Migration to InvenioRDM: A Case Study

Guillaume Viger, Austin Sharp, Eric Newman, and Matt Carson
Galter Library, Northwestern University

RDM

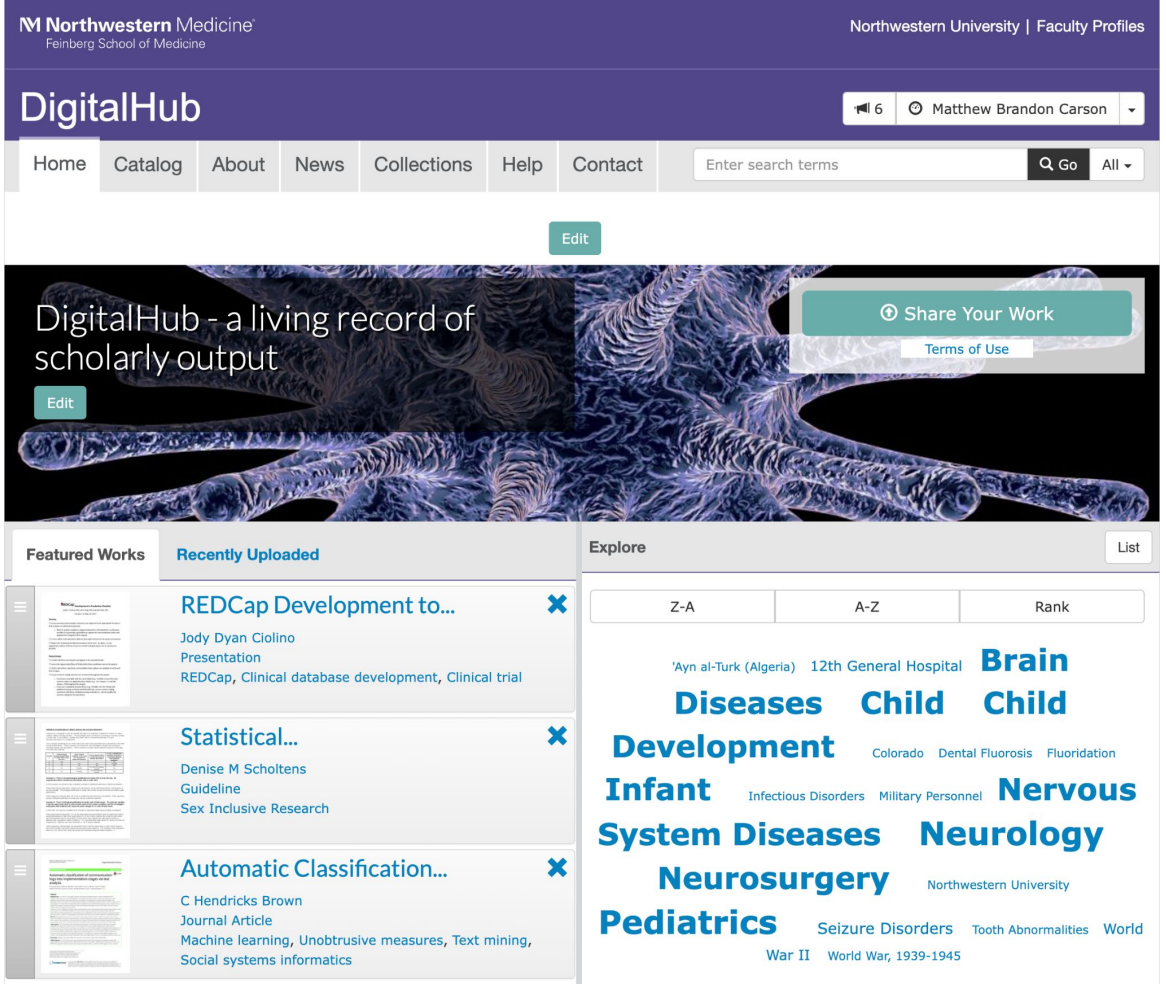
Prism

Overview

- Motivation
- How did we do it?
- Lessons learned (so far)
- Current status and challenges

Why Migrate?

- To enhance data management for sharing, reproducibility, and open access
- Years of customizations made upgrading problematic; chance to explore new options
- Grant funding acquired
- Conversations with COAR and CERN began



Why InvenioRDM?

CERN

zenodo

Free Open Source Software

Invenio is Free Open Source Software supported by a committed community of multidisciplinary institutions.

Code | Docs | Examples

Friendly and Responsive Community

Although Invenio was born at CERN, its community is growing bigger every day. Talk to the team now in our chatroom or forum.

