Transforming a Digital Collection into a Data Collection

Houghton Library's Slavery, Abolition, Emancipation, and Freedom Collection

HKS Data+Donuts Talk

2024-05-03

Description

At this session of Data + Donuts, Boyd will discuss the Slavery, Abolition, Emancipation, and Freedom (SAEF) Data Collection, an open access research dataset representing over 1200 items in Houghton Library's SAEF collection. The SAEF Data Collection is Harvard Library's first major endeavor to represent a significant print collection as data, supporting research methodologies involving computational analysis of images, Optical Character Recognition (OCR) text, and machine-generated transcriptions of manuscript materials. To create the collection, Boyd helped lead a joint team from Houghton Library, Harvard Library's Open Scholarship and Research Data Services, Harvard Library's Imaging Services, and the Harvard Dataverse data curation and repository development teams.

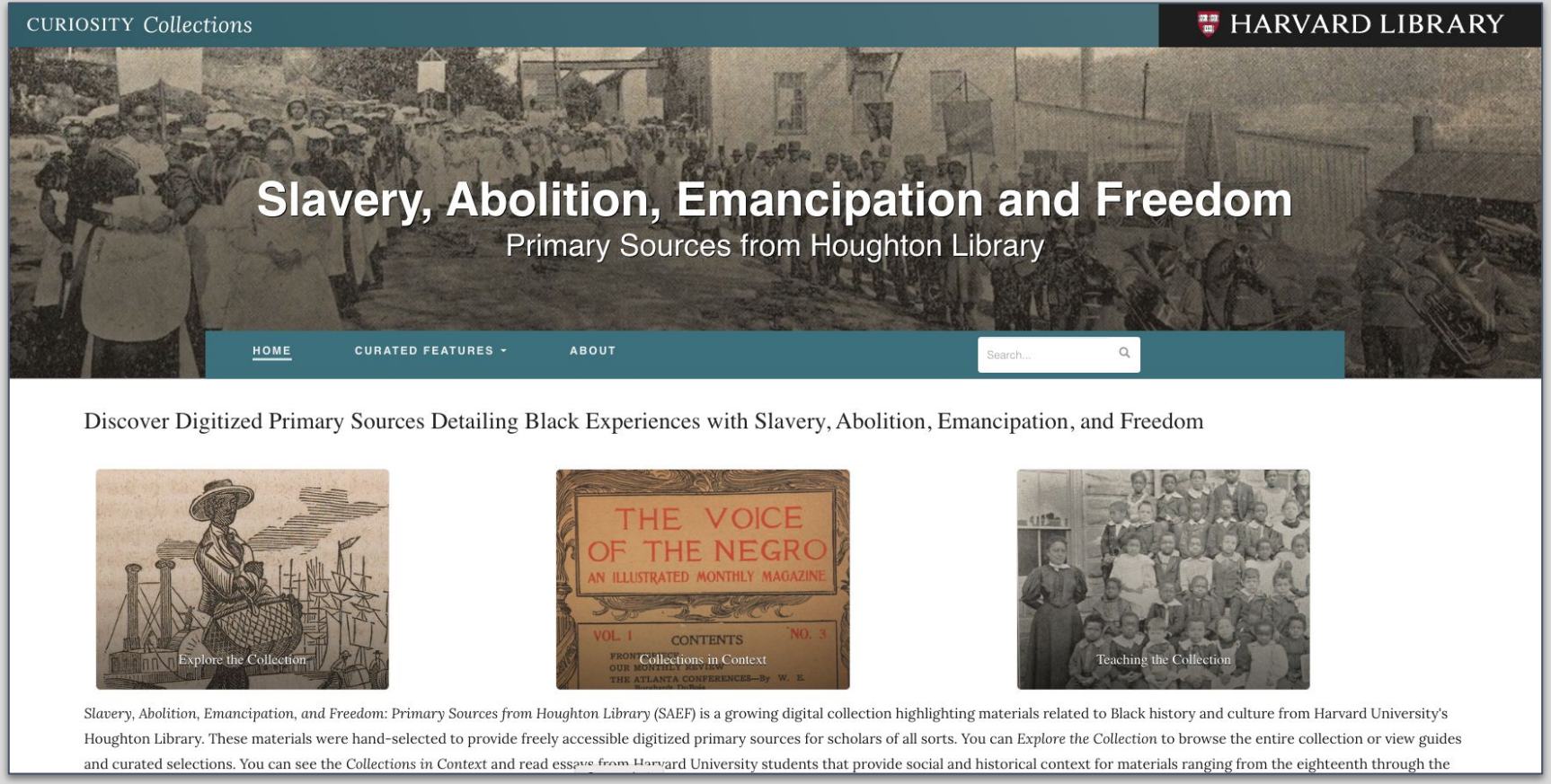
Objective

Describe a project to transform a library digital collection into a research data collection

Agenda

1. Houghton Library's Slavery, Abolition, Emancipation, and Freedom (SAEF) Collection
2. Advancing Open Knowledge Project
3. Research Data Sharing
4. Transforming the SAEF Digital Collection
5. Lessons Learned
6. Collection Tour

Photo by Andrew Neel on Unsplash



https://curiosity.lib.harvard.edu/slavery-abolition-emancipation-and-freedom

Slavery, Abolition, Emancipation & Freedom project and collection

- **Project:** Digitize existing Houghton Library materials relating to African American history and culture
- **Curator**: Christine Jacobson (formerly, Dorothy Berry)
- **Collection**: 1,200+ curated items w/more to come (incl., letters, drawings, novels, poetry, broadsides)
- **Funding**: Support by Harvard Library Advancing Open Knowledge Grant

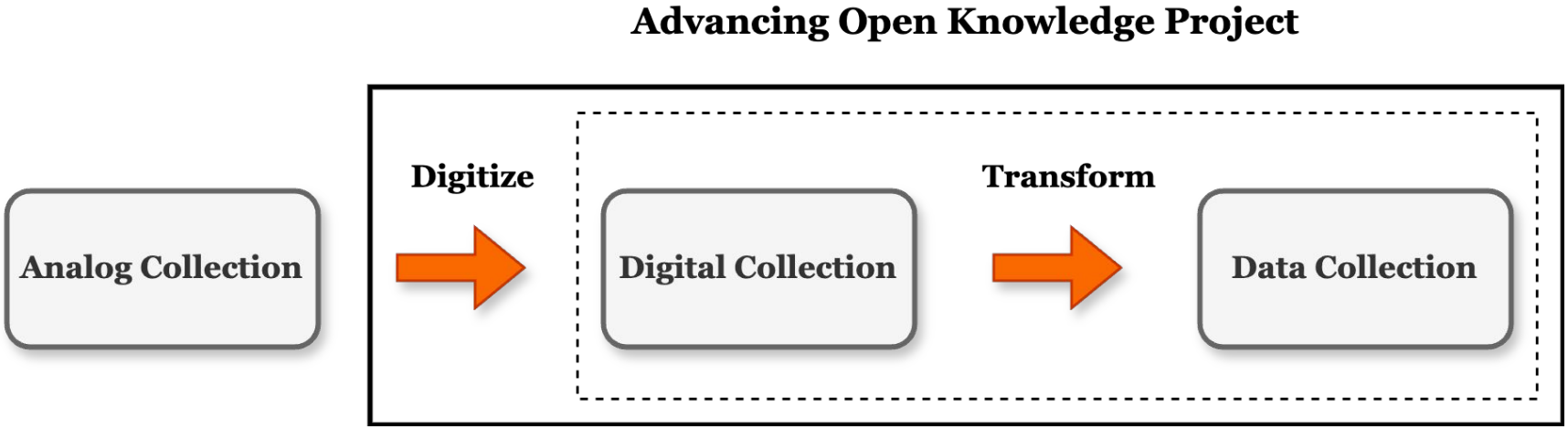
Advancing Open Knowledge Project

"Harvard Library's Advancing Open Knowledge Grants Program seeks to advance open knowledge and foster innovation to further equity, diversity, inclusion, belonging and antiracism."

Multi-org collaboration to:

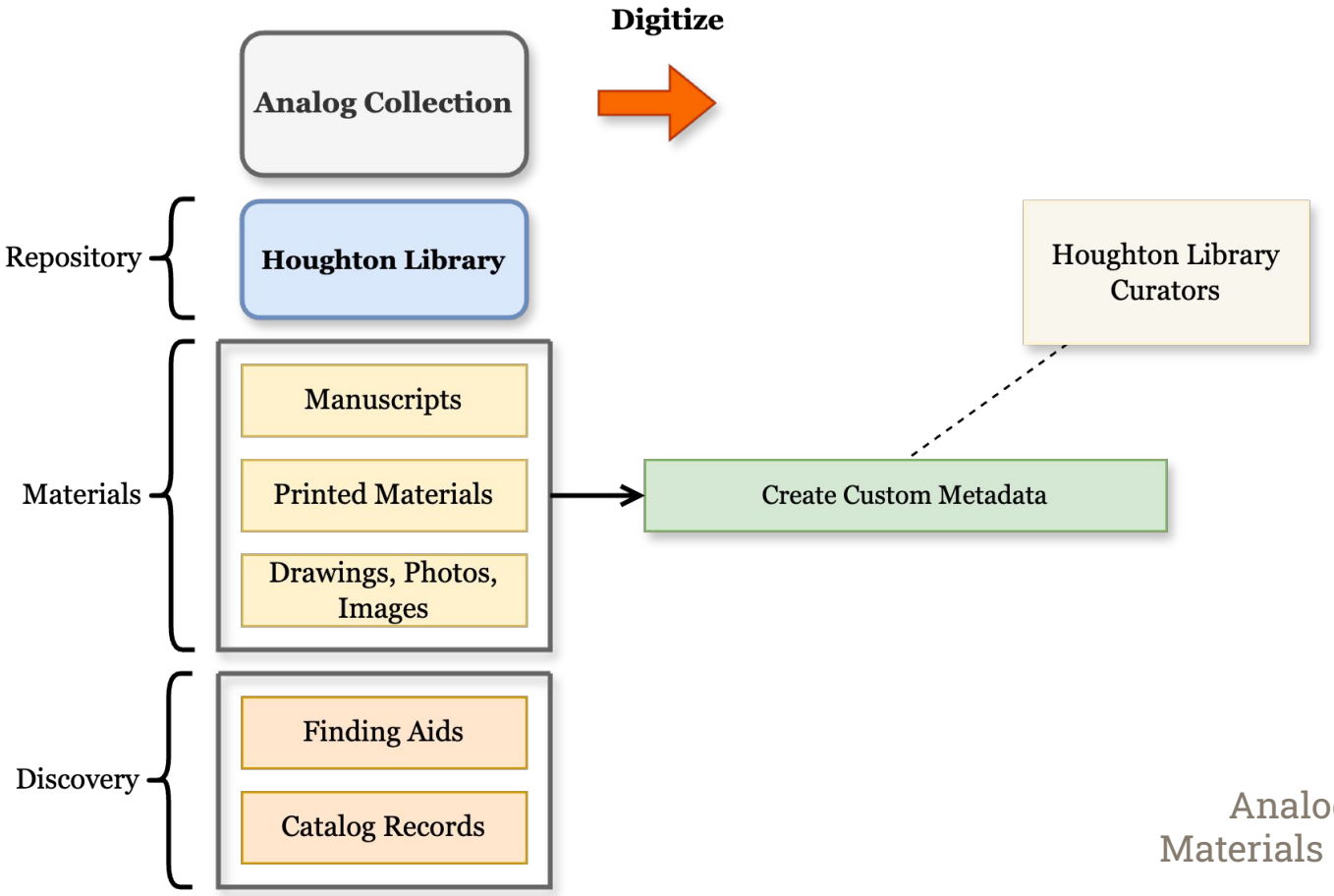
1. Digitize Houghton SAEF materials, create custom metadata, and create online collection (2020)
2. Generate transcriptions of handwritten manuscripts using ML (2020-2021)
3. Transform digitized materials into publicly available datasets (2021-2023)

