

How to Curate Research Data

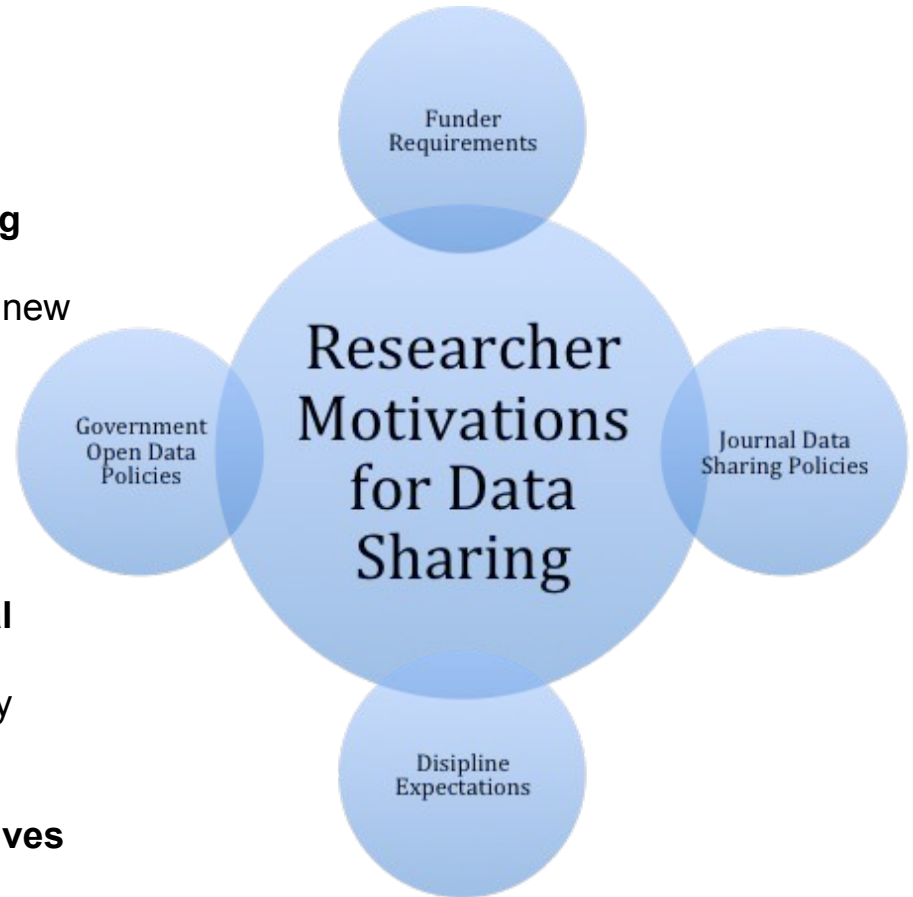
An 8-step guide with incentives to collaborate

