

Creating Effective Data Management Plans

Katie Coburn

UC Merced Library

kcoburn@ucmerced.edu

What is Data?




What is Data? White House:

“Research data is defined as the **recorded factual material** commonly accepted in the scientific community as **necessary to validate research findings**”

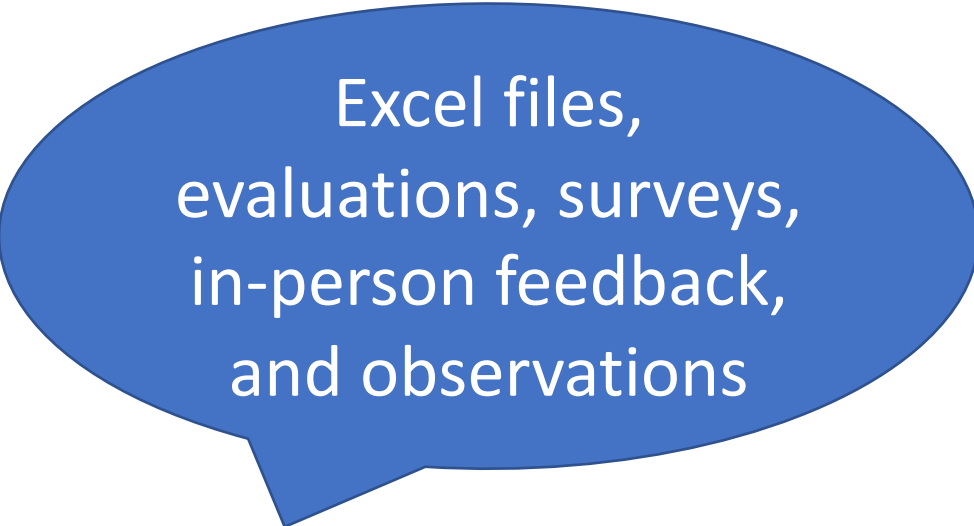
[OMB Circular A-110](#) (now 2 CFR, Ch. II, §215.36(d)(2)(i), and codified in 5 U.S.C. 552(a)(4)(A)

What is Data?

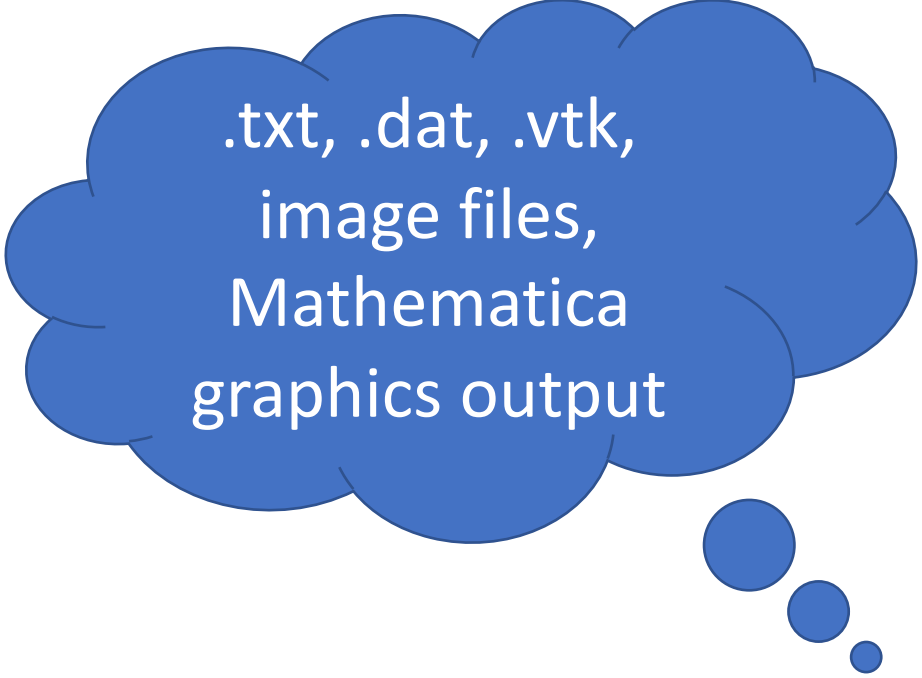
Your Data:



Images, movies,
graphs, .csv files,
text, slides



Excel files,
evaluations, surveys,
in-person feedback,
and observations



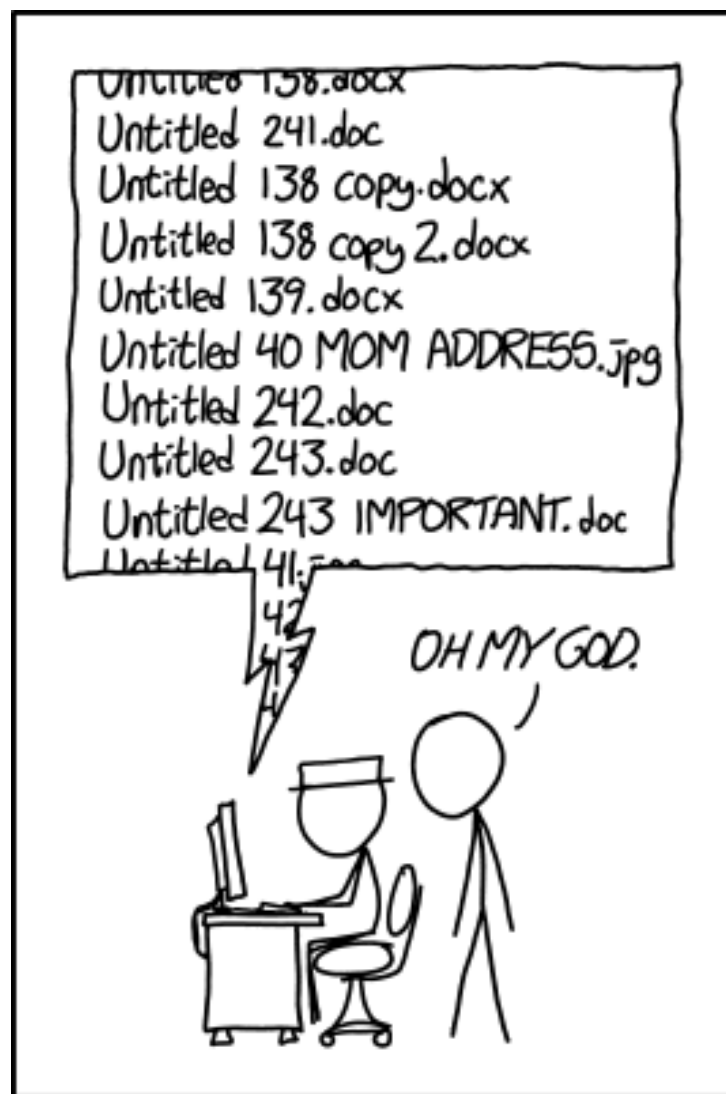
.txt, .dat, .vtk,
image files,
Mathematica
graphics output

Why Manage Data?

