AN OPEN SOURCE DDI-BASED DATA CURATION SYSTEM FOR SOCIAL SCIENCE DATA

NADDI 2014. Vancouver, Canada

2 Partners, a Consultant, and a Software Developer

ipa
INNOVATIONS FOR POVERTY ACTION

Yale
ISPS

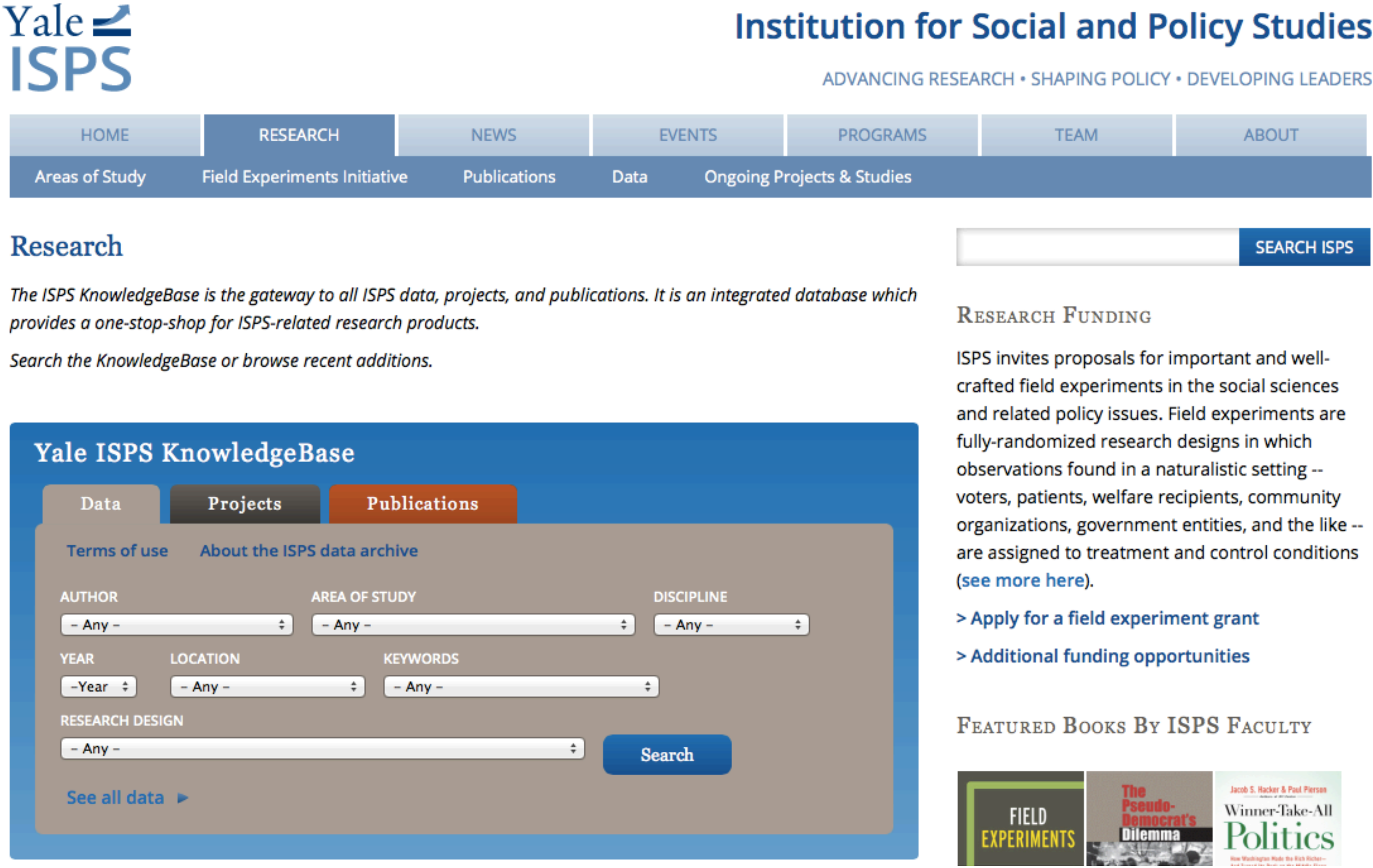
Digital Lifecycle Research & Consulting

colectica

STRATEGIC PLAN ›› 2013–2018

MORE EVIDENCE, LESS POVERTY

ipa
INNOVATIONS FOR
POVERTY ACTION



Institution for Social and Policy Studies

ADVANCING RESEARCH • SHAPING POLICY • DEVELOPING LEADERS

Research

The ISPS KnowledgeBase is the gateway to all ISPS data, projects, and publications. It is an integrated database which provides a one-stop-shop for ISPS-related research products.

Search the KnowledgeBase or browse recent additions.

Yale ISPS KnowledgeBase

Terms of use   About the ISPS data archive

AUTHOR: – Any –

AREA OF STUDY: – Any –

DISCIPLINE: – Any –

YEAR: –Year

LOCATION: – Any –

KEYWORDS: – Any –

RESEARCH DESIGN: – Any –

See all data ▸

Research Funding

ISPS invites proposals for important and well-crafted field experiments in the social sciences and related policy issues. Field experiments are fully-randomized research designs in which observations found in a naturalistic setting -- voters, patients, welfare recipients, community organizations, government entities, and the like -- are assigned to treatment and control conditions (see more here).

> Apply for a field experiment grant

> Additional funding opportunities

Featured Books By ISPS Faculty

The Repository as Data (Re) User: Hand Curating for Replication

Limor Peer, PhD
Yale Universitv. Institution for Social and Policv Studies

How does the ISPS Data Archive re-use data?

A key data curation task is appraisal and selection, with re-appraisal after initial selection. (DCC)

A well-curated archive ensures that, “data are accessible to designated users for first time use and reuse.” (DCC)

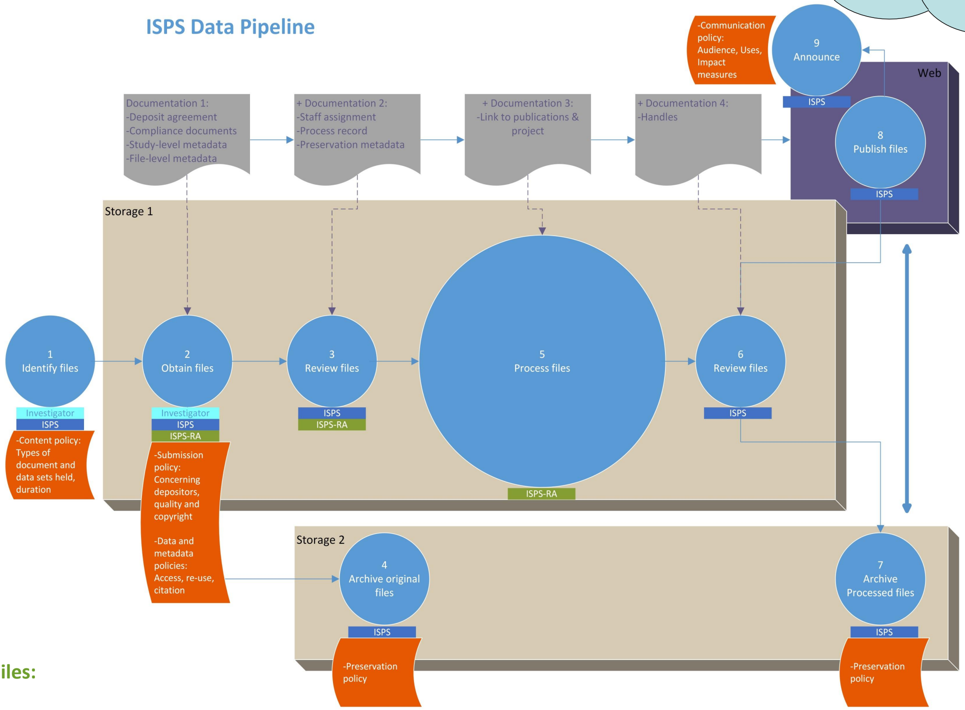
We argue that, in a replication archive, a key criterion for re-appraisal is whether the data and code reproduce the published results.

