How to Curate Research Data

An 8-step guide with incentives to collaborate

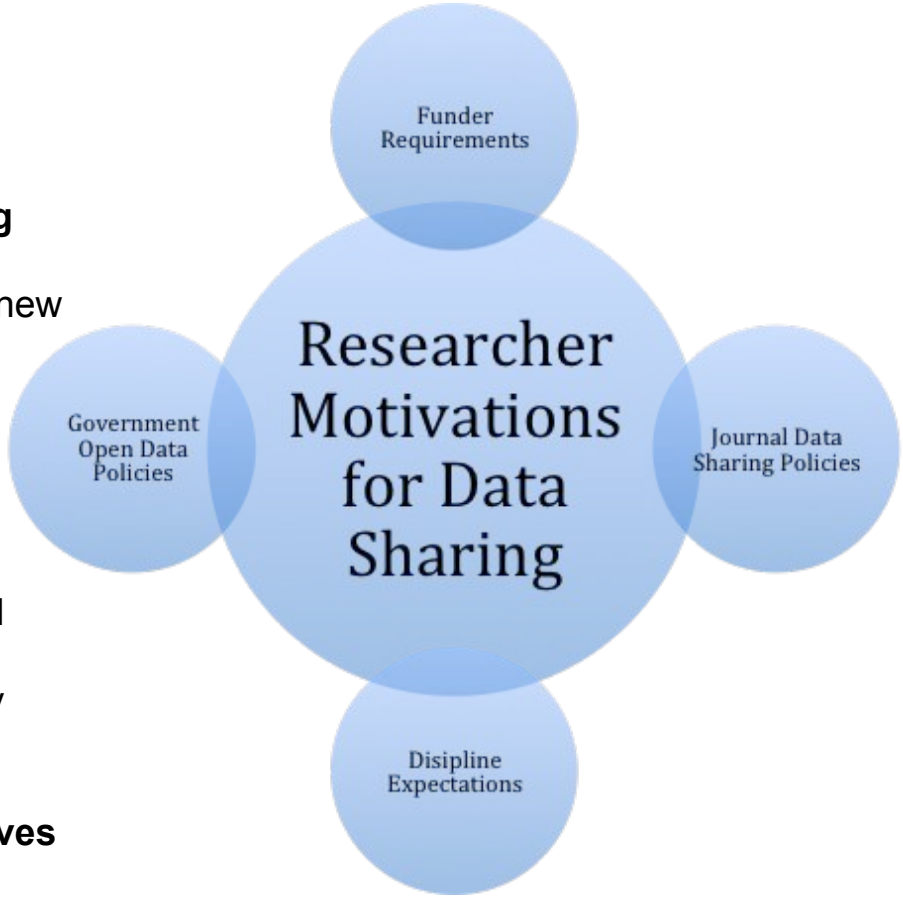
Lisa Johnston

Research Data Management/Curation Lead,
University of Minnesota

LIBRARIES
UNIVERSITY OF MINNESOTA

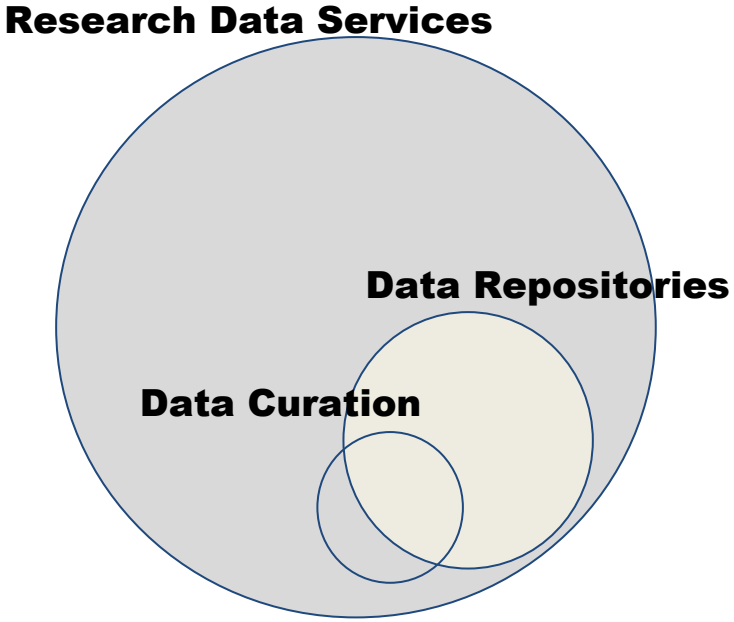
Drivers for data sharing

- **Funders (Federal/private) require data sharing**
  - Public access
  - Return on $$ investment ⇒ others can do new research
- **Journal data sharing policies**
  - Increase transparency
  - Facilitate reproducibility
- **Researcher/disciplinary culture shift in digital age**
  - Ease of sharing ⇒ culture of reproducibility
  - Citation impact, reputation building
- **(parallel effort) Government open data initiatives**
  - Democratize scientific knowledge/results
  - Release the potential of $$ data



LIBRARIES
UNIVERSITY OF MINNESOTA

Data curation is one part of research data services



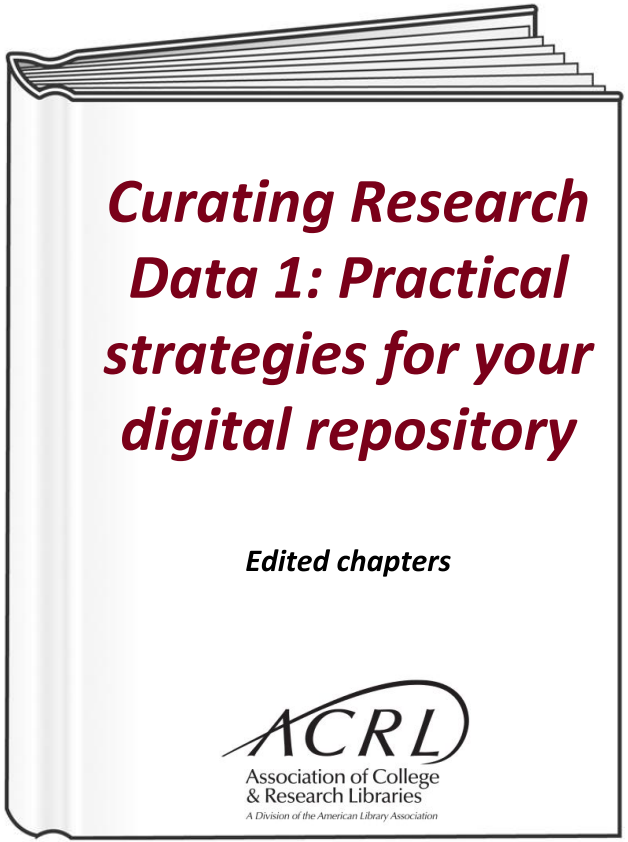
Goal of data curation ⇒ Prepare and securely store research data in ways that