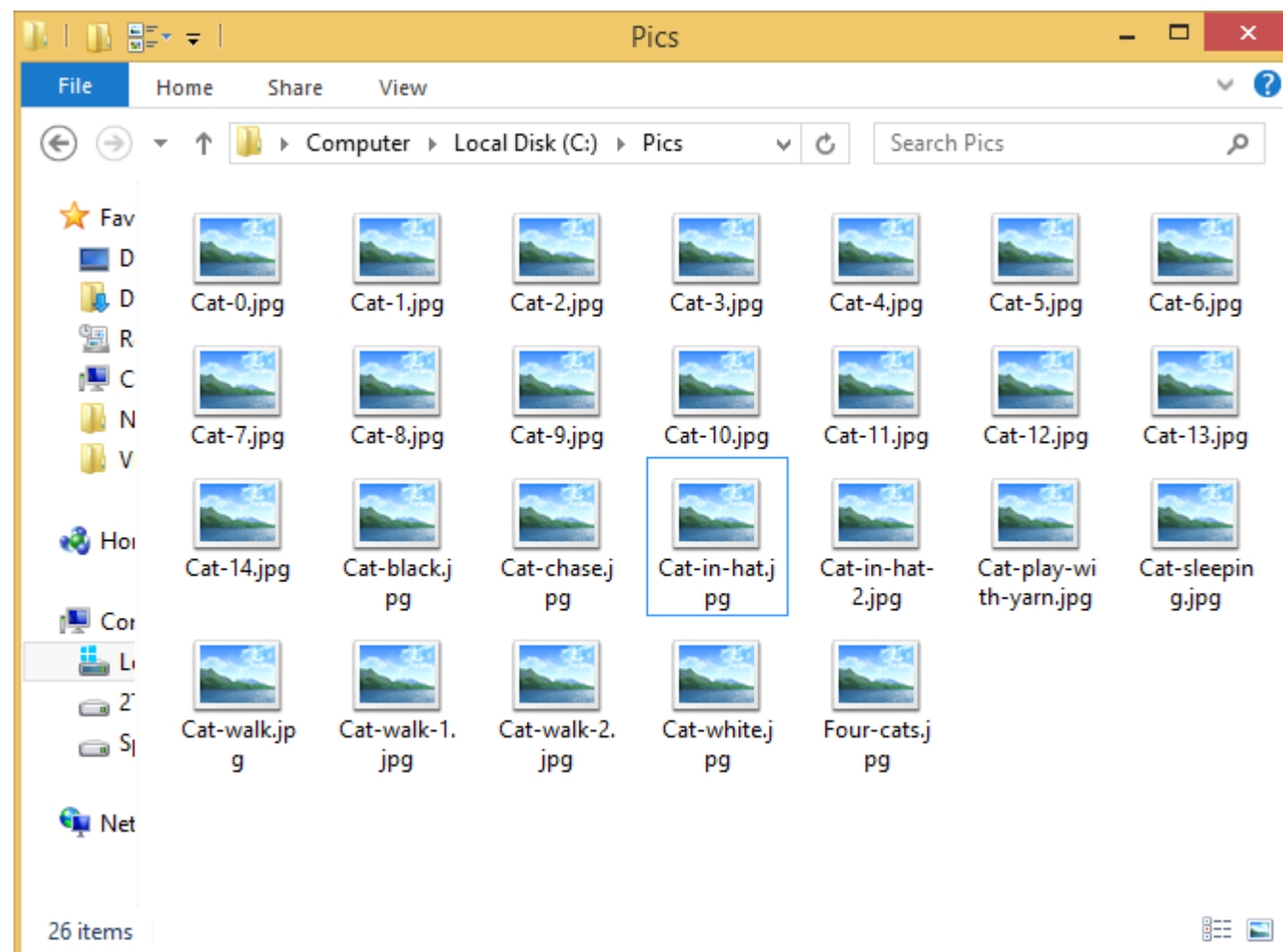
- Make your own research easier
- Stop yourself from drowning in irrelevant details
- Save data for later
- Avoid accusations of fraud or bad science
- Share your data for reuse
- Get credit for your data

Elements of Data Management

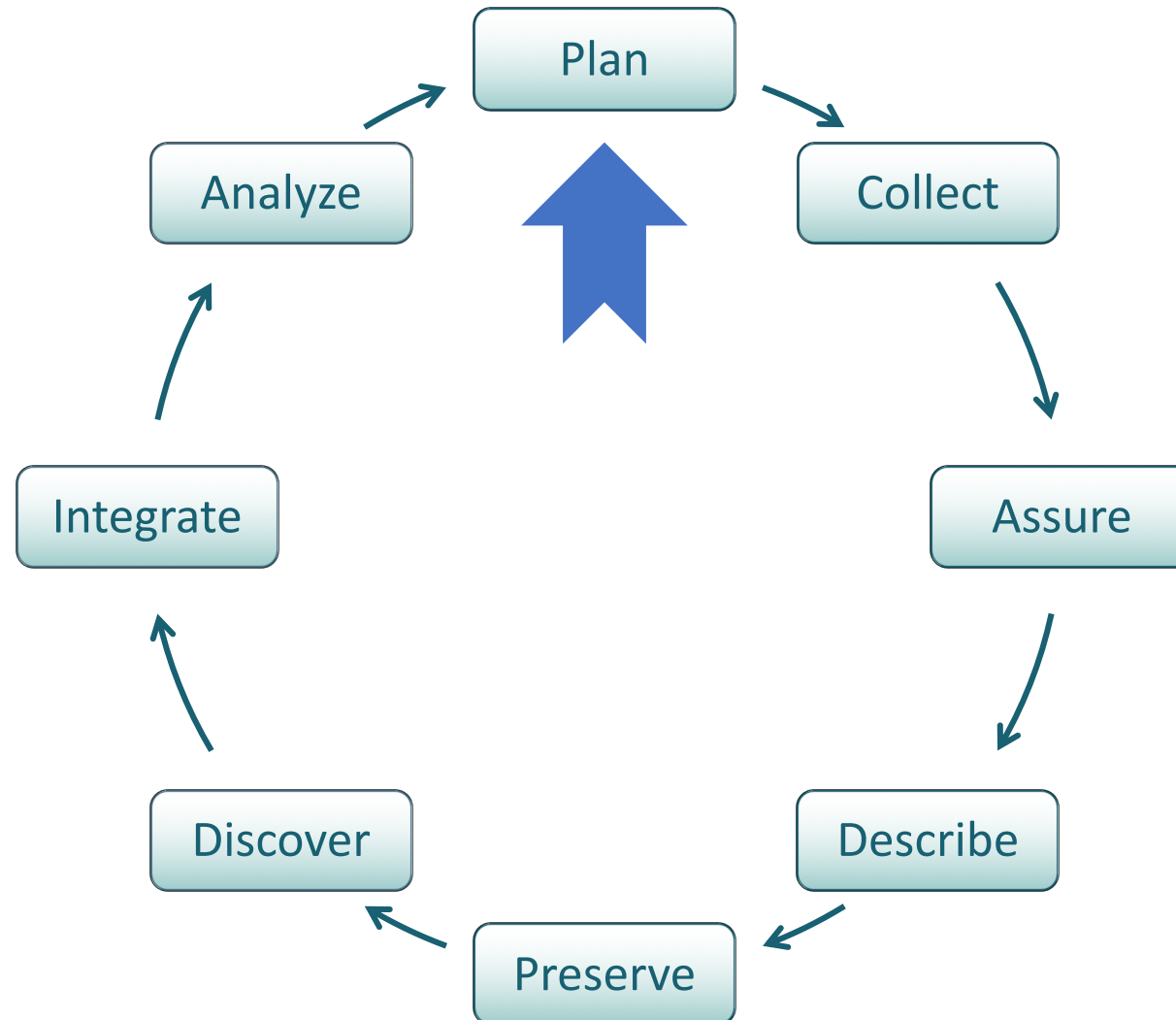
- Data inventory
- Data dictionaries, codebooks
- File naming conventions
- Directory structure; README file
- Version control
- Storage and back-up procedures
- Responsibility assignment
- Legal agreements; policies for access, use, retention