So, in addition to traditional curatorial tasks, dedicated data curation staff replicate analyses and validate published results for each study before publishing the files online.

In practice, this has implications for: Resources, Expertise, and Relationships.

How does replication drive curation at the ISPS Data Archive?

ISPS Data Pipeline



Documentation 1:
-Deposit agreement
-Compliance documents
-Study-level metadata
-File-level metadata

+ Documentation 2:
-Staff assignment
-Process record
-Preservation metadata

+ Documentation 3:
-Link to publications & project

+ Documentation 4:
-Handles

Storage 1

1 Identify files
Investigator
ISPS
-Content policy: Types of document and data sets held, duration

2 Obtain files
Investigator
ISPS
ISPS-RA
-Submission policy: Concerning depositors, quality and copyright
-Data and metadata policies: Access, re-use, citation

3 Review files
ISPS
ISPS-RA

5 Process files
ISPS-RA

6 Review files
ISPS

Storage 2

4 Archive original files
ISPS
-Preservation policy

7 Archive Processed files
ISPS
-Preservation policy

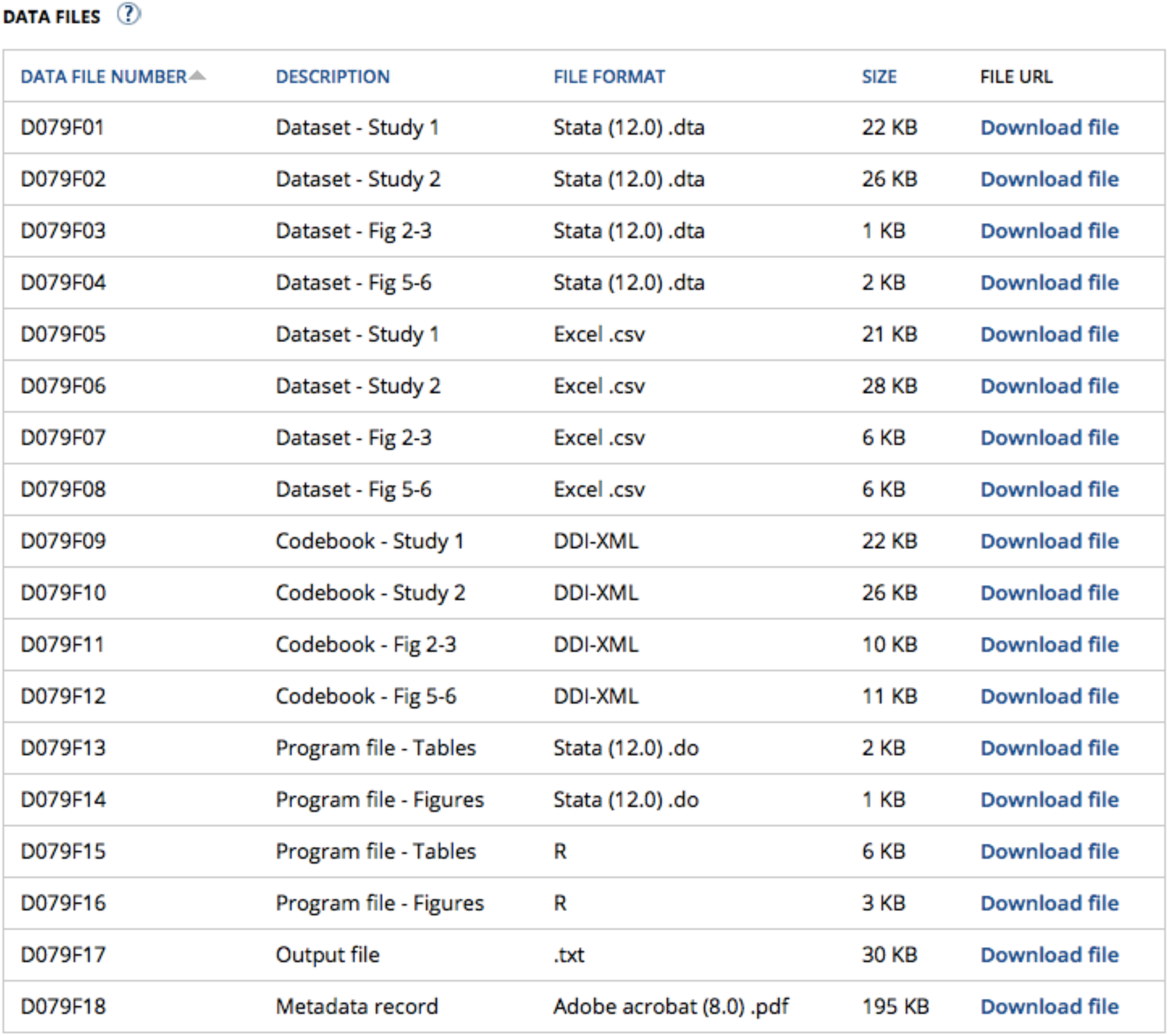
Web

8 Publish files
ISPS

9 Announce
ISPS
-Communication policy: Audience, Uses, Impact measures

Process Files:

1. Assign staff to study and files
2. Move original files to Archive space
3. Make copies of processed files and move to collaborative space
4. Identify related publication and project
5. Rename all copied files for public dissemination according to ISPS Data Archive naming conventions
6. Check and complete variable-level metadata for each data file
7. Compare variable information, check for additional variables and recoded variables, check variable/value labels
8. Check all files for confidential and other sensitive information
9. Run the statistical code and check against published results
10. Re-write statistical code in R and check replication
11. Communicate with PI as needed
12. Create new DDI-XML file with variable-level information
13. Create additional files by converting to readable formats (e.g., ASCII, PDF)
14. Update study- and file-level metadata record
15. Update tracking documents: process record / general study database / status document