HL Advancing Open Knowledge Grants

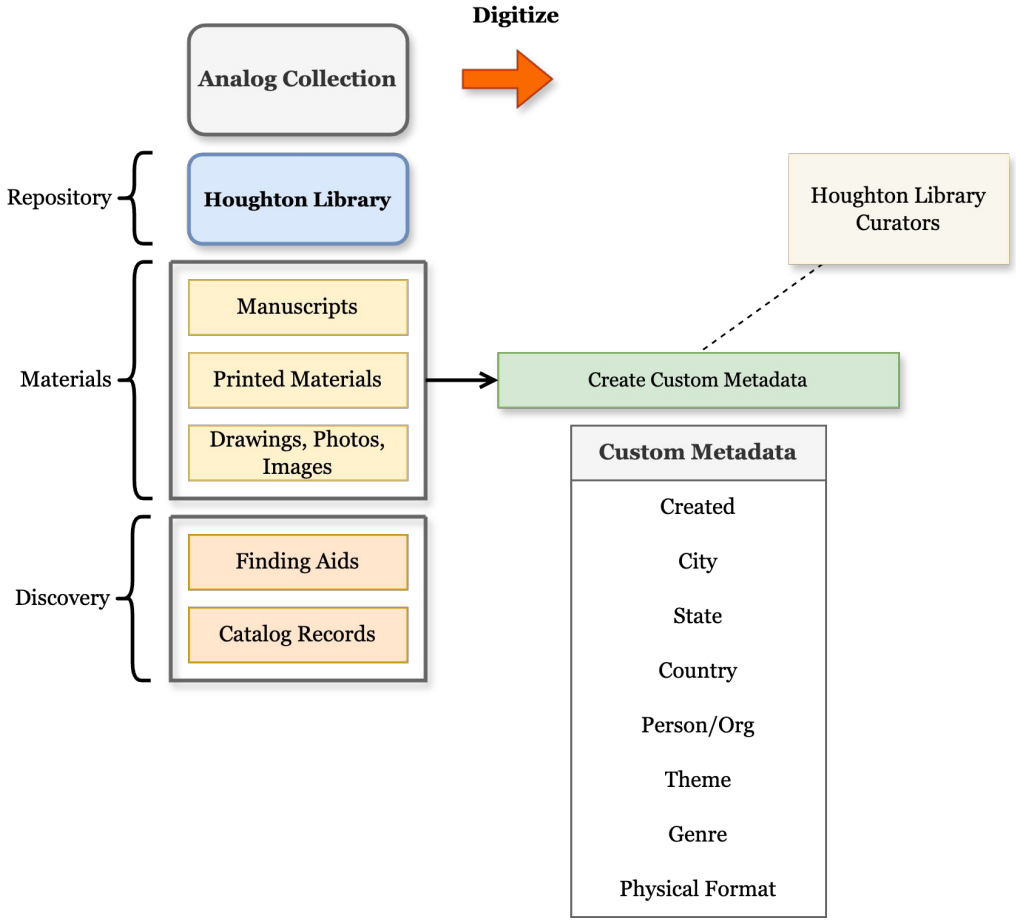


Project Overview

Analog Collection Materials

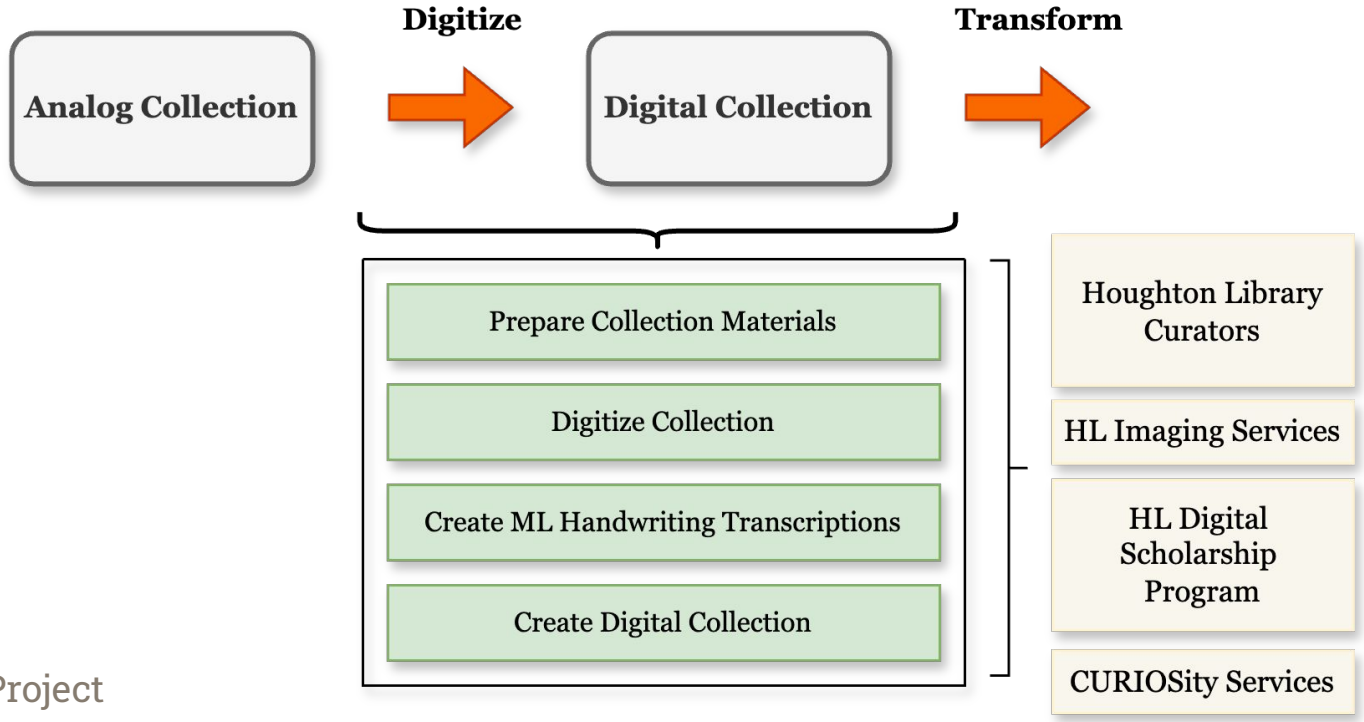


Analog Collection
Materials & Discovery

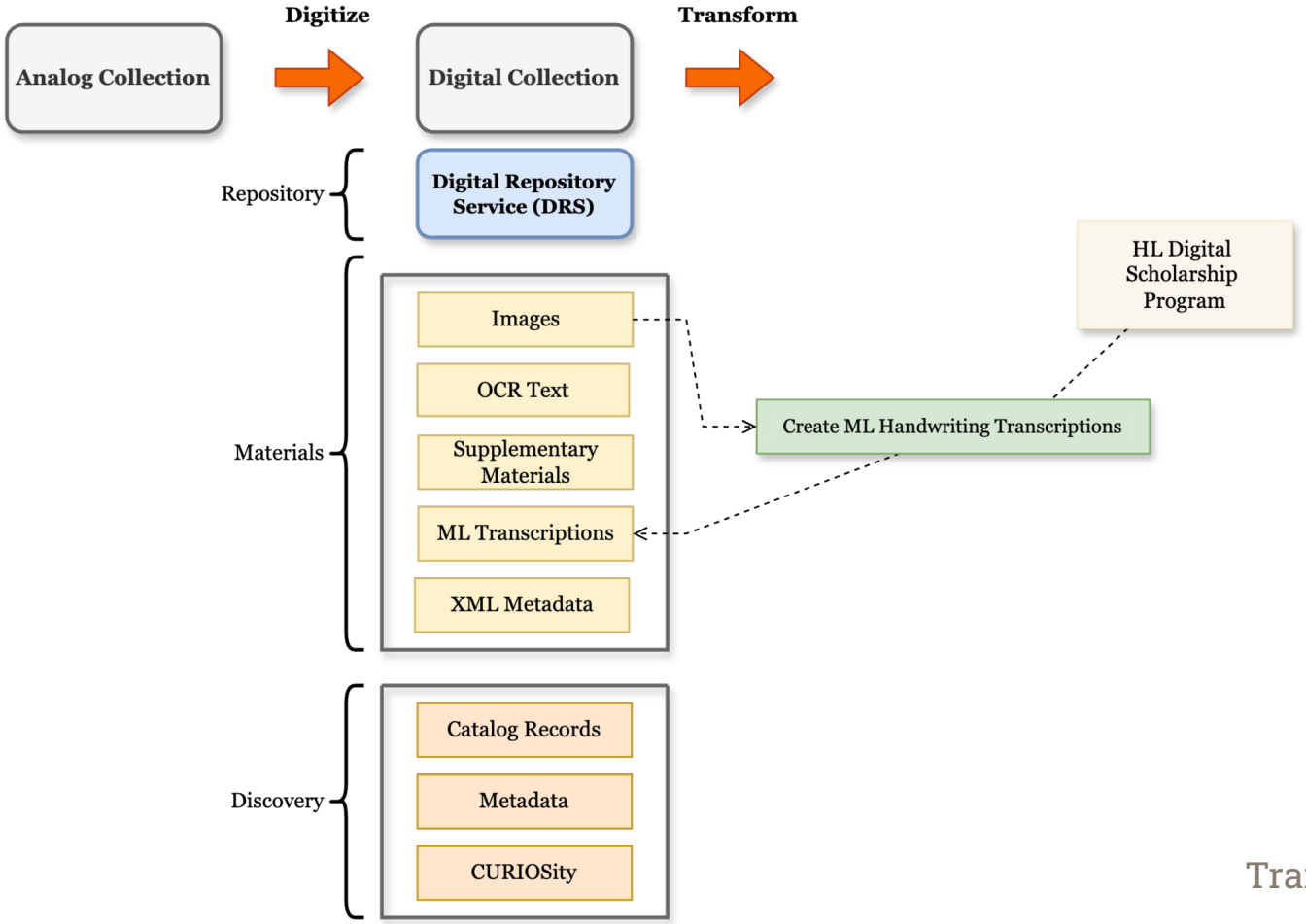


Custom Metadata Details

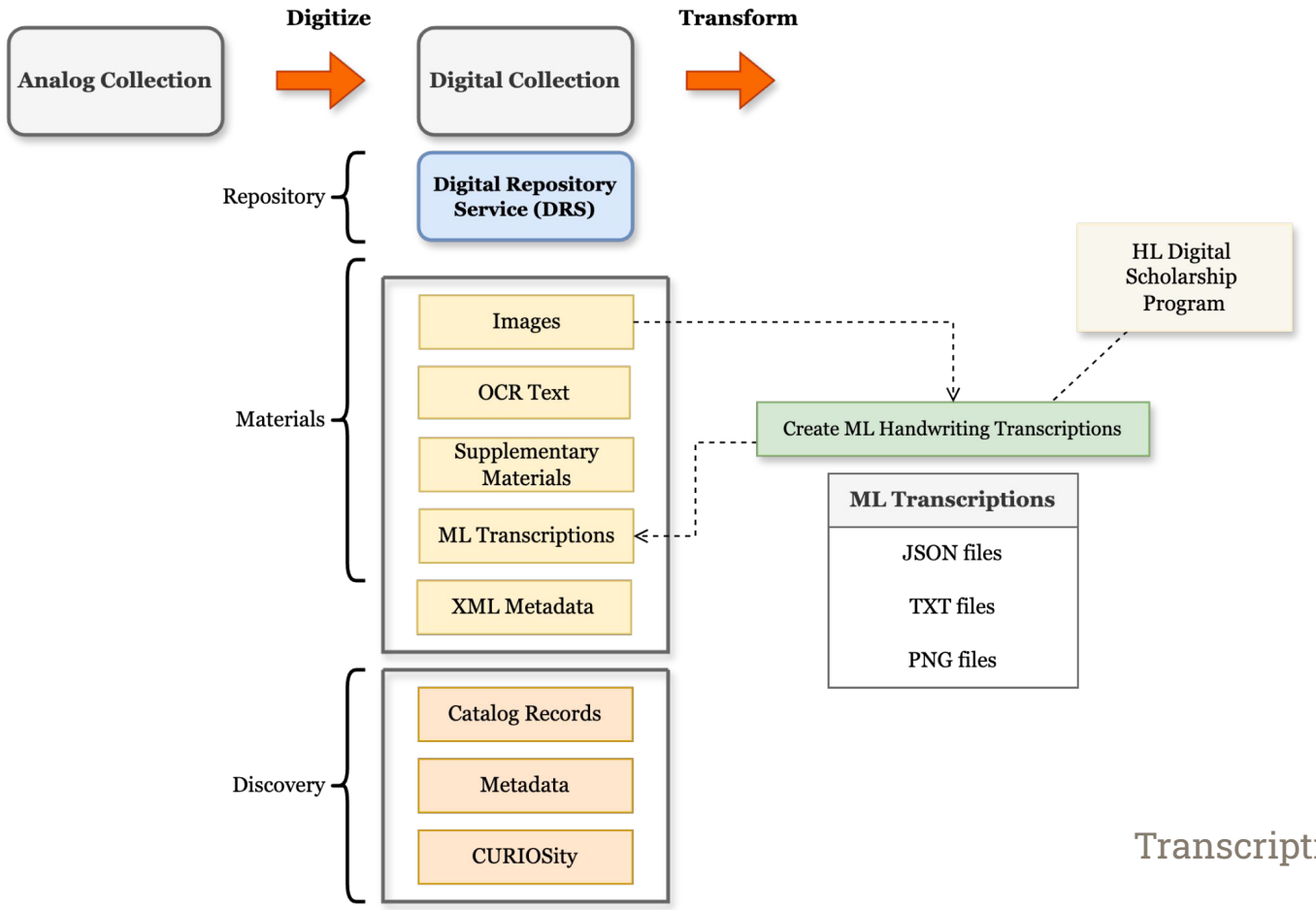
Digital Collection Materials



First Stage Project Components

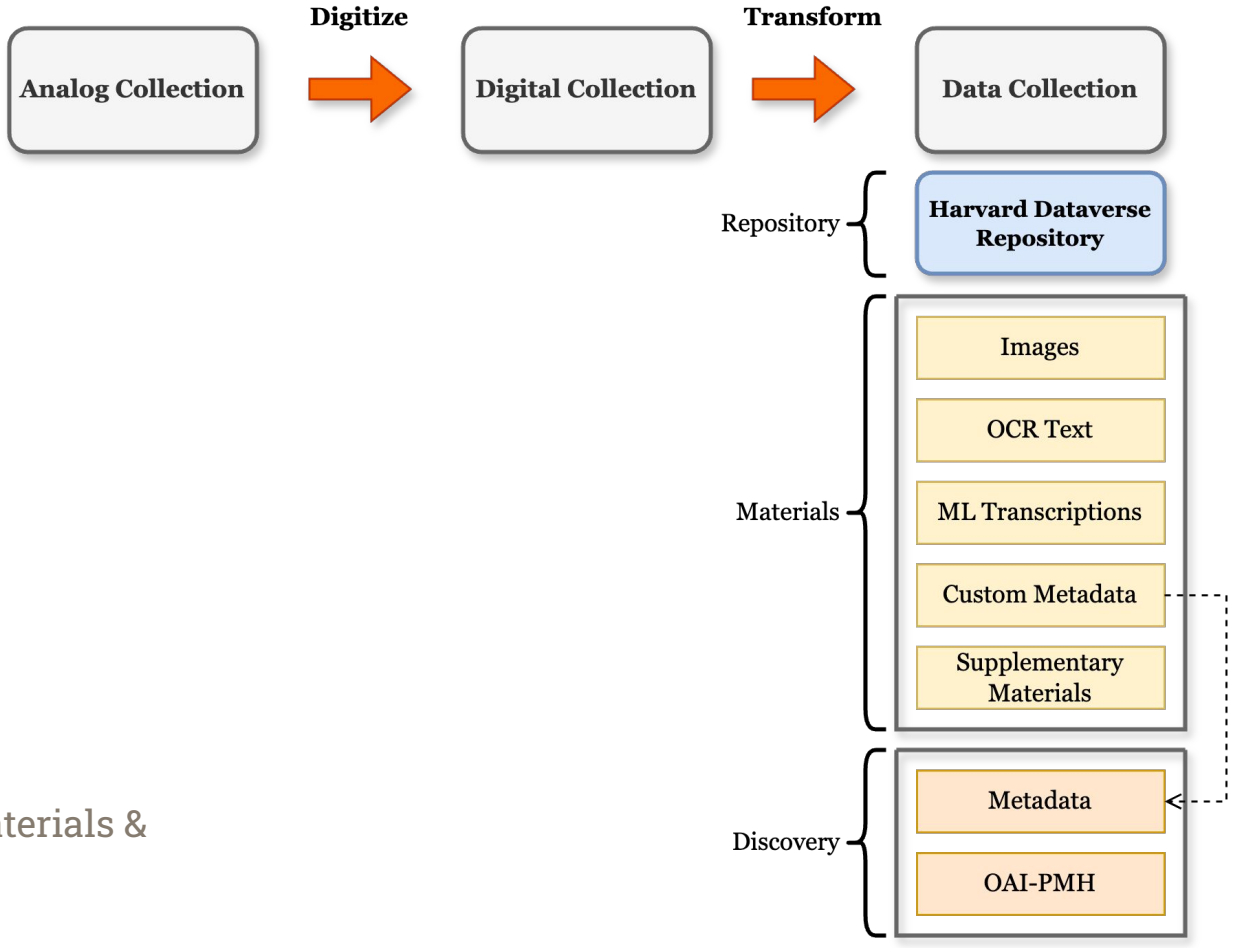


Create ML Transcriptions

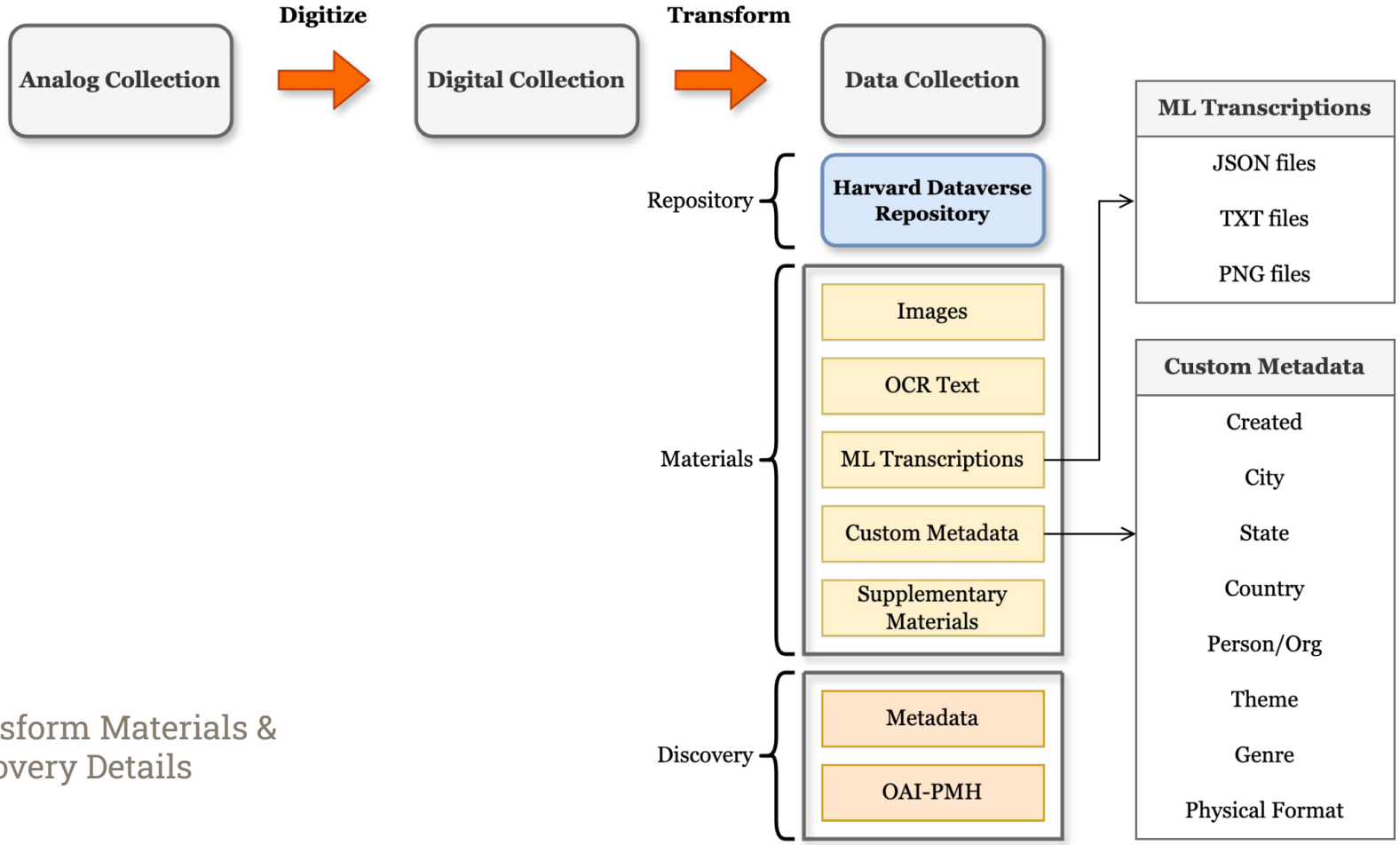


Create ML Transcriptions Details