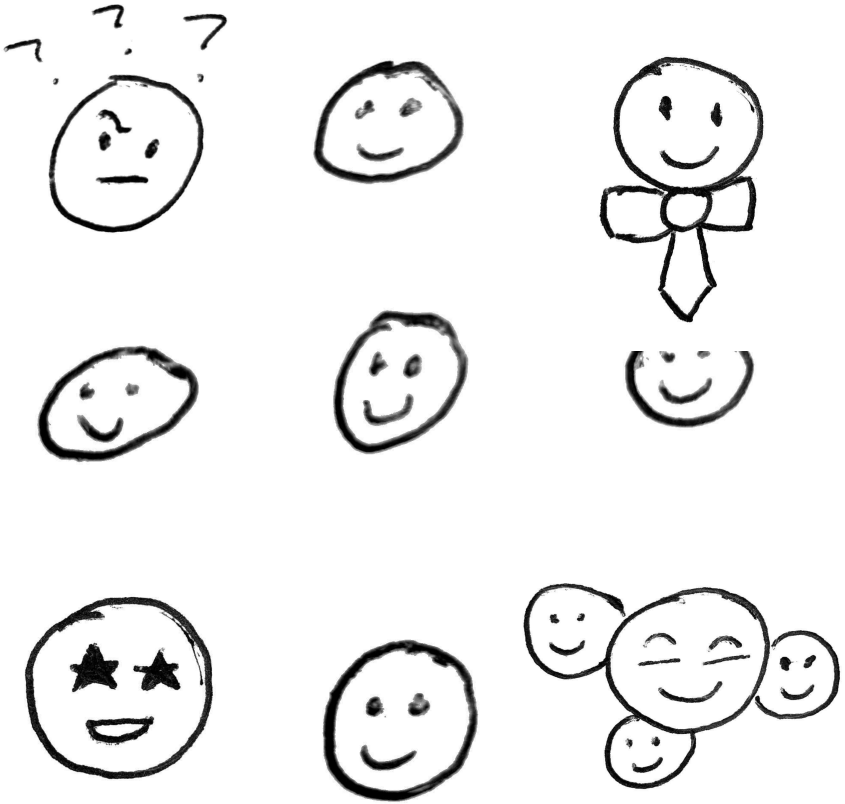
Chatroom | Forum | Get Involved | Events

The "Safe bet"

Invenio community has been around for 20 years. Solid services have been built on top of it to ensure long-term confidence.

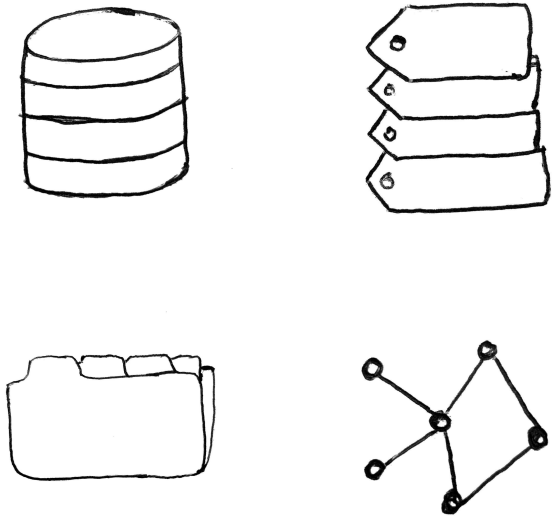
Live services | Products

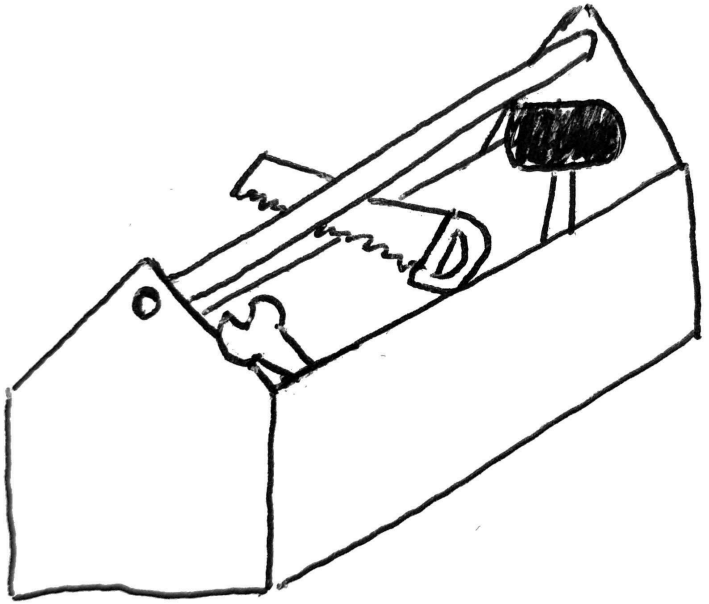
Who?



