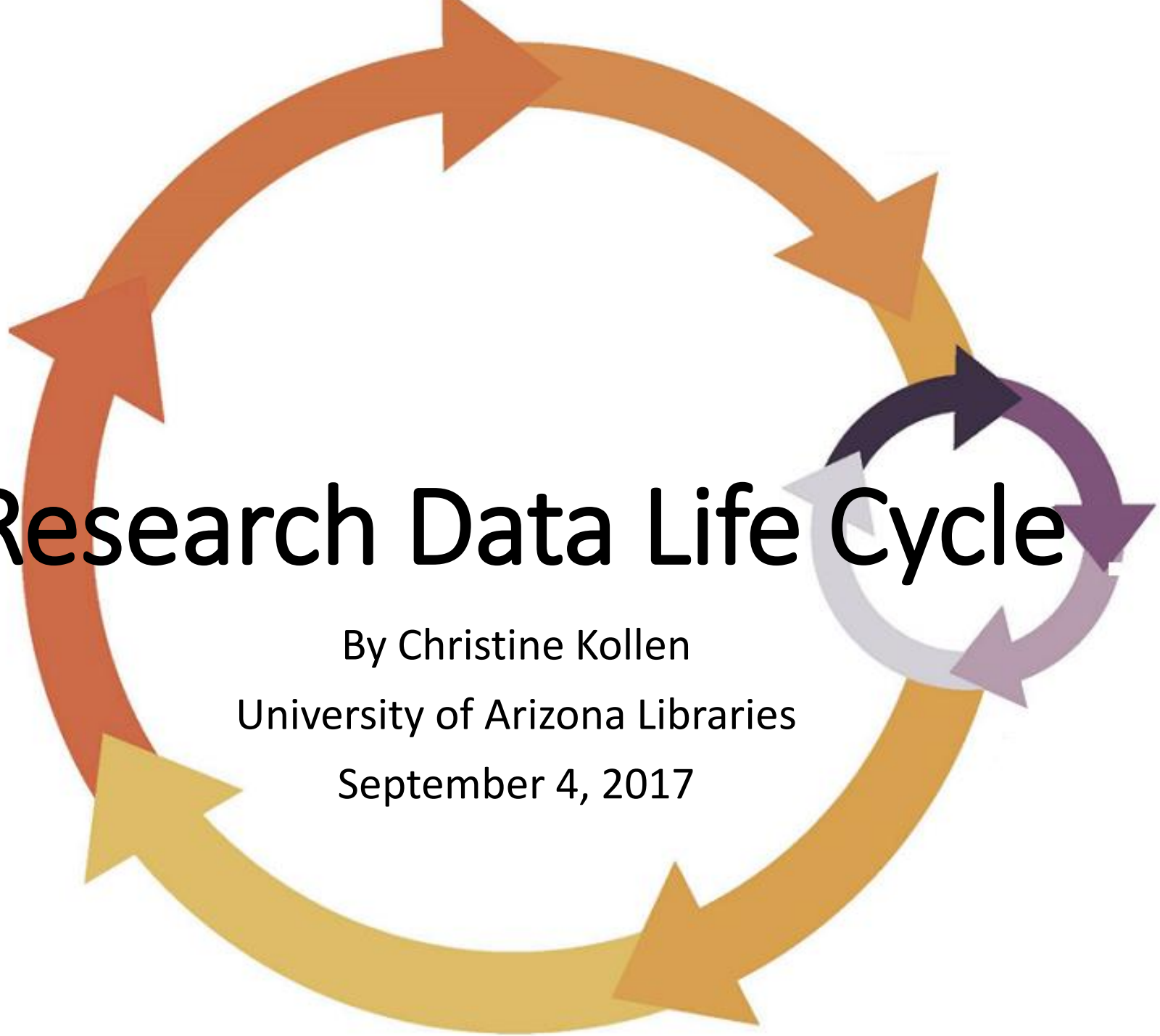


Research Data Life Cycle

By Christine Kollen

University of Arizona Libraries

September 4, 2017

Agenda

- Introduction
- Ready Set, Data
- Issues in Research Data Management
- Research Data Life Cycle
- Data Sharing and Access - benefits
- Long-lived Data
- Metadata and Documentation
- Data Management Plans
- What are publishers and funders saying about data?