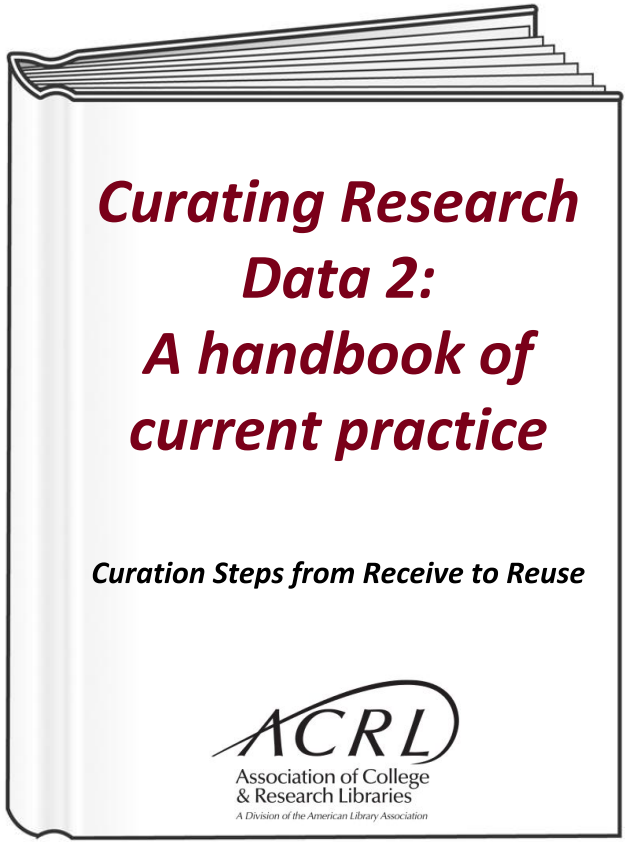
1. **make it useful beyond its original purpose**
2. **ensure completeness for validation and replication**
3. **facilitate long-term discoverability, access, and persistence**

Data curation steps include = *quality assurance*
file integrity checks
documentation review
metadata creation
file transformations
metadata brokerage….

Note: The RDA Data Foundations and Terminology working group has a growing dictionary of data related terms that is searchable at http://smw-rda.esc.rzg.mpg.de/index.php/Main_Page

LIBRARIES
UNIVERSITY OF MINNESOTA





(Forthcoming publications will be available from American Library Association (www.alastore.ala.org) and Amazon.com in November 2016.)

LIBRARIES
UNIVERSITY OF MINNESOTA

Step 0: Establish Your Data Curation Service

Curating Research Data: A handbook of current practice

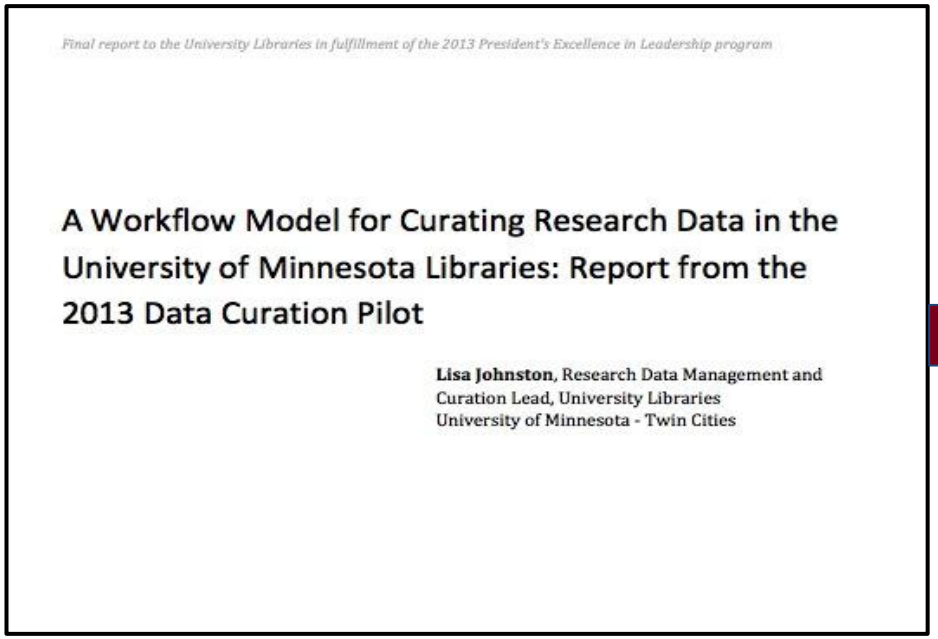
Citation: “Amish Barn Raising in Otsego County.” WBNG. http://media.wbng.com/images/600*394/DSCN7181.JPG.

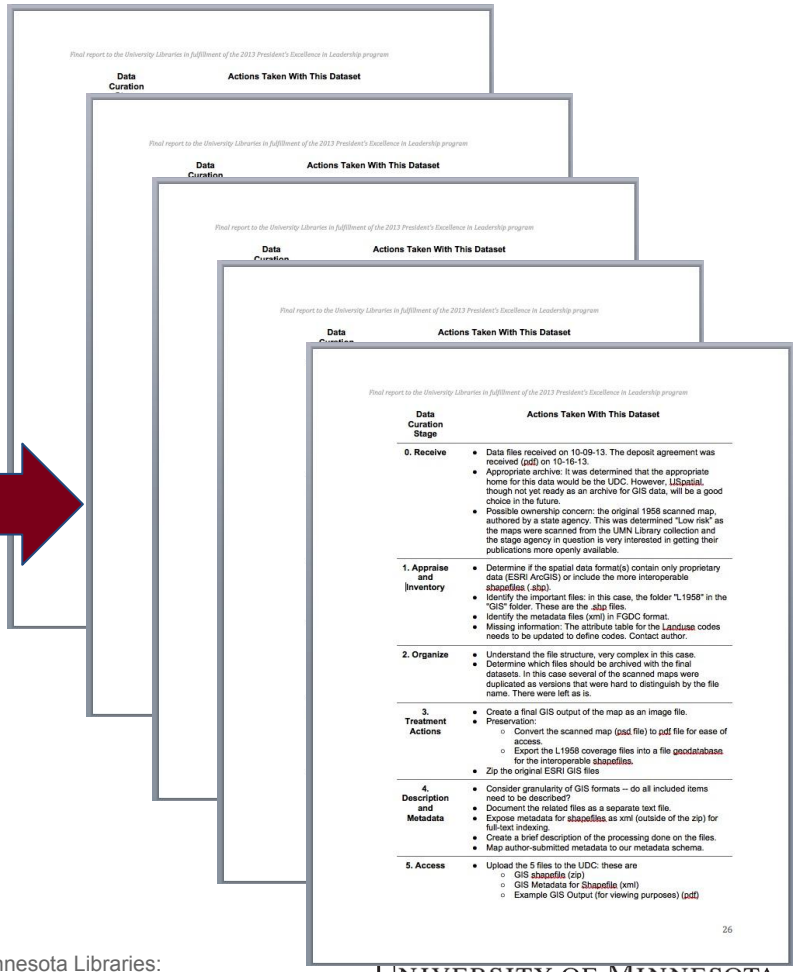
Sub Steps

- Define Mission and Scope
- Develop Policy and Procedure
- Identify Your Target Audience
- Understand the Costs
- Invest in Staff Resources
- Build/Acquire the Technological Infrastructure

LIBRARIES
UNIVERSITY OF MINNESOTA

Example from Preliminary Step 0





Citation: Johnston, Lisa R. (2014). A Workflow Model for Curating Research Data in the University of Minnesota Libraries: Report from the 2013 Data Curation Pilot. University Digital of Minnesota Conservancy. http://hdl.handle.net/11299/162338.

Example from Preliminary Step 0

University Digital Conservancy Home / University of Minnesota - Twin Cities / Data Repository for U of M (DRUM) / View Item

"Laundry Soap" from the Ojibwe Conversations Archives Project

Tainter, Rose; Kingbird-Porter, Margaret; Hermes, Mary (2014)

Persistent link to this item
http://dx.doi.org/10.13020/D6H596
http://purl.umn.edu/160534
http://hdl.handle.net/11299/160534

Services
Full Metadata (xml)
View Usage Statistics

Published Date
2013-11-22

Author Contact
Hermes, Mary (mhermes@umn.edu)

Type
Dataset
Video or Animation
Human Subjects Data

Abstract
Two first speakers of Ojibwe discuss and debate laundry soap in a video recorded in Hayward WI in March 2009.

