Optimizing data ingest: Experiences from ICPSR

Jared Lyle
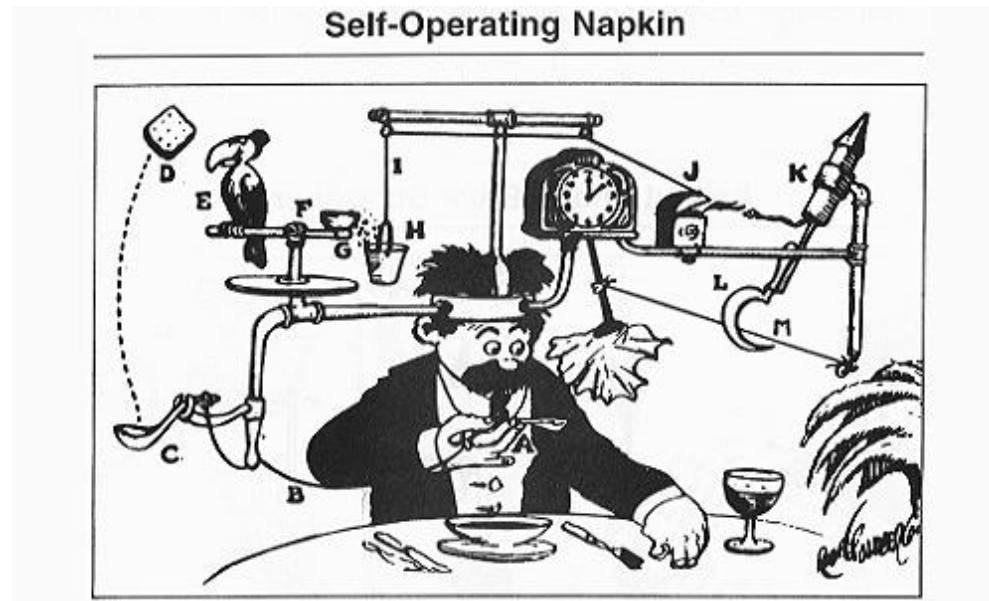Abay Israel
Justin Noble

Outline

1. Data ingest goals
2. Brief history of ICPSR's data ingest process
3. Ingest processes from other repositories
4. June 2017 improvements and enhancements
5. Lessons learned & future improvements

Data Ingest Goals

Transfer:

- Content

- Metadata

- Legal permissions

Content

Self-Operating Napkin

Atari's "Star Trek" instructions:

Insert Quarter.
Avoid Klingons.

-See Isaacson's *Steve Jobs*

Metadata

https://www.adn.com/arts/2017/01/20/artists-take-over-the-asylum-with-igca-member-show/

WHO'S VIEWED YOUR PROFILE

0 Your profile has been viewed by 0 people in the past 90 days.

0 You have shown up in search results 0 times in the past 90 days.

PROFILE STRENGTH

Beginner

Improve Your Profile Strength ▸

https://blog.kissmetrics.com/the-progress-bar/

https://blog.kissmetrics.com/the-progress-bar/

Legal permissions

- Do they have authority to deposit the content with you?
- Can you then modify, reformat, preserve, describe, and redisseminate?
- Are there any human disclosure issues?

Brief history of ICPSR's data ingest process

ICPSR

ICPSR Data Deposit Form

Please complete this form to provide ICPSR with information about the substantive and technical characteristics of your data collection. Completing the form ensures that your collection will be accurately and thoroughly described for potential secondary analysts in the social science community. It is vital that the information solicited on this form be provided as completely and accurately as possible.

The deposit form allows ICPSR staff to prepare a study description for the online searchable databases that ICPSR maintains as well as for various ICPSR printed publications. This information also permits the creation of authoritative bibliographic citations.

Finally, the data deposit form grants permission for ICPSR to archive and distribute your data collection. Please sign the form in the box below to attest to the four statements above the signature line.

1. Descriptive title of data collection (if acronym is used, give complete meaning), including time period(s) and geographic location(s) that the data cover:

2. Principal Investigator(s) and affiliation(s) at time of data collection (for multiple investigators, give proper name order):

Please sign below. By signing, you give assurance of the following:

- I give permission for this data collection to be redisseminated by the Inter-university Consortium for Political and Social Research under its standard terms of use.

- I have implicit or explicit copyright to this work and have the right to make it publicly available through ICPSR.

- In preparing this data collection for public archiving and distribution, I have removed all information directly identifying the research subjects in these data, and I have used due diligence in preventing information in the collection from being used to disclose the identity of research subjects. I further agree to release and hold harmless ICPSR (including staff and the ICPSR Council) and the University of Michigan from any and all liability from claims arising out of any legal action concerning identification of research subjects, breaches of confidentiality, or invasions of privacy by or on behalf of said subjects.

