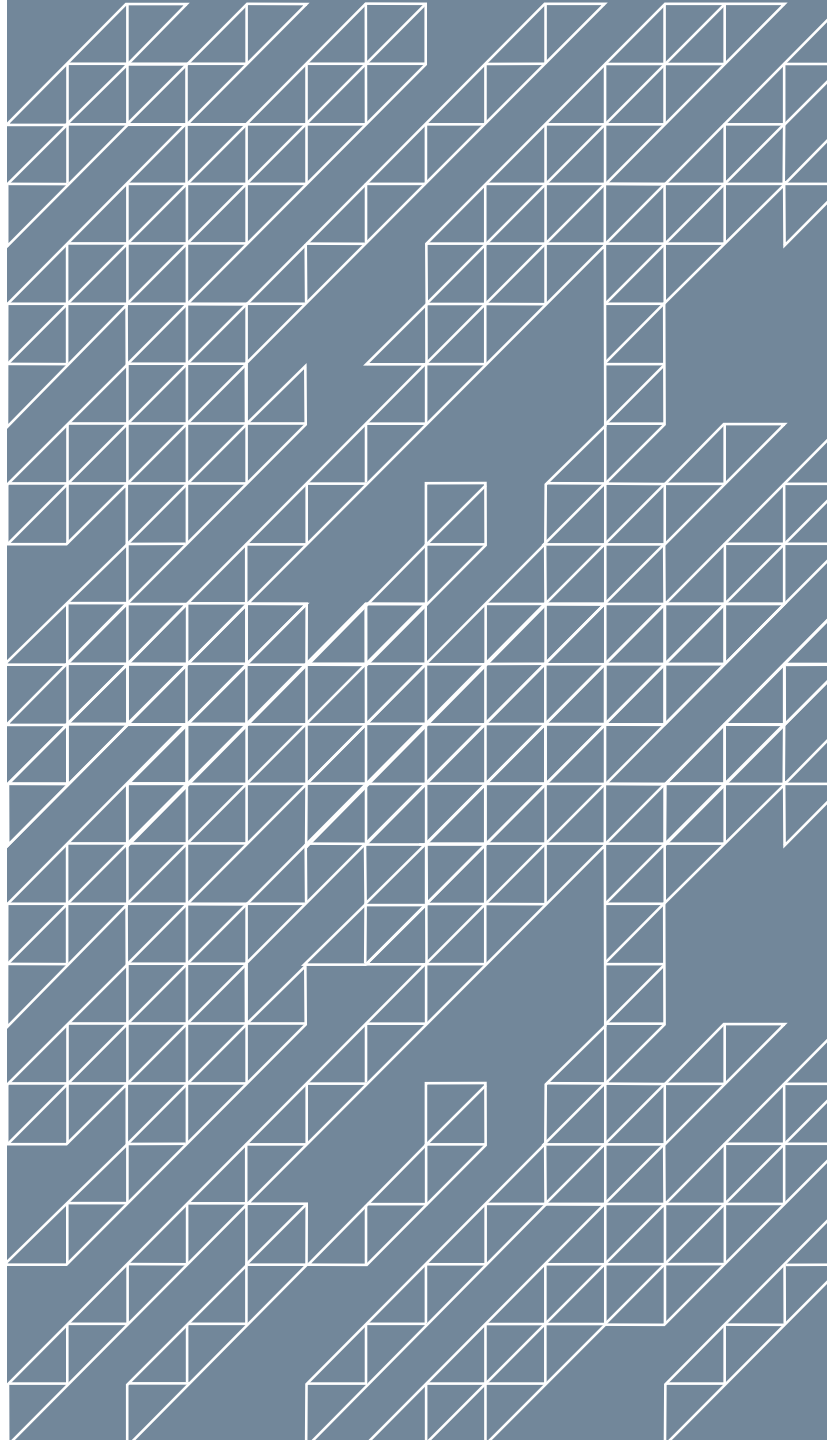
Preparing Data for Reuse

Practical Approaches to Enhancing Data Reusability

Briana Wham,
Data Learning Center Manager & Research Data Librarian



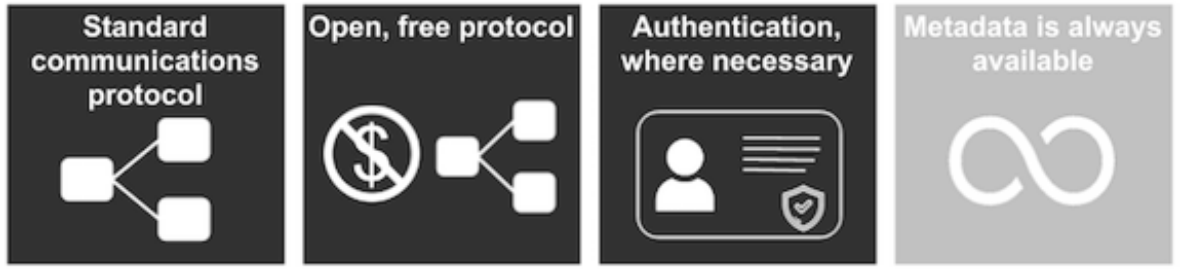
Learning Objectives

- Understand why preparing data for reuse is essential
- Explore data repositories
- Learn key components of data documentation
- Investigate and determine documentation (README and data dictionary) needs