DATA FILES

DATA FILE NUMBER	DESCRIPTION	FILE FORMAT	SIZE	FILE URL
D079F01	Dataset - Study 1	Stata (12.0) .dta	22 KB	Download file
D079F02	Dataset - Study 2	Stata (12.0) .dta	26 KB	Download file
D079F03	Dataset - Fig 2-3	Stata (12.0) .dta	1 KB	Download file
D079F04	Dataset - Fig 5-6	Stata (12.0) .dta	2 KB	Download file
D079F05	Dataset - Study 1	Excel .csv	21 KB	Download file
D079F06	Dataset - Study 2	Excel .csv	28 KB	Download file
D079F07	Dataset - Fig 2-3	Excel .csv	6 KB	Download file
D079F08	Dataset - Fig 5-6	Excel .csv	6 KB	Download file
D079F09	Codebook - Study 1	DDI-XML	22 KB	Download file
D079F10	Codebook - Study 2	DDI-XML	26 KB	Download file
D079F11	Codebook - Fig 2-3	DDI-XML	10 KB	Download file
D079F12	Codebook - Fig 5-6	DDI-XML	11 KB	Download file
D079F13	Program file - Tables	Stata (12.0) .do	2 KB	Download file
D079F14	Program file - Figures	Stata (12.0) .do	1 KB	Download file
D079F15	Program file - Tables	R	6 KB	Download file
D079F16	Program file - Figures	R	3 KB	Download file
D079F17	Output file	.txt	30 KB	Download file
D079F18	Metadata record	Adobe acrobat (8.0) .pdf	195 KB	Download file

2 Research Organizations

Institution for Social and Policy Studies (Yale)

- Data preparation at end of research project
- Replication
- Field Experiments
- Linked publications, data, and code

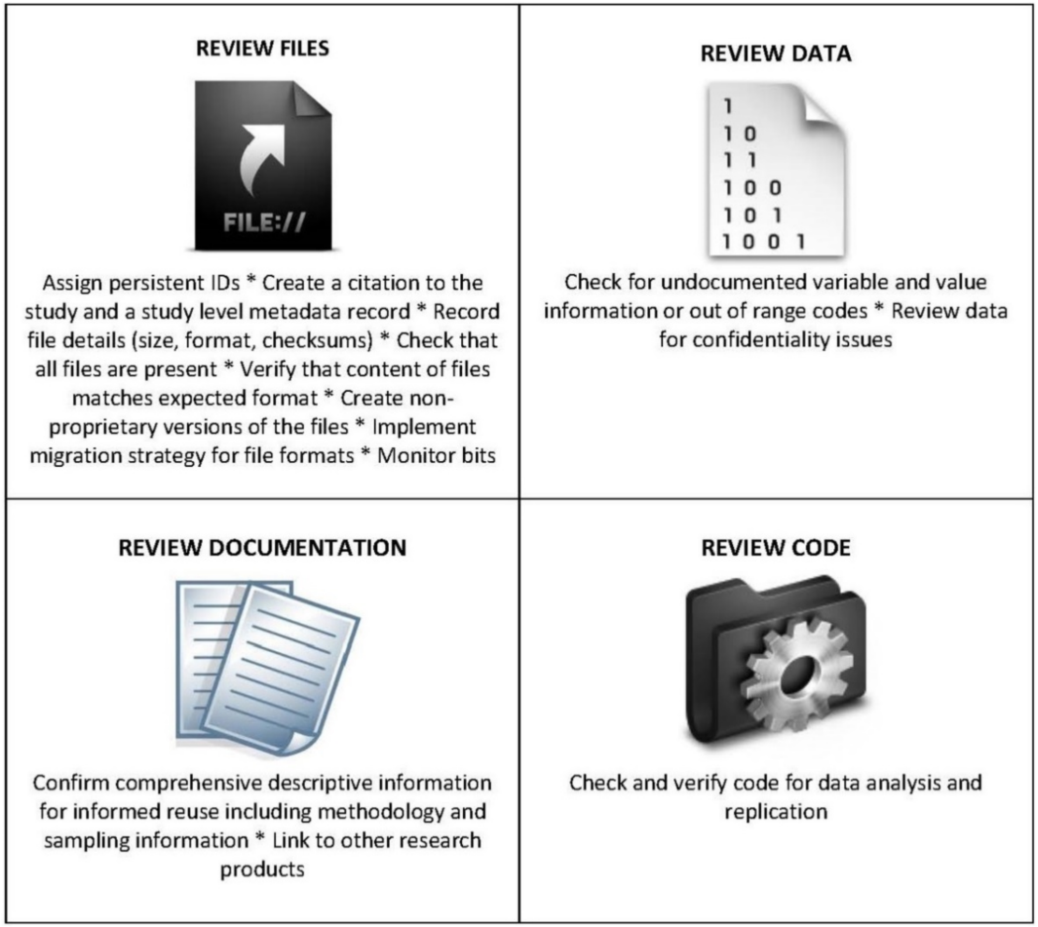
Innovations for Poverty Action

- Data preparation before analysis and at end of research project
- Project hosting from distributed research sites
- Lifecycle data management

ISPS and IPA Requirements

- Curation workflow management (dashboard)
- Track changes to files (provenance)
- Integrate metadata production with data and code review and cleaning
- Preservation metadata and formats
- Secure storage and access
- Smooth transition to public dissemination of content
- Preference for open source solutions

Data Quality Review



REVIEW FILES

Assign persistent IDs * Create a citation to the study and a study level metadata record * Record file details (size, format, checksums) * Check that all files are present * Verify that content of files matches expected format * Create non-proprietary versions of the files * Implement migration strategy for file formats * Monitor bits

REVIEW DATA

Check for undocumented variable and value information or out of range codes * Review data for confidentiality issues

REVIEW DOCUMENTATION

Confirm comprehensive descriptive information for informed reuse including methodology and sampling information * Link to other research products