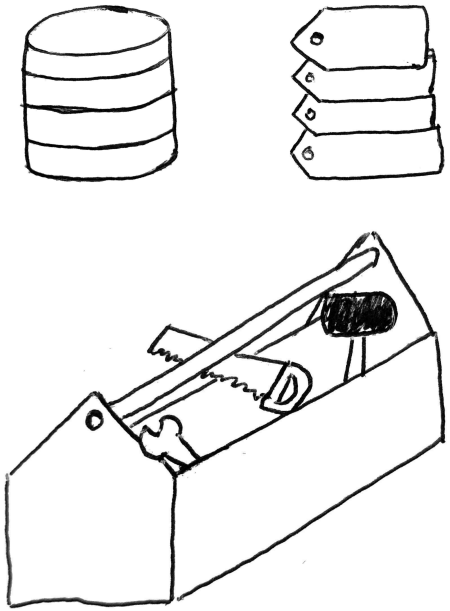
Goal

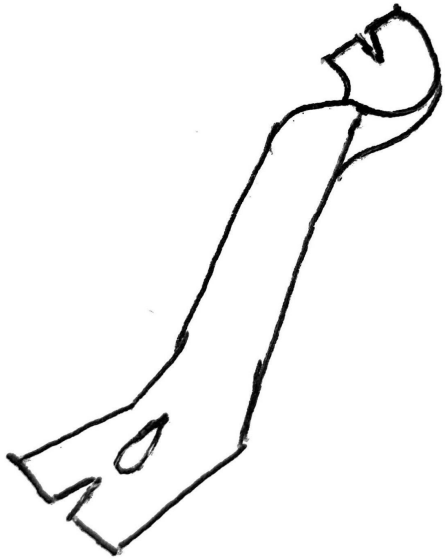
Exporting the Data



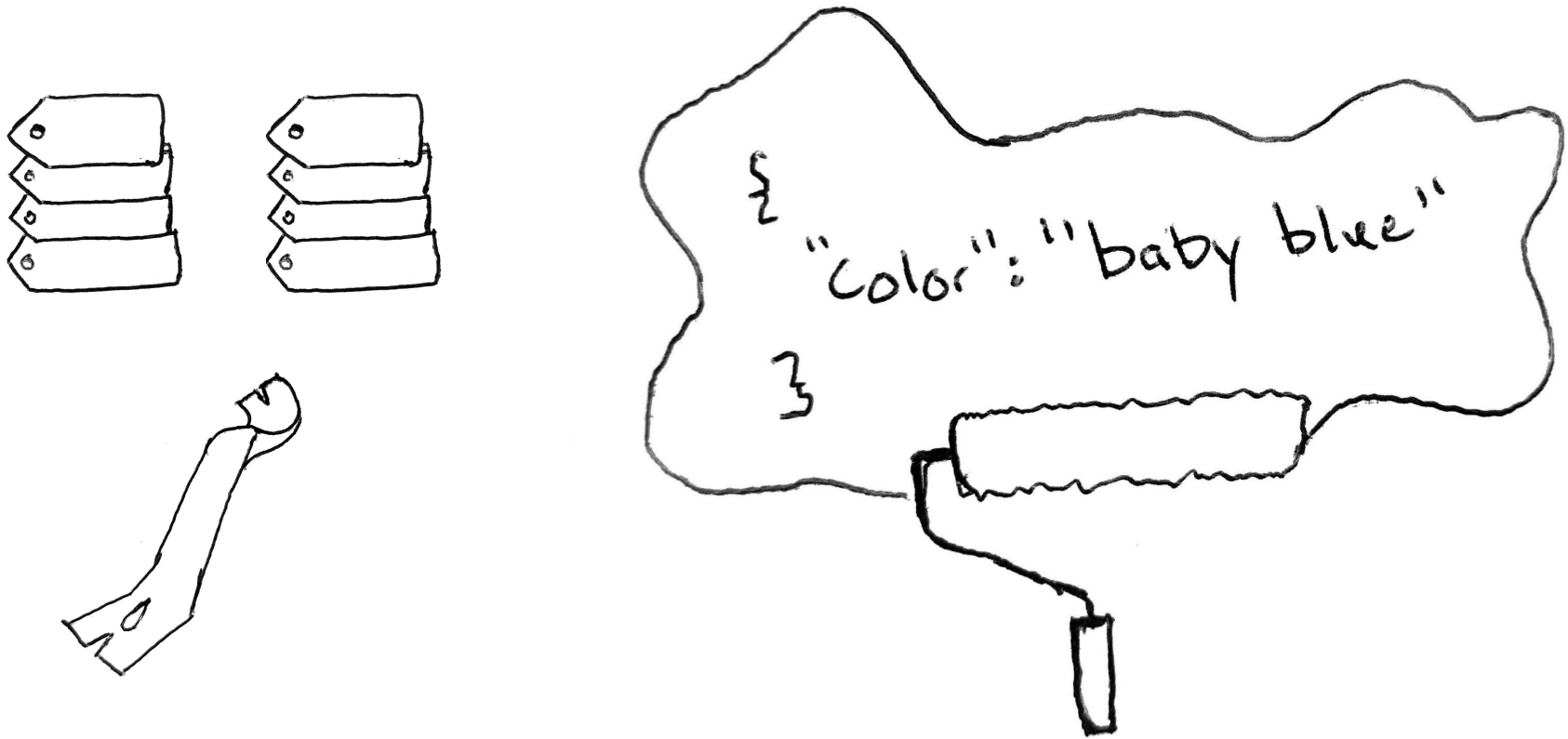