Data Should Be FAIR

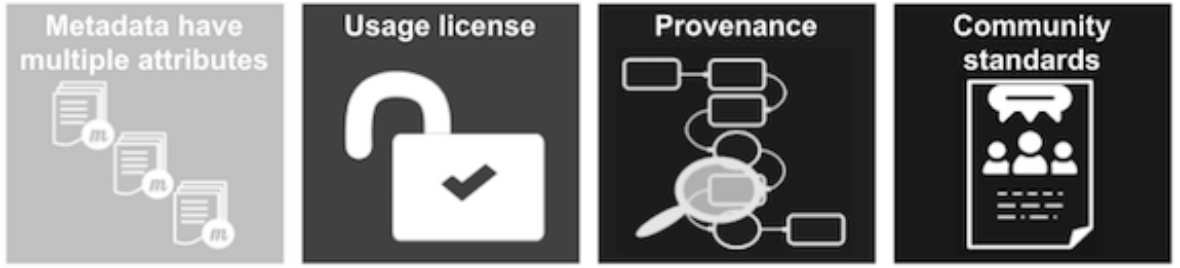
Findable

Accessible



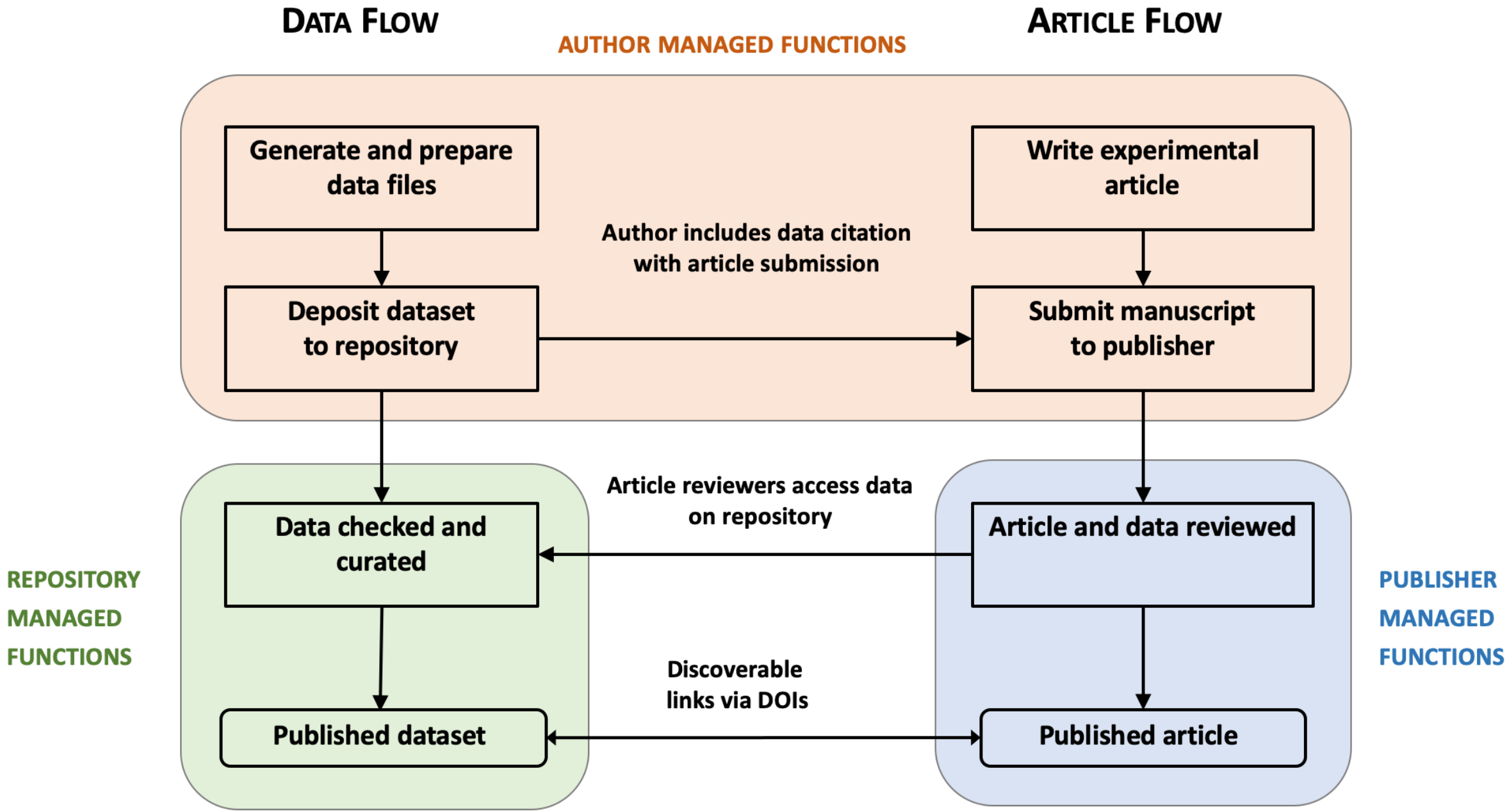
Interoperable

Reusable



"FAIR principles" (https://www.force11.org/group/fairgroup/fairprinciples); image: ARDC

What issues have you faced reusing data?



https://data.research.cornell.edu/data-management/sharing/fair/

FAIR Checklist for Dataset/Files

- ❑ Is the dataset in a trusted data repository?
- ❑ Does the dataset have a registered DOI?
- ❑ Are data files in standard and/or commonly available open formats (as much as possible)?
- ❑ Are the data and/or metadata retrievable via an API and/or discoverable through an open search protocol?

https://data.research.cornell.edu/data-management/sharing/fair/

What do repositories offer?

Types of Data Repositories

Disciplinary

Set up to accommodate the data needs of a specific research community

Institutional

Support and promote the research outputs of an institution and set up to accept various data types, formats, and disciplinary focuses.

Generalist

Accept data regardless of type, format, content, or disciplinary focus.