Introduction

- Workshop will cover main components of the research data life cycle
- Some of the content, exercises, and discussion topics were taken from ANDS 23 (Research Data) Things program and UCSD Library 23 (Research Data) Things program

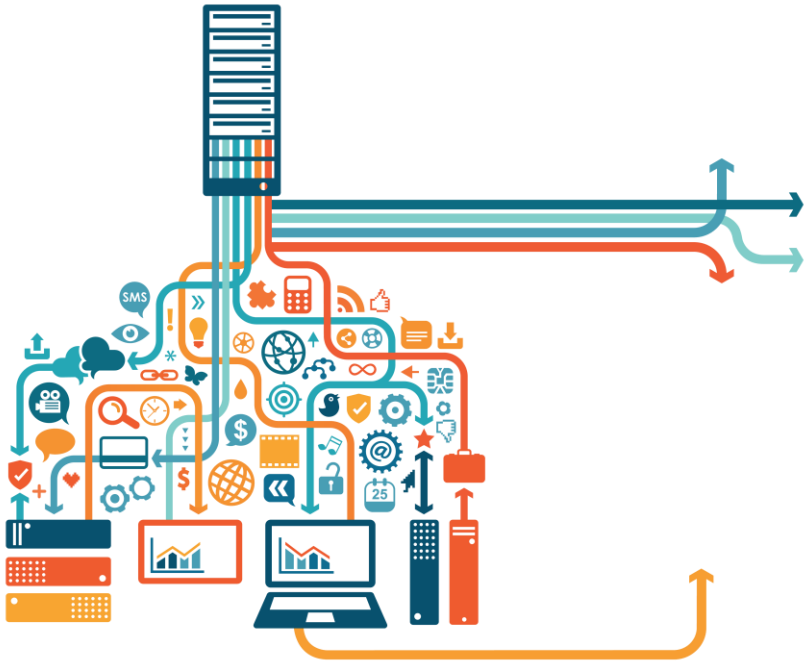
- Introduce yourself – what is your name, institution, and position? What do you hope to learn from this workshop?

Ready Set Data!

What is Research Data?

"the recorded factual material commonly accepted in the scientific community as necessary to validate research findings, but not any of the following: preliminary analyses, drafts of scientific papers, plans for future research, peer reviews, or communications with colleagues."

(OMB Circular 110).



What is Research Data?

Observational data – captured in real time, usually irreplaceable

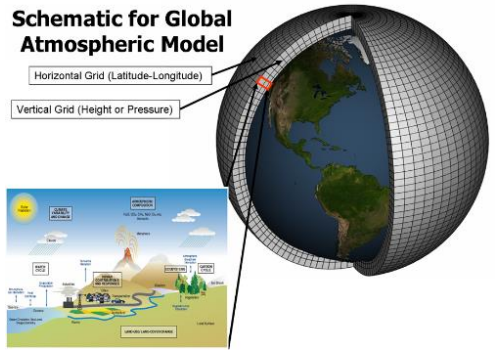
- Sensor readings
- Images of the physical world
- Survey data
- Telemetry

Experimental – from lab equipment, reproducible but expensive

- Gene sequences
- Images of water flowing through a flume
- Chromatograms

Simulation – test models, models and metadata more important than raw data

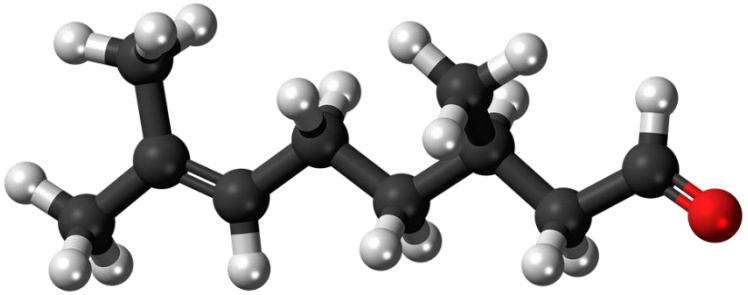
- Climate models
- Economic models



What is Research Data? (continued)