License
Attribution-NonCommercial-ShareAlike 4.0 International License

Suggested Citation
Tainter, Rose; Kingbird-Porter, Margaret; Hermes, Mary. (2014). "Laundry Soap" from the Ojibwe Conversations Archives Project. Retrieved from the Data Repository for the University of Minnesota, http://dx.doi.org/10.13020/D6H596.

Show full item record

View/Download file

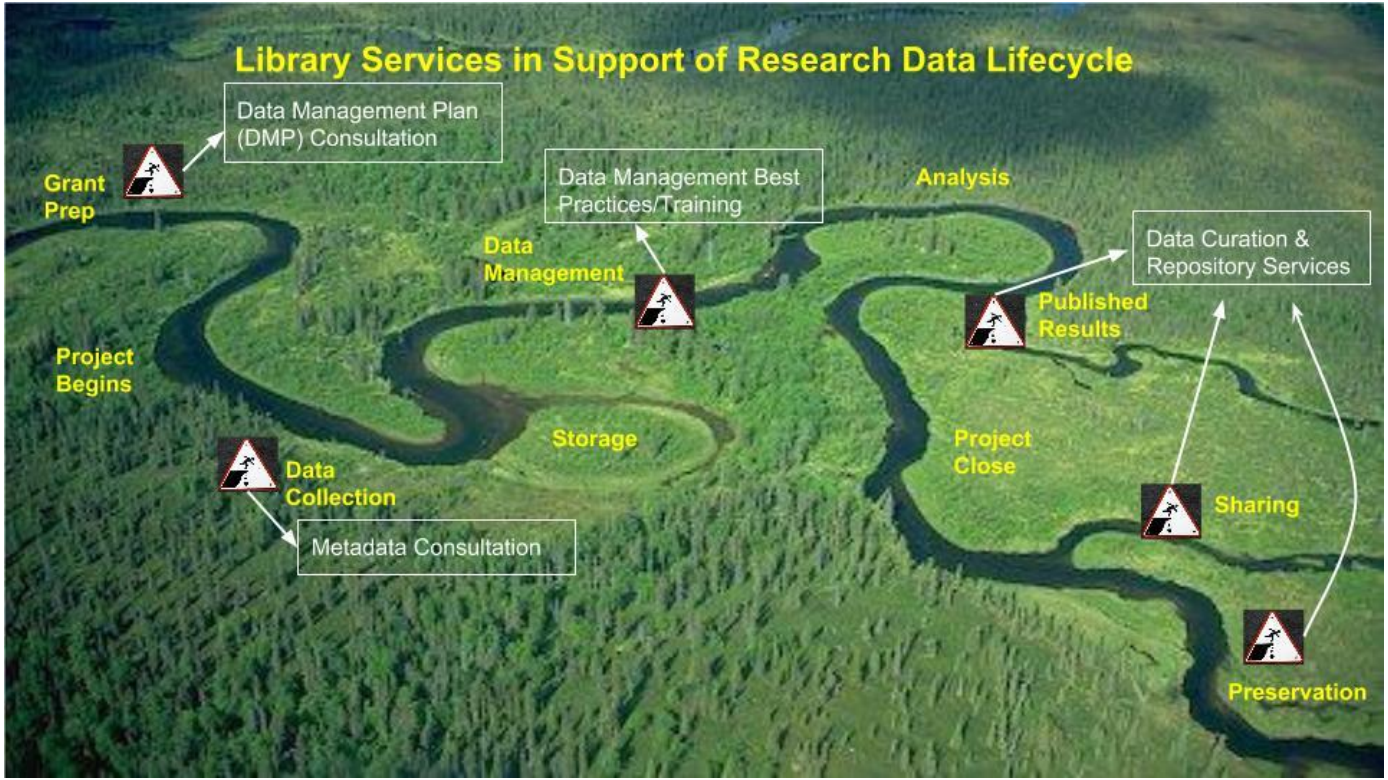
File View/Open	Description	Size	Format
archivedVersion.zip	Annotated (ELAN) Video Files	14.82Kb	application/zip

DSPACE

Citation: Tainter, Rose; Kingbird-Porter, Margaret; Hermes, Mary. (2014). "Laundry Soap" from the Ojibwe Conversations Archives Project. Retrieved from the Data Repository for the University of Minnesota, http://dx.doi.org/10.13020/D6H596.

LIBRARIES
UNIVERSITY OF MINNESOTA

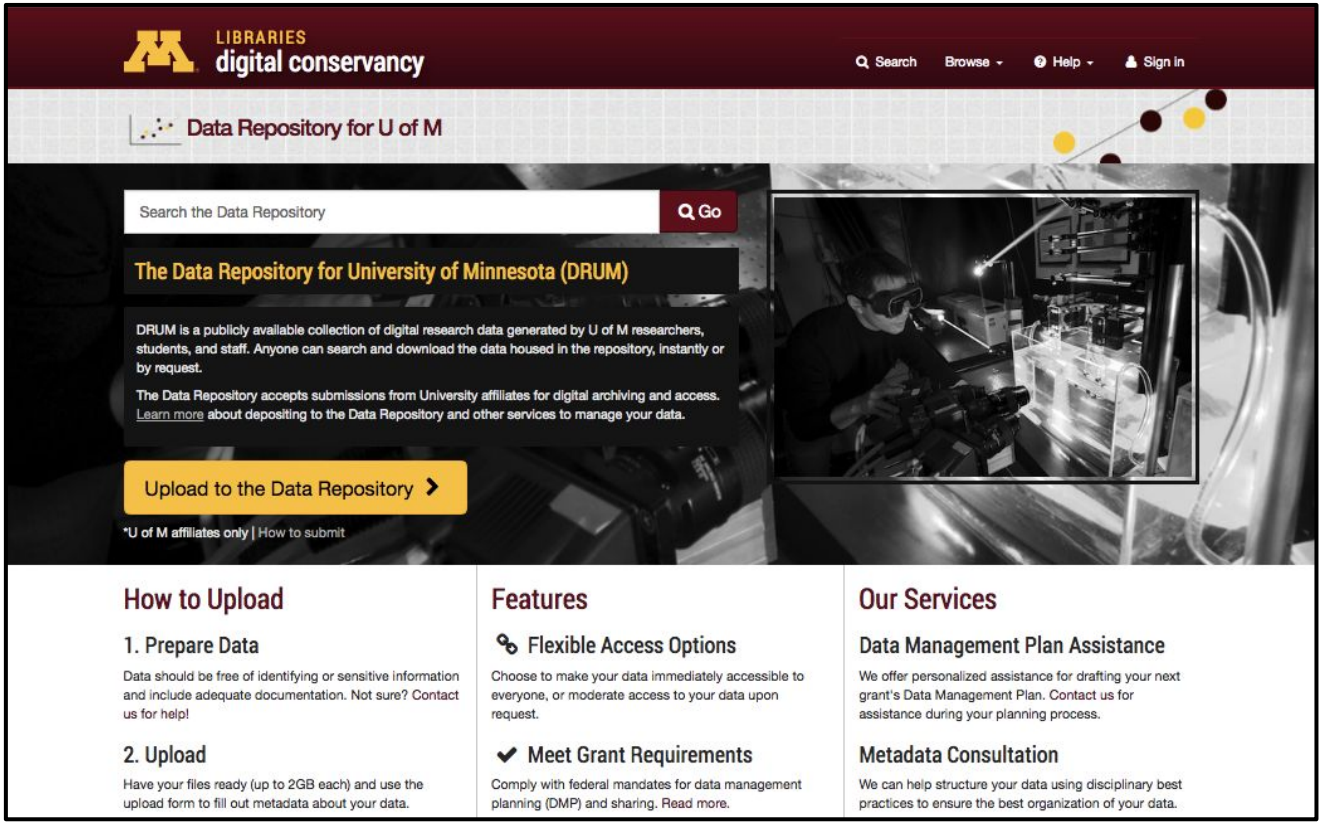
Launched new services across the research data life-cycle



Citation: “The Supporting Documentation for Implementing the Data Repository for the University of Minnesota (DRUM): A Business Model, Functional Requirements, and Metadata Schema” at http://hdl.handle.net/11299/171761.