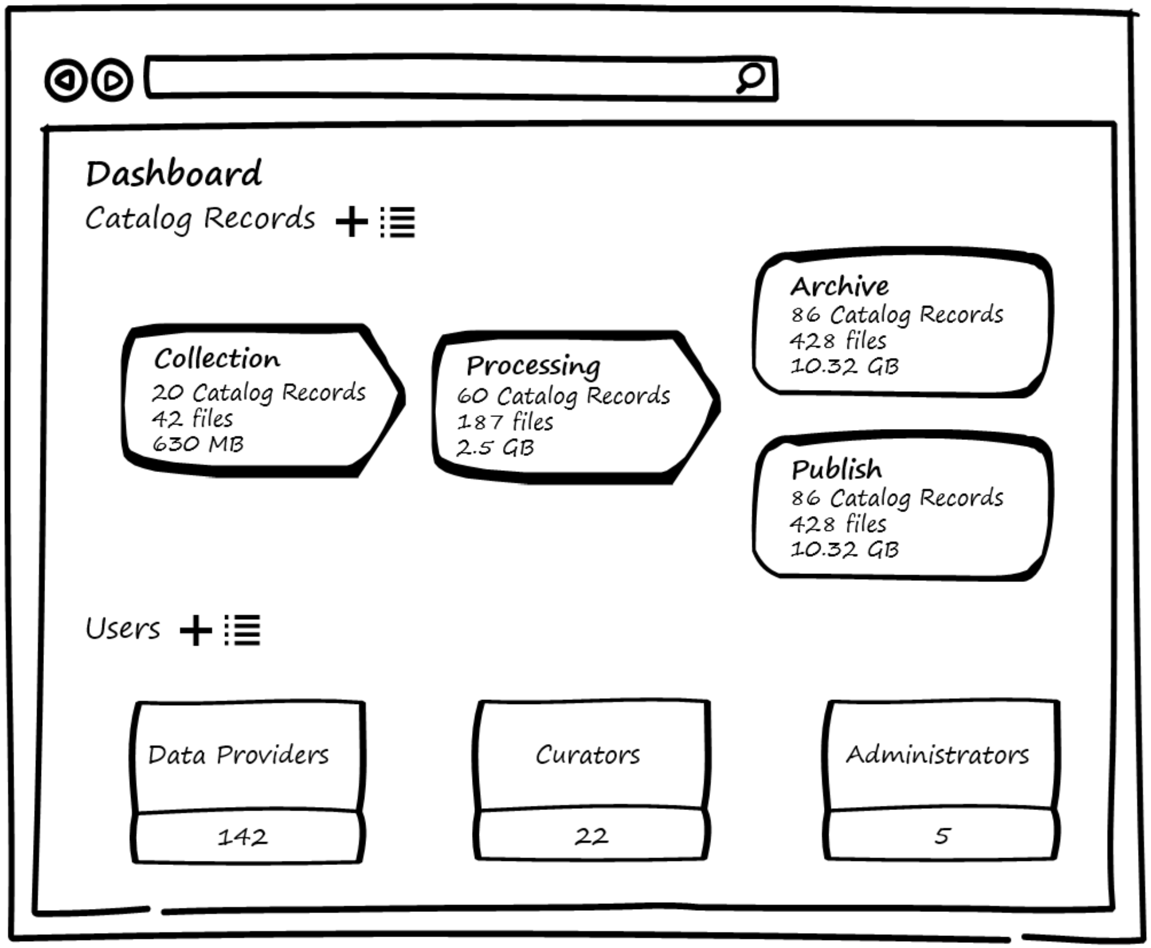
REVIEW CODE

Check and verify code for data analysis and replication

Source: Peer, Green, and Stephenson. 2014. Committing to Data Quality Review. International Journal of Digital Curation. Forthcoming.

Preprint: http://isps.yale.edu/sites/default/files/files/CommitingToDataQualityReview_idcc14-PrePrint.pdf

