse:

java -jar DVUploader-v1.1.0.jar -key=<api key> -did=<dataset doi> -server=<server URL> <dir/file(s)>

- Many options including:
  - -recurse – go into subdirectories
  - -ex=<pat> - ignore files with the specified pattern
  - -verify – check fixity
  - -limit=<x> – max number to process
  - -directupload – use new direct-upload-to-S3-capability
- Useful For:
  - Automating uploads
  - Many files
  - Incremental uploads (will upload new files in a dir)
- Upcoming
  - proof-of-concept code to re-create a dataset from an archival bag

Globus Integration

- Ability to add files (or just file metadata, leaving file at source) to Dataverse
- Developed as an external tool to be integrated into the Dataverse upload workflow



TRSA

- Trusted Remote Storage Agents
  - Agent - Dataverse can communicate with this
  - Storage - especially for sensitive or big data
  - Remote - Dataverse does not control access
  - Trusted - service agreement guarantees

Signed URLs for secure API access from External Tools

The Dataverse External Tool mechanism has been sending the user's API key to the tool

- Allows the tool to read/write on the user's behalf, but
- API key is visible in browser and using tools on draft datasets shouldn't be done on public machines
- Tool is trusted to only do the intended operations and only access the intended dataset/datafile

Signed URLs - a much more secure alternative allowing the tool to make specific API calls for a limited time

- URLs are cryptographically signed and validated when used - will be rejected if modified or used after the configured time period
- Allowed URLs and valid durations (usually seconds to hours) are defined as part of configuring the tool (i.e. admins control what's allowed)
- URLs are also limited by the permissions of the user running the tool
- Tools can be configured to avoid having any URL visible in the browser, or be limited to having one very short-lived URL (e.g. 30 seconds)

SPA Re-architecture Project

- Single Page Application (frontend)
  - **modernize** the application
  - **separate backend and frontend** to increase **interoperability**
  - Dataverse backend becomes an **API-first application**
  - **extend modularization** of backend and frontend
  - **speed up development** and implementation of new UI/UX ideas
  - Native **accessibility** (A11y) and **internationalization** (i18n) support
- **empower the community**

Persistent problems require community solutions & reimagining curation best practices

Community frameworks

1. What is a realistic, sustainable storage limit for institutional repositories? What level/tier do we set limits at?
2. How long should we preserve Big Data?
3. What does stricter appraisal look like?
4. What are the ongoing stewardship requirements?
5. How does data management planning differ?
6. How to communicate better with researchers about the realities of sharing and preserving/archiving Big Data

Thanks!

katherine_mika@harvard.edu

gdurand@iq.harvard.edu