LIBRARIES
UNIVERSITY OF MINNESOTA

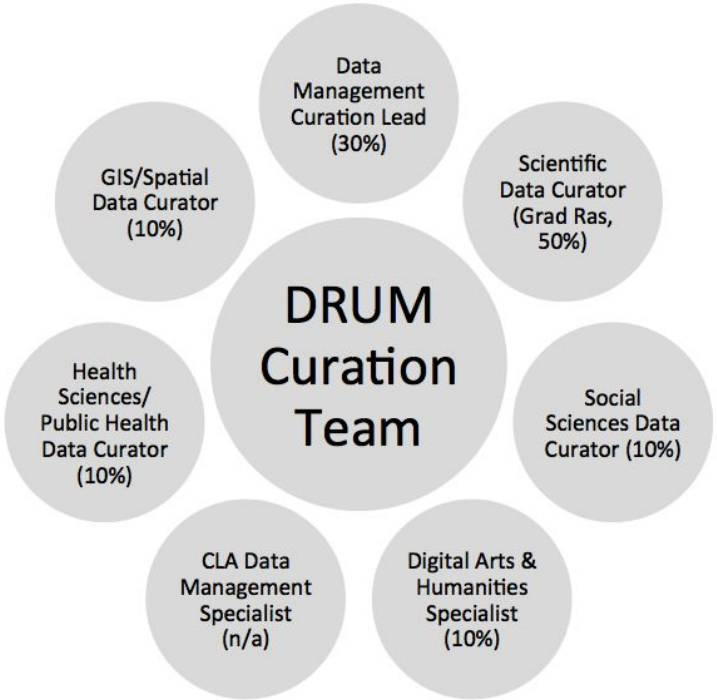
Data Repository for the University of Minnesota (DRUM)

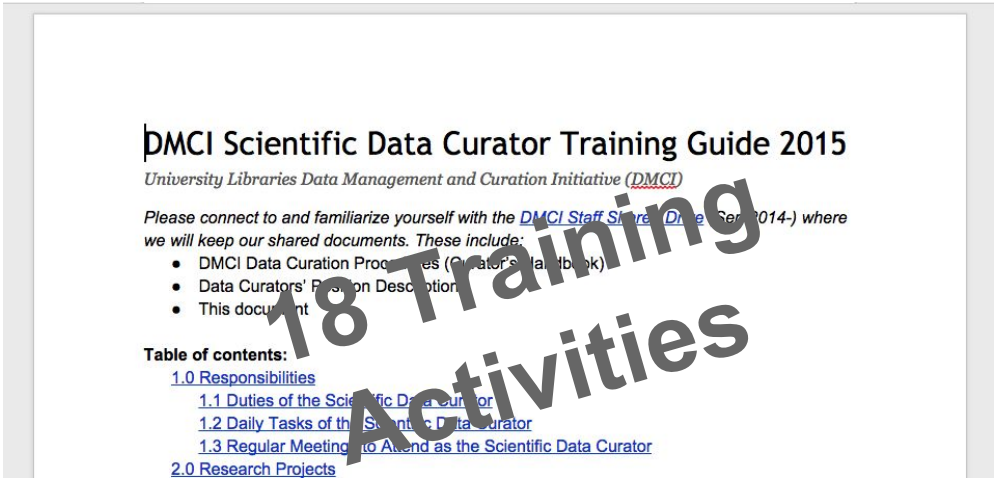


http://z.umn.edu/drum

LIBRARIES
UNIVERSITY OF MINNESOTA

DRUM Staffing Model




DMCI Scientific Data Curator Training Guide 2015

University Libraries Data Management and Curation Initiative (DMCI)

Please connect to and familiarize yourself with the DMCI Staff Share Drive (Sept 2014-) where we will keep our shared documents. These include:

- DMCI Data Curation Proc[illegible]es (Curator's Handbook)
- Data Curators' P[illegible]on Descriptions
- This docu[illegible]nt

Table of contents:

1.0 Responsibilities
1.1 Duties of the Scientific Data Curator
1.2 Daily Tasks of the Scientific Data Curator
1.3 Regular Meetings to Attend as the Scientific Data Curator
2.0 Research Projects

18 Training Activities

Model published: "The Supporting Documentation for Implementing the Data Repository for the University of Minnesota (DRUM): A Business Model, Functional Requirements, and Metadata Schema" at http://hdl.handle.net/11299/171761.

LIBRARIES
UNIVERSITY OF MINNESOTA

Step 1: Receive the Data

Curating Research Data: A handbook of current practice

Sub Steps

- Recruit Data for Your Service
- Negotiate Deposit
- Obtain Author Deposit Agreements
- Facilitate Transfer of the Data
- Obtain Metadata and Documentation
- Receive Notification of Data Arrival

Image: https://www.appointment-plus.com/images/blog/dock-employee-using-scheduling-software.jpg.

LIBRARIES
UNIVERSITY OF MINNESOTA

Example from Step 1: Receive Data

ICPSR
A PARTNER IN SOCIAL SCIENCE RESEARCH
INTER-UNIVERSITY CONSORTIUM FOR POLITICAL AND SOCIAL RESEARCH

Restricted Data Use Agreement
for Confidential Data from
the National Addiction and HIV Data Archive Program

I. Definitions

A. "Investigator" is the person primarily responsible for analysis and other use of Confidential Data obtained through this Agreement.

B. "Research Staff" are all persons at the Investigator's institution, excluding the Investigator, who will have access to Confidential Data obtained through this Agreement.

C. "Institution" is the university or research institution at which the Investigator will conduct research using Confidential Data obtained through this Agreement.