Types of Data Repositories

Disciplinary

Qualitative Data Repository

ICPSR

Vivli
CENTER FOR GLOBAL CLINICAL RESEARCH DATA

NCBI

Databrary

dbGaP

NIMH Data Archive
NDA | ABCD | CCF | NIAAA$_{DA}$ | OAI

Institutional

ScholarSphere

PennState University Libraries

datacommons@psu
data. energy.environment.society.technology

Generalist

DRYAD

zenodo

figshare

Types of Data Repositories

Public-Public (Open access)	Public-Private (Mediated open access)	Private-Private (Closed access)
Metadata is fully discoverable	Metadata is fully discoverable	Metadata is not publicly available
Data are accessible and immediately downloadable	Mediated access to data via data custodian	Data not discoverable or available to third parties
Preferred option for non-sensitive data from completed projects	Good option for sensitive or confidential data	Safest option for highly-sensitive data

Table originally from Curtin University, Research Data Team, http://libguides.library.curtin.edu.au

Desirable Characteristics of Data Repositories

Desirable Characteristics of Data Repositories for Federally Funded Research

Unique Persistent Identifiers	Long-Term Sustainability	Metadata	Curation and Quality Assurance
Free and Easy Access	Broad and Measured Reuse	Clear Use Guidance	Security and Integrity
Risk Management/ Confidentiality	Common Format	Provenance	Retention Policy

For more see NOT-OD-21-016 – Supplemental Information to the NIH Policy for Data Management and Sharing: Selecting a Repository for Data Resulting from NIH-Supported Research

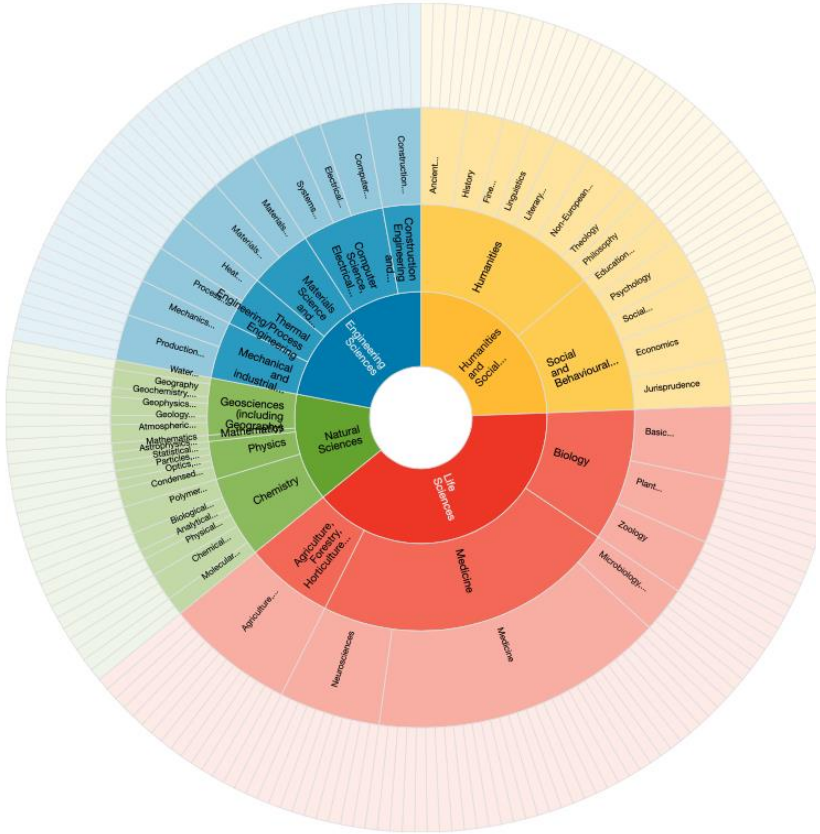
re3data.org

REGISTRY OF RESEARCH DATA REPOSITORIES

Browse by subject

Graphical | Text

click to zoom into subjects or to select a bottommost subject in the hierachy as filter for the re3data search page
ctrl + click on a top subject to select it as filter