Build flexible data curation workflows



Dashboard

Catalog Records + ☰

Collection
20 Catalog Records
42 files
630 MB

Processing
60 Catalog Records
187 files
2.5 GB

Archive
86 Catalog Records
428 files
10.32 GB

Publish
86 Catalog Records
428 files
10.32 GB

Users + ☰

Data Providers	Curators	Administrators
142	22	5

Neat Features

□ Built on DDI 3.2

□ Web-based

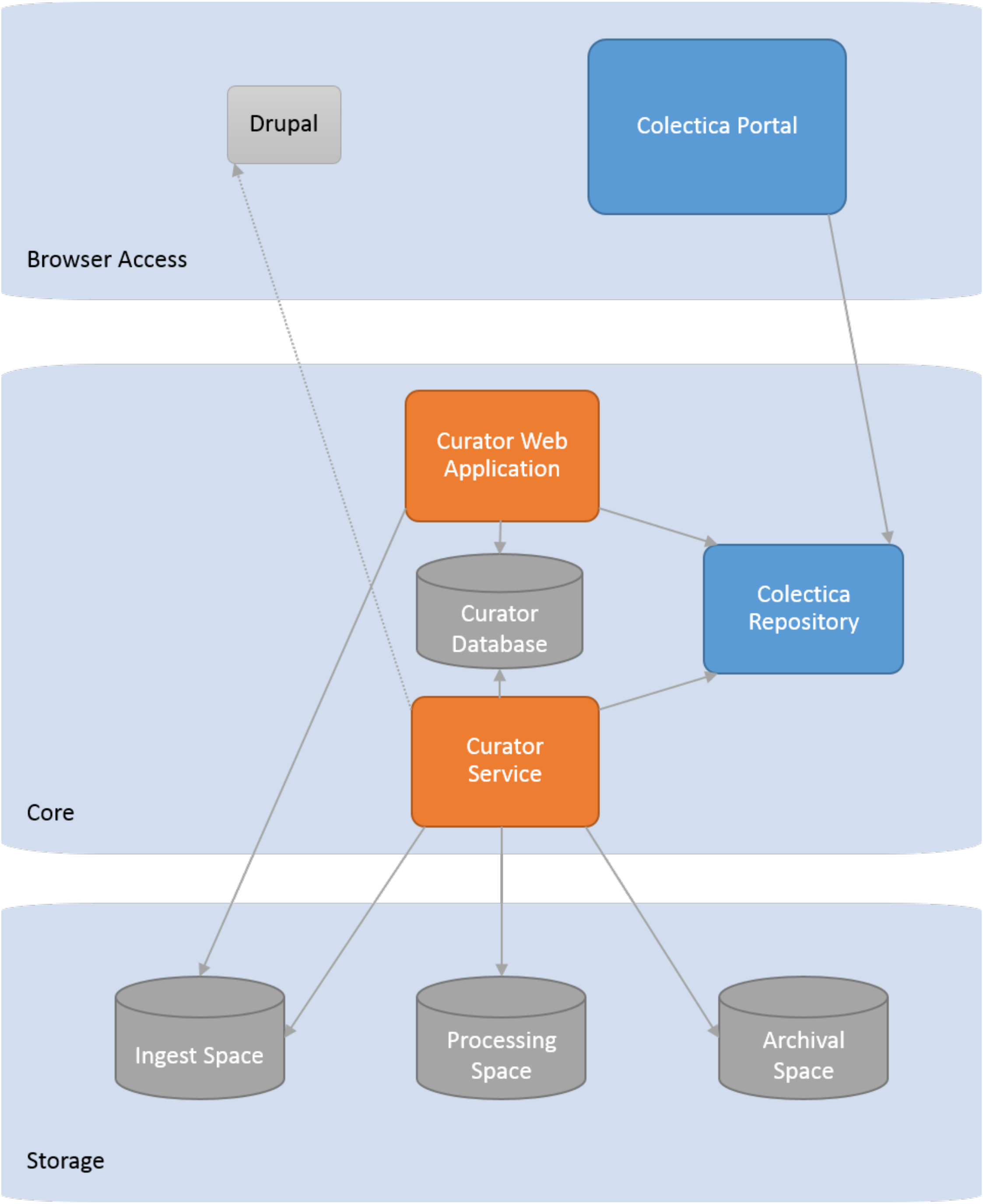
□ Open Source

<ddi>

open source initiative®

Builds on Existing Tools

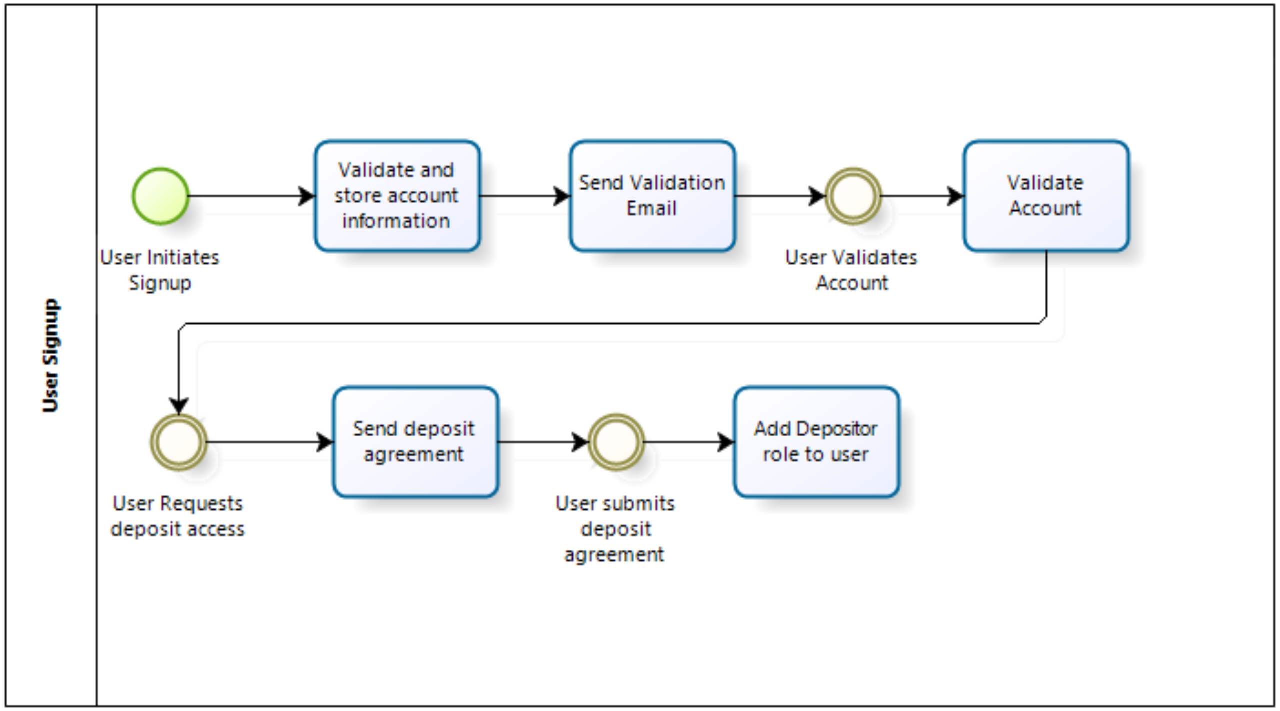
StatTransfer


Drupal
Colectica Portal
Browser Access
Curator Web Application
Curator Database
Colectica Repository
Curator Service
Core
Ingest Space
Processing Space
Archival Space
Storage

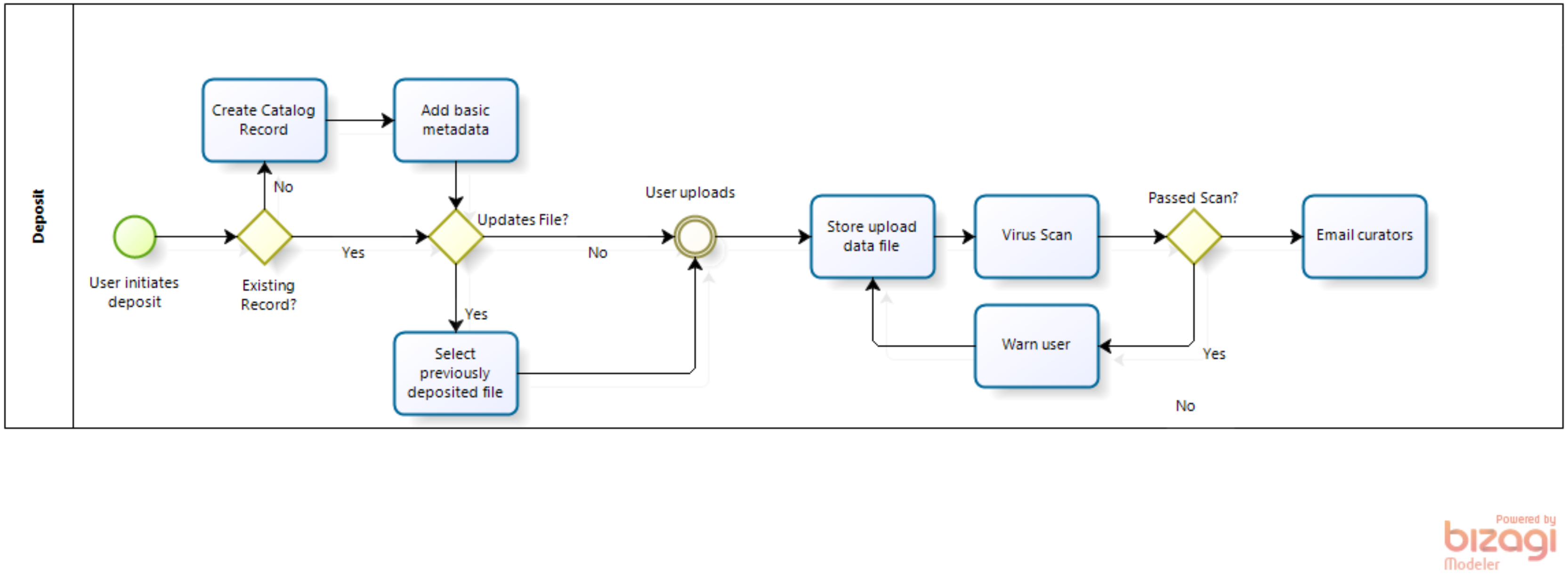
User Roles

- Depositor
- Curator
- Administrator
- Machines
- Researchers

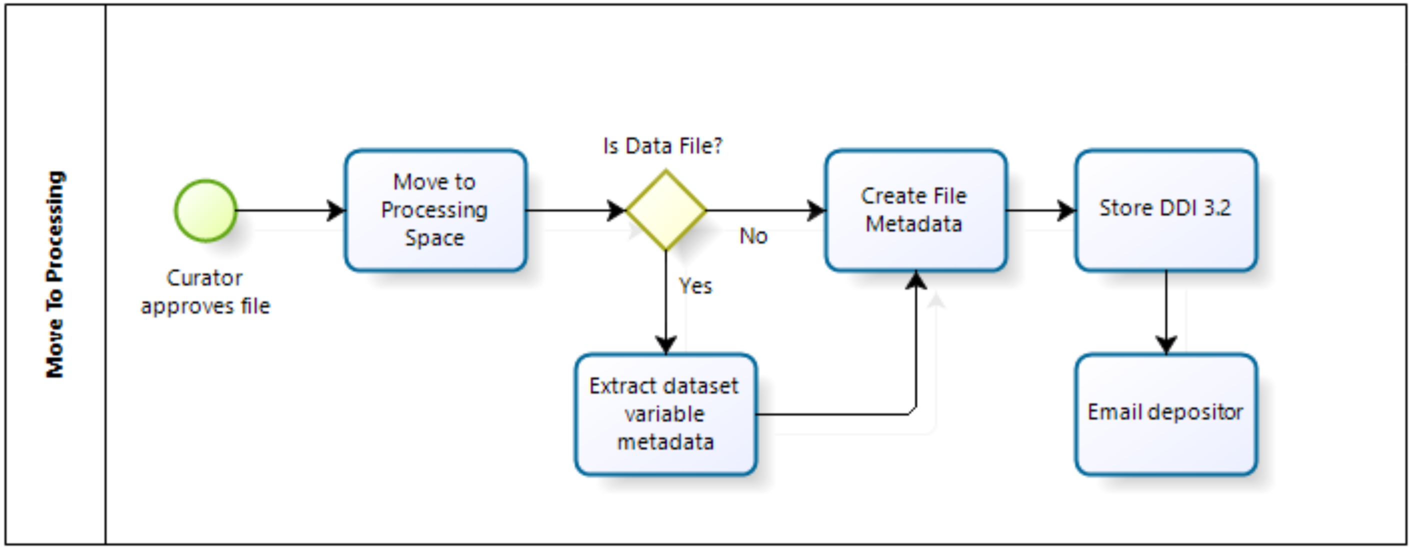
User Signup


User Signup
User Initiates Signup
Validate and store account information
Send Validation Email
User Validates Account
Validate Account
User Requests deposit access
Send deposit agreement
User submits deposit agreement
Add Depositor role to user