Lisa Johnston

Research Data Management/Curation Lead,
University of Minnesota

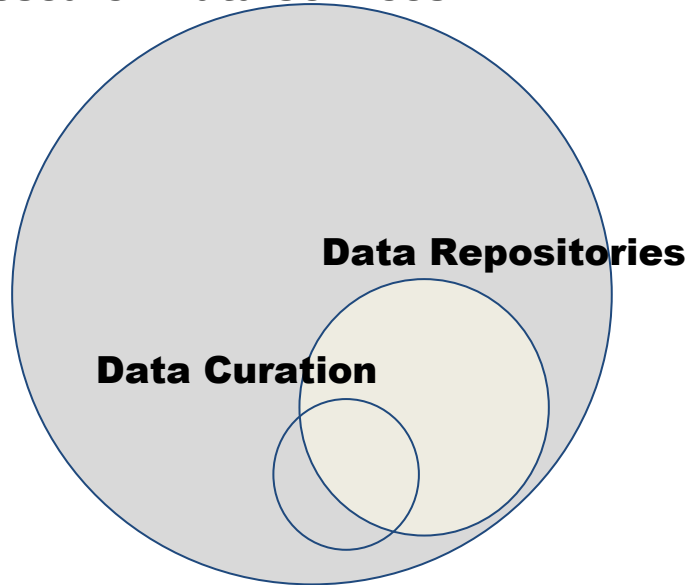
Drivers for data sharing

- **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



Data curation is one part of research data services

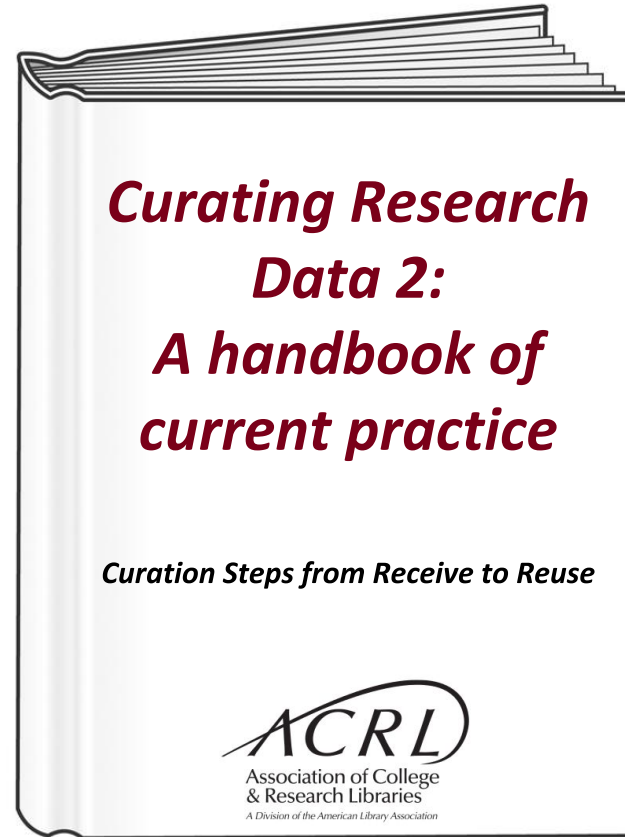
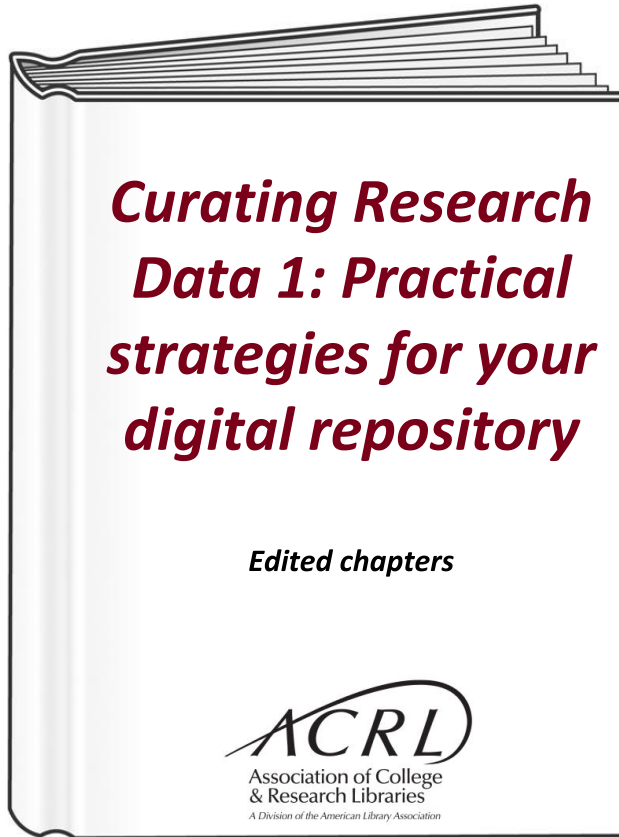
Research Data Services



Goal of data curation \Rightarrow 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....



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

Step 0: Establish Your Data Curation Service

Curating Research Data: A handbook of current practice



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

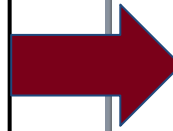
Citation: "Amish Barn Raising in Otsego County." WBNG. http://media.wbng.com/images/600*394/DSCN7181.JPG.

Example from Preliminary Step 0

Final report to the University Libraries in fulfillment of the 2013 President's Excellence in Leadership program

A Workflow Model for Curating Research Data in the University of Minnesota Libraries: Report from the 2013 Data Curation Pilot

Lisa Johnston, Research Data Management and Curation Lead, University Libraries
University of Minnesota - Twin Cities



Final report to the University Libraries in fulfillment of the 2013 President's Excellence in Leadership program

Data Curation

Actions Taken With This Dataset

Final report to the University Libraries in fulfillment of the 2013 President's Excellence in Leadership program

Data Curation

Actions Taken With This Dataset

Final report to the University Libraries in fulfillment of the 2013 President's Excellence in Leadership program

Data Curation

Actions Taken With This Dataset

Final report to the University Libraries in fulfillment of the 2013 President's Excellence in Leadership program