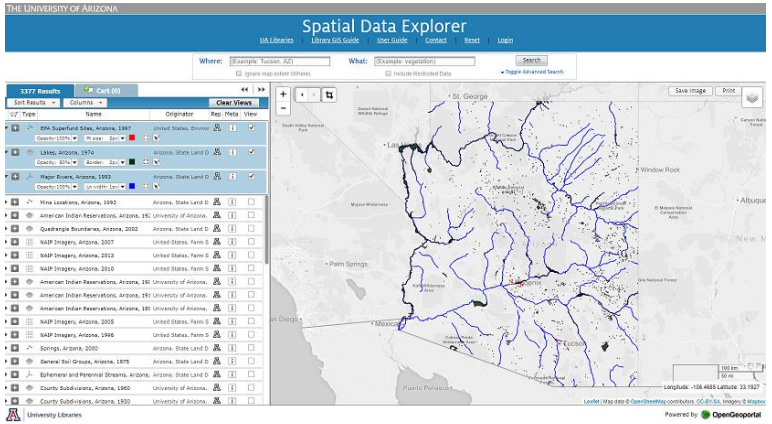
Derived/Compiled – reproducible, but expensive

- Text and data mining
- Compiled databases
- 3D models



Reference or canonical – collection of smaller datasets

- Gene sequence databanks
- Chemical structures
- Spatial Data portals, such as Spatial Data Explorer



Exercise 1 -- Ready Set Data

Go online and look at one of the following repositories:

- Harvard DataVerse - https://dataverse.harvard.edu/
- Global Change Master Directory – http://gcmd.nasa.gov -- Search for data
- Dryad Digital Repository -- http://datadryad.org/
- NeuroMorpho – http://neuromorpho.org – Search metadata or keyword
- Qualitative Data Repository -- https://qdr.syr.edu/ -- Discover → Search Data

Questions?

1. How does this data differ from what you are familiar with? By format, size, access?

2. Does the collection have a code book or data dictionary? Other documentation?

3. Explore the metadata representing the collection. Besides the title and description, what other elements are described?

Issues in Research Data Management

YouTube video from NYU Health Sciences Library – what happens when a researcher does not manage their data (4:40 minutes)

Data Management Best Practices

UA Data Management Resources on best practices:

Data Organization - http://data.library.arizona.edu/data-management-tips/data-organization

Discussion

1. Are there file naming conventions in your discipline?
2. It is important to share data in a non-proprietary format that uses open data standards. Which of the following formats are preferred when sharing data?