Data Collection Materials



Transform Materials & Discovery

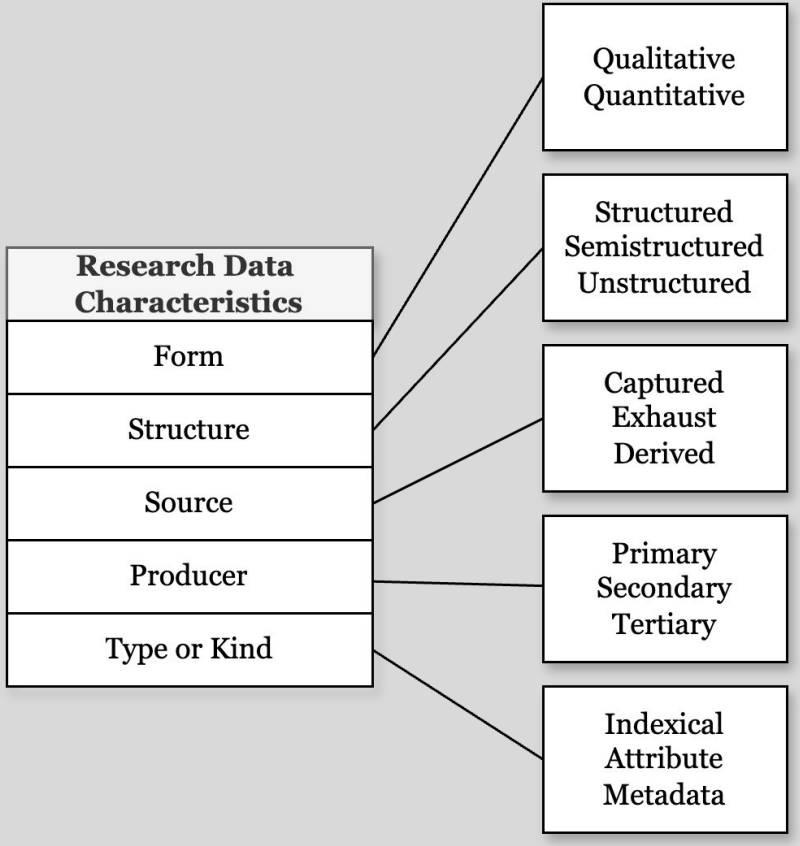


Transform Materials & Discovery Details

Research Data Sharing

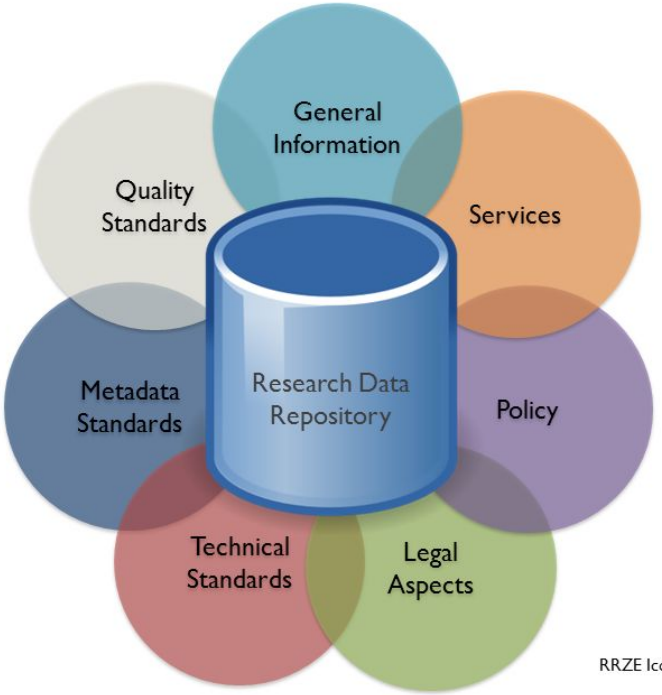
Research data

- Information and samples observed, collected, or created during a research project
- Analyzed to produce results or validate scientific claims
- Serve as evidence in a scholarly argument
- Relational, multiple components needed to make sense of whole
- Portable
- Takes many forms