Filter

- Subjects ⊞
- Content Types ⊞
- Countries ⊞
- AID systems ⊞
- API ⊞
- Certificates ⊞
- Data access ⊞
- Data access restrictions ⊞
- Database access ⊞
- Database access restrictions ⊞
- Database licenses ⊞
- Data licenses ⊞
- Data upload ⊞
- Data upload restrictions ⊞
- Enhanced publication ⊞
- Institution responsibility type ⊞
- Institution type ⊞
- Keywords ⊞
- Metadata standards ⊞
- PID systems ⊞
- Provider types ⊞
- Quality management ⊞
- Repository languages ⊞
- Software ⊞
- Syndications ⊞
- Repository types ⊞
- Versioning ⊞

Sustainable File Formats For Long-Term Storage & Preservation

Veerle Van den Eynden, Louise Corti, Matthew Woollard, Libby Bishop and Laurence Horton
2011

TYPE OF DATA	RECOMMENDED FILE FORMATS FOR SHARING, RE-USE AND PRESERVATION
Quantitative tabular data with extensive metadata a dataset with variable labels, code labels, and defined missing values, in addition to the matrix of data	SPSS portable format (.por) delimited text and command ('setup') file (SPSS, Stata, SAS, etc.) containing metadata information some structured text or mark-up file containing metadata information, e.g. DDI XML file
Quantitative tabular data with minimal metadata a matrix of data with or without column headings or variable names, but no other metadata or labelling	comma-separated values (CSV) file (.csv) tab-delimited file (.tab) including delimited text of given character set with SQL data definition statements where appropriate
Geospatial data vector and raster data	ESRI Shapefile (essential: .shp, .shx, .dbf ; optional: .prj, .sbx, .sbn) geo-referenced TIFF (.tif, .tfw) CAD data (.dwg) tabular GIS attribute data
Qualitative data textual	eXtensible Mark-up Language (XML) text according to an appropriate Document Type Definition (DTD) or schema (.xml) Rich Text Format (.rtf) plain text data, ASCII (.txt)
Digital image data	TIFF version 6 uncompressed (.tif)
Digital audio data	Free Lossless Audio Codec (FLAC) (.flac)
Digital video data	MPEG-4 (.mp4) motion JPEG 2000 (.jp2)
Documentation	Rich Text Format (.rtf) PDF/A or PDF (.pdf) OpenDocument Text (.odt)

FAIR Checklist for Data Documentation

- ❑ Are all associated data files unambiguously named and described including file types, software requirements, and/or conversion information in the metadata?
- ❑ Does the metadata include useful disciplinary notation and terminology?
- ❑ Does the metadata include machine-readable standards where available?
- ❑ Are related articles referenced and linked in the metadata?
- ❑ Is a citation format for the dataset provided?
- ❑ Are any license terms, attribution, or terms of use clearly indicated?
- ❑ Is the metadata exportable in a machine-readable structured text-based format?

https://data.research.cornell.edu/data-management/sharing/fair/

Data Documentation

- the context of data collection: project history, aim, objectives and hypotheses
- data collection methods: sampling, data collection process, instruments used, hardware and software used, scale and resolution, temporal and geographic coverage and secondary data sources used
- dataset structure of data files, study cases, relationships between files
- data validation, checking, proofing, cleaning and quality assurance procedures carried out
- changes made to data over time since their original creation and identification of different versions of data files
- information on access and use conditions or data confidentiality

Veerle Van den Eynden, Louise Corti, Matthew Woollard, Libby Bishop and Laurence Horton 2011

Documentation Formats

Metadata

Highly-structured data laid out in fields, often with controlled vocabularies in each field.

Data Dictionaries

Defines variables without cluttering datasets.

README.txt files

Simple text files that provide support for navigating folders and understanding the contents of your files.