Deposit Files



Move to Processing



Example Processing Steps

- Check for missing variable labels
  - Add the labels
- Review data for personally identifiable information
  - Mark as non-public, or remove
- Add survey questionnaire to the file set
- Review and verify data processing code

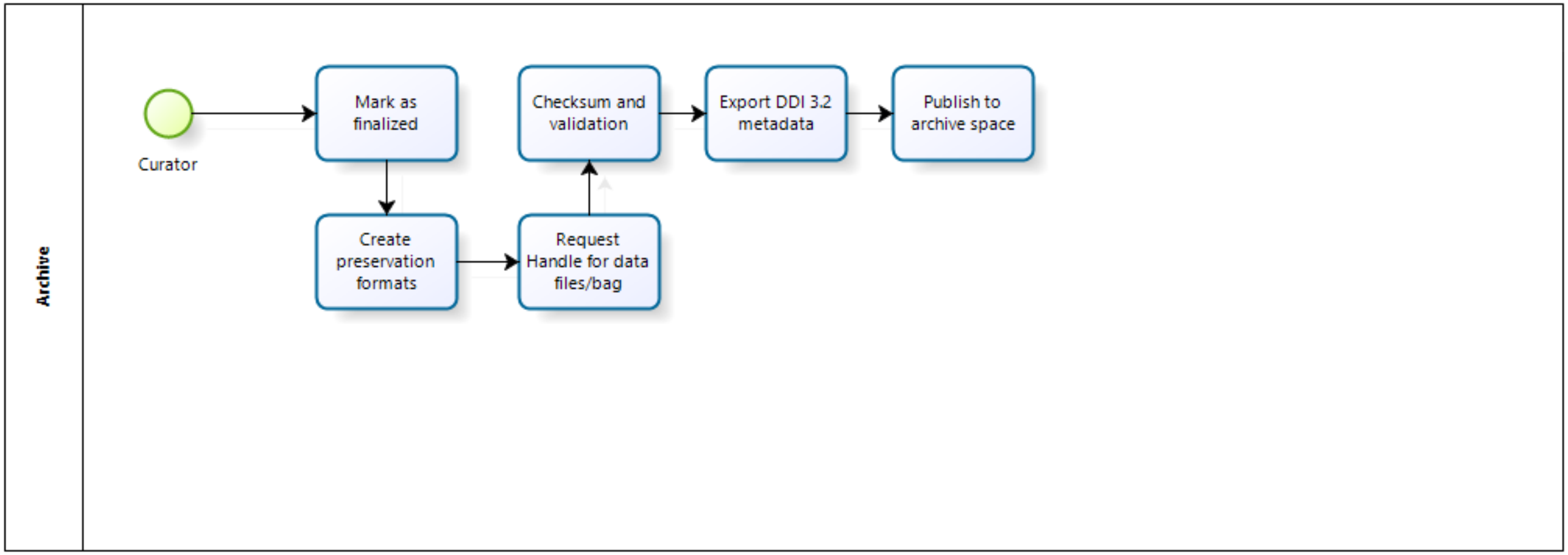
Processing: Example 1

- Goal: Ensure no missing variable labels
- Current Approach
  - Use Stata to open .dta file
  - Manually scan for missing labels
  - Use Stata to edit and save new copy of .dta file
  - Use Excel to make changes to metadata and "process record"

Processing: Example 1

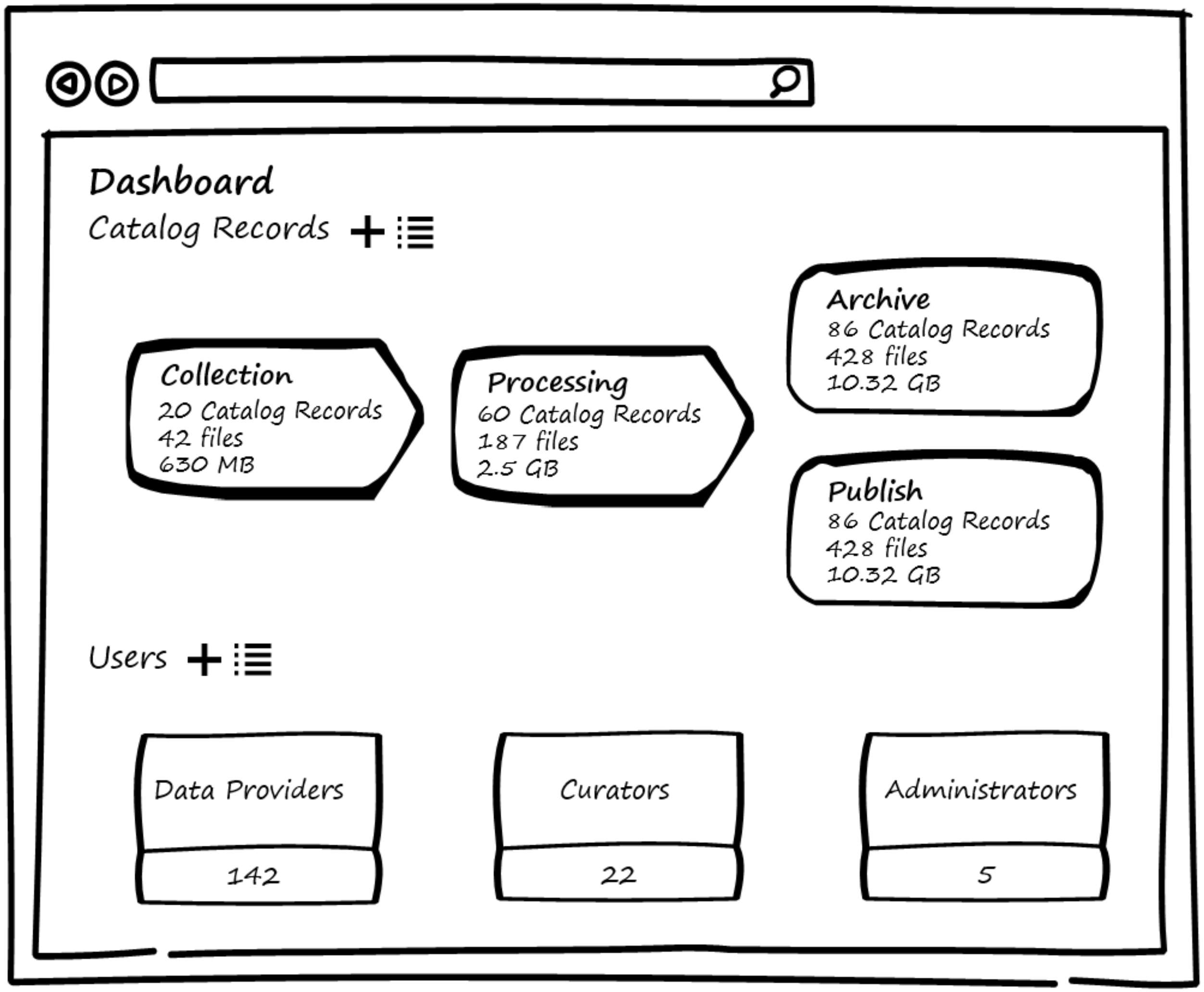
- Goal: Ensure no missing variable labels
- New Approach
  - Curator opens Web application
  - Curator sees a list of variables with missing labels
  - Curator adds labels as appropriate
  - The system logs this information and generates a new .dta file