Data Curation

Actions Taken With This Dataset

Final report to the University Libraries in fulfillment of the 2013 President's Excellence in Leadership program

Data Curation Stage

Actions Taken With This Dataset


- | Data Curation Stage | Actions Taken With This Dataset |
|------------------------------------|---|
| 0. Receive | <ul style="list-style-type: none">Data files received on 10-09-13. The deposit agreement was received (pdf) on 10-16-13.Appropriate archive: It was determined that the appropriate home for this data would be the UDC. However, LISpatial, though not yet ready as an archive for GIS data, will be a good choice in the future.Possible ownership concern: the original 1958 scanned map, authored by a state agency. This was determined "Low risk" as the maps were scanned from the UMN Library collection and the state agency in question is very interested in getting their publications more openly available. |
| 1. Appraise and Inventory | <ul style="list-style-type: none">Determine if the spatial data format(s) contain only proprietary data (ESRI ArcGIS) or include the more interoperable shapfiles (.shp).Identify the important files: in this case, the folder "L1958" in the "GIS" folder. These are the .shp files.Identify the metadata files (.xml) in FGDC format.Missing information: The attribute table for the Lapanduse codes needs to be updated to define codes. Contact author. |
| 2. Organize | <ul style="list-style-type: none">Understand the file structure, very complex in this case.Determine which files should be archived with the final datasets. In this case several of the scanned maps were duplicated as versions that were hard to distinguish by the file name. There were left as is. |
| 3. Treatment Actions | <ul style="list-style-type: none">Create a final GIS output of the map as an image file.Preservation:<ul style="list-style-type: none">Convert the scanned map (.shp file) to pdf file for ease of access.Export the L1958 coverage files into a file geodatabase for the interoperable shapfiles.Zip the original ESRI GIS files. |
| 4. Description and Metadata | <ul style="list-style-type: none">Consider granularity of GIS formats – do all included items need to be described?Document the related files as a separate text file.Expose metadata for shapfiles as xml (outside of the zip) for full-text indexing.Create a brief description of the processing done on the files.Map author-submitted metadata to our metadata schema. |
| 5. Access | <ul style="list-style-type: none">Upload the 5 files to the UDC: these are<ul style="list-style-type: none">GIS shapfile (.zip)GIS Metadata for Shapfile (.xml)Example GIS Output (for viewing purposes) (pdf) |

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)



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](#)

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

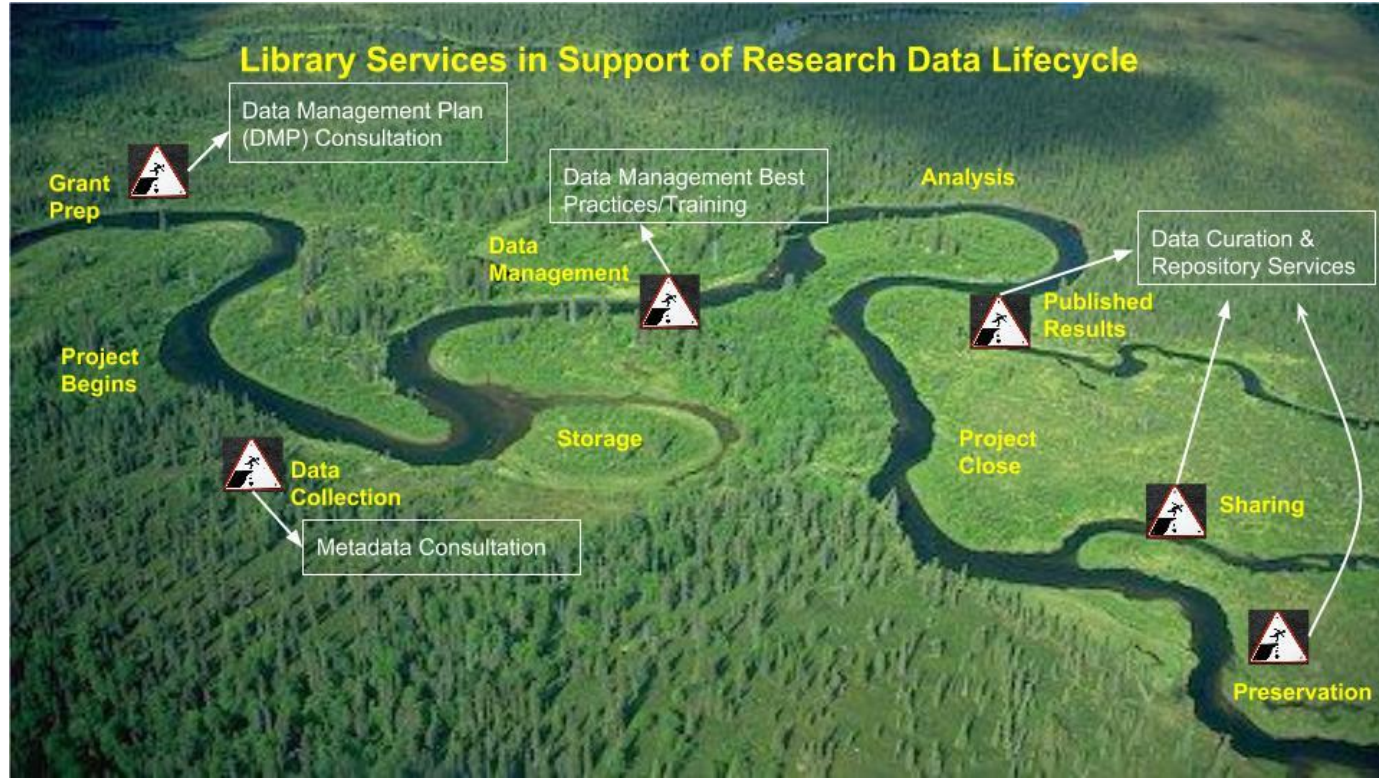
Services
[Full Metadata \(xml\)](#)
[View Usage Statistics](#)