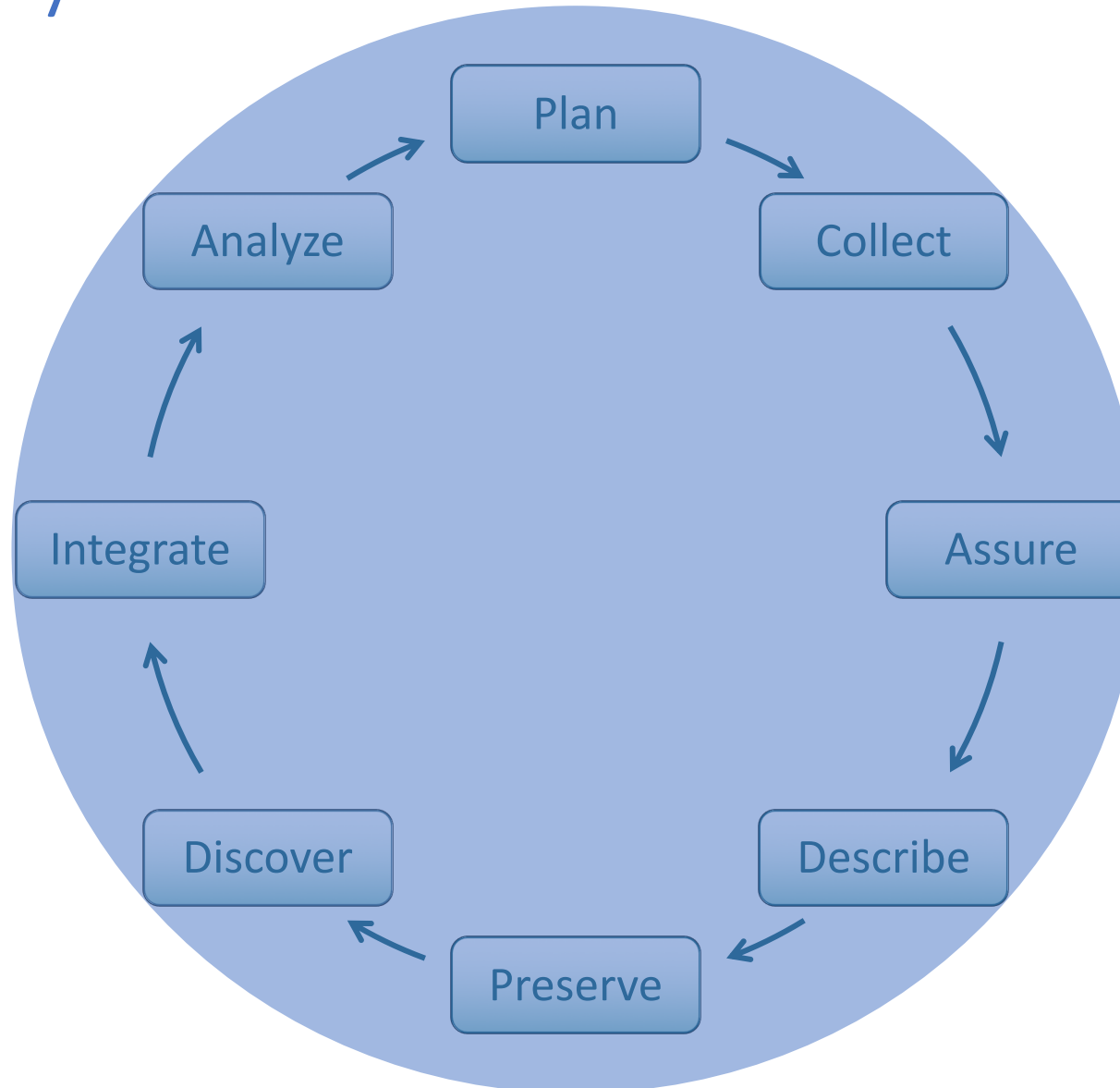
PROTIP: NEVER LOOK IN SOMEONE ELSE'S DOCUMENTS FOLDER.