Microsoft Word	PDF/A
Microsoft Excel	CSV (comma-separated values)
GIF/JPEG	TIFF

The Research Data Management Lifecycle

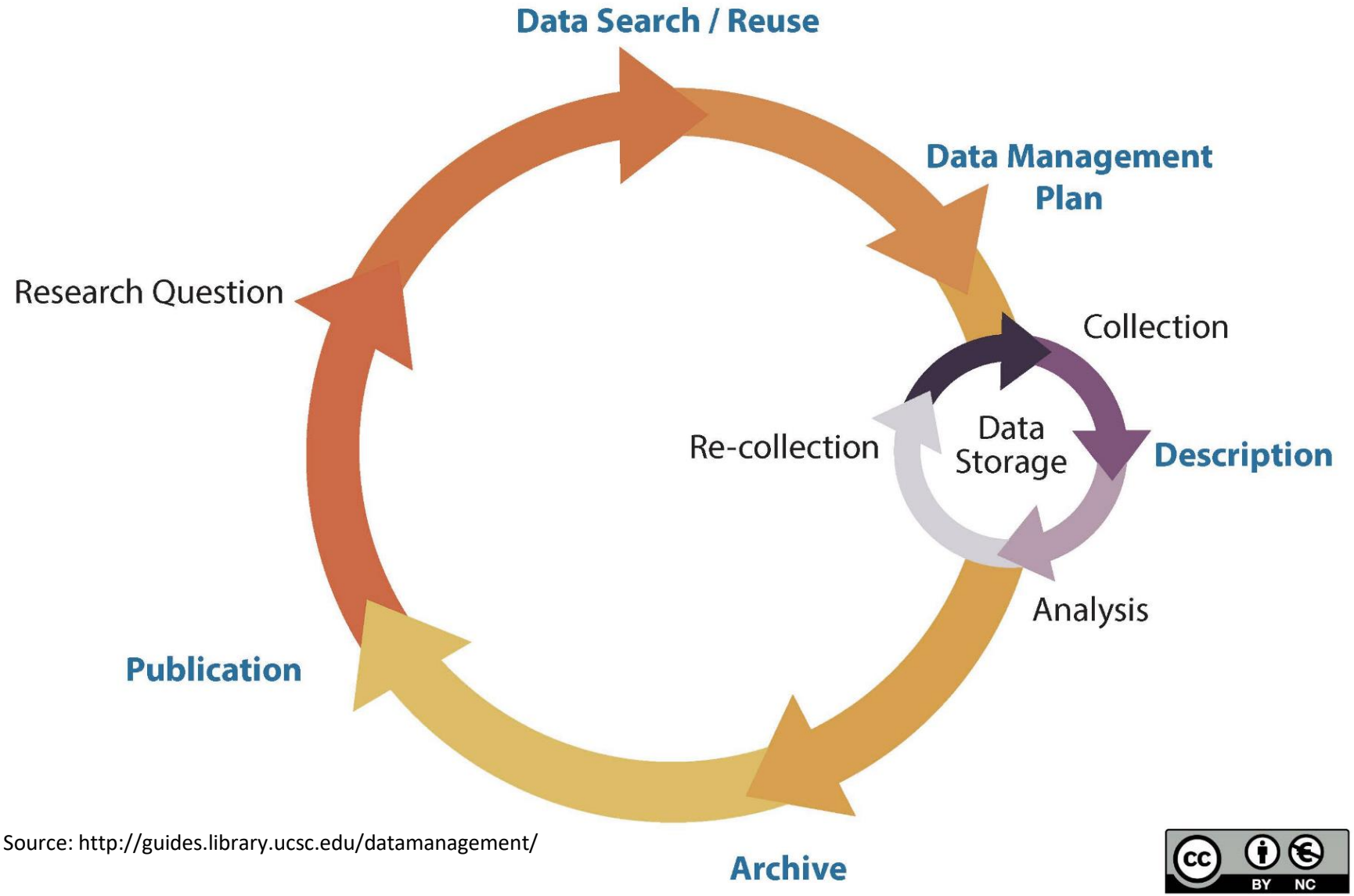


Image Source: http://guides.library.ucsc.edu/datamanagement/

Research Data Life Cycle

- Data often has a longer lifespan than the research project that creates them
- Other projects may analyze or add to the data; reused by other researchers
- Funders and journal editors are beginning to require that researchers make the underlying data accessible for the long term

Finding data repositories

"The resulting data ecosystem, therefore, appears to be moving away from centralization, is becoming more diverse, and less integrated…" (M.D. Wilkinson)

- Numerous repositories
- Scales range from institutional (campus repositories) to globally-scoped repositories
- Accept a wide range of data types and formats
- Little attempt to integrate or harmonize deposited datasets
- Few requirements for the descriptors of a dataset

Data to use in your Research

What does the data repository landscape look like? Let's explore the Registry of Research Data Repositories, http://re3data.org to explore data repositories by academic discipline.

Go to http://re3data.org and click on Browse > By Country > click on Mexico

Look at a few – EcoCyc, California Coastal Atlas, International Maize & Wheat Improvement

- What content types are available?
- How do you access the database?
- Is it available for data upload or are there restrictions?
- Do they assign a persistent identifier (such as a DOI)? What metadata schema do they use?
- What metadata schema do they use?