Metadata

Minimal

Enable basic discovery and access

- Creator
- Title
- Publisher
- Date
- Embargo term
- License
- Access terms and conditions

Cox and Verbaan 2018

Metadata

Minimal

Enable basic discovery and access

- Creator
- Title
- Publisher
- Date
- Embargo term
- License
- Access terms and conditions

General Contextual

May be of an administrative nature or relate to project methodologies

- Funder
- Grant #
- Project information
- Data generation process
- Geographical location of data collection
- Date range of data collection

Cox and Verbaan 2018

Metadata

Minimal

Enable basic discovery and access

- Creator
- Title
- Publisher
- Date
- Embargo term
- License
- Access terms and conditions

General Contextual

May be of an administrative nature or relate to project methodologies

- Funder
- Grant #
- Project information
- Data generation process
- Geographical location of data collection
- Date range of data collection

Discipline/ Data Specific

Enables re-use

- Machine settings
- Experimental conditions

**Can be difficult to capture in structured metadata fields unless depositing in a discipline specific repository or community standard exists -> information can be put in data documentation

Cox and Verbaan 2018

Metadata

https://fairsharing.org/

http://www.dcc.ac.uk/drupal/resources/metadata-standards

http://rd-alliance.github.io/metadata-directory/

Minimal

Enable basic discovery and access

- Creator
- Title
- Publisher
- Date
- Embargo term
- License
- Access terms and conditions

General Contextual

May be of an administrative nature or relate to project methodologies

- Funder
- Grant #
- Project information
- Data generation process
- Geographical location of data collection
- Date range of data collection

Discipline Specific

Enables re-use

- Machine settings
- Experimental conditions

*Example Metadata schema: Darwin Core (biodiversity taxa data), Ecological Metadata Language, ISO1911 & ISO 19139 (geospatial data) , DDI (social science, survey data), MIxS (Genomic data)

*Example standards: Biological classification system, NCBI Taxonomy

Cox and Verbaan 201

Data Dictionary

Common information included:

- Variable name
- Variable definition
- How the variable was measured
- Data units
- Data formats
- Minimum and maximum values
- Coded values and their meanings
- Representation of null vs. NA values
- Precision of measurement
- Known issues with the data (missing values, bias, etc.)
- Relationship to other variables
- Other important notes about the data

Table 1. List of taxa included in the data files, including taxonomic code used in all data files (Taxon), Latin name and common name of each taxon[4, 5].

count	Taxon	Latin_Name	Common_Name
1	CMED	Cheirogaleus medius	Fat-tailed dwarf lemur
2	DMAD	Daubentonia madagascariensis	Aye-aye
3	EALB	Eulemur albifrons	White-fronted brown lemur
4	ECOL	Eulemur collaris	Collared brown lemur
5	ECOR	Eulemur coronatus	Crowned lemur
6	EFLA	Eulemur flavifrons	Blue-eyed black lemur
7	EFUL	Eulemur fulvus	Common brown lemur
8	EMAC	Eulemur macaco	Black lemur
9	EMON	Eulemur mongoz	Mongoose lemur
10	ERUB	Eulemur rubriventer	Red-bellied lemur
11	ERUF	Eulemur rufus	Red-fronted brown lemur
12	ESAN	Eulemur sanfordi	Sanford's brown lemur
13	EUL	Eulemur	Eulemur hybrid
14	GMOH	Galago moholi	Mohol bushbaby
15	HGG	Hapalemur griseus griseus	Eastern lesser bamboo lemur
16	LCAT	Lemur catta	Ring-tailed lemur
17	LTAR	Loris tardigradus	Slender loris
18	MMUR	Mircocebus murinus	Gray mouse lemur
19	MZAZ	Mirza coquereli	Northern giant mouse lemur
20	NCOU	Nycticebus coucang	Slow loris
21	NPYG	Nycticebus pygmaeus	Pygmy slow loris
22	OGG	Otolemur garnettii garnettii	Northern greater galago
23	PCOQ	Propithecus coquereli	Coquerel's sifaka
24	PPOT	Perodicticus potto	Potto
25	VAR	Varecia	Varecia hybrid
26	VRUB	Varecia rubra	Red ruffed lemur
27	VVV	Varecia variegata variegata	Black-and-white ruffed lemur