Using the Tools

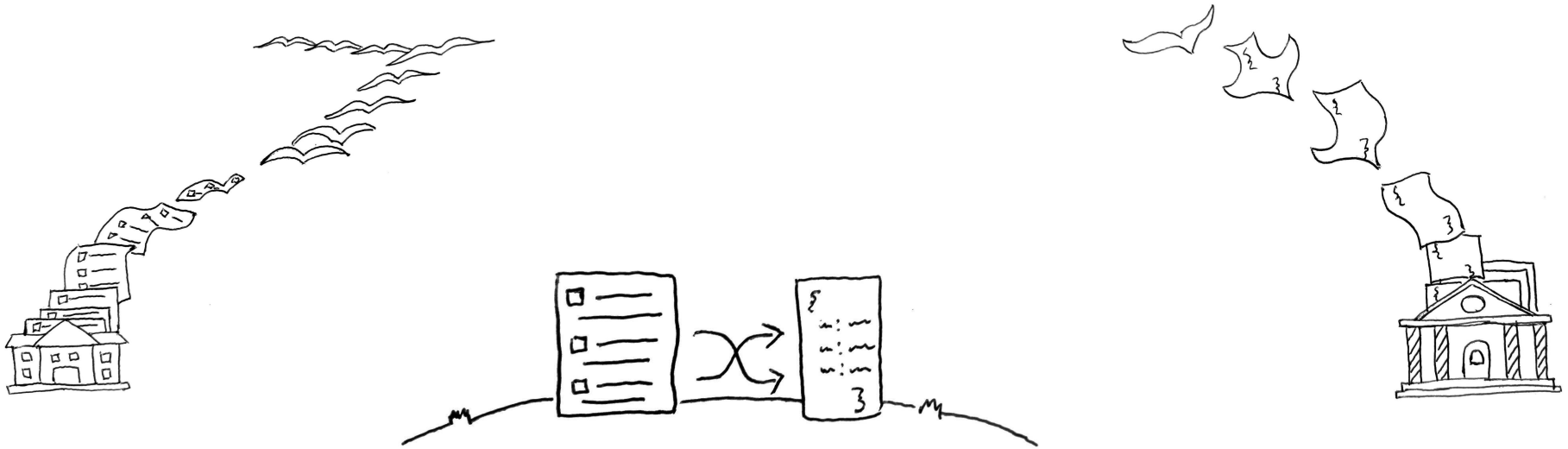




Transforming the Data



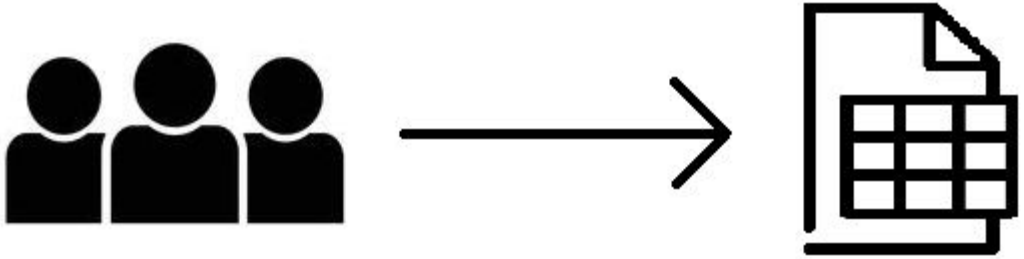
Getting There



How did we do it?