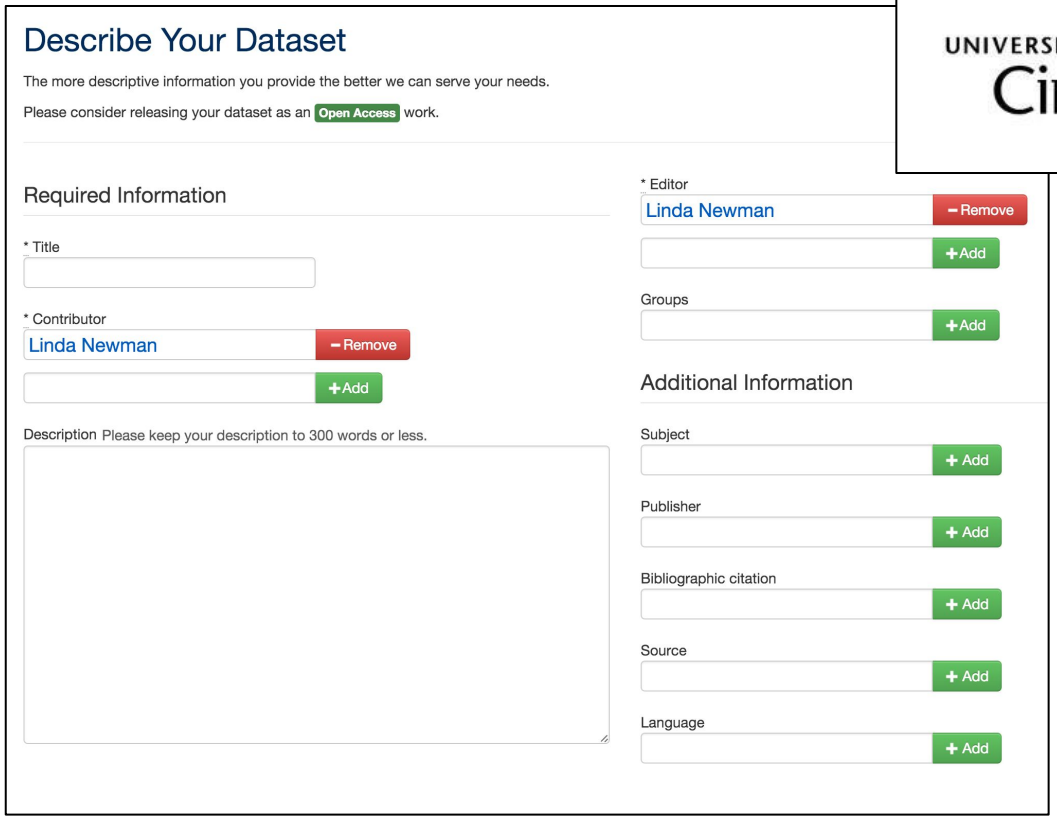
D. "Representative of the Institution" is a person authorized to enter into legal agreements on behalf of Investigator's Institution.

E. "Confidential Data" consist of identifiable private information, linkable to a specific individual

ONLY

Citation: Kaye Marz. "Case Study—Legal Agreements for Acquiring Restricted-Use Research Data" Curating Research Data Volume 2: A Handbook of current practice.

Example from Step 1: Receive Data



Citation: Amy Koshoffer, Carolyn Hansen, and Linda Newman. ”Case Study—Challenges with Quality of Data Set Metadata in a Self-Submission Repository Model.” Curating Research Data Volume 2: A Handbook of current practice.

Step 2: Appraise and Select

Curating Research Data: A handbook of current practice

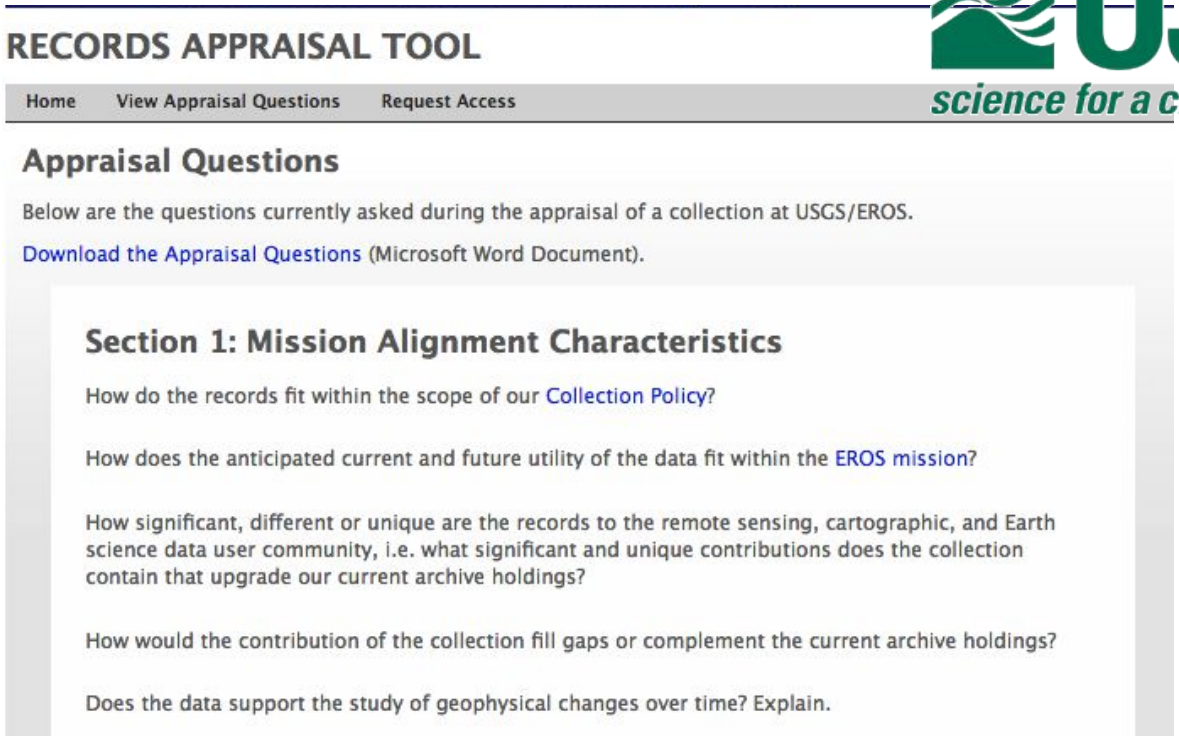
Sub Steps

- Appraise the Files
- Consider Any Risk Factors
- Inventory the Submission
- Select (or reject)
- Assign the Submission

Image: http://michaelhyatt.com/wp-content/uploads/2010/12/iStock_000004729175Small.jpg

LIBRARIES
UNIVERSITY OF MINNESOTA

Example from Step 2: Appraise and Select



RECORDS APPRAISAL TOOL

Home View Appraisal Questions Request Access

Appraisal Questions

Below are the questions currently asked during the appraisal of a collection at USGS/EROS.

Download the Appraisal Questions (Microsoft Word Document).

Section 1: Mission Alignment Characteristics

How do the records fit within the scope of our Collection Policy?

How does the anticipated current and future utility of the data fit within the EROS mission?

How significant, different or unique are the records to the remote sensing, cartographic, and Earth science data user community, i.e. what significant and unique contributions does the collection contain that upgrade our current archive holdings?

How would the contribution of the collection fill gaps or complement the current archive holdings?

Does the data support the study of geophysical changes over time? Explain.

USGS science for a changing world

Citation:John Faundeen. "Case Study—Scientific Records Appraisal Process: US Geological Survey." Curating Research Data Volume 2: A Handbook of current practice.