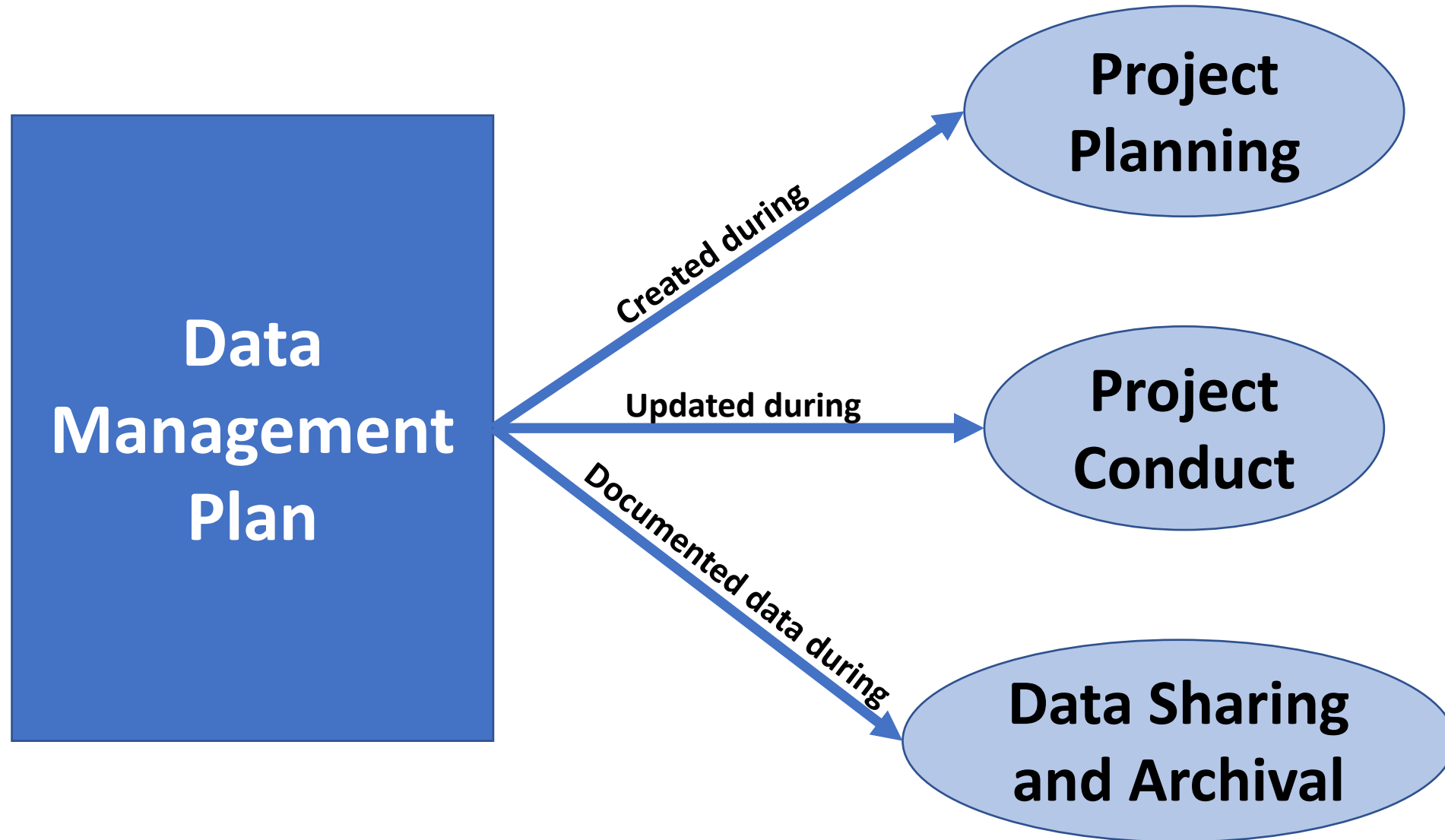


Data Life Cycle



Data Life Cycle





A DMP is a living document

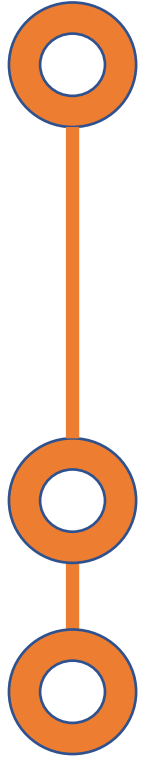
- An effective DMP demonstrates rigor, capacity
- Keep your plan current
- Incorporate changes
- Use as a guide for daily activities

Why Manage Data?

- Make your own research easier
- Stop yourself from drowning in irrelevant details
- Save data for later
- Avoid accusations of fraud or bad science
- Share your data for reuse
- Get credit for your data
- ... **Because many funders require you to**

Why Manage Data?

- Make your own research easier
- Stop yourself from drowning in irrelevant details
- Save data for later
- Avoid accusations of fraud or bad science
- Share your data for reuse
- Get credit for your data
- **... Because many funders require you to**
 - **... and have for a while (10+ years).**



NIH Data Sharing Policy (2003)

applies to applicants seeking \$500,000 or more in direct costs in any year of the proposed research

NSF Data Sharing Policy (2011)

OSTP Memo “Increasing Access to the Results of Federally Funded Scientific Research” (2013)

Federal agencies with research funding budgets of over \$100 million release policies 2014-2015: <http://datasharing.sparcopen.org/data>

NSF Data Sharing Policy

Investigators are expected to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data, samples, physical collections and other supporting materials created or gathered in the course of work under NSF grants. Grantees are expected to encourage and facilitate such sharing. See [Proposal & Award Policies & Procedures Guide \(PAPPG\) Chapter XI.D.4.](#)

<https://www.nsf.gov/bfa/dias/policy/dmp.jsp>

NSF Data Management Plan Requirements

Proposals submitted or due on or after January 18, 2011, must include a supplementary document of no more than two pages labeled "Data Management Plan". This supplementary document should describe **how the proposal will conform to NSF policy on the dissemination and sharing of research results**. See [PAPPG Chapter II.C.2.j](#) for full policy implementation.

FAIR Guiding Principles

- **Findable**
- **Accessible**
- **Interoperable**
- **Reusable**

Wilkinson, M. D. *et al.* [The FAIR Guiding Principles for scientific data management and stewardship](#). *Sci. Data* 3:160018
doi: 10.1038/sdata.2016.18 (2016).

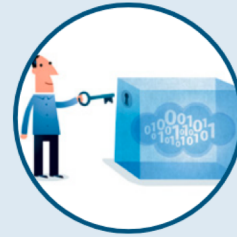
Boeckhout, Martin *et al.* [The FAIR Guiding Principles for data stewardship: fair enough?](#) *European Journal of Human Genetics*
doi:10.1038/s41431-018-0160-0 (2018).

FAIR Guiding Principles



Data and supplementary materials have sufficiently rich metadata and a unique and persistent identifier.

FINDABLE



Metadata and data are understandable to humans and machines. Data is deposited in a trusted repository.

ACCESSIBLE



Metadata use a formal, accessible, shared, and broadly applicable language for knowledge representation.

INTEROPERABLE



Data and collections have a clear usage licenses and provide accurate information on provenance.

REUSABLE

What is a DMP?

- #1: Types of data produced
- #2: Data and metadata standards
- #3: Policies for access and sharing
- #4: Policies for reuse, redistribution, etc.
- #5: Plans for archiving and preservation

The following slides include bad examples of DMP text. [Go here for good examples \(and more detail\).](#)

#1 Types of Data: Prompts

- What is the nature of your data?
- How will your data be created or captured?
- Will you use existing data? If so, what data and why?
- What data will be shared (if any) and why?



How will your data be created or captured?

- Good answers:

- **Specific** about stages of the data
- Corresponding methods and tools
- For every type of data
- Name the file types; Excel workbook or .csv, etc.
- Detail, detail, detail

- Bad answers:

- Vague

Bad Example:

“The measurement data are generated by various laboratory electronic measurement instruments . Furthermore, photographs of the fabricated circuits will be a part of the data generated. The circuit simulation data are produced by computers and are in the form of graphs and tables describing various performance aspects of the circuits being analyzed.”

Will you use existing data?

- Good answers:
 - **Specific**
 - Clearly states details about existing data sets to be used
 - Website address?
 - Source?
 - Relevant variables?
 - Clearly explains the relationship of the new data being produced to the existing data

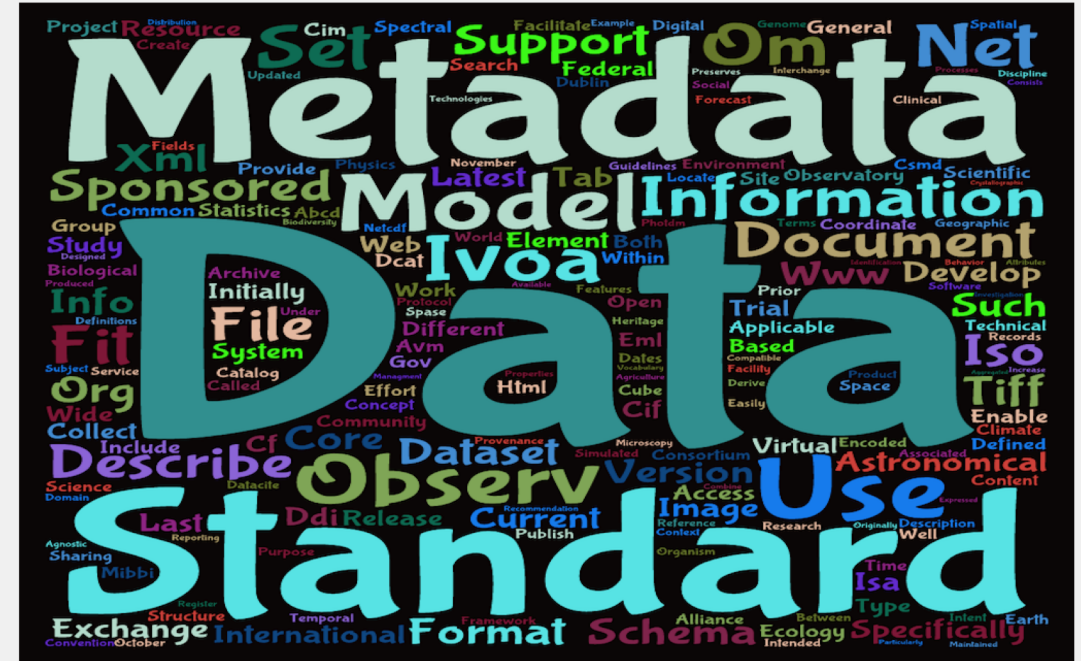
- Bad answers:
 - Vague

Bad Example:

“Our proposed work does not collect any new observations; we only use existing observations and those in the process of being collected through previously funded NSF projects.”