View/Download file

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



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>.

Data Repository for the University of Minnesota (DRUM)

The screenshot shows the DRUM website interface. At the top, there is a dark red header with the University of Minnesota Libraries Digital Conservancy logo and navigation links for Search, Browse, Help, and Sign In. Below the header, a breadcrumb trail reads "Data Repository for U of M". A search bar with a "Go" button is present. The main content area features a dark background with a search bar, a description of DRUM, and an "Upload to the Data Repository" button. A photograph of a person using a microscope is visible on the right. Below this, three columns describe "How to Upload", "Features", and "Our Services".

LIBRARIES digital conservancy

Search Browse Help Sign In

Data Repository for U of M

Search the Data Repository **Go**

The Data Repository for University of Minnesota (DRUM)

DRUM is a publicly available collection of digital research data generated by U of M researchers, students, and staff. Anyone can search and download the data housed in the repository, instantly or by request.

The Data Repository accepts submissions from University affiliates for digital archiving and access. [Learn more](#) about depositing to the Data Repository and other services to manage your data.

Upload to the Data Repository >

*U of M affiliates only | [How to submit](#)

How to Upload

- 1. Prepare Data**
Data should be free of identifying or sensitive information and include adequate documentation. Not sure? Contact us for help!
- 2. Upload**
Have your files ready (up to 2GB each) and use the upload form to fill out metadata about your data.

Features

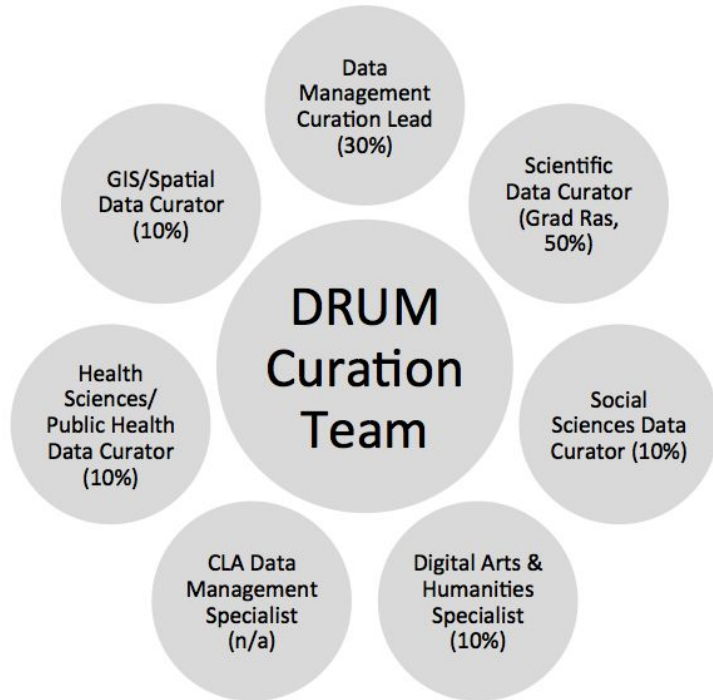
- Flexible Access Options**
Choose to make your data immediately accessible to everyone, or moderate access to your data upon request.
- Meet Grant Requirements**
Comply with federal mandates for data management planning (DMP) and sharing. [Read more.](#)

Our Services

- Data Management Plan Assistance**
We offer personalized assistance for drafting your next grant's Data Management Plan. Contact us for assistance during your planning process.
- Metadata Consultation**
We can help structure your data using disciplinary best practices to ensure the best organization of your data.