Archive



Archive
Curator
Mark as finalized
Create preservation formats
Request Handle for data files/bag
Checksum and validation
Export DDI 3.2 metadata
Publish to archive space

Dashboard



Dashboard

Catalog Records

Collection
20 Catalog Records
42 files
630 MB

Processing
60 Catalog Records
187 files
2.5 GB

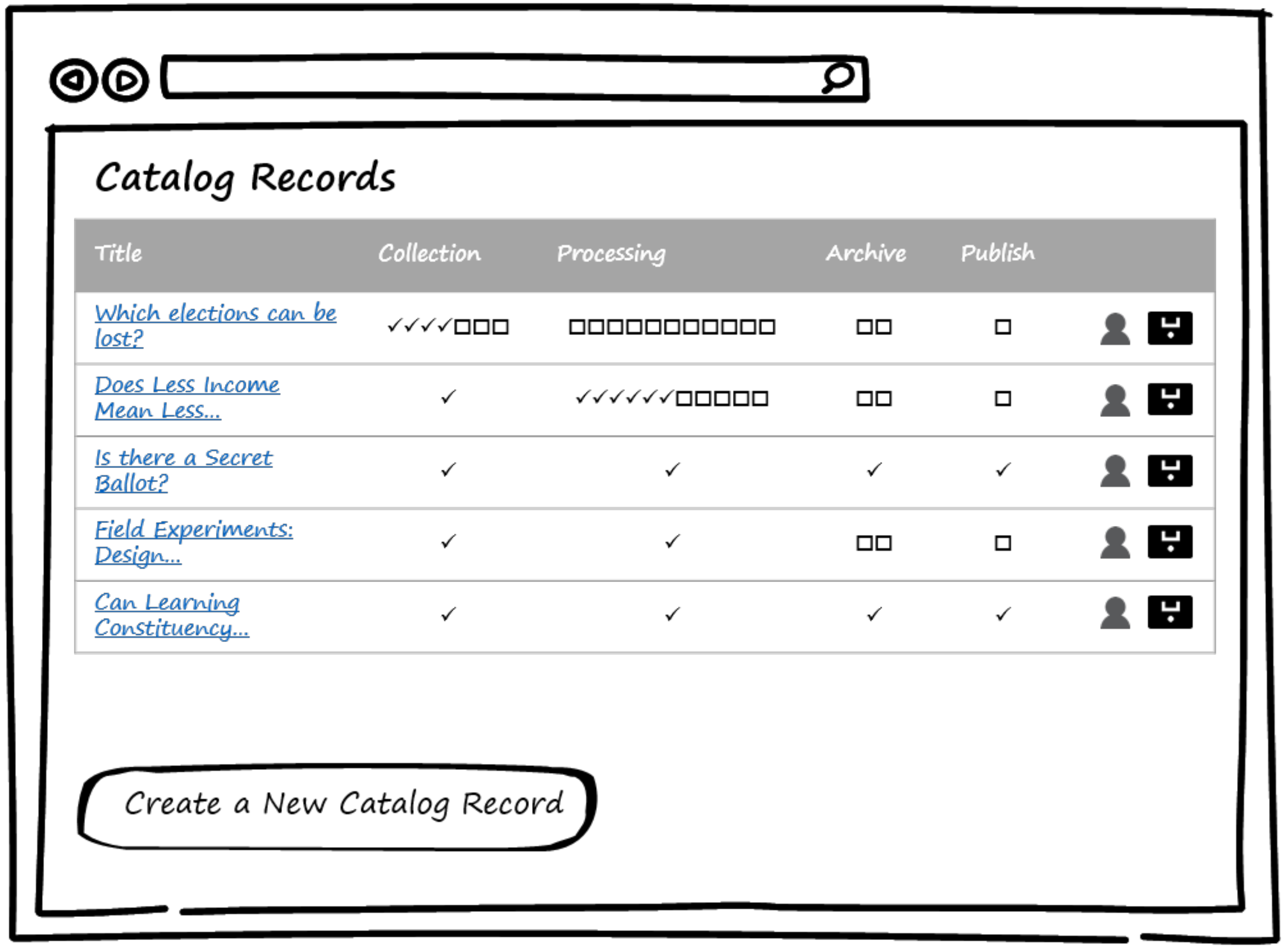
Archive
86 Catalog Records
428 files
10.32 GB