#2 Data and Metadata Standards: Prompts

- Which file formats will you use and why?
- What metadata standards and data documentation will you apply?
- If there are no standard formats, how will you make sure your data is accessible?
- What metadata are needed?
- Who is responsible?



Metadata Standards Directory Working Group

The RDA Metadata Standards Directory Working Group is supported by individuals and organizations involved in the development, implementation, and use of metadata for scientific data. The overriding goal is to develop a collaborative, open directory of metadata standards applicable to scientific data can help address infrastructure challenges.

Directory of metadata standards by subject area

Which file formats will you use and why?

- Good answers:
 - **Specific**
 - Consider potential users of your data & their needs
 - Think about sustainability of files, etc.
- References to best practices (for your use):
 - [Producing PDF files](#)
 - [Producing ZIP and TAR files](#)
 - [Producing dataset files](#)
 - [Producing image files](#)
 - [Producing digital video files](#)
 - [Producing digital audio files](#)
 - [Producing Microsoft Office files](#)

- Bad answers:
 - Vague

Bad Example

“The format of the data will be specific to the format used by the particular software in which it was created. For data generated from instruments, the output will often be in a proprietary ASCII format; in some cases non-proprietary text format will be available.”

Which metadata standards will you apply?

- Good answers:
 - **Specific**
 - Enable others to discover your data
 - Use machine-readable files that address who, what, when, where, why, & how
- References to best practices (for your use):
 - [Best practices involving metadata](#)

- Bad answers:
 - Vague
 - Inappropriate ("Available on request" defeats the purpose)

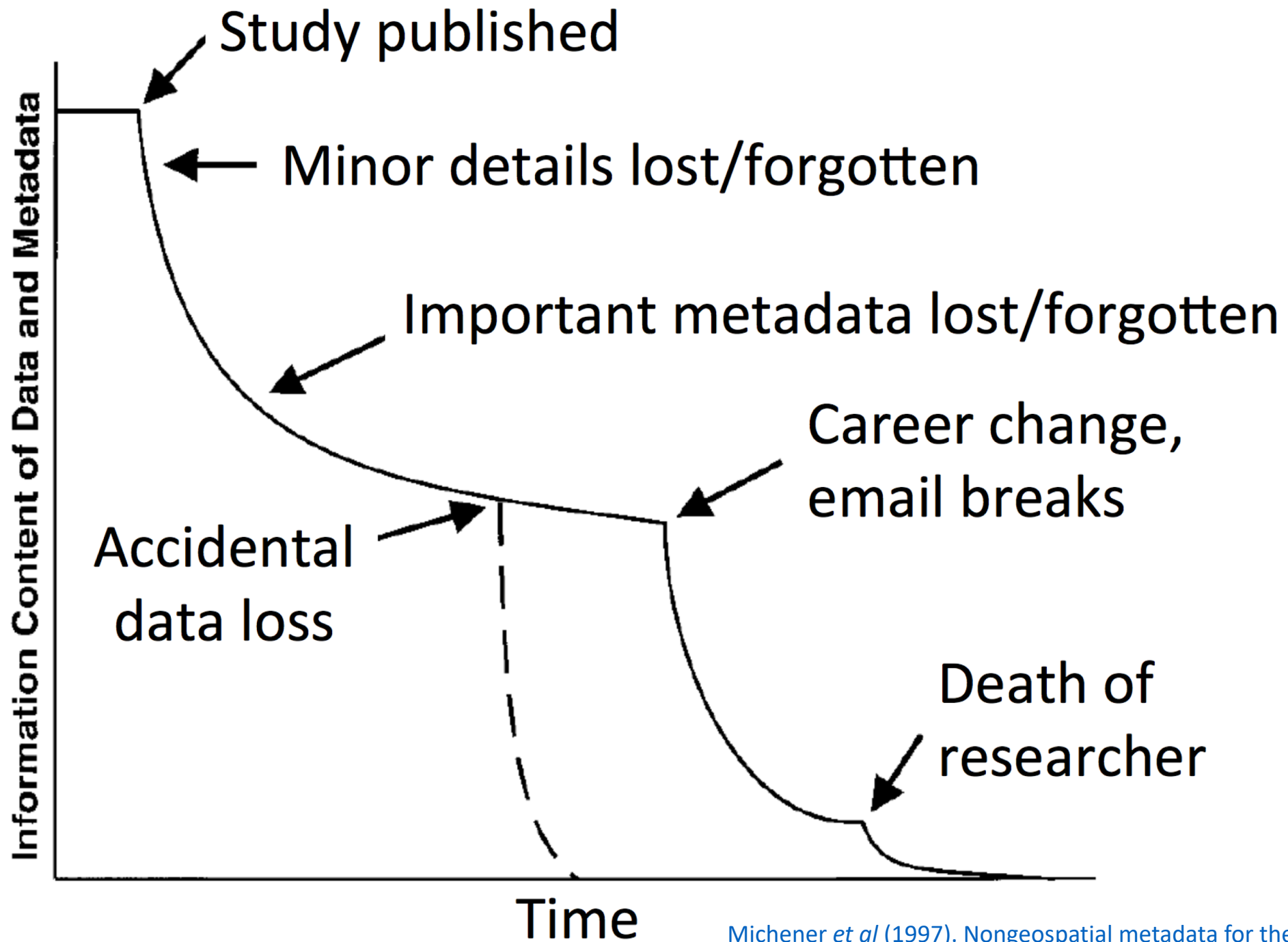
Bad Example

"In addition to data, metadata concerning how the data were generated (software, hardware, dates, measurement protocols, etc.) will be maintained and disseminated on request. Metadata will be stored in Microsoft Word documents."