- If applicable: The Institutional Review Board (IRB) on my campus granted permission for the data collection effort underlying the file(s) being deposited. *(Please attach a copy of the consent form.)*

Printed name and title

Institutional affiliation

Signature   Date

Pre-2007

ICPSR Deposit Form and Agreement - Deposit #23342

You are signed in as lyle@umich.edu Depositing to TEST

Select Deposit		Deposit Access and Notifications	
0. Arrival Information	I. Citation	II. Description	III. Methodology/Sampling
IV. Update Existing Collection	V. Restrictions	VI. Upload Files	VII. Review and Submit

Contact ICPSR staff (734) 615-7957 or deposit@icpsr.umich.edu if you need help completing this form.

Skip to different step-- **does not save any changes made to this page**:

[I. Citation ▾]

I. Citation

a. Principal Investigator. For multiple investigators, list names in order they should appear:

PI	Institutional Affiliation

Use the following text field to list any other PIs and their affiliations:

[   ]

b. Title. Provide a descriptive title of the data collection, including the time period(s) and geographic location(s) that the data cover:

[   ]

2007

Upload \ Describe \ Restrictions \ Verify \ Agreement

Upload

Please upload all files relevant to this deposit.

If you have multiple files to submit, you may bundle your files using WinZip (or its equivalents), tar, and/or GNU's gzip.

Note about large files - There is a 2GB limit on file uploads.

File Name	Size	Format	Checksum
upload_abc.txt	355 bytes	SPSS System File, Variables: 129, Cases: 272	16516516516

Delete +

All materials that cannot be uploaded, including hardcopy only documentation and/or removable media, should be sent to the following address:

ICPSR Acquisitions
Attn: ICPSR
P.O. Box 1248
Ann Arbor, MI 48106-1248

For UPS/FEDEX shipments:
ICPSR Acquisitions
Attn: ICPSR
330 Packard Street
Ann Arbor, MI 48104

< [ ] >

2010 Mock-up

Data Deposit Form

Welcome! Please select the archive to which you would like to deposit your data collection. If you need help completing this form, contact ICPSR Staff at 734-615-7957 or deposit@icpsr.umich.edu. NIJ grant holders can contact National Archive of Criminal Justice Data National Institute of Justice Data Resources Program staff at 800-999-0960 (9 a.m. to 5 p.m. EST) or nacjd@icpsr.umich.edu.

General Archive ▾

Your Information

If you are a frequent depositor, you may prefer to login or create a new MyData account.

Your First Name	Your Last Name	Your Email Address

Describe the Collection

This section collects basic details about the data collection. On the next page of the deposit form, you will be able to describe this data collection in more detail.

Title of the Data Collection

Descriptive titles typically include the time period(s) and geographic location(s) that the data cover.

Principal Investigator(s)