Step 3: Processing and Treatment Actions for Data

Curating Research Data: A handbook of current practice

Sub Steps

- Secure the Files
- Start a Curation Log
- Inspect the File Representation and Organization
- Inspect the Data
- Work with the Author to Enhance the Data Submission (readme.txt)
- Consider File Formats
- Arrangement and Description

Image: Thumbnail used by the Data Repository for the University of Minnesota (DRUM)

LIBRARIES
UNIVERSITY OF MINNESOTA

Examples from Example Step 3: Processing

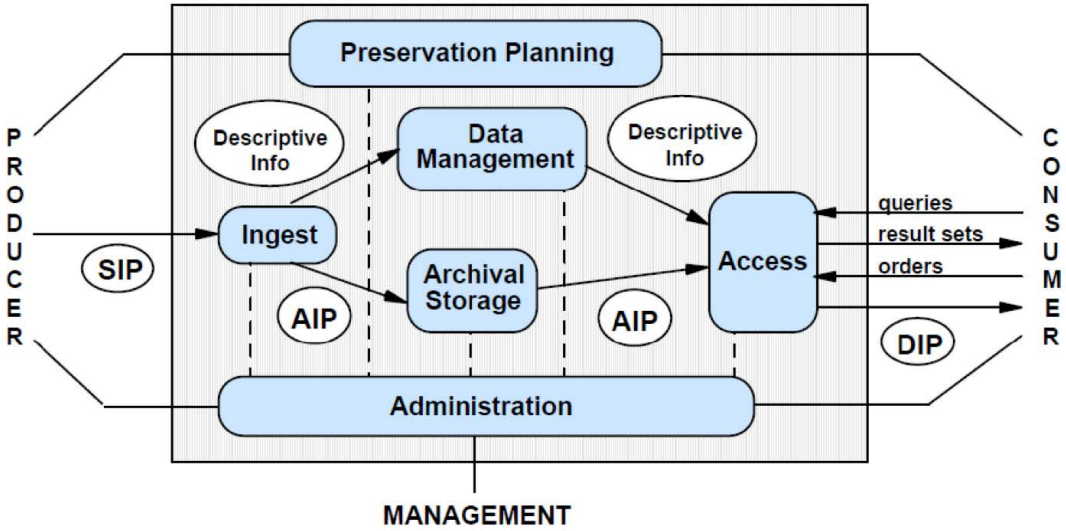
```
This readme.txt file was generated on <YYYYMMDD> by <Name>

GENERAL INFORMATION

1. Title of Dataset:
2. File Information:
   A. Filename:
   B. Short description:
   C. Filename:
   D. Short description
   E. Filename:
   F. Short description:
   G. If data set includes multiple files related to one another,
include relationship here:

3. Principal Investigator Contact Information
   A. Name:
   B. Institution:
   C. Address:
   D. Email:

4. Associate or Co-investigator Contact Information
   A. Name:
   B. Institution:
   C. Address:
   D. Email:
```

RStudio®

Citation: Readme.txt template. http://z.umn.edu/readme; “Case Study—Preserving 3D Data Sets: Workflows, Formats, and Considerations” by the Archaeology Data Service; “Case Study—Helpful Commands for Exporting Metadata from Statistical Software Packages SSPS, Stata, and R” by Alicia Hofelich Mohr both in Curating Research Data Volume 2: A Handbook of current practice.

Step 4: Ingest and Store Data in the Repository

Curating Research Data: A handbook of current practice



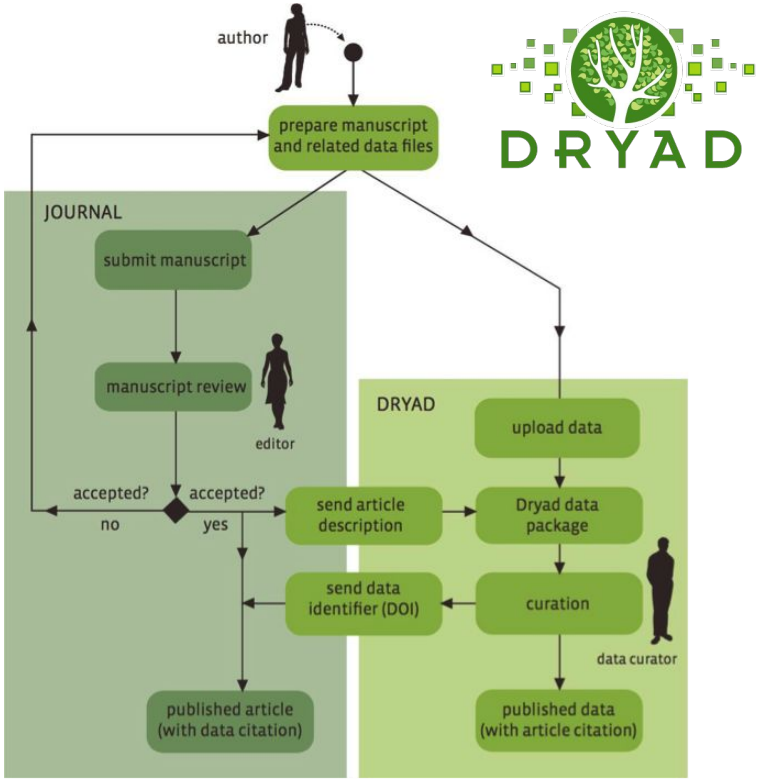
Sub Steps

- Ingest the Data Files
- Store the Assets Securely
- Develop Trust in Your Repository

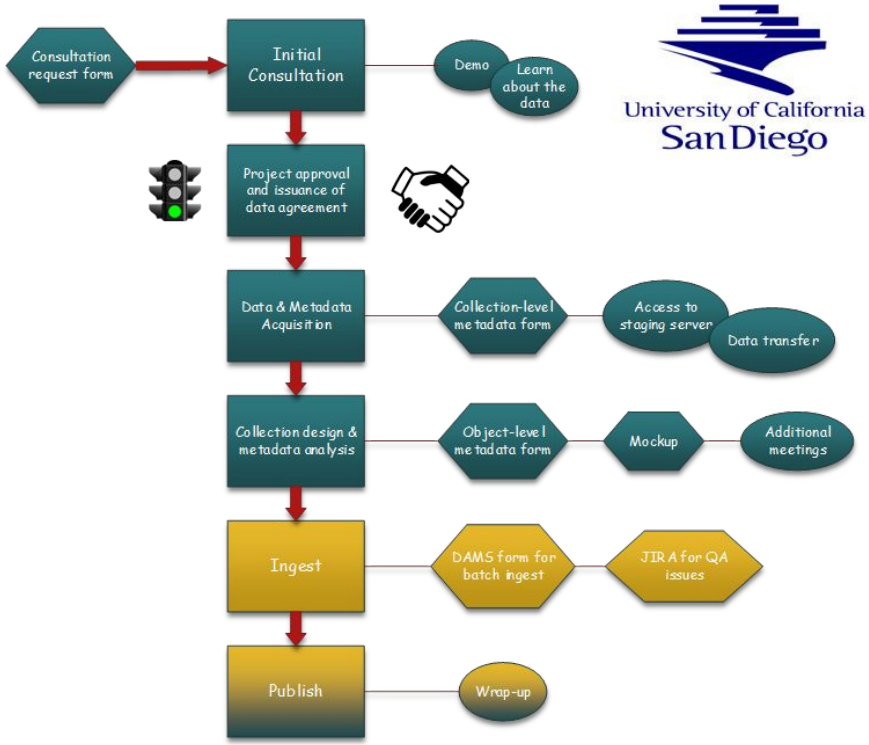
Image: CCSDS. "Reference Model for an Open Archival Information System (OAIS), Recommended Practice." CCSDS 650.0-M-2 (Magenta Book). Issue 2, June 2012. http://public.ccsds.org/publications/archive/650x0m2.pdf.

LIBRARIES
UNIVERSITY OF MINNESOTA

Examples from Step 4: Ingest and Store



Citation: Erin Clary and Debra Fagan."Case Study—Dryad Curation Workflows." Curating Research Data Volume 2: A Handbook of current practice.



Citation: Juliane Schneider, Arwen Hutt, and Ho Jung Yoo. "Case Study—Standardization and Automation of Ingest Processes in a Fully Mediated Deposit Model." Curating Research Data Volume 2: A Handbook of current practice.