#3 Policies for Access and Sharing: Prompts

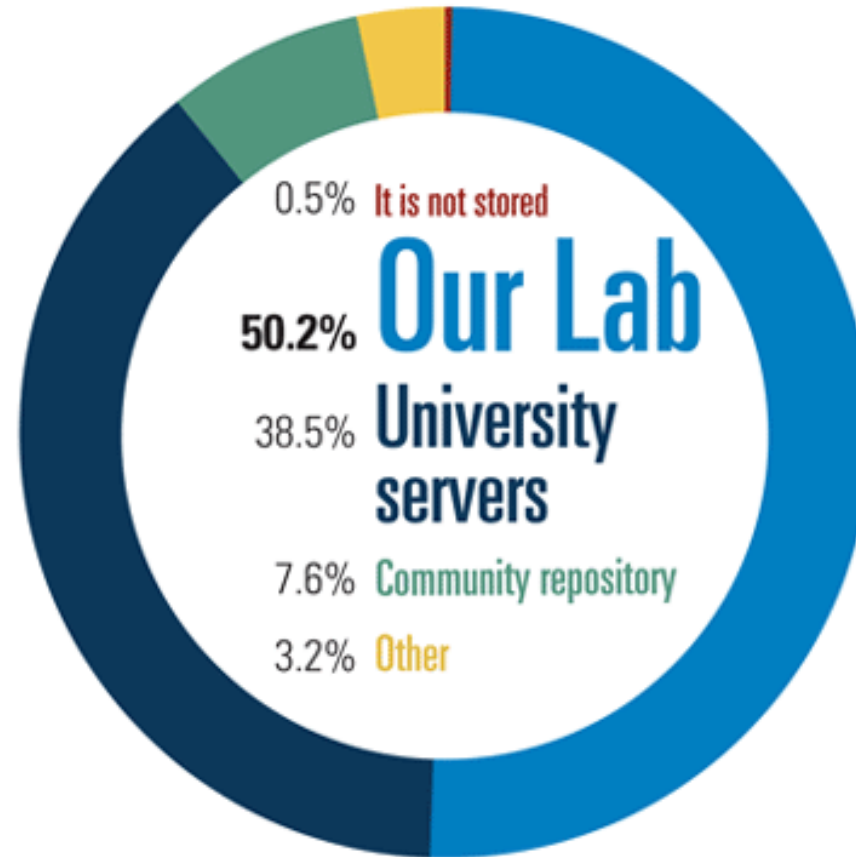
- Who may be interested in your data?
 - [“Long tail”](#)
 - 2 weeks from now? 2 years? 20 years?
- What resources will you need to share your data?
- Will you use an embargo period? Why?
- How can others access your data?
- Are there restrictions?
 - Publisher
 - IRB
 - ORED





Where do you archive most of the data generated in your lab or for your research?

“Even within a single institution **there are no standards for storing data**, so each lab, or often each fellow, uses ad hoc approaches.”



Who might access your data and how?

- Good answers:
 - Specific
 - Reduce the effort necessary to access
 - Thoroughly describe any restrictions and how they will be handled
 - Describe and justify any embargo period(s)





“hypotheses come and go, but
data remain”

Santiago Ramón y Cajal, 1897

Search 319 datasets

Search

Why share your data?



Your journal requires it.



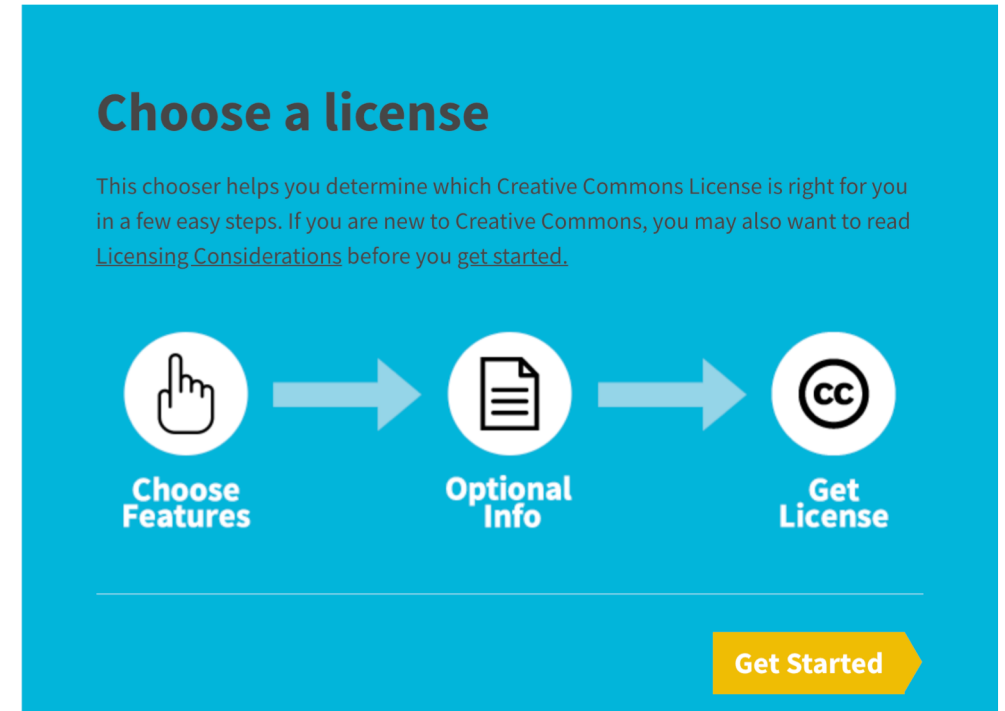
Your funder requires it.



It's the right thing to do.

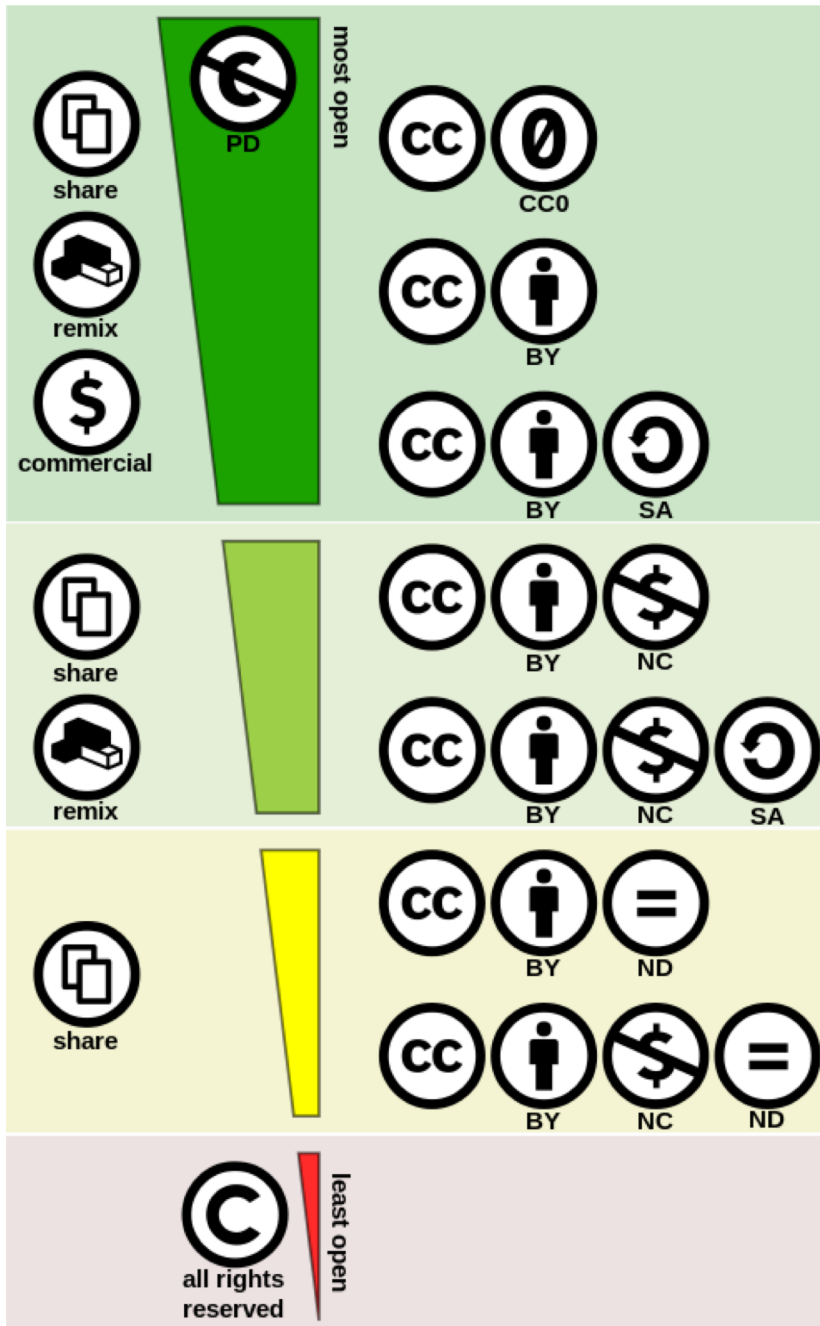
#4 Policies for reuse, redistribution: Prompts