Please list the Principal Investigators in order of importance to the study. If the Principal Investigator is an organization, complete just the affiliation field.
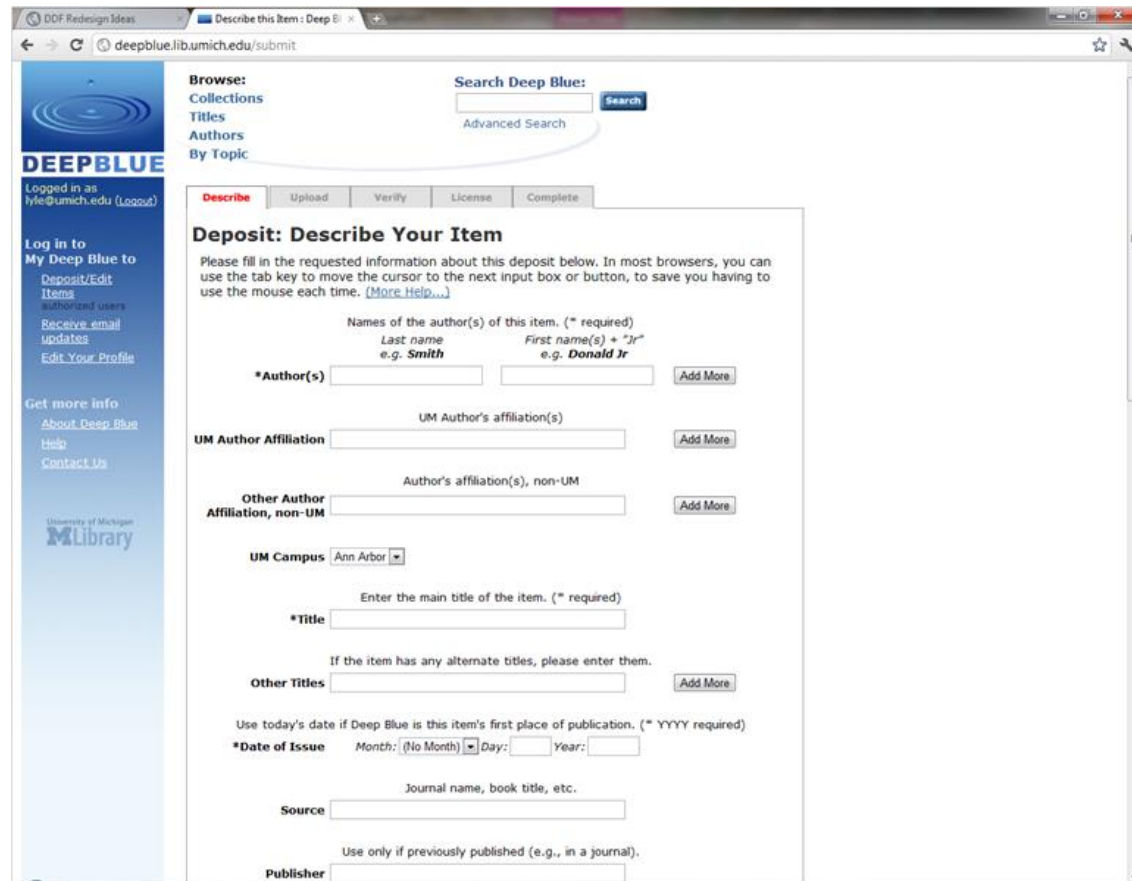
PI First Name	PI Last Name	PI Affiliation

ADD ANOTHER ROW

2012

Ingest processes from other repositories

DeepBlue

Dryad

Figshare

Zenodo

Web

June 2017 improvements and enhancements

- **Deposit workspace** that allows depositors to incrementally work on a submission as opposed to needing to upload and describe all files at once

- **Collaboration** – Users can share projects with other collaborators and grant various permission levels to these collaborators. This allows for multiple researchers or PIs to easily collaborate on a submission of materials to ICPSR.

- **Folder structure** – For each project, a depositor has the flexibility to organize his/her submission into folders and sub-folders within the project workspace.

- **File-level** metadata – When depositing materials, researchers will be able to provide both study-level descriptive information and metadata as well as provide information to describe individual files within a submission.

- **Project communication log** – Within each project, there is a project communication log that allows for the data depositor and ICPSR curators to send messages to each other about the data collection.

+ CREATE NEW PROJECT

Search within workspace, enter 3 or more characters

Test APSA Data Deposit ✎ Edit Project Header

Hide inactive

🏠 Workspace

— 💼 OPENICPSR-100501

👤 Shared with me

+ 💼 OPENICPSR-100647

| 📁 Create Folder | ⊕ Upload Files | ⊕ Import From Zip | more ▾ |

🗑 DELETE SELECTION

☐	Name ⊡	Type ⊡	Size ⊡	Last Modified ⊡	Actions
☐	📕 APSA_Codebook.pdf	application/pdf	3 MB	3/10/2017 8:17:26 AM	
☐	📊 APSA_code.sps	text/x-spss-syntax	3 KB	8/18/2016 5:10:44 AM	
☐	📊 APSA_data.sav	application/x-spss-s...	433 KB	3/9/2017 9:49:35 AM	

Project

+ Expand All

+ Project Description

+ Scope of Project

+ Methodology

+ Related Publications

Project Communication Log   **+** add entry   👁 watch

OPENICPSR-100501

📢 **PUBLISH PROJECT**

👤 Share Project

👤 Change Owner

☰ View Log

⚠ Report a Problem

Storage

Space Used: 3 MB

Sub Folders: 0

Files: 3

Deposit System improved other ICPSR technologies

- openICPSR, DataLumos, Census
- The Preview functionality on the study homepages
- Help icons through all our pages
- The future ICPSR Metadata editor

Constant Improvement: The Iterative Process

Internal and external feedback

- Multiple methods to receive user feedback
    - Targeted forms and surveys
    - Website links (Report a problem/Give feedback)
    - Use cases with examples and instructions
    - Open forum and brainstorming sessions
- Various audiences/users
    - ICPSR Staff
    - Depositors and Principal Investigators
    - Researchers, Staff, and Students
    - Experienced and novice users

Prioritization, Feasibility and Usability

Human Centered Approach

- Building user stories to drive design
  - As a [user], I would like [feature] so that [reason]
- The user experience is a core component to any new feature
- All improvements must be ADA compliant
- Never underemphasize the need for proper testing!