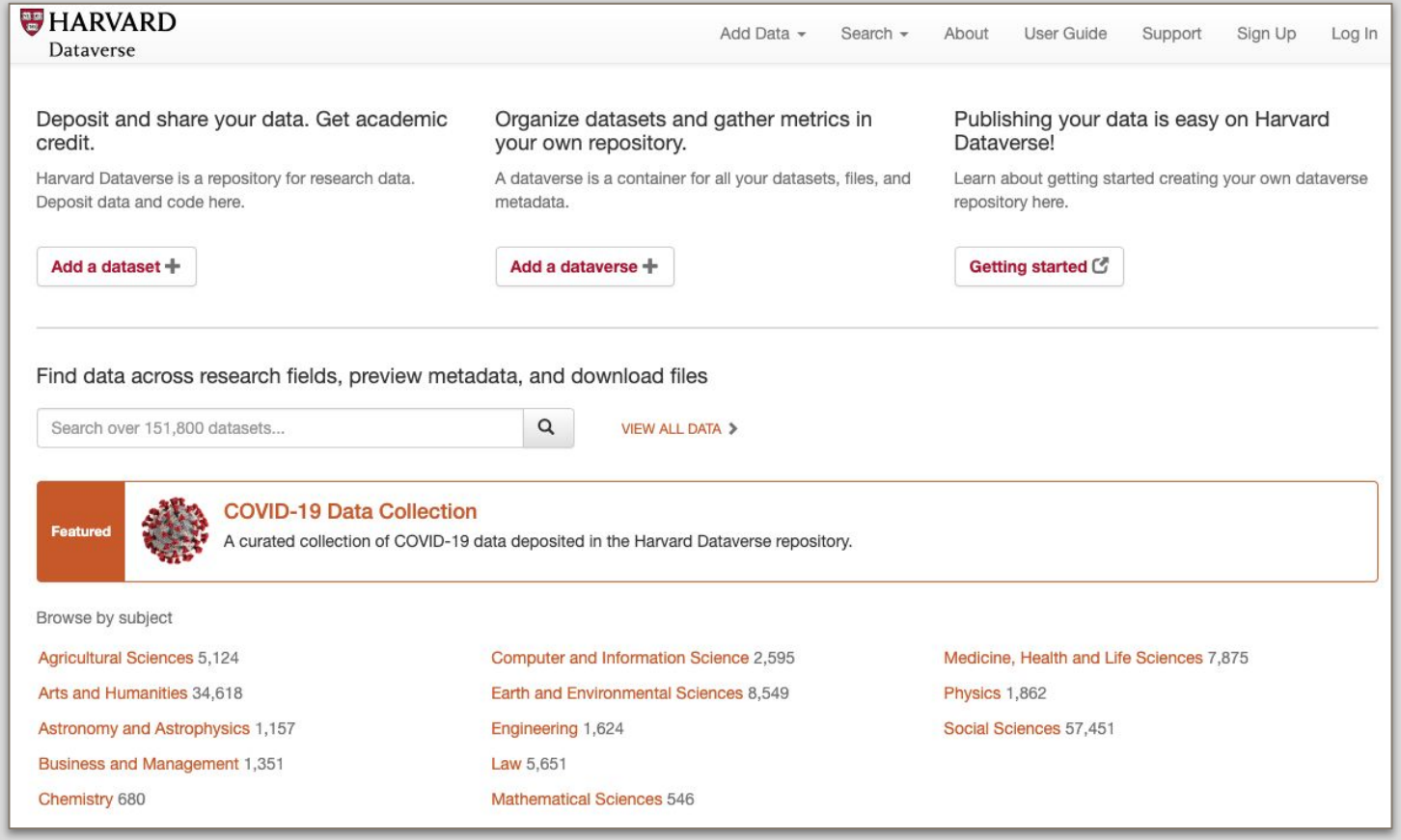


Research data repository

- Database of well-described. Well-documented, and stewarded research data datasets.
- Research data repositories focus on providing the sustainable infrastructure for the long term storage and access to, specifically, research data.



RRZE Icon Set (CC: BY-SA)



https://dataverse.harvard.edu

Research data sharing

- Deposit data in a repository
- Choose data licenses and terms of use
- Apply metadata to make published data more findable
- Write good documentation so shared data is reusable
- Steward shared data over its useful term

Photo by Kelly Sikkema on Unsplash

Credit: K. Mika

FAIR Guiding Principles

"All research objects should be Findable, Accessible, Interoperable and Reusable (FAIR) both for machines and for people."

Wilkinson, et al. (2016). The FAIR guiding principles for scientific data management and stewardship.

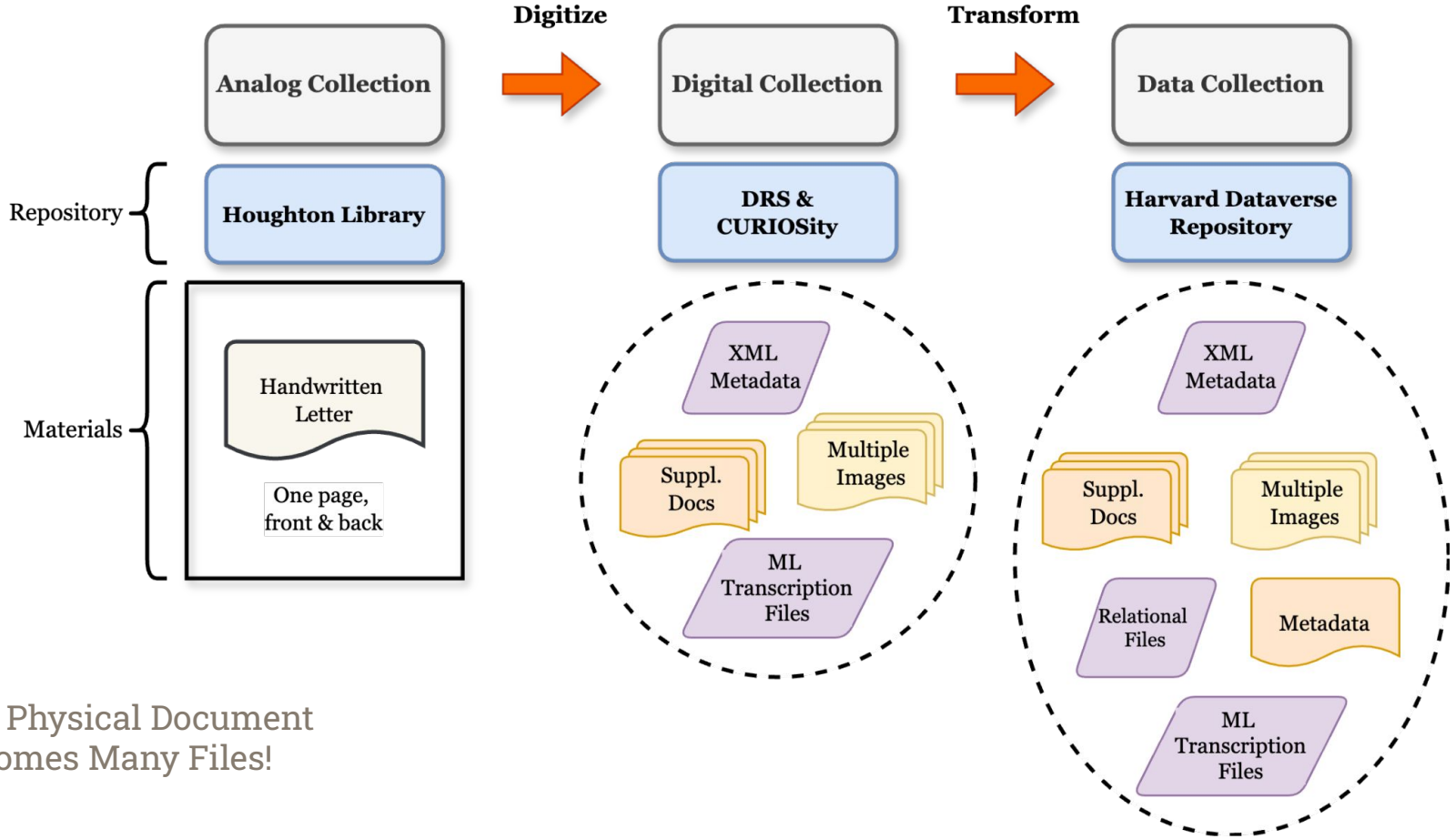
Image credit: BioSistemika

FAIR data, research data sharing & data repositories

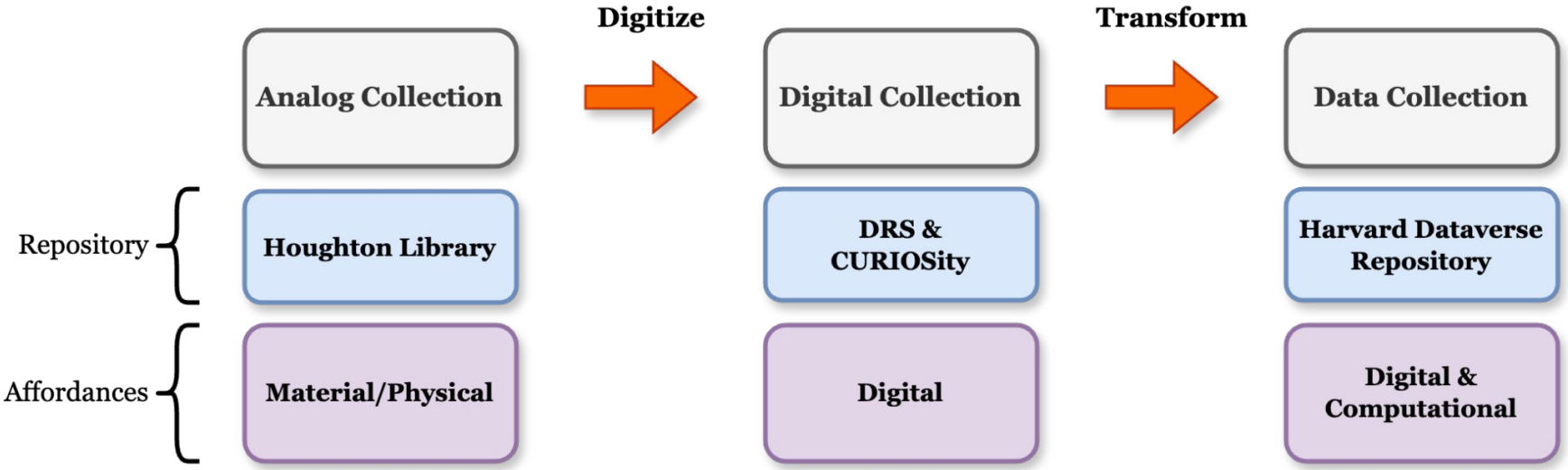
Findable	Accessible	Interoperable	Reusable
• Well-described using standards • Discoverable via aggregators & catalogs	• Persistent identifier • Citable • Reliable storage	• Standard & open formats • Observe good preservation practices	• Clear terms of use • Clear documentation • Versioning

Metadata	Technical Infrastructure	Context, Policy & Rights

Analog v. digital v. data collections



One Physical Document
Becomes Many Files!