Step 5: Descriptive Metadata

Curating Research Data: A handbook of current practice

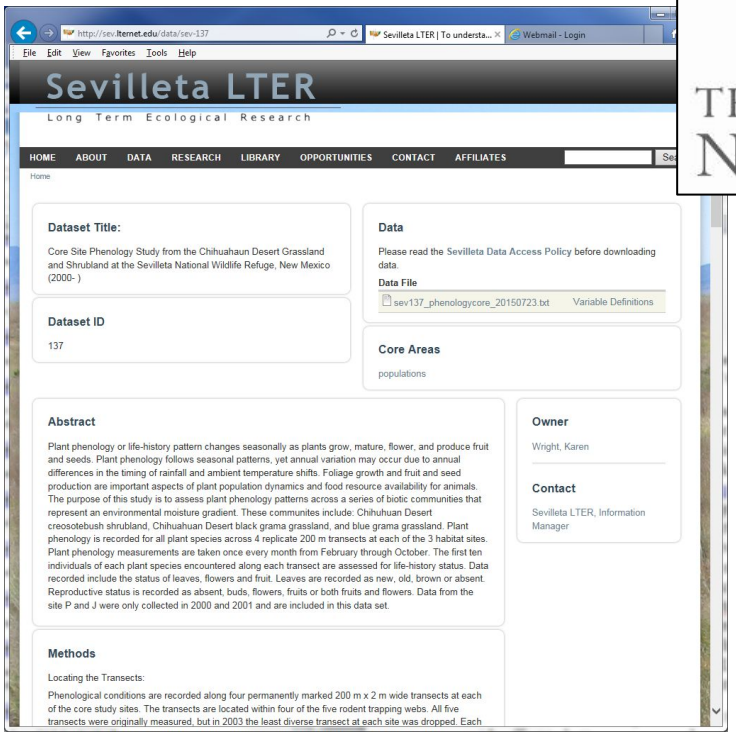
Sub Steps

- Create and Apply Descriptive Metadata
- Consider Metadata Standards for Disciplinary Data

Image: foggyray90. “Infinite Regress - A man paints himself painting himself.” flicker. https://c1.staticflickr.com/9/8566/16499327408_68d2b97d79_b.jpg.

LIBRARIES
UNIVERSITY OF MINNESOTA

Example from Step 5: Descriptive Metadata



THE UNIVERSITY of NEW MEXICO

Citation: Jon Wheeler, Mark Servilla, and Kristin Vanderbilt. “Case Study—Beyond Discovery: Cross-Platform Application of Ecological Metadata Language in Support of Quality Assurance and Control.” Curating Research Data Volume 2: A Handbook of current practice.

Step 6: Access

Curating Research Data: A handbook of current practice

Sub Steps

- Determine Appropriate Access Conditions
- Apply the Terms of Use and Any Relevant Licenses and Copyrights for the Data
- Contextualize the Data
- Enhance the Submission to Increase Exposure and Discovery
- Apply Any Necessary Access Controls
- Ensure Persistent Access (e.g., DOIs)
- Release Data for Access and Notify Author

Image: Wikimedia Commons: "HK PolyU Hung Hom Bay Campus 8 Hung Lok Road HKCC Library entrance gates Mar-2013.JPG."

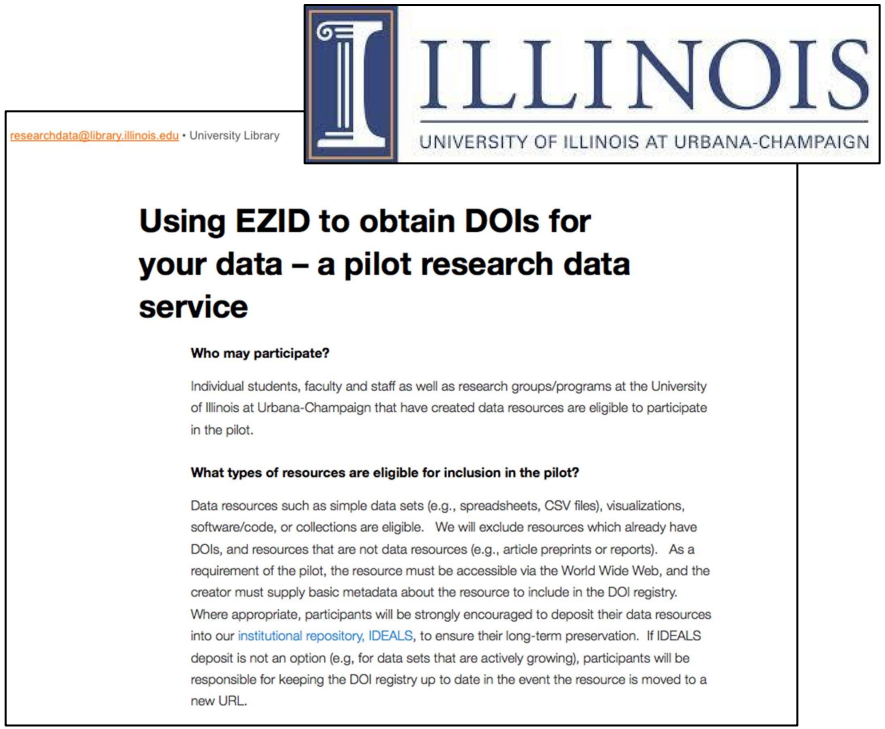
LIBRARIES
UNIVERSITY OF MINNESOTA

Example from Step 6: Access

EZID

Long-term identifiers made easy

A SERVICE OF THE UNIVERSITY OF CALIFORNIA CURATION CENTER — UC3



researchdata@library.illinois.edu • University Library

ILLINOIS

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Using EZID to obtain DOIs for your data – a pilot research data service

Who may participate?

Individual students, faculty and staff as well as research groups/programs at the University of Illinois at Urbana-Champaign that have created data resources are eligible to participate in the pilot.

What types of resources are eligible for inclusion in the pilot?

Data resources such as simple data sets (e.g., spreadsheets, CSV files), visualizations, software/code, or collections are eligible. We will exclude resources which already have DOIs, and resources that are not data resources (e.g., article preprints or reports). As a requirement of the pilot, the resource must be accessible via the World Wide Web, and the creator must supply basic metadata about the resource to include in the DOI registry. Where appropriate, participants will be strongly encouraged to deposit their data resources into our institutional repository, IDEALS, to ensure their long-term preservation. If IDEALS deposit is not an option (e.g, for data sets that are actively growing), participants will be responsible for keeping the DOI registry up to date in the event the resource is moved to a new URL.

Citation: Susan M. Braxton, Bethany Anderson, Margaret H. Burnette, Thomas G. Habing, William H. Mischo, Sarah L. Shreeves, Sarah C. Williams, and Heidi J. Imker. “Case Study—A Participant Agreement for Minting DOIs for Data Not in a Repository.” Curating Research Data Volume 2: A Handbook of current practice.