Agile Process at ICPSR
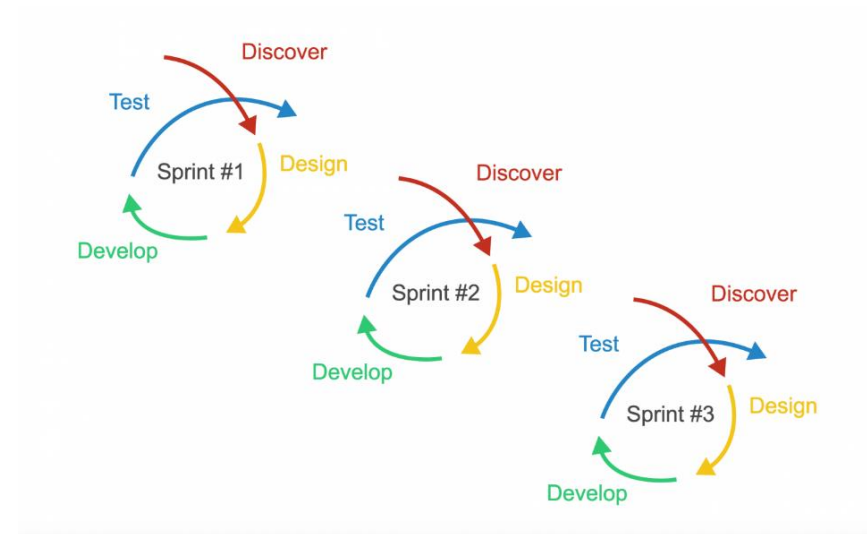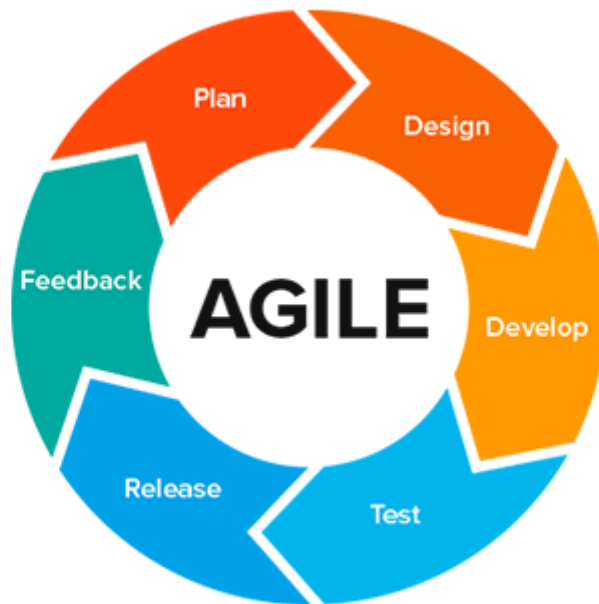
Feedback leading to new features
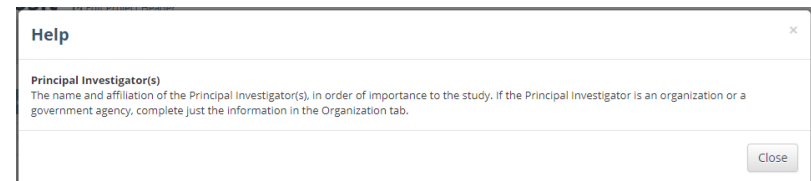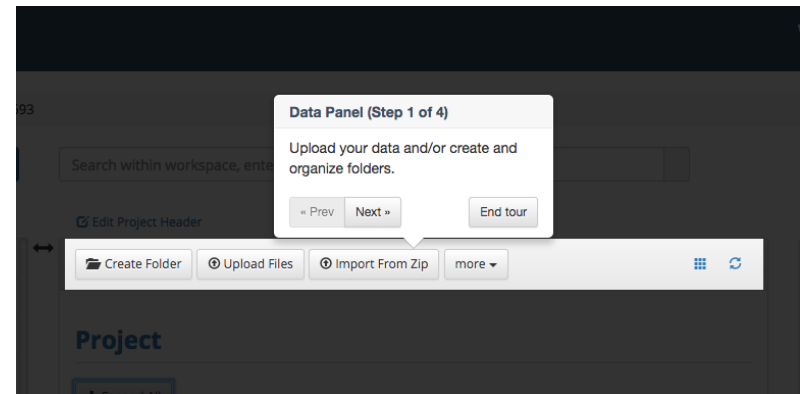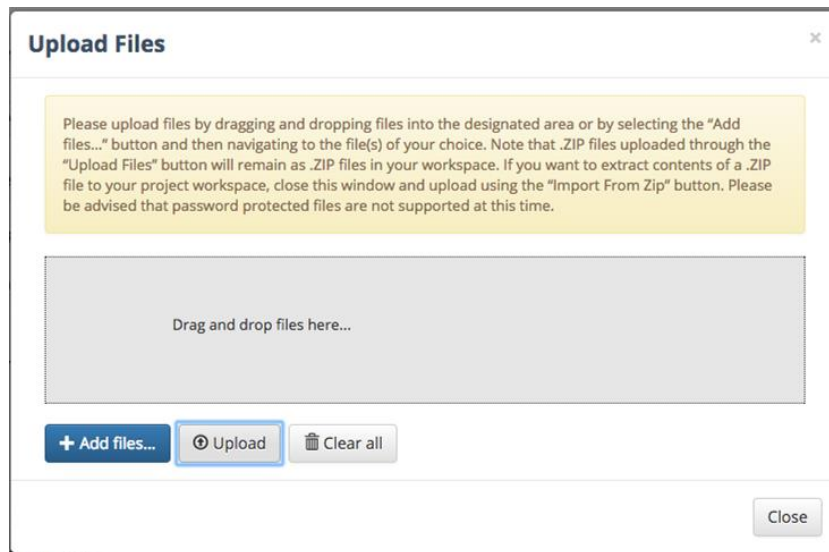
Lessons learned