Sharing Data

Open Data

- Is freely available on the internet
- Permits any user to download, copy, analyze, re-process, or use for any other purpose
- Is without financial, legal or other technical barriers

Benefits

- Accelerates the pace of discovery
- Grows the economy
- Improves the integrity of the scientific and scholarly record
- Becoming recognized by many in the research community, important part of the research enterprise

Sharing Data (continued)

- Shared Data – data available to a specific group of people for a specific purpose.
- Closed Data – data that only those within an organization can see

Sharing Data – attitudes of researchers

Wiley's Researcher Data Insights Survey (2016) found that:

- Globally – 69% share their data; 31% do not
- Data is shared:
  - 41% as supplementary material in a journal
  - 29% - personal institutional or project website
  - 25% Institutional data repository
  - 10% Disciplinary data repository
  - 6% General purpose data repository (figshare, Dryad)
- Motivations for sharing – increase impact & visibility, public benefit, transparency and re-use, journal requirement

Source: Researcher Data Sharing Insights - https://hub.wiley.com/community/exchanges/discover/blog/2017/04/19/open-science-trends-you-need-to-know-about

Data Restrictions and Protection

- Appropriate protection of privacy
- Security of data
- Confidentiality/HIPAA or FERPA
- Intellectual Property and Copyright
- Embargo
- Other rights or requirements?

Long-lived Data: Curation and Preservation

- Data Curation - “active and ongoing management of data through its life cycle of interest and usefulness to scholarship, science, and education…”
  (UI Graduate School of Library and Information Science)
- Data Preservation – “series of managed activities necessary to ensure continued access to digital materials for as long as necessary.” (Digital Preservation Handbook)

Data curation is the process of making data FAIR:

- **Findable**
- **Accessible**
- **Interoperable**
- **Reuseable**

Archiving for preservation and long-term access

What happens to data once a research project is complete?

- How long should the data be retained?
- What format?
- Migration schedule?
- Plans for archiving other research products, physical samples and derivatives

Metadata and Documentation

- Metadata is structured information - describes content, quality, format, location and contact information
- Similar to descriptive cataloguing of library resources
- Metadata schema are sets of metadata elements (or fields) for describing a particular type of information resource
- Most familiar those used in library catalogs and publications repositories such as MARC and Dublin Core

Metadata and Documentation (continued)

- Also important to provide documentation so that your data will be understood and interpreted correctly
- How your data was created, context for the data, structure and any manipulations or analysis that have been done
- Can be as basic as a readme text file
  - See Easy Data Management: Add README.txt file - http://databrarians.org/2016/05/easy-data-management-add-a-readme-txt-to-your-project-folders/

Exercise 2 - Metadata and Documentation

Look at a good quality metadata record:

Long-term variation of surface phytoplankton chlorophyll a in the Southern Ocean during 1965-2002

Questions:

1. Why do you think this record is considered high quality?
2. What metadata fields help discovery and reuse of the data?
3. Why is metadata often neglected?

Data Management Plan (DMP)