Different Affordances

Affordances

"the range of functions and constraints that an object provides for, and places upon, structurally situated subjects" (Davis & Chouinard, 2017, p. 241)

Photo by Leonel Fernandez on Unsplash

Affordances

Analog Collection	Digital Collection	Data Collection
• Engagement with physical materials • Supports analysis of physical properties • Limited access	• Engagement with digital surrogates • Supports analysis of digital properties • (Potentially) Wide access	• Digital affordances • Computational engagement (incl. combining, computing on contents & reuse)

Affordances

Analog Collection	Digital Collection	Data Collection
• Engagement with physical materials • Supports analysis of physical properties • Limited access	• Engagement with digital surrogates • Supports analysis of digital properties • (Potentially) Wide access	• Digital affordances • Computational engagement (incl. combining, computing on contents & reuse)

Using affordances, examples

Analog Collection	Digital Collection	Data Collection
• Spectral age analysis • Chemical ink dating • Pollen analysis	• Image manipulation (resize, copy, enlarge, resample, remix) • Automatic OCR & transcription • Share widely	• Use as secondary data • Combine with other datasets • Compute on digital content (ex. ML, text & data mining, variable extraction)

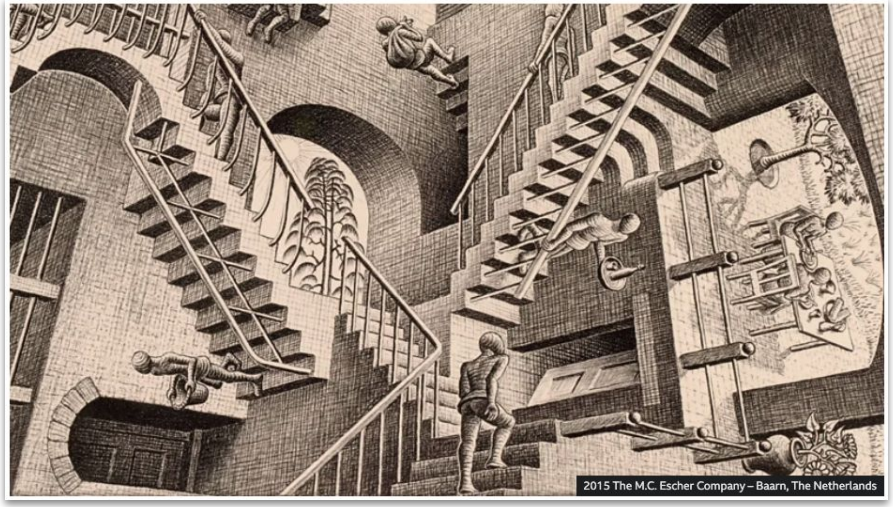
FAIR data collection user expectations, examples

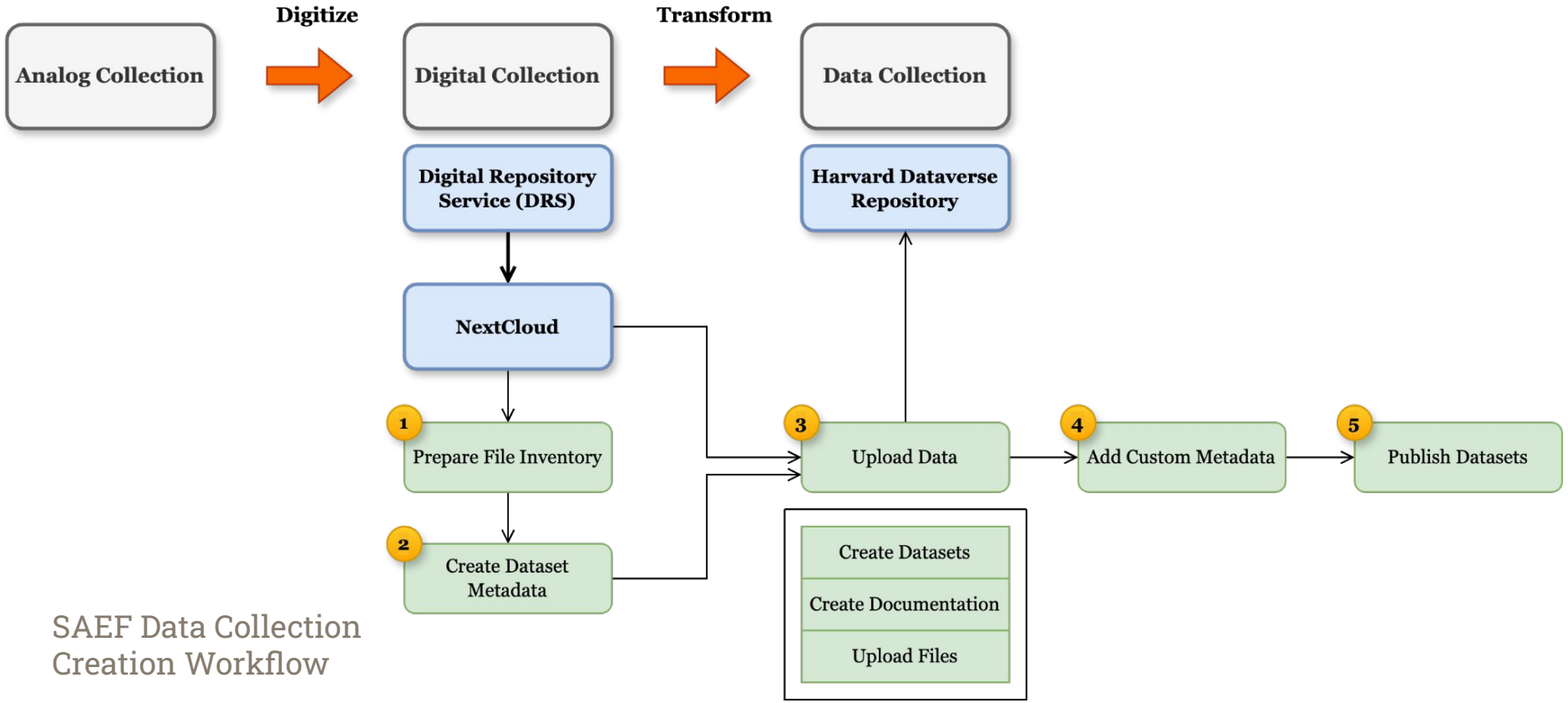
(Datasets are) Easy to find via search engines	**Findable**
Granular metadata to support search, filtering	
Access using APIs, not GUIs	**Accessible**
Citable, persistent unique identifiers (DOIs)	
Reliable access	
Standard file formats	**Interoperable**
Coherent file organization	**Reusable**
Clear documentation & terms of use	

Transforming the Digital Collection

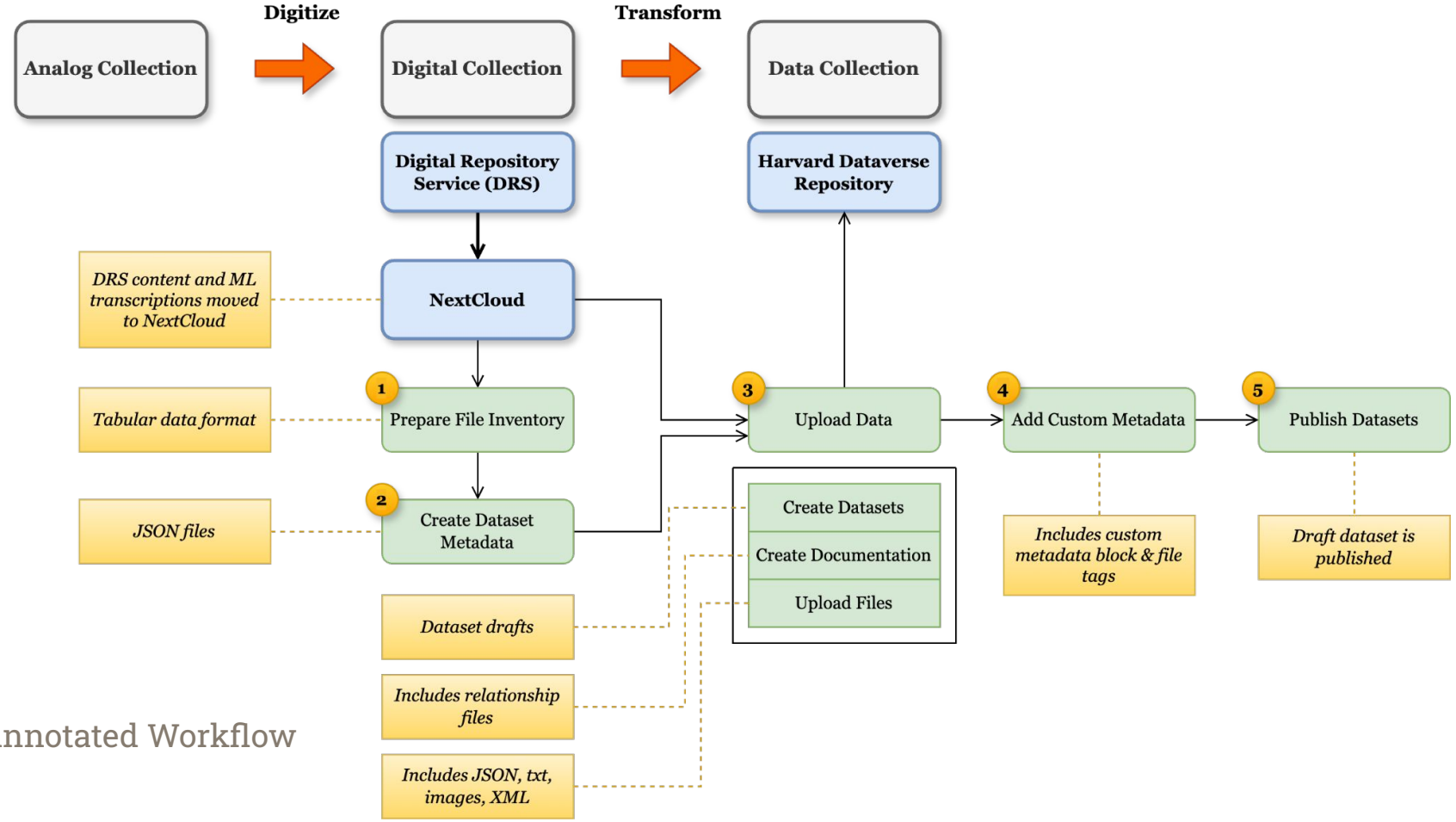
Preview: challenges

- File management (70,000+)
- Maintaining & representing file relations
- Transfering the files (51GB)
- Making curatorial choices
- Creating & populating the datasets
- Assigning the metadata (incl. custom metadata)