Lesson 1

Positive enhancements can have unintended and sometimes negative consequences

+ CREATE NEW PROJECT

Search within workspace, enter 3 or more characters

Test APSA Data Deposit ✐ Edit Project Header

Hide inactive

🏠 Workspace

— 💼 OPENICPSR-100501

👤 Shared with me

+ 💼 OPENICPSR-100647

📁 Create Folder   ⊕ Upload Files   ⊕ Import From Zip   more ▾   ⊞   ⟳

🗑 DELETE SELECTION

☐	Name ⊟	Type ⊟	Size ⊟	Last Modified ⊟	Actions
☐	📕 APSA_Codebook.pdf	application/pdf	3 MB	3/10/2017 8:17:26 AM	
☐	📊 APSA_code.sps	text/x-spss-syntax	3 KB	8/18/2016 5:10:44 AM	
☐	📊 APSA_data.sav	application/x-spss-s...	433 KB	3/9/2017 9:49:35 AM	

Project

+ Expand All

+ Project Description

+ Scope of Project

+ Methodology

+ Related Publications

Project Communication Log   + add entry   👁 watch

OPENICPSR-100501

📢 PUBLISH PROJECT

👤 Share Project

👤 Change Owner

☰ View Log

⚠ Report a Problem

Storage

Space Used: 3 MB

Sub Folders: 0

Files: 3

Lesson 2

Importance of development based on user needs

+ CREATE NEW PROJECT

Search within workspace, enter 3 or more characters

Hide inactive

🏠 Workspace

— 💼 OPENICPSR-100501

👤 Shared with me

+ 💼 OPENICPSR-100647

Test APSA Data Deposit ✏️ Edit Project Header

📁 Create Folder | ⊕ Upload Files | ⊕ Import From Zip | more ▾

🔳 🔄

🗑 DELETE SELECTION

☐	Name ⊟	Type ⊟	Size ⊟	Last Modified ⊟	Actions
☐	📕 APSA_Codebook.pdf	application/pdf	3 MB	3/10/2017 8:17:26 AM	
☐	📊 APSA_code.sps	text/x-spss-syntax	3 KB	8/18/2016 5:10:44 AM	
☐	📊 APSA_data.sav	application/x-spss-s...	433 KB	3/9/2017 9:49:35 AM	

Project

+ Expand All

+ Project Description

+ Scope of Project

+ Methodology

+ Related Publications

Project Communication Log   **+ add entry**   👁 watch

OPENICPSR-100501

📢 PUBLISH PROJECT

👤 Share Project

👤 Change Owner

☰ View Log

⚠ Report a Problem

Storage

Space Used: 3 MB

Sub Folders: 0

Files: 3

Lesson 3

A thoughtfully developed, user-friendly data deposit system is great, but it's just one component of an repository's acquisition strategy

Future improvements

- Streamlined path from web site to deposit
- Navigation improvements
- Simplified workflow
- Help text with detailed examples
- Productivity enhancements

Thank you!

lyle@umich.edu