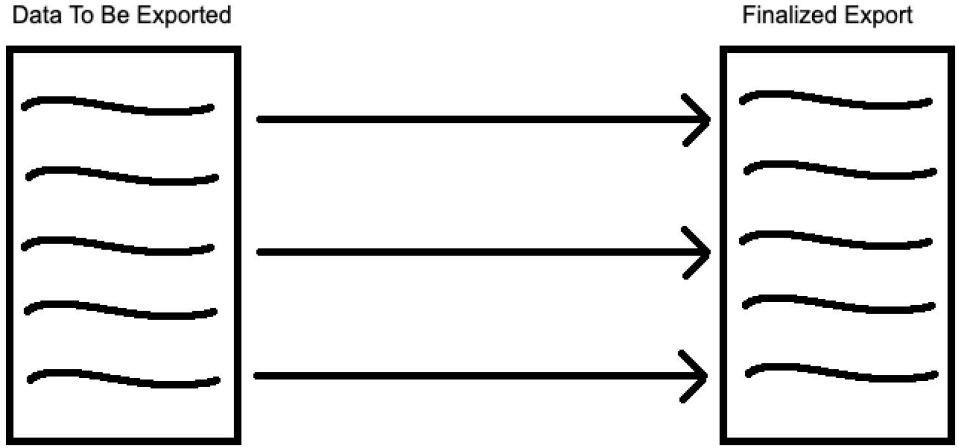
Export: Metadata Mapping

- Organized an internal working group specifically for metadata
- Composition: metadata librarians, repository stakeholder representatives, and developers
- Came up with a table of correspondences



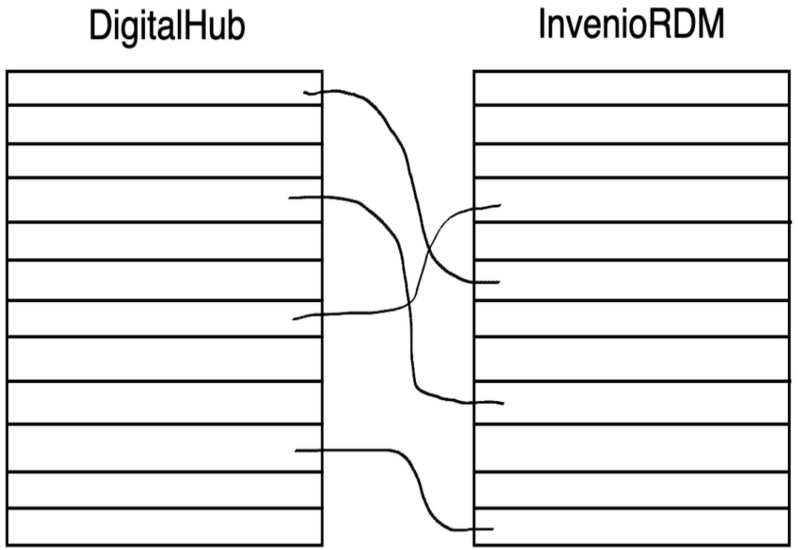
Export: Metadata Mapping

- Mapping informed serialization rules from DigitalHub to Prism
- Used InvenioRDM metadata reference for details of structure
- Having export structure the data and do the correspondences, we could parallelize the development
- Feedback from import
- [InvenioRDM Documentation]

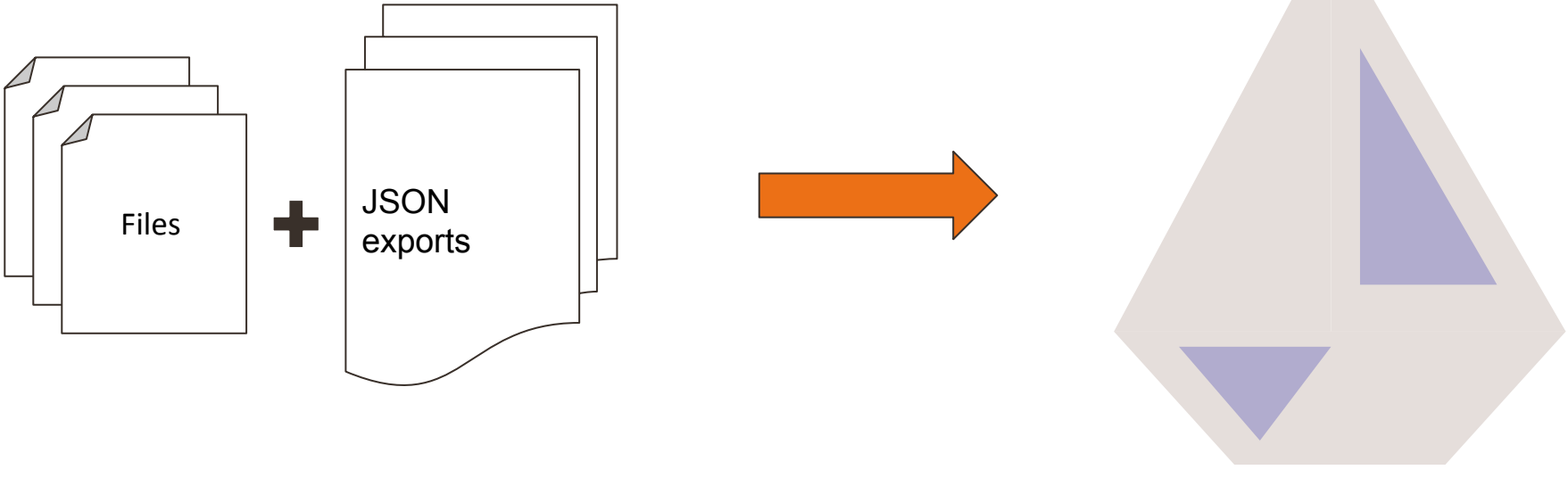


Export: Metadata Mapping Tips

- Existing Code From Sufia could be used as a jump start
- Beware of quirks unique to ActiveFedora