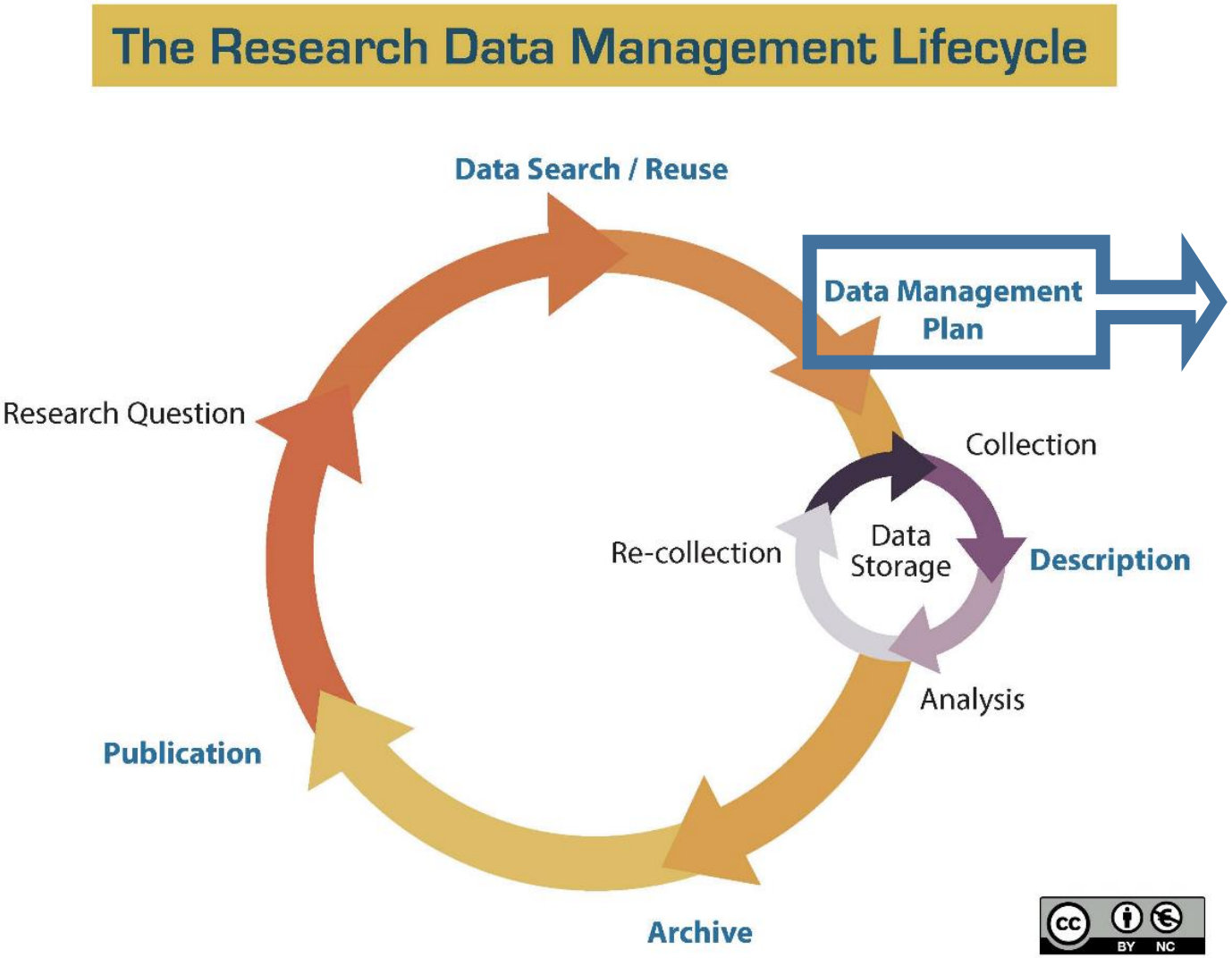


Image Source: http://guides.library.ucsc.edu/datamanagement/

Documents the lifecycle of your data and provides details on data collection for storage, access, sharing, and reproducibility of your results.

This can ensure the availability and accessibility of your research results after your project is complete.

Exercise 3 – Data Management Plans (DMP)

Go DMPTool at https://dmptool.org

Questions:

1. Review 2-3 Data Management Plan samples. You will find them under Public DMPs on the main screen.
   What are 2 to 3 pieces of information that are essential to a DMP? Why?
2. Log in to your DMPTool account and review the NIH Generic and the NSF Generic templates. What are the strengths and weaknesses of the NIH template and the NSF template?

What are publishers and funders saying about data?