<http://z.umn.edu/drum>

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 Shared Drive](#) (GPO 1014-) where we will keep our shared documents. These include:

- DMCI Data Curation Procedures (Curator's Manual book)
- Data Curators' Position Description
- This document

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](#)

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

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

Example from Step 1: Receive Data



Describe Your Dataset

The more descriptive information you provide the better we can serve your needs.

Please consider releasing your dataset as an **Open Access** work.

Required Information

* Title

* Contributor
 - Remove
 + Add

Description Please keep your description to 300 words or less.

* Editor
 - Remove
 + Add

Groups
 + Add

Additional Information

Subject
 + Add

Publisher
 + Add

Bibliographic citation
 + Add

Source
 + Add

Language
 + Add

Step 2: Appraise and Select

Curating Research Data: A handbook of current practice



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

Sub Steps

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

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.

Step 3: Processing and Treatment Actions for Data

Curating Research Data: A handbook of current practice

Submission
under
curatorial
review

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

Examples from Example Step 3: Processing

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

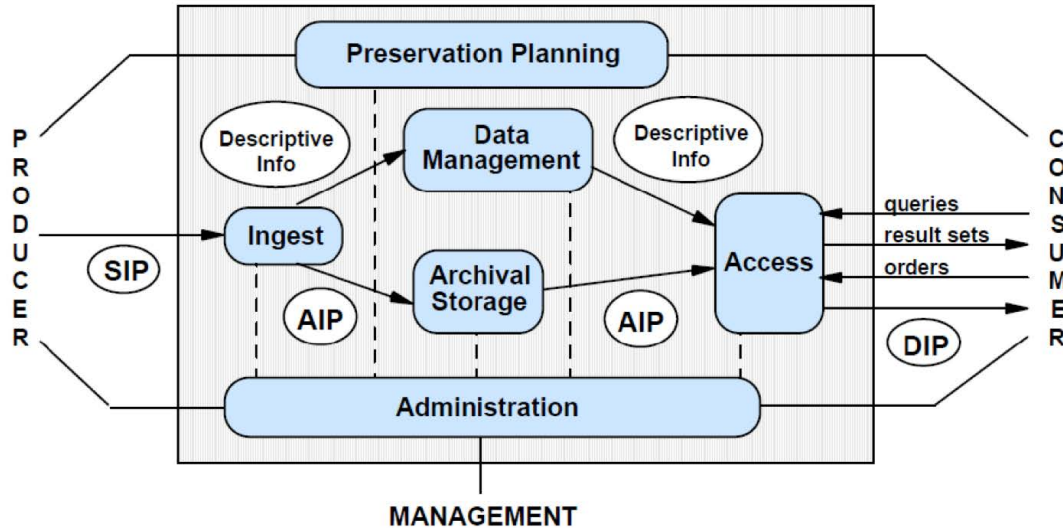
GENERAL INFORMATION

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:



Step 4: Ingest and Store Data in the Repository

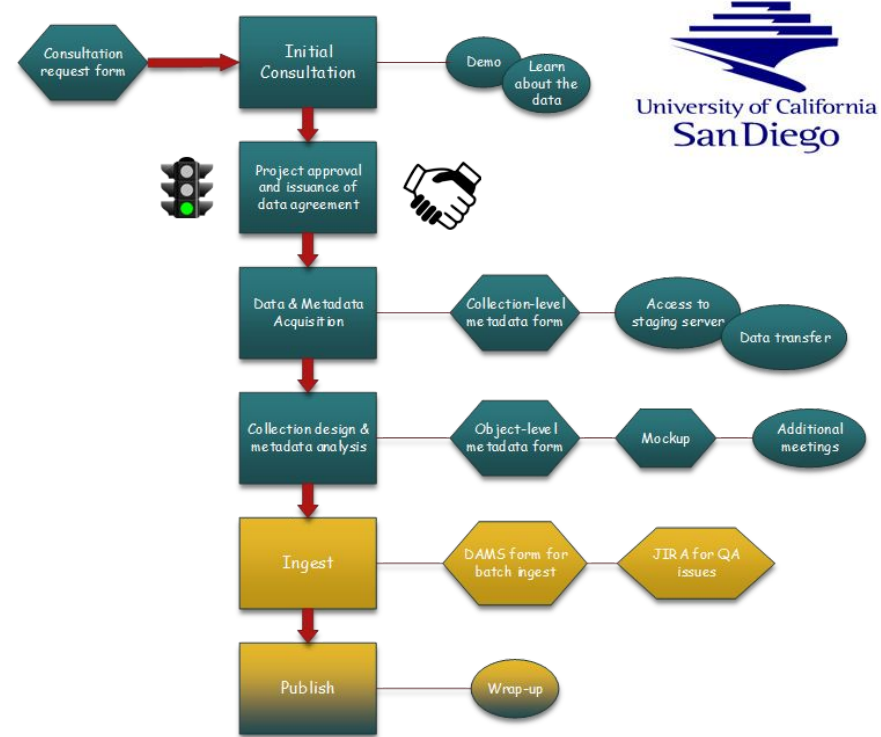
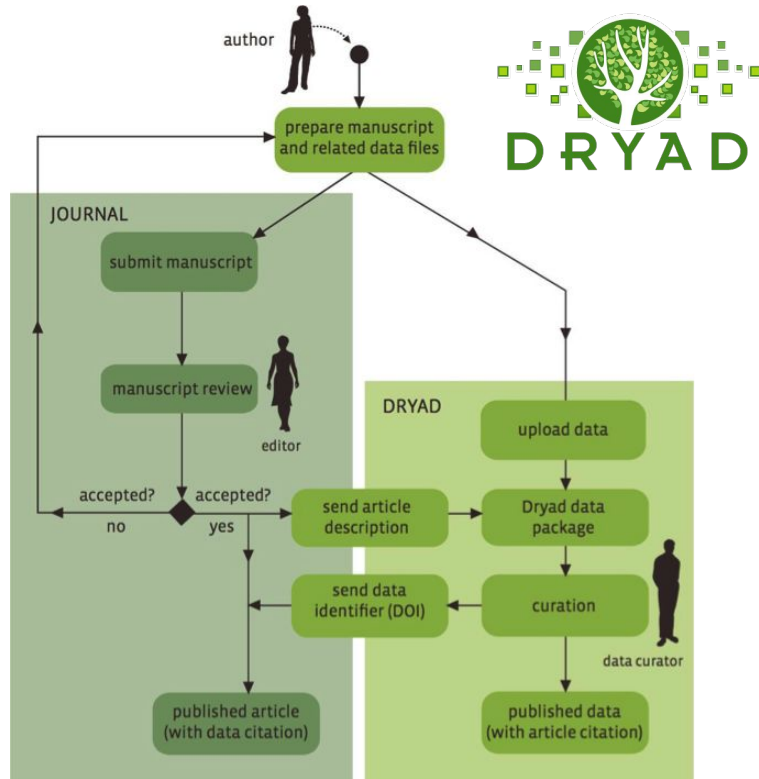
Curating Research Data: A handbook of current practice



Sub Steps

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

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

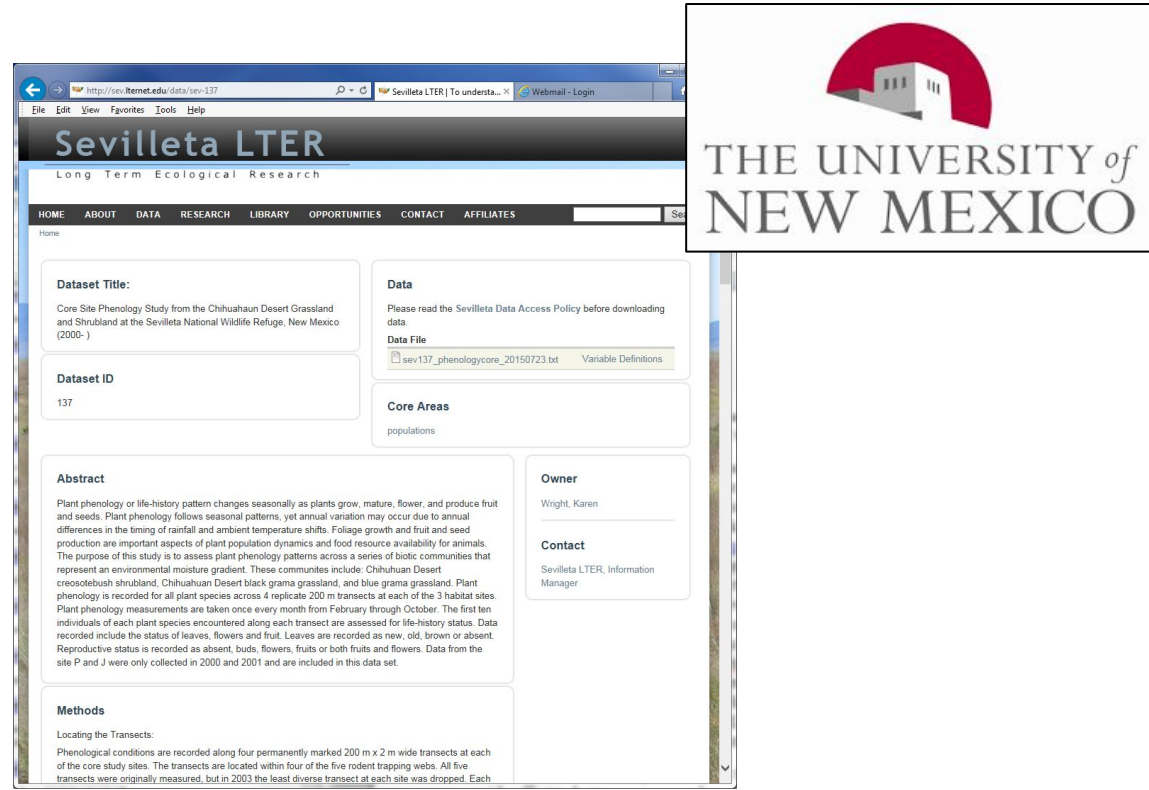
Curating Research Data: A handbook of current practice