Import



Import - Two Approaches Considered

REST API

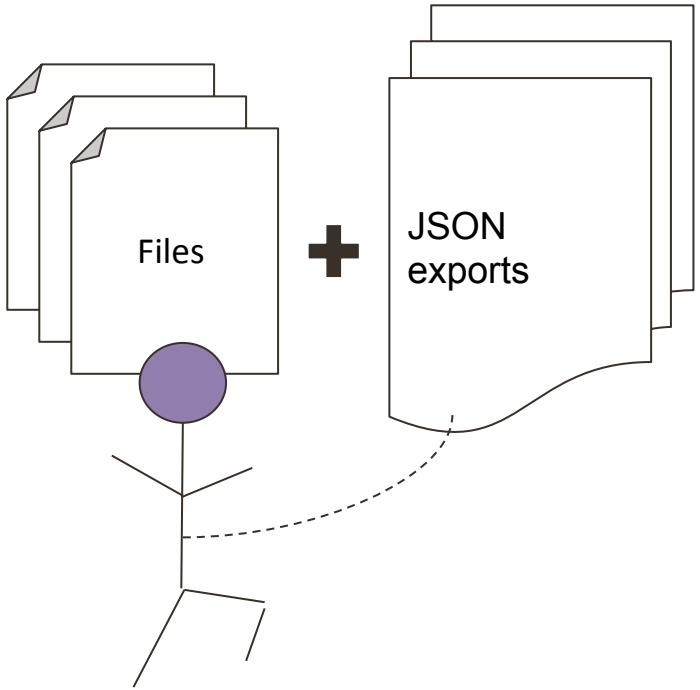
- \+ Stable
- \+ Less development effort
- − Limited to released features
- − Slow

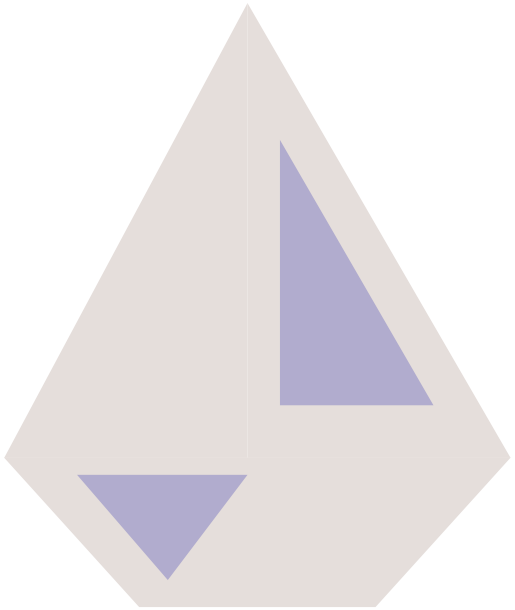
Programmatic API

- − Less stable
- − More development effort
- \+ Unconstrained by released features
- \+ Fast

Import - Deciding factor

Programmatic API

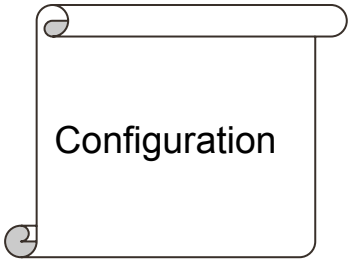




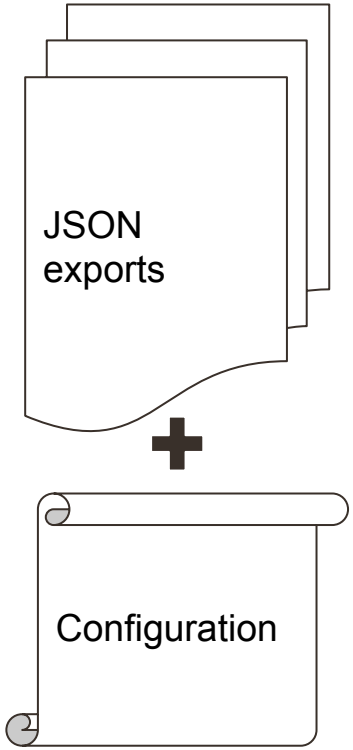
Prism

Import extension

```
pipenv run invenio galter-import --config import-config.json
```