Table 5: DLC Weight File variable descriptions

count	Weight File Variable Name	Weight File Variable Definition
1	Taxon	Taxonomic code: In most cases, comprised of the first letter of the genus and the first three letters of the species; if taxonomic designation is a subspecies, comprised of the first letter of genus, species, and subspecies, and hybrids are indicated by the first three letters of the genus. See table 1 for details.
2	Hybrid	Hybrid status: N=not a hybrid. S=species hybrid. B=subspecies hybrid. If sire is one of multiple possible and animal could be a hybrid, it is designated a hybrid.
3	DLC_ID	Specimen ID: Unique DLC number assigned at accession of animal
4	Sex	Sex: M=male. F=Female. ND=Not determined

Briney 2015

https://doi.org/10.5061/dryad.fj974

Data Dictionary

Variable	Description	Data Type	Prescribed Values/ Format	Unit of Measurement	Sample Data
phase	important stages of the experiment when plot size, crop rotation or treatments changed	categorical	1-5	-	3
year	year planted and harvested	date	YYYY	-	1955
plot	label indicating plot number, North/South location, and A/B/C/D subplot (see Data Sources section above for an important note about plot names)	categorical	3NA; 3NB; 3NC; 3ND; 3SA; 3SB; 3SC; 3SD; 4NA; 4NB; 4NC; 4ND; 4SA; 4SB; 4SC; 4SD; 5NA; 5NB; 5NC; 5ND; 5SA; 5SB; 5SC; 5SD	-	4SD
plot_num	plot number only	categorical	3-5	-	4
plot_dir	N/S and E/W location	categorical	NE; NW; SE; SW	-	SE
rotation	number of years in the crop rotation schedule for this plot	categorical	1-3	-	2
corn	flag making it easy to group corn in rotation and continuous corn	T/F	TRUE; FALSE	-	TRUE
crop	crop planted this year with separate values for corn in rotation (C) and continuous corn (CC)	categorical	A [alfalfa]; C [corn]; CC [continuous corn]; H [hay]; O [oats]; S [soybeans]	-	C
variety	crop variety name	text	-	-	Illinois 1570
all_corn	whether this was a year when corn was planted in all plots	T/F	TRUE; FALSE	-	TRUE
yield_bush	yield for all crops except hay	numerical	-	bushels/acre	95.6
yield_ton	yield for hay	numerical	-	tons/acre	3.48
treated	whether this plot was treated that year	T/F	TRUE; FALSE	-	TRUE

Demo - Creating a Data Dictionary in R

Writing a README – Best Practices

- Create readme files for logical “clusters” of related files/data
- Name the README so that it is easily associated with the data file(s) it describes
- Write your README document as plain text file, often formatted with markdown (.md)
- Format multiple README files identically
- Use standardized date formats
- Follow the scientific conventions for your discipline for taxonomic, geospatial, and geologic names and keywords

Writing a README – Recommended Content

General Information

- **Provide a title for the dataset**
- **Name/institution/address/email information for**
  - **Principal investigator or person responsible for collecting the data**
  - Associate or co-investigators
  - Contact person for questions
- **Date of data collection**
- **Information about geographic location of data collection**
- Keywords
- Language
- Information about funding sources

https://data.research.cornell.edu/data-management/sharing/readme/

Writing a README – Recommended Content

Data and file overview

- **For each filename, a short description of what data it contains**
- Format of the file if not obvious from the file name
- The relationship between files if multiple files relate to one another
- **Date that the file was created**
- Date(s) that the file(s) was updated (versioned) and the nature of the update(s), if applicable
- Information about related data collection that is not in the described dataset

https://data.research.cornell.edu/data-management/sharing/readme/

Writing a README – Recommended Content

Sharing and access information

- **Licenses or restrictions placed on the data**
- Links to publications that cite or use the data
- Recommended citation for the data

https://data.research.cornell.edu/data-management/sharing/readme/

Writing a README – Recommended Content

Methodological information

- **Description of methods for data collection or generation** (include links or references to publications or other documentation containing experimental design or protocols used)
- **Description of methods used for data processing**
- Any software or instrument-specific information needed to understand or interpret the data, including software version numbers
- Describe any quality-assurance procedures performed on the data
- Definitions of codes or symbols
- People involved with sample collection, processing, analysis, and/or submission

https://data.research.cornell.edu/data-management/sharing/readme/

Writing a README – Recommended Content

Data-specific information

*Repeat this section as needed for each dataset (or file, as appropriate)