Sub Steps

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

Example from Step 5: Descriptive Metadata



The screenshot displays the Sevilleta LTER website interface. The browser address bar shows the URL <http://sev.lternet.edu/data/sev-137>. The page title is "Sevilleta LTER" with the subtitle "Long Term Ecological Research". The navigation menu includes links for HOME, ABOUT, DATA, RESEARCH, LIBRARY, OPPORTUNITIES, CONTACT, and AFFILIATES. The main content area is divided into several sections:

- Dataset Title:** Core Site Phenology Study from the Chihuahuan Desert Grassland and Simulband at the Sevilleta National Wildlife Refuge, New Mexico (2000-)
- Dataset ID:** 137
- Abstract:** Plant phenology or life-history pattern changes seasonally as plants grow, mature, flower, and produce fruit and seeds. Plant phenology follows seasonal patterns, yet annual variation may occur due to annual differences in the timing of rainfall and ambient temperature shifts. Foliage growth and fruit and seed production are important aspects of plant population dynamics and food resource availability for animals. The purpose of this study is to assess plant phenology patterns across a series of biotic communities that represent an environmental moisture gradient. These communities include: Chihuahuan Desert creosotebush shrubland, Chihuahuan Desert black grama grassland, and blue grama grassland. Plant phenology is recorded for all plant species across 4 replicate 200 m transects at each of the 3 habitat sites. Plant phenology measurements are taken once every month from February through October. The first ten individuals of each plant species encountered along each transect are assessed for life-history status. Data recorded include the status of leaves, flowers and fruit. Leaves are recorded as new, old, brown or absent. Reproductive status is recorded as absent, buds, flowers, fruits or both fruits and flowers. Data from the site P and J were only collected in 2000 and 2001 and are included in this data set.
- Methods:** Locating the Transects: Phenological conditions are recorded along four permanently marked 200 m x 2 m wide transects at each of the core study sites. The transects are located within four of the five rodent trapping webs. All five transects were originally measured, but in 2003 the least diverse transect at each site was dropped. Each
- Data:** Please read the Sevilleta Data Access Policy before downloading data.
Data File: sev137_phenologycore_20150723.txt Variable Definitions
- Core Areas:** populations
- Owner:** Wright, Karen
- Contact:** Sevilleta LTER, Information Manager

The University of New Mexico logo is displayed in the top right corner of the page.

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



Example from Step 6: Access



Long-term identifiers made easy

A SERVICE OF THE UNIVERSITY OF CALIFORNIA CURATION CENTER — UC3



researchdata@library.illinois.edu • University Library

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.