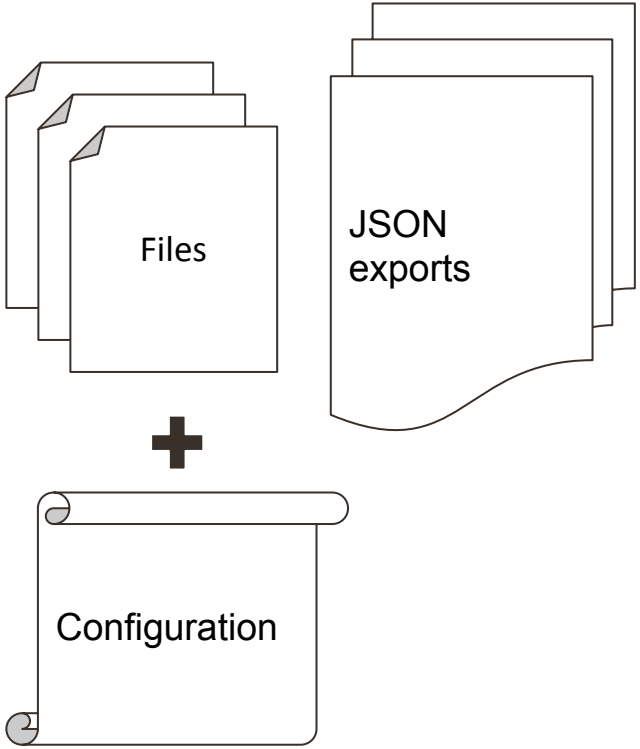
Import Communities



For each community
- Prepare it
- Import users
- Import metadata

Prism

Import Records



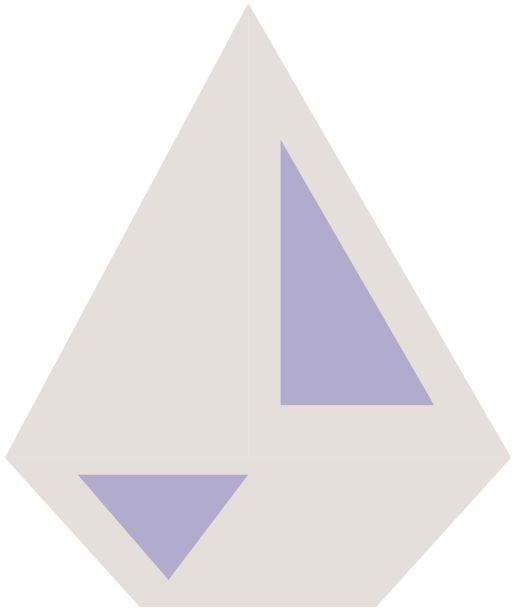
For each record
- Prepare it
- Import users
- Import metadata
- Assign to community
- Import file