Data sharing policies are becoming increasingly common

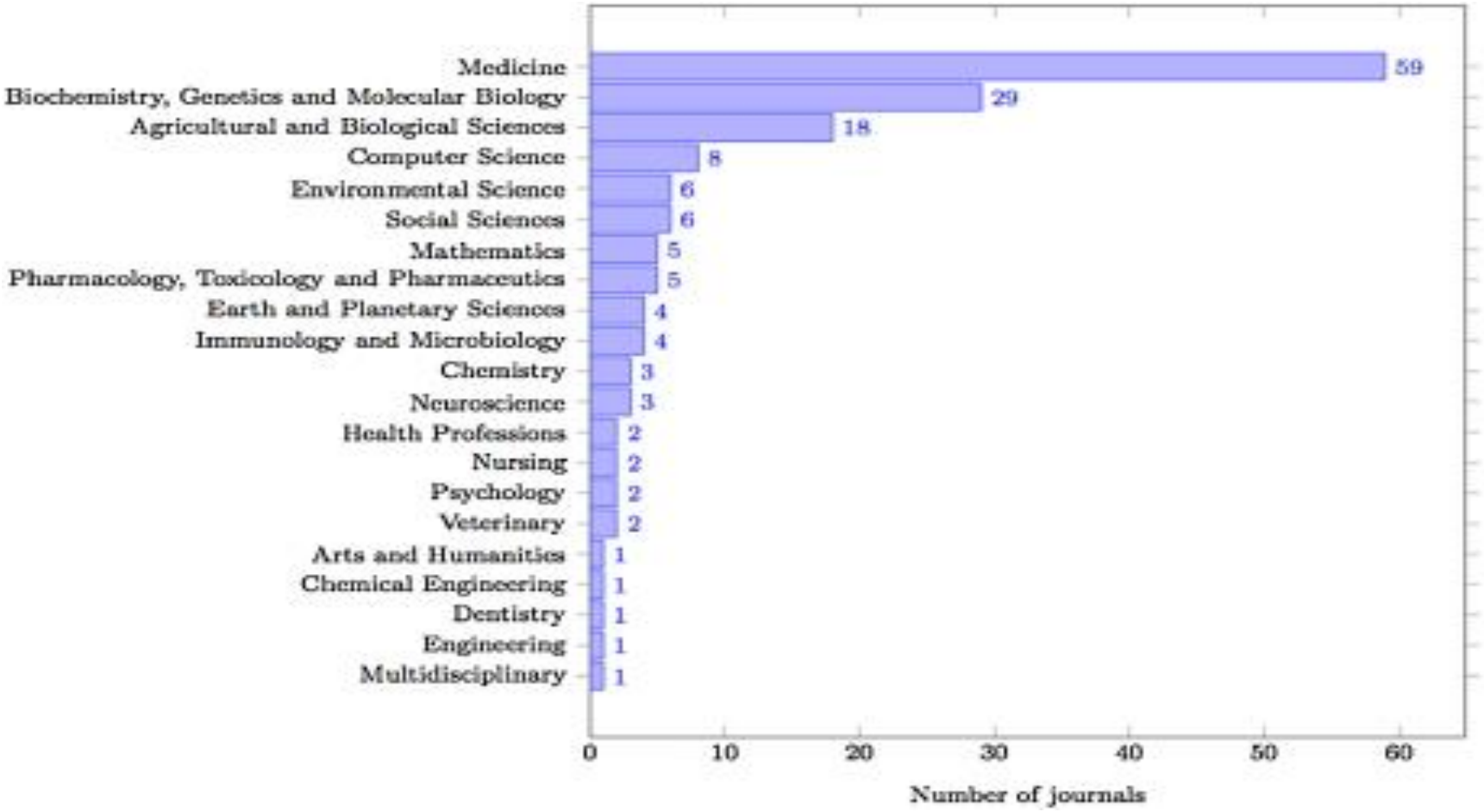
- Journal editors are asking authors to make the data underlying an article available. In addition, new forms of data publishing are emerging → data journal
- Funders, US federal agencies and some foundations are requiring researchers to:
  - Submit a data management plan as part of the grant proposal or funding request
  - Deposit dataset(s) supporting published research results in a public data repository

What is a data journal?

- A journal that publishes “data papers” describing datasets in rich detail so they can be found and used by other researchers.
- The data paper contains a link to the entire set of data, which is usually published in a public data repository. The dataset has a persistent identifier, usually a DOI.
- Two types of journals: hybrid and pure
  - Hybrid journals publish regular papers and data papers
  - Pure journals publish only data papers
- Pure data journals “explicitly provide peer review prior to ‘publication’ of the data.” But the quality of that peer review varies. (Todd Carpenter)
- Hybrid journals may or may not peer review the dataset, although they usually review the accompanying metadata for completeness.