- Count of number of variables and rows
- **Variable list, including full names and definitions of column headings for tabular data**
- **Units of measurement**
- **Definitions for codes or symbols used to record missing data**

https://data.research.cornell.edu/data-management/sharing/readme/

Demo – README Template

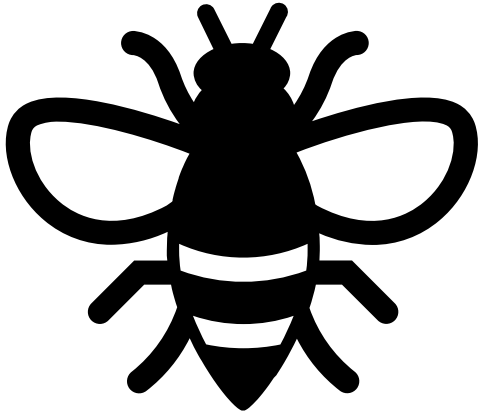
Activity – Can you Reuse This Data?

Imagine you're a researcher from another institution. You found this dataset and want to use it. Can you understand it well enough?

Dataset Description:

This dataset was deposited by an entomologist to their local data repository; it has been slightly simplified for the purposes of this activity.

It contains experimental measurements documenting how honey bee colonies exposed to high temperatures maintain and meet the water needs of the colony to survive.



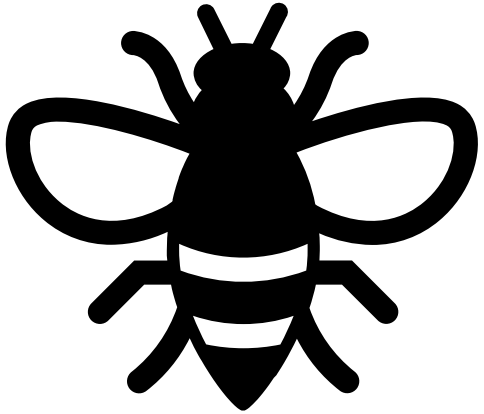
Access Activity Materials: https://tinyurl.com/2e5834as

Activity – Can you Reuse This Data?

10 min

1. Review the dataset and the bare-bones README

2. Consider:
    i. What's missing or unclear?
    ii. What assumptions did you have to make?
    iii. What information would help you trust and reuse this data?



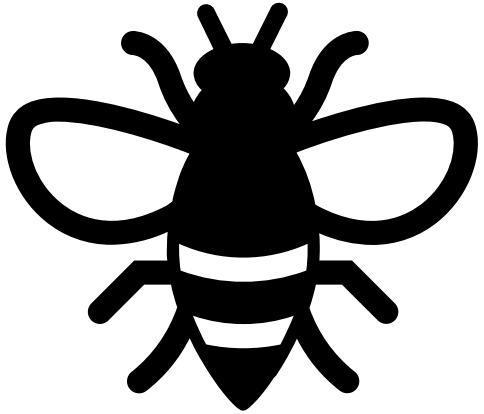
Access Activity Materials: https://tinyurl.com/2e5834as

Activity – Can you Reuse This Data?

Access Updated data & README: https://tinyurl.com/5fpnkmd2

10 min

1. What's different?
2. What information was added?
3. How does this improved README support reuse?



Activity – Finding the Right Home for Your Data

Imagine you are preparing to share your own dataset. How would you choose a repository and documentation format?