Then

For each record
- Publish

Prism

Review Records



Prism

Librarians review representative samples

- All fields
- Edge cases
- User interface

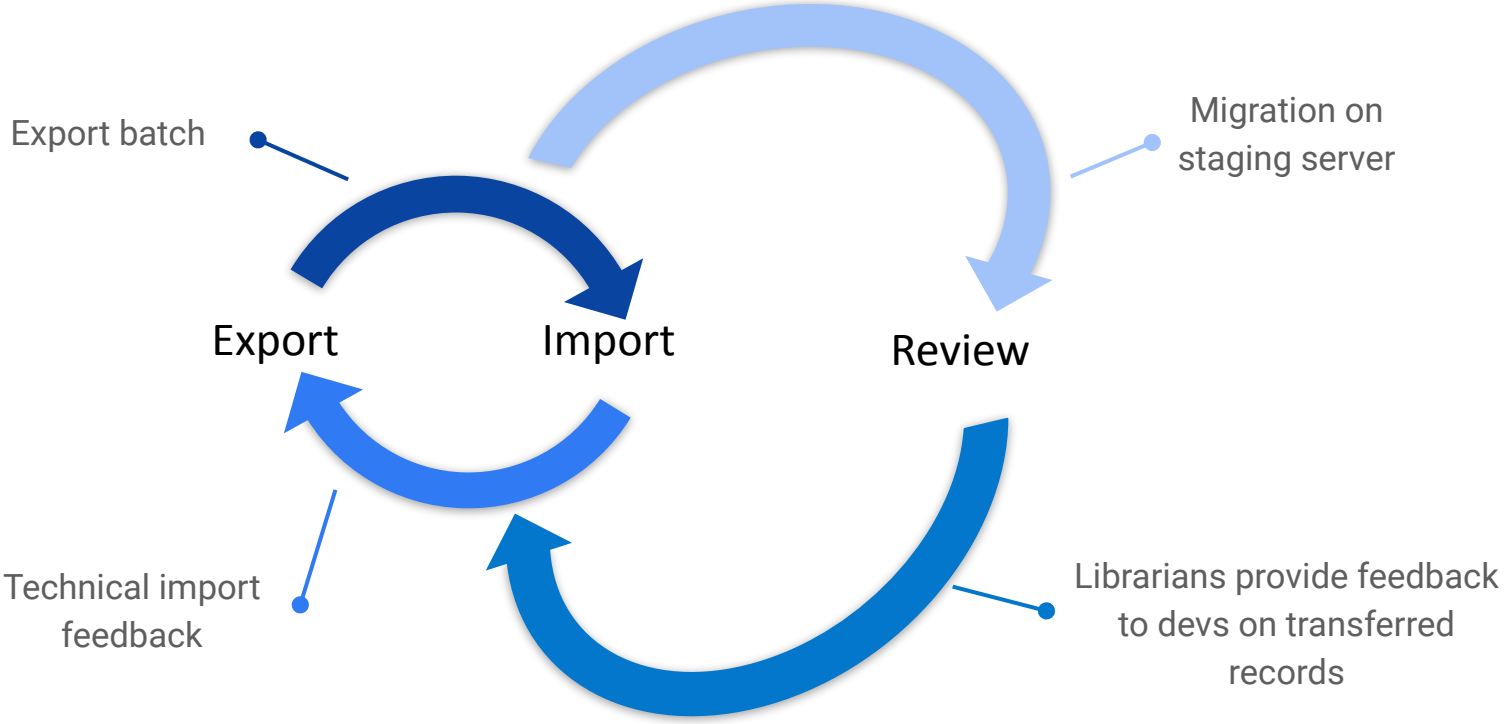
Export

DigitalHub

Review Records Feedback

Review	Collection	Title->metadata.title	Alternate Title(s)->metadata.additional_titles.tit	Resource type(s)->metadata.resource_type.title.en	Keyword->metadata.subjects.subject	Rights->metadata.rights.title.en	Creator->metadata.creators.person_or_org.nam
KG	Gutzman, Karen E. Enhancing the Visibility and Impact of Your Research. DigitalHub. Galter Health Sciences Library, 2014	OK	NA	OK	OK	OK	OK
KG	Galter Library Memorandum of Understanding	OK	NA	Form in Prism but "Other Published Works" in I	OK	OK	OK but no mention of Eileen as Depositor
KG	Mason, Michael L., 1895-1963. Michael Livingood Mason. DigitalHub. Galter Health Sciences Library & Learning Center, 1940	ok	na	ok	ok	ok	Not OK - Missing dates for "Mason, Michael L., 1895
KG	Bennett, Thomas (Medical educator). Letter to Dr. N.s. Davis. DigitalHub. Galter Health Sciences Library & Learning Center, 1867.	ok	na	ok	ok	ok	ok
KG	Intro to the NIH Public Access Policy. Video transcript - Part 1 of the NIH Public Access Policy video guides	ok	na	ok		ok	ok
	Gutzman, Karen E. Northwestern Elements Training. DigitalHub.						

Migration Summary: An Iterative Process



Lessons learned

Challenges

- Communicating field characteristics:
  - cardinality, substructure, meaning
- Working on migration as field structures and programmatic APIs were changing
- Details matter: docs don’t always reflect reality

Dealing with Challenges

- Fields: show on the interface
  - use a different medium than just words

- Tests allowed us to detect upstream changes
  - churn still present, but changes automatically detected

Lessons Learned… so far

- Tooling useful
- Programmatic API import did pay off
  - Same development effort
- Very iterative process
  - Don’t make perfect enemy of the good
    - “ [...] very clean system and the data is visually organized in a really approachable way. I also appreciated how responsive it was compared to DigitalHub

Current Status

Current Status

- Nearly all record metadata ingested with no issues; a few outliers
- Librarians reviewing metadata of migrated records in detail
- Challenges on the horizon
  - Performing the file import will slow feedback cycle
  - Most challenging: DOI transfer
    - Not possible yet with InvenioRDM
    - Upstream development will be needed

With thanks…

Teams

- The Invenio team @ CERN & RDM collaborators (here)
- Galter Health Sciences Library & Learning Center

Guillaume Viger
Sara Gonzales
Karen Gutzman
Josh Elder
Austin Sharp
Eric Newman
Mahonry Estrada
Gretchen Neidhardt
Matt Carson

Guillaume Viger

Sara Gonzales

Karen Gutzman

Josh Elder

Austin Sharp

Eric Newman

Mahonry Estrada

Gretchen Neidhardt

Matt Carson

Kristi Holmes

Support

Work presented here is supported in part by:

CERN Knowledge Transfer Fund

NUCATS: UL1TR001422 (NCATS)

all of the InvenioRDM project partners

Questions?