Step 7: Preservation for the Long Term

Curating Research Data: A handbook of current practice

Sub Steps

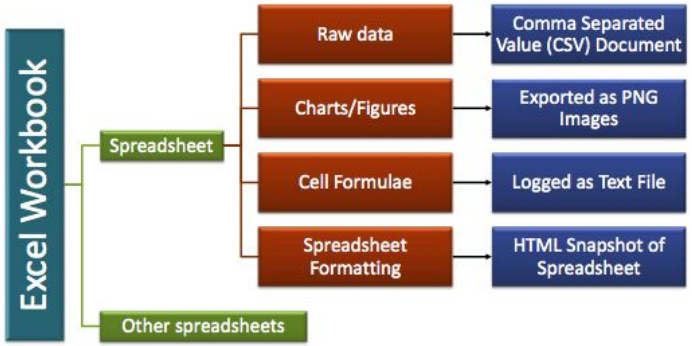
- Plan for Long-Term Reuse
- Monitor Preservation Needs and Take Action

Image: Wikicommons https://commons.wikimedia.org/wiki/File:NORADCommandCenter.jpg.

LIBRARIES
UNIVERSITY OF MINNESOTA

Example from Step 7: Preservation

Free tool: **Excel Archival Tool**
https://github.com/mcgrory/ExcelArchivalTool

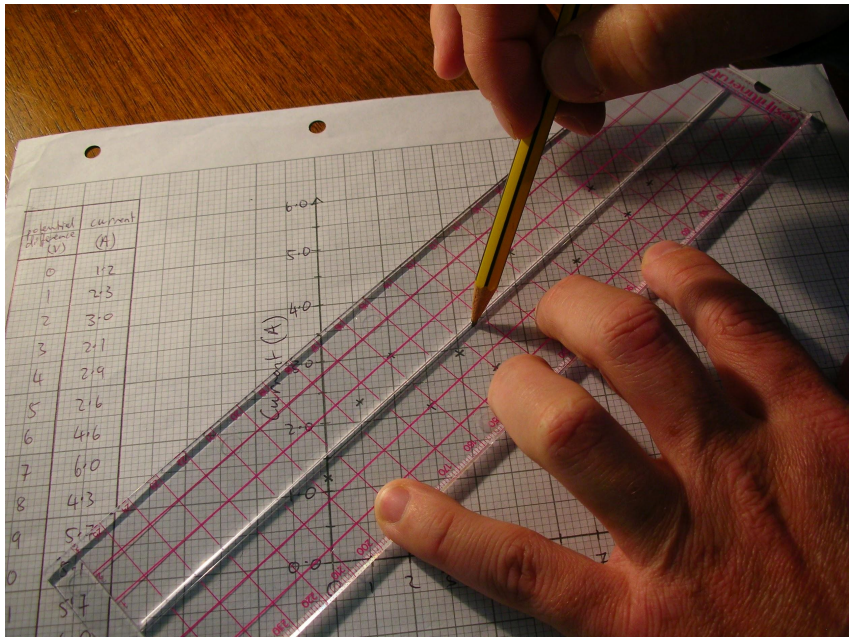


File View/Open	Description
Data For Gamma-toxin.xlsx	Experiment Data Readings
Archived_Data_for_Gamma_Toxin.zip	Archival Version of Data for Gamma-toxin.xlsx
Combined Figures.pzf	Data Analysis and Figures in Prism
Combined_Figures.xml	Archival Version of Combined Figures.pzf

Citation: McGrory, John. (2015). Poster for "Excel Archival Tool: Automating the Spreadsheet Conversion Process". Retrieved from the University of Minnesota Digital Conservancy, http://hdl.handle.net/11299/171966.

Step 8: Reuse

Curating Research Data: A handbook of current practice



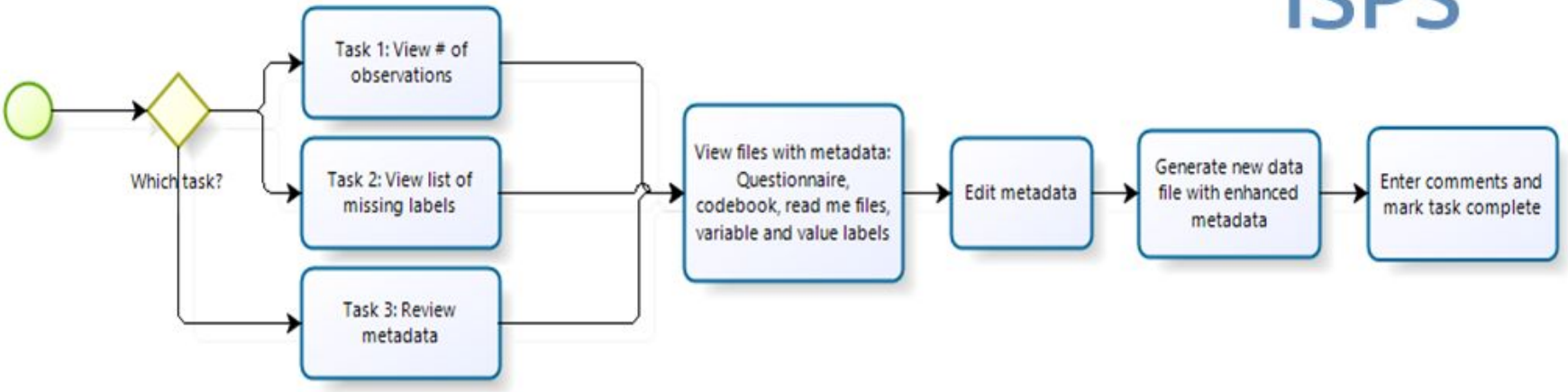
Sub Steps

- Monitor Data Rese
- Consider Post-Publication Review Techniques
- Provide Ongoing Support as Long as Necessary
- Cease Data Curation

Image: http://my.bestfitlineruler.com/wp-content/uploads/2009/05/drawing-the-bfl1.jpg

LIBRARIES
UNIVERSITY OF MINNESOTA

Example from Step 8: Reuse



Citation: Limor Peer. “Case Study—Enabling Scientific Reproducibility with Data Curation and Code Review.” Curating Research Data Volume 2: A Handbook of current practice.

Data Curation ⇒ How to scale in an IR setting?

Collaboration is key

Multiple data curation experts are needed to effectively curate the diverse data types an institutional repository typically receives.

Data curation expertise needed:

- File format-- GIS, spreadsheet/tabular, statistical/survey, video/audio, computer code
- Discipline-specific-- genomic sequence, chemical spectra, biological image
- Frequency-- Centers of excellence, departmental focus

Building the Data Curation Network

The Data Curation Network will enable academic institutions to better support researchers that are faced with a growing number of requirements to ethically share their research data.

We will

Phase 1: Develop a plan for implementing a “network of expertise” model for data curation staff across institutions
- Includes the projected staffing, costs, skills sets, and demand necessary for implementation

Phase 2: Pilot the model across our six institutions

Phase 3: Grow and sustain the Network beyond original institutions

Data Curation Network Partners

UNIVERSITY OF MINNESOTA

Washington University in St. Louis

PennState

Cornell University

UNIVERSITY OF MICHIGAN

ILLINOIS

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

The Data Curation Network project is supported by a generous grant from the ALFRED P. SLOAN FOUNDATION.

Our Phase 1 objectives

- **Underway** → Monitor the demand for curation services at each of our institutions. Our baseline report now available on our website.

- **Fall 2016** → Seek input from researchers to better understand how data curation services fit into their research workflow and data management needs through informal engagement activities held in parallel on each of our campuses.

- **Future** → Pilot curation workflows, survey curation staff, and establish metrics for how to assess the impact of curated data vs non curated.

Follow our progress!

https://sites.google.com/site/DataCurationNetwork

#DataCurationNetwork

DCN Project Team: Lisa R. Johnston (PI), Jake Carlson, Cynthia Hudson--Vitale, Heidi Imker, Wendy Kozlowski, Rob Olendorf, and Claire Stewart