Data Journals

116 data journals published by 15 different publishers, by subject. (Figure 2)



Candela, L., Castelli, D., Manghi, P. and Tani, A. (2015), Data journals: A survey. J Assn Inf Sci Tec, 66: 1747–1762. doi:10.1002/asi.23358

Examples of Data Journals

- *Geoscience Data Journal*
  http://onlinelibrary.wiley.com/journal/10.1002/(ISSN)2049-6060

- *Earth System Science Data*
  http://earth-system-science-data.net/

- *GigaScience* – big data from life & biomedical sciences; open-access, open-data, open peer-review
  https://academic.oup.com/gigascience

- *Scientific Data (Nature.com)*
  http://www.nature.com/sdata/

Journal or Funder Recommended Repositories

- Nature.com recommended: https://www.nature.com/sdata/policies/repositories
- PlosOne recommended: http://journals.plos.org/plosone/s/data-availability#loc-recommended-repositories
- Biosharing.org: https://biosharing.org/
- NIH-supported: https://www.nlm.nih.gov/NIHbmic/nih_data_sharing_repositories.html
- UA Library Data Management Data Repositories

Exercise 4 – Publisher and data repository policies

Choose one of the following:

- PLOS One data policy – http://journals.plos.org/plosone/s/data-availability
  - Review their policy
- Dryad – data repository that integrates data and articles -- http://datadryad.org/pages/journalLookup and look up a journal you are familiar with and see what advice it gives on submitting data and link. Also look at FAQ - http://datadryad.org/pages/faq

Questions:

1. Was the policy you looked at clear?
2. Did you understand what you needed to do?

Conclusion

Important to think through the Research Data Life Cycle

- Identify potential data to reuse
- Develop your data management plan
- Collect, analyze, and reanalyze data, with organized protocols for data management, data storage and back-up
- Archive data – migrate to suitable format, finalize metadata and documentation
- Publication – distribute, share and promote data

Resources

"23 (Research Data) Things." *Australian National Data Service.* August 3, 2017. http://www.ands.org.au/partners-and-communities/23-research-data-things

"23 (Research Data) Things." *University of California San Diego Library*. June 19, 2017. https://ucsdlib.github.io/23-Research-Data-Things/

Carpenter, Todd. "What Constitutes Peer Review of Data? A Survey of Peer Review Guidelines." August 4, 2017. https://scholarlykitchen.sspnet.org/2017/04/11/what-constitutes-peer-review-research-data/

"Data Management Planning Tool, DMPTool." *University of California, California Digital Library, California Curation Center*. July 14, 2017. https://dmptool.org

"Data Management Resources." *University of Arizona Libraries*. July 14, 2017. http://data.library.arizona.edu

*"*Guiding Principles for Findable, Accessible, Interoperable and Re-Usable Data Publishing Version B1.0." *FORCE 11*. July 14, 2017. https://www.force11.org/fairprinciples.

NYU Health Sciences Library. "Data Sharing and Management Snafu in 3 Short Acts." July 14, 2017. https://www.youtube.com/watch?v=66oNv_DJuPc.

"Open/Closed/Shared: the world of data." Open Data Institute. July 14, 2017. https://vimeo.com/125783029

"Open Data." *Scholarly Publishing and Academic Resources Coalition*. July 14, 2017. https://sparcopen.org/open-data/

Vocile, Bobby. "Open Science Trends You Need to Know About." *Wiley Exchanges. April 20, 2017. https://hub.wiley.com/community/exchanges/discover/blog/2017/04/19/open-science-trends-you-need-to-know-about*.

Wilkinson, M.D., *et. al. "*The FAIR Guiding Principles for scientific data management and stewardship". *Scientific Data* 3:160018. https://www.nature.com/articles/sdata201618

Witt, Michael. "23 Things: Libraries for Research Data." *Research Data Alliance*. July 14, 2017. https://www.rd-alliance.org/group/libraries-research-data-ig/outcomes/23-things-libraries-research-data-supporting-output

Questions?

Contact:

kollen@email.arizona.edu or 520-305-0495