- Will you permit reuse or redistribution? Will you allow commercial use?
- How will you make your data available for re-use?
- What license(s) might you use to guide reuse of your data?



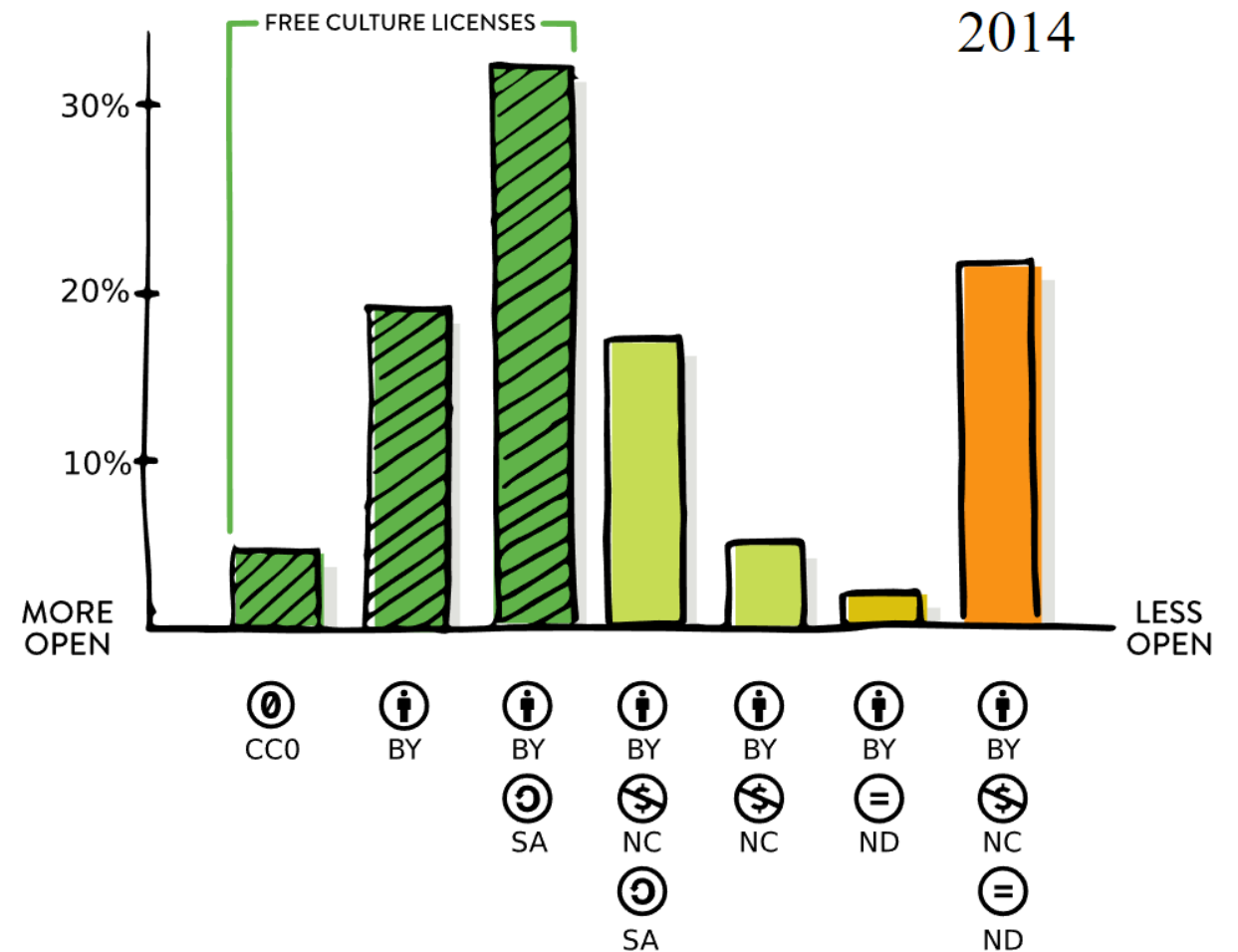
<https://wiki.creativecommons.org/wiki/Data>

[Pick a License, Any License](#) (J. Atwood, on software)



- [Choose](#) a license; [Creative Commons](#)?

- What restrictions do you want or need to place?



#5 Plans for archiving, preservation: Prompts

- What data will be preserved for the long term? For how long?
- Where will data be preserved?
- What data transformations need to occur before preservation?
- What metadata will be submitted with the datasets?



Where will data be preserved?

- Browse [database of subject-specific repositories](#)
- Use [Repository Finder](#)
- Use [Dash](#) ([or Dryad](#))
- Whatever suits your needs (& any privacy requirements, funder needs)
- Explain your decision



Tools for Writing a DMP

- Use [DMPTool](#)
- [Make an appointment with a Data Curation Expert through LibCal](#)
- Online library [resources & links](#)

Welcome

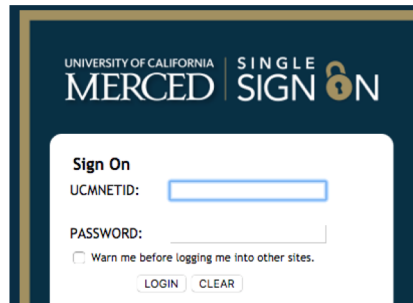
Get started

Create data management plans that meet institutional and funder requirements.