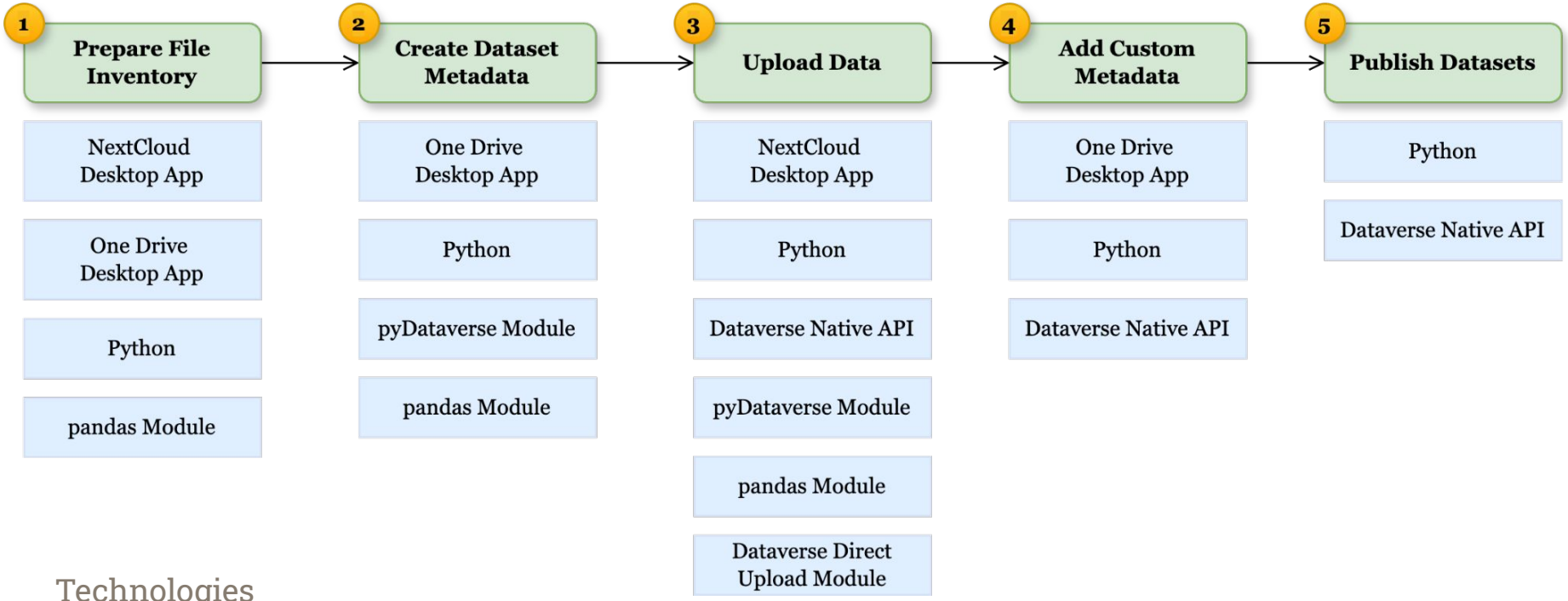




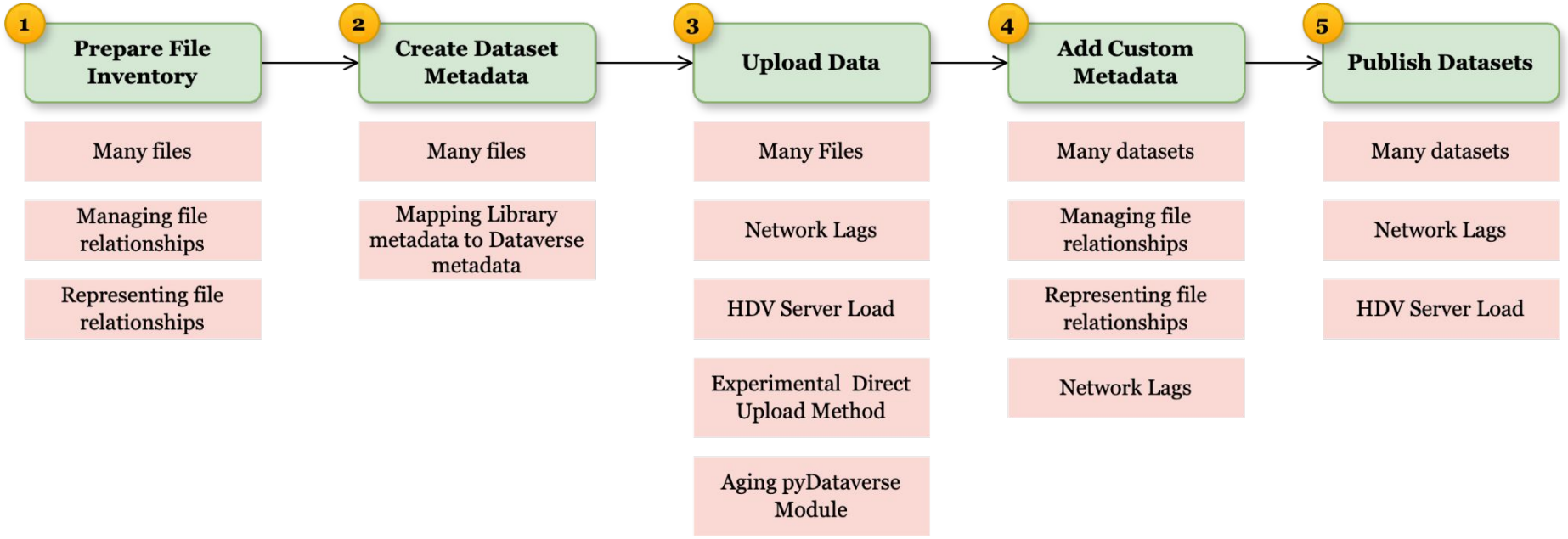
SAEF Data Collection Creation Workflow



Annotated Workflow



Technologies


1
Prepare File Inventory
Many files
Managing file relationships
Representing file relationships
2
Create Dataset Metadata
Many files
Mapping Library metadata to Dataverse metadata
3
Upload Data
Many Files
Network Lags
HDV Server Load
Experimental Direct Upload Method
Aging pyDataverse Module
4
Add Custom Metadata
Many datasets
Managing file relationships
Representing file relationships
Network Lags
5
Publish Datasets
Many datasets
Network Lags
HDV Server Load

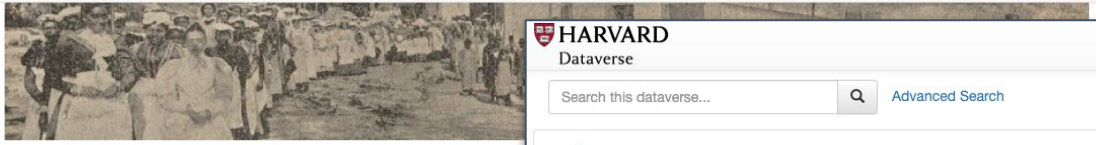
Challenges

The SAEF Data Collection

https://dataverse.harvard.edu/dataverse/SAEF

HARVARD Dataverse

Add Data | Search | About | User Guide | Support | Ceilyn Boyd 25



Slavery, Abolition, Emancipation, and Freedom Collection

Discover D… Freedom

(Harvard University)

Harvard Dataverse > Houghton Library Dataverse Collection >

Project Background

Houghton Library, Harvard University's largest rare books and manuscripts repository, i… collections related to Black history range from the 18th century through today, but have… together a curated collection of materials ranging from the Early Republic through Reco…

In the summer of 2020, under the leadership of Digital Collections Program Manager D… curating a digital collection of materials relating to African American history and culture… 21st century book arts, and while our digitization has run the gamut, we have not histor… related to Black experiences from the 18th through early 20th century.

License

These datasets are shared under the: Open Data Commons – Public Domain Dedicatio…

Data Access

We recommend using the Dataverse API to access the datasets and files in this collecti… detailed inventories and persistent identifiers for the datasets and files in the collection.

Data Reuse

We recommend you review the dataset documentation and best practices for using this…

Researchers who reuse and create new data based on this collection are encouraged to deposit their content in the Harvard Dataverse.