Step 7: Preservation for the Long Term

Curating Research Data: A handbook of current practice



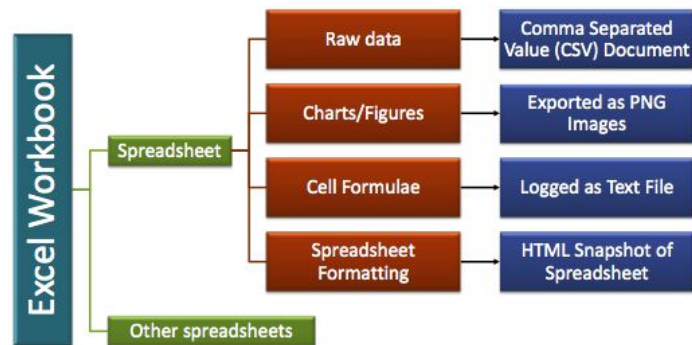
Sub Steps

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

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

Step 8: Reuse

Curating Research Data: A handbook of current practice

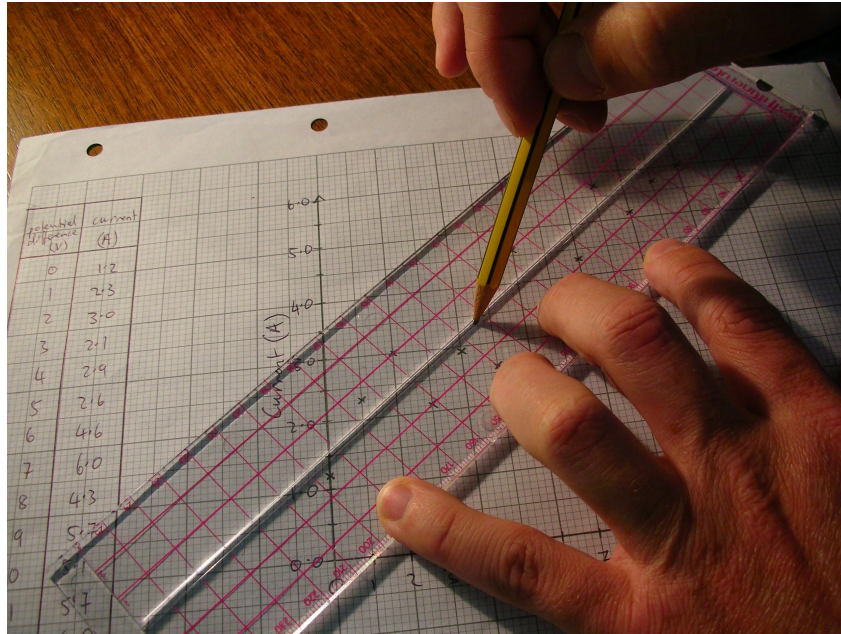
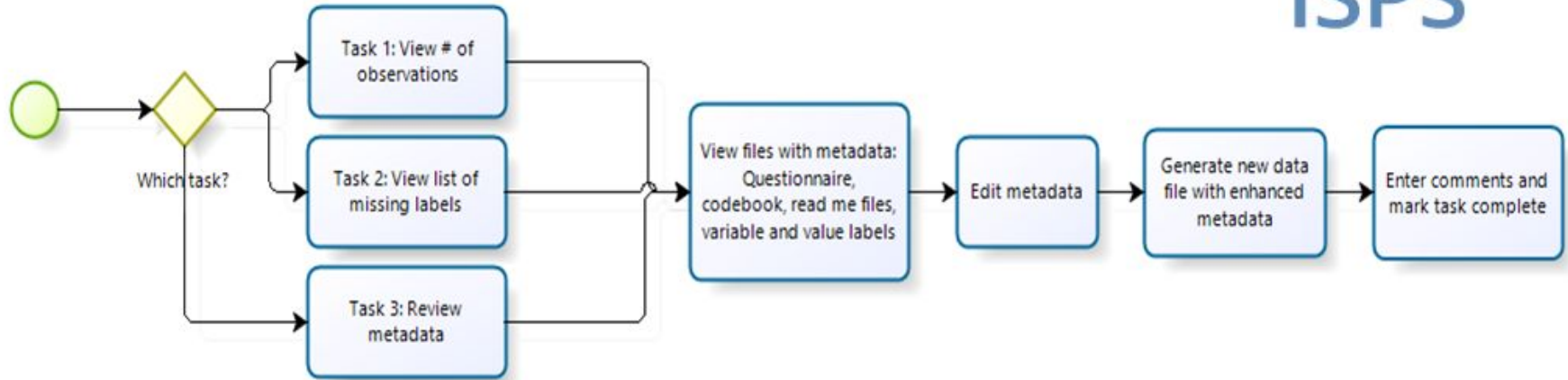


Image: <http://my.bestfitlineruler.com/wp-content/uploads/2009/05/drawing-the-bf11.jpg>

Sub Steps

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

Example from Step 8: Reuse



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