Features of DMPTool

- UCM is affiliated; can use your sign-on



- Connect your [ORCID ID](#)
- View plans shared by UCM
- Templates of funders (NSF)
- Work with collaborators
- Step-by-step prompts w/ resources
- Download your plan



[Register for an ORCID ID!](#)

Discussion

Contact Us



Melinda Laroco Boehm, Ph.D.
Senior Research Development Officer
Office of Research Development
mboehm2@ucmerced.edu
(209) 382-4301



Emily S. Lin
*Head, Digital Curation and
Scholarship
Library*
elin@ucmerced.edu
(209) 658-7146



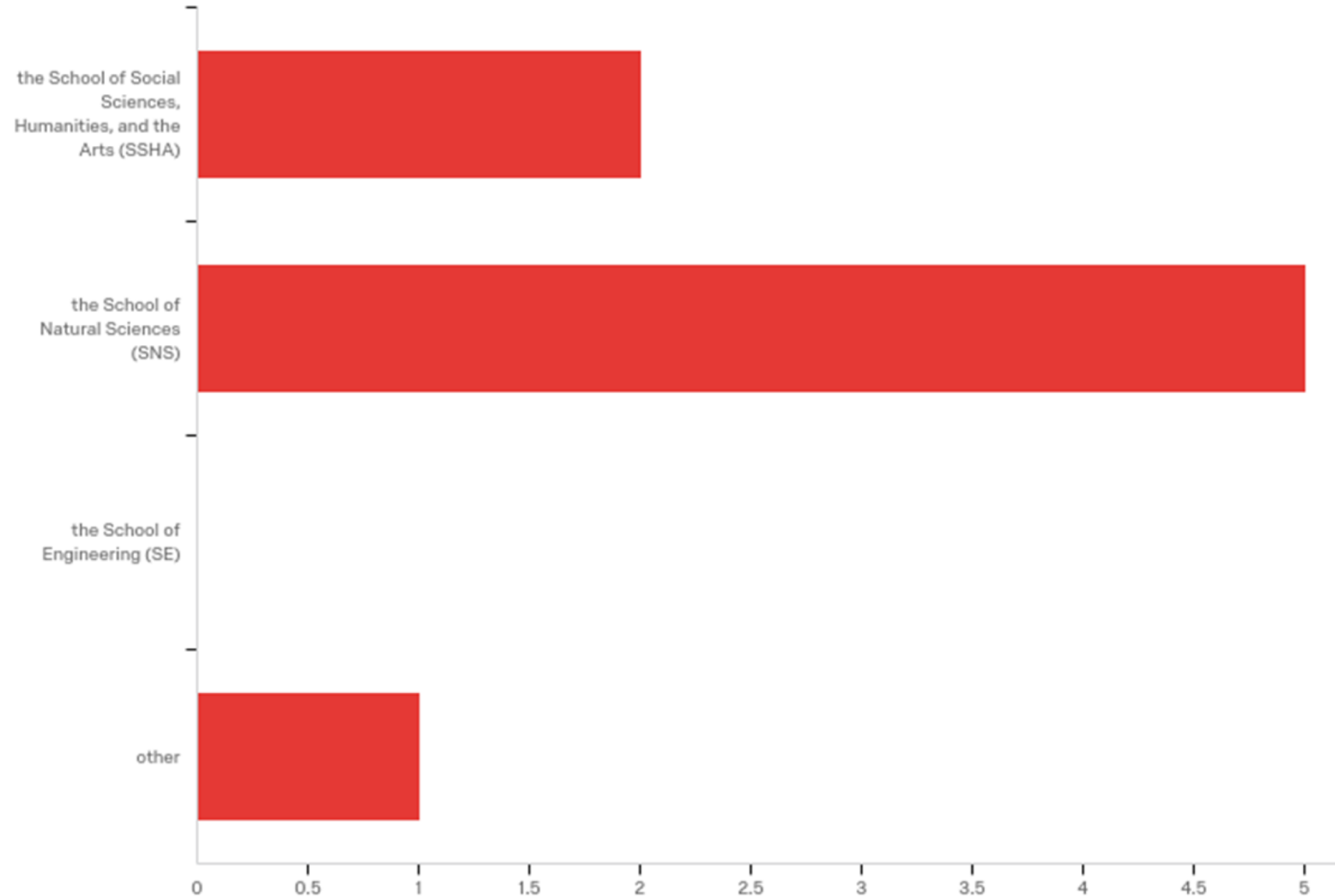
Katie Coburn
*CLIR Postdoctoral Fellow: Data
Curation Specialist*
Library
kcoburn@ucmerced.edu
(209) 205-6178



Jerrold Shiroma
Digital Scholarship Librarian
Library
jshiroma@ucmerced.edu
(209) 756-0237

Introductions

Answer	%	Count
graduate student	12.50%	1
faculty or staff member	75.00%	6
postdoctoral fellow	12.50%	1
other	0.00%	0
Total	100%	8



Survey Results (Your Practices)

What is the primary type of data that you work with?

(Note: Data can be almost ANYTHING, from text files to CSV files to photos to physical items.)

large metagenomic datasets, txt, csv, fasta, fastq

txt, csv and tif files

Audio files (recordings of interviews which are generally transcribed but need to be stored in original recorded format)

Text files (transcriptions of audio recordings as well as public material such as newspaper articles)

Excel files

STATA files

SPSS files

NVIVO files (converted text documents)

Sometimes but rarely pictures

digital data, point clouds, .shp files

.txt/.dat, .vtk, image files, Mathematica graphics output

Publicly open computational package code

images, movies, graphs, csv files (numerical data), text, slides,

Excel files, evaluations, surveys, in person feedback, and observations.

Survey Results (Your Practices)

On a regular basis (daily, weekly, or monthly -- whatever metric makes sense), approximately how much data do you work with?

daily - 10G

Daily

No clue how to begin to answer this. Not a ton - certainly not "big data".

In recent studies, I have worked with:

- Transcribed interviews of 15-45 minutes (the same 24 interview transcriptions in Word, Excel, and NVIVO formats)
- Quantitative data from the same 24 participants' answers to a closed-ended survey
- Excel and STATA files containing responses from closed-ended survey approximately 300 participants (15-20 minute survey)

Up to 100GB/week

When a paper is published

10GB/week

Every day, I review and monitor spreadsheets and programming evaluations

Survey Results (Your Practices)

How do you currently organize your data?

(That is, do you use Dropbox, Box, a physical server? How do you name your files? Etc.).

Data is organized on lab computers and personal laptop in folders/ sub folders by date and experimental technique. File names incorporate date and sample type.

Dropbox for me

Box for my team

I have a complicated file structure that makes sense to me but seems to confound students. I also name my files and save as with new name when I make a major revision. I understand this is also confusing to students who are used to working on online files that autosave and autotrack all changes but I can't wrap my brain around that.

Box, cloud server

Computer cluster storage + external hard drives.

Files named by date, simulation run-label, and a short data type descriptor.

Some data stored on Dropbox.

Github

box for each student

Box, Webforms, spreadsheets

Survey Results (Your Practices)

What is your data backup strategy?

multiple cloud services and a physical storage

All data is saved in three locations: lab computer, personal laptop and a flash drive

All of my work is on Dropbox folders that are auto-synced on my laptop and the cloud.
I also have Crashplan installed and I assume it runs.
Work that my students and staff do is required to be on UC Merced Box.com and they should be auto-syncing

when a step is complete we store them in our cloud server

Redundant external hard drives and computer cluster storage. Some data stored on Dropbox.

Github is an online platform (meaning that a certain level of reliability is expected).

box, crash plan

None.

Survey Results (Your Practices)

What are your plans for data sharing (if you are able to/interested in sharing your data)?

dash!

More journals and funders are requesting this so I'd like to figure out how that would work for the kind of data I have.

At the moment we have digitally published two collections with UCSD

Box, Github

People can download codes freely from Github

none - want to share all data across the group

NA