HARVARD Dataverse

Add Data | Search | About | User Guide | Support | Ceilyn Boyd 25

Search this dataverse... | Advanced Search | + Add Data

- Dataverses (0)
- Datasets (1,229)
- Files (78,567)

Publication Year
2023 (1,229)

Subject
Arts and Humanities (1,229)
Social Sciences (1)

Author Name
Houghton Library (1,229)

Author Affiliation
Harvard University (1,229)

Keyword Term
Abolitionist Apologetics (1)
Abolitionist politics (1)
Abolitionists (1)
African American troops (1)
African Methodist Episcopal Church (1)
More...

Data Type
Institutional papers (429)
Pamphlets (325)
Personal papers (157)
Broadsides (60)
Publications (40)
More...

Geographic Coverage Country / Nation
United States (764)

1 to 10 of 1,229 Results | Sort

hou00124c01193
Feb 17, 2023
Houghton Library, 2023, "hou00124c01193", https://doi.org/10.7910/DVN/RFT3ME, Harvard Dataverse, V1, UNF:6:CSVbS6h4faygN3mrkiZYbw== [fileUNF]
Higginson, Thomas Wentworth, 1823-1911. Post-war correspondence, 1866-1875.

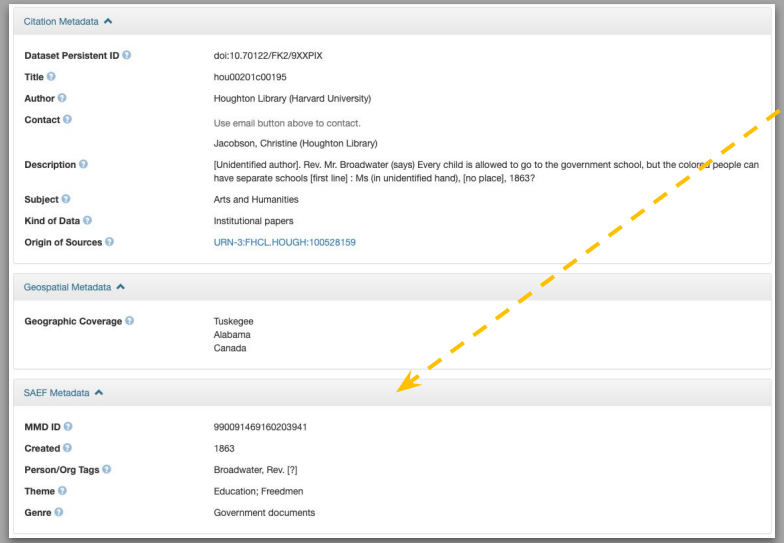
hou00124c01199
Feb 17, 2023
Houghton Library, 2023, "hou00124c01199", https://doi.org/10.7910/DVN/PL7O3V, Harvard Dataverse, V1, UNF:6:zLhq9pStlpW10Thif6lv3g== [fileUNF]
United States. Army. South Carolina Volunteers, 1st. Correspondence, miscellaneous, 1863-1866.

hou00201c00078
Feb 17, 2023
Houghton Library, 2023, "hou00201c00078", https://doi.org/10.7910/DVN/I0AV04, Harvard Dataverse, V1, UNF:6:G2qOsbYon2xwHfkx59MOqw== [fileUNF]
Owen, Robert Dale, 1801-1877. Memo to S.G. (Samuel Gridley) Howe, New York, New York, 1863 July 17.

modbm_ac7_j6501_797a
Feb 17, 2023
Houghton Library, 2023, "modbm_ac7_j6501_797a", https://doi.org/10.7910/DVN/SAUQLI, Harvard Dataverse, V1, UNF:6:xieDvnBV6YODOi//Cg6NEQ== [fileUNF]
The address of Abraham Johnstone, a black man : who was hanged at Woodbury, in the county of Glocester, and state of New Jersey, on Saturday the the [sic] 8th day of July last; to the people of colour. To which is added his dying confession or declaration also, a copy of a letter...

modbm_us_5261_197_024
Feb 17, 2023
Houghton Library, 2023, "modbm_us_5261_197_024", https://doi.org/10.7910/DVN/JJ4RJ8, Harvard Dataverse, V1, UNF:6:1SVsgdFiUDGJzgUW3UvJxQ== [fileUNF]
The humanitarian side of religion

Dataset Metadata



Custom metadata block supports discovery by project metadata

Custom metadata applied at file level

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

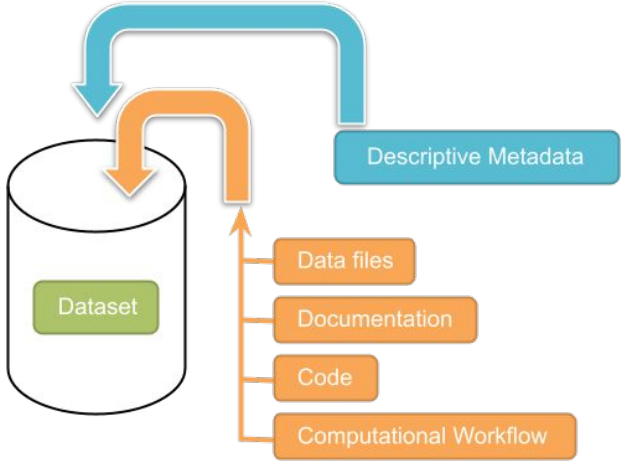
City:Washington | Country:United States | Created:1864 | Data | Genre:Correspondence | Person/Org:Balch, F. V. (Francis Vergnies) | Physical Format:Institutional papers | State:District of Columbia | Theme:Freedmen | UID:hou00201c00009

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

Data collection overview

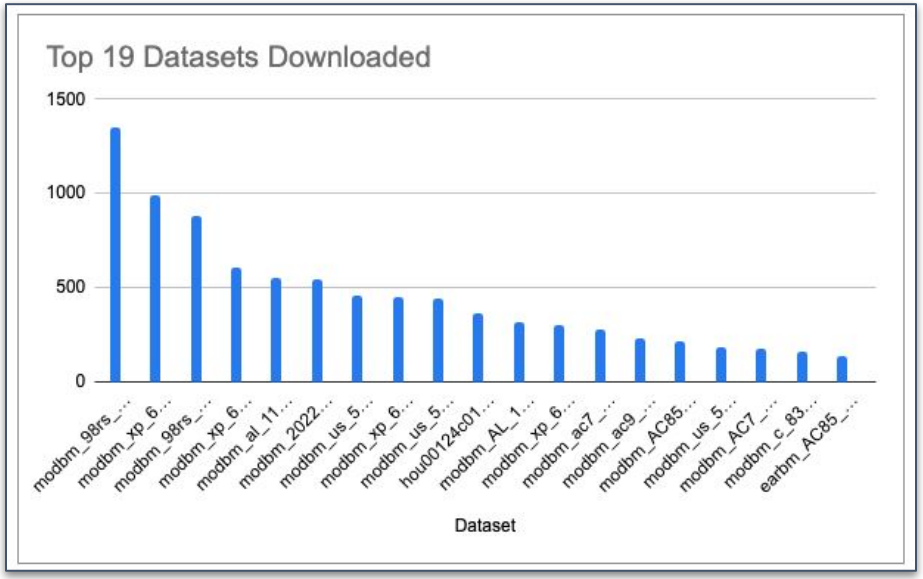
- 1,229 datasets
- 78,567 files
- File types
  - Image (40,626)
  - Text (35,274) (incl. json, xml)
  - Tabular Data (1,982)
  - Data (684)
  - Document (1)
- 51GB of content



Source: Dataset + File management

Data collection usage

- 19319 total downloads
- 803 (65%) datasets downloaded, 1354 times, 7% of all downloads
- Most frequent download: modbm_98rs_99_pds
- "Twentieth century Negro literature, or, A cyclopedia of thought on the vital topics relating to the American Negro by one hundred of America's greatest Negroes" ; edited and arranged by D.W. Culp. Toronto, Canada ; Naperville, Ill. : J.L. Nichols & Co., 1902. 98RS-99. Houghton Library, Harvard University, Cambridge, Mass.



Source: saef_downloads.csv

Data curation

- One dataset per digital object
- Created custom SAEF metadata block
  - Original source: Custom SAEF Project Metadata
- File tags corresponding to custom metadata
- Collection README
- Collection inventory
- Collection relationship files

Lessons Learned

Takeaways

- Good file management is key
- Create and work with tabular inventories of data and metadata to save time
- Use pandas to wrangle tabular files
- Use direct upload method to avoid server poor performance via the Dataverse API
- Upload in small batches, publish separately
- Create and process log files to aid in failure recovery
- Use scripts vs. notebooks where possible
- Schedule large uploads for late-night/weekends to minimize server impact

Collection Tour

https://dataverse.harvard.edu/dataverse/SAEF

🎉 Acknowledgements

- Dorothy Berry, *National Museum of African American History and Culture*
- Christine Jacobson, *Houghton Library*
- Bill Comstock, *Imaging Services*
- Matt Cook, *Harvard Library Digital Scholarship Program*
- Sonia Barbosa, Leonid Andreev, Phil Durbin, Julian Gautier, *IQSS/Dataverse*
- Jim Myers, *Global Dataverse Community Consortium (GDCC)*

Q/A & Discussion

Thank you!

Contact me:

- Ceilyn, cboyd@g.harvard.edu

Resources

SAEF Custom Metadata

- https://public.flourish.studio/visualisation/8301893/

NextCloud

- https://nextcloud.com/
- https://wiki.harvard.edu/confluence/display/LibraryStaffDoc/NextCloud+File+Staging

GitHub repo

- https://github.com/cmbz/hl-saef

Pandas

- https://pandas.pydata.org/

Dataverse Native API

- https://guides.dataverse.org/en/latest/api/index.html

The Redesigned pyDataverse

- https://github.com/gdcc/pyDataverse
- PyDataverse Working Group

DVUploader

- https://github.com/gdcc/python-dvuploader

Related: Harvard Library's Historic Datasets Project (digitizing analog tables)