Publish
86 Catalog Records
428 files
10.32 GB

Users

Data Providers	Curators	Administrators
142	22	5

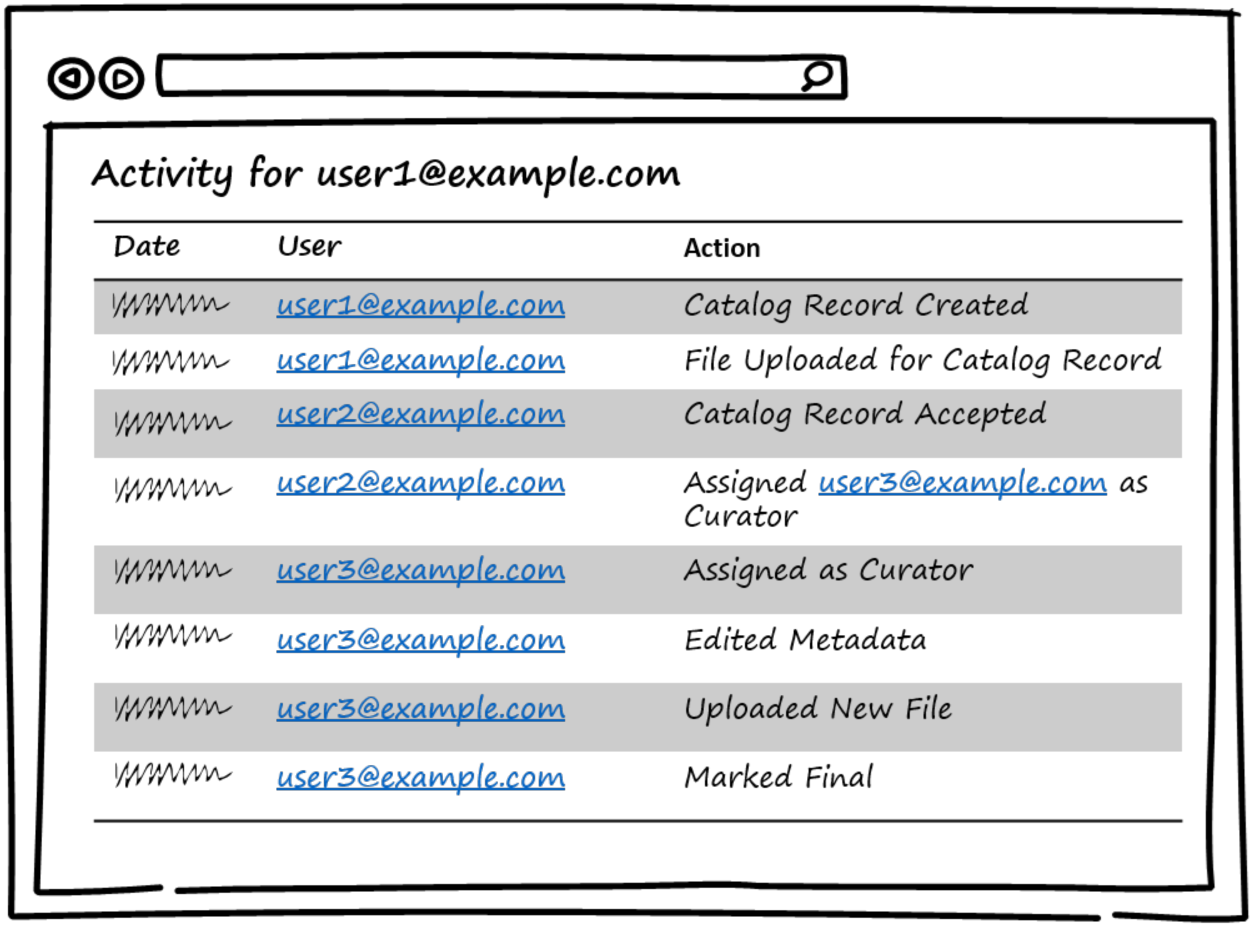
Status



Catalog Records

Title	Collection	Processing	Archive	Publish	
Which elections can be lost?	✓✓✓✓□□□	□□□□□□□□□□□	□□	□	👤 💾
Does Less Income Mean Less...	✓	✓✓✓✓✓✓□□□□□	□□	□	👤 💾
Is there a Secret Ballot?	✓	✓	✓	✓	👤 💾
Field Experiments: Design...	✓	✓	□□	□	👤 💾
Can Learning Constituency...	✓	✓	✓	✓	👤 💾

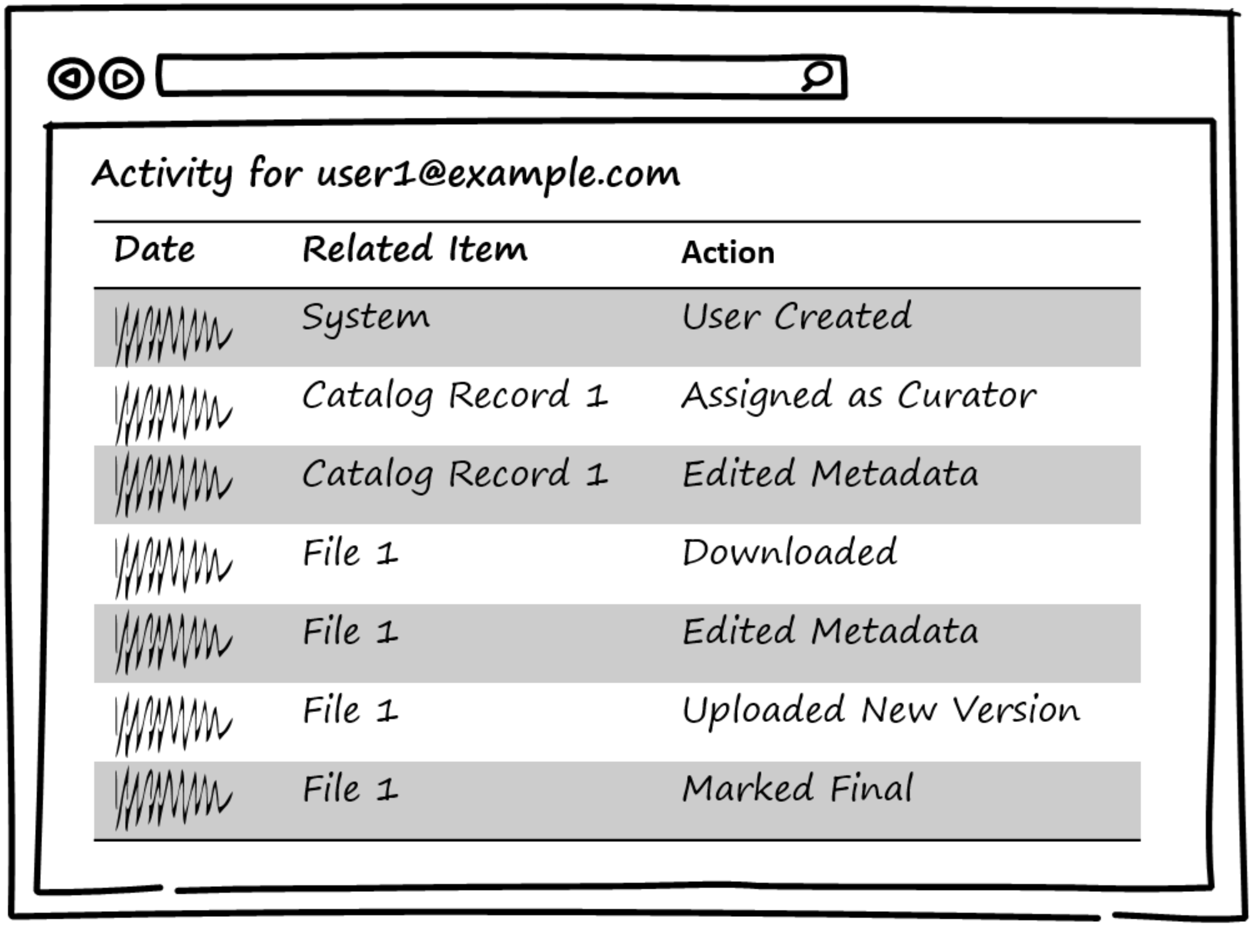
History by Item



Activity for user1@example.com

Date	User	Action
	user1@example.com	Catalog Record Created
	user1@example.com	File Uploaded for Catalog Record
	user2@example.com	Catalog Record Accepted
	user2@example.com	Assigned user3@example.com as Curator
	user3@example.com	Assigned as Curator
	user3@example.com	Edited Metadata
	user3@example.com	Uploaded New File
	user3@example.com	Marked Final

History by User



Activity for user1@example.com

Date	Related Item	Action
	System	User Created
	Catalog Record 1	Assigned as Curator
	Catalog Record 1	Edited Metadata
	File 1	Downloaded
	File 1	Edited Metadata
	File 1	Uploaded New Version
	File 1	Marked Final

Data Migration

- Automatically migrate existing data archive into the Curator system

Timeline

- Now: Design
- April – June: Development
- July+: Ongoing development and maintenance

colectica

Thank you

Contributor	**Organization**	**Email**
Ann Green	Independent Consultant	green.ann@gmail.com
Jeremy Iverson	Colectica	jeremy@colectica.com
Niall Keleher	Innovations for Poverty Action	nkeleher@poverty-action.org
Limor Peer	Yale University	limor.peer@yale.edu
Dan Smith	Colectica	dan@colectica.com

colectica.com