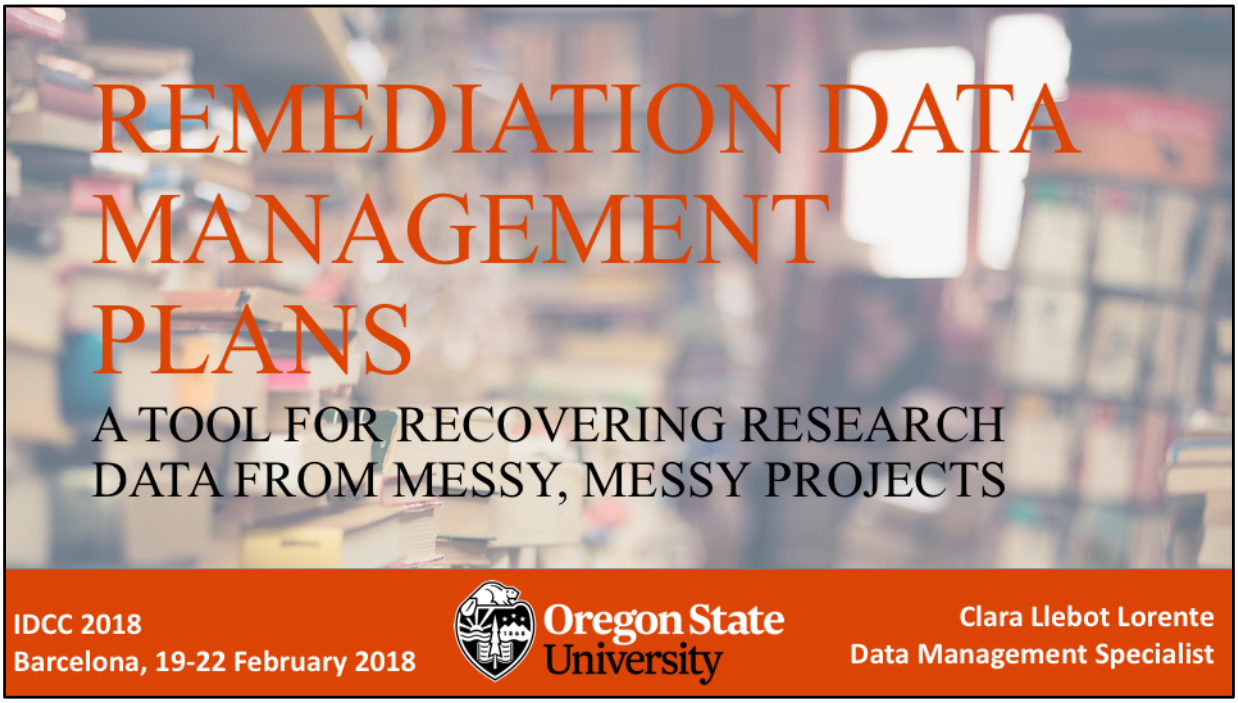


Image credit: by EliFrancis in pixabay.com Published in the public domain.

Image credits: images from the Watershed Research Cooperative. http://watershedresearch.org

Citations:
Goal of WRC from watershedresearch.org
Proposal statements from Souder, J., Hatten, J., Ganio, L. & Bladon, K. (2016). *From chaos to consistency: Moving towards data stewardship and sharing for the watershed research cooperative.* (Project proposal to the Fish and Wildlife Habitat in Managed Forests Research Program)

I started working as a data management specialist at Oregon State University in 2016. Before being a data management specialist I was a postdoc in oceanography, so the first few months were all about learning about the job and about what data management is in the context of this university. It was during this period when Steve, the other data librarian at OSU, asked me to go to a meeting with him. Part of what I do in my job, and what Steve used to do before I started, is to have consultations with researchers who have data management challenges. This was a consultation with a researcher from the Watershed Research Initiative. Is anybody familiar with the Watershed Research Initiative? This project is part of the college of forestry. The goal

of the project, in a nutshell, is to study the environmental effects environmental effects caused by contemporary forest management activities at a watershed scale. There are three watersheds, and they do different treatments in different areas of the forest, and take data of several variables about hydrology, climate, fish, nutrients in the river, sediments, vegetation… They started collecting data in 2002, with no plan for managing their data. It was a very large project, with millions of dollars in funding, and they collected a lot of data. Now, 15 years later, it is time to publish the findings, and combine different datasets to synthesize what they have learned. And they realized, and that was why we were having the consultation, that they do not know what data they have. They don't know what shape it is in. They don't know who to call when they have questions about someone's else's data, because it is not clear who is responsible for what. Most of the data is completely undocumented. More than 50 researchers have participated in the project. Some have left, some have even died. The result of the meeting was an agreement that the Watershed Research Cooperative would pay the library so that the Data Management Specialist (that's me!) would write a data management plan to try to put some order in the mess.

Context of Data Management Plans in the US

- geographically disperse, interdisciplinary, data based science
- Public access to research
- accountability for federal granted projects

EXECUTIVE OFFICE OF THE PRESIDENT
OFFICE OF SCIENCE AND TECHNOLOGY POLICY
WASHINGTON, D.C. 20502

February 22, 2013

MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES

FROM: John P. Holdren
Director

SUBJECT: Increasing Access to the Results of Federally Funded Scientific Research

1. Policy Principles

The Administration is committed to ensuring that, to the greatest extent and with the fewest constraints possible and consistent with law and the objectives set out below, the direct results of federally funded scientific research are made available to and useful for the public, industry, and the scientific community. Such results include peer-reviewed publications and digital data.

Scientific research supported by the Federal Government catalyzes innovative breakthroughs that

"The Office of Science and Technology Policy (OSTP) hereby directs each Federal agency with over $100 million in annual conduct of research and development expenditures **to develop a plan to support increased public access to the results of research funded by the Federal Government.** "

Image credit: Jasiek Krzysztofiak in nature.com
http://www.nature.com/news/specials/global/index.html

Context DMPs

Elements of a Data Management Plan

1. Data types and formats
2. Roles and responsibilities
3. Data organization
4. Data documentation
5. Storage, backup and data security
6. Archiving, preserving and sharing

A DMP is a **plan**. Its role is to **prevent** bad data management practices and encourage good data stewardship.

Context DMPs

OREGON STATE UNIVERSITY 4

Image credits:
Garage: by Pexels in pixabay.com Published in the Public Domain.
Question mark: by TeroVesalainen in pixabay.com. Published in the Public Domain.

But what about data that has already been collected? Data that is already messy, like the WRC? We don't want these data to end up in a drawer of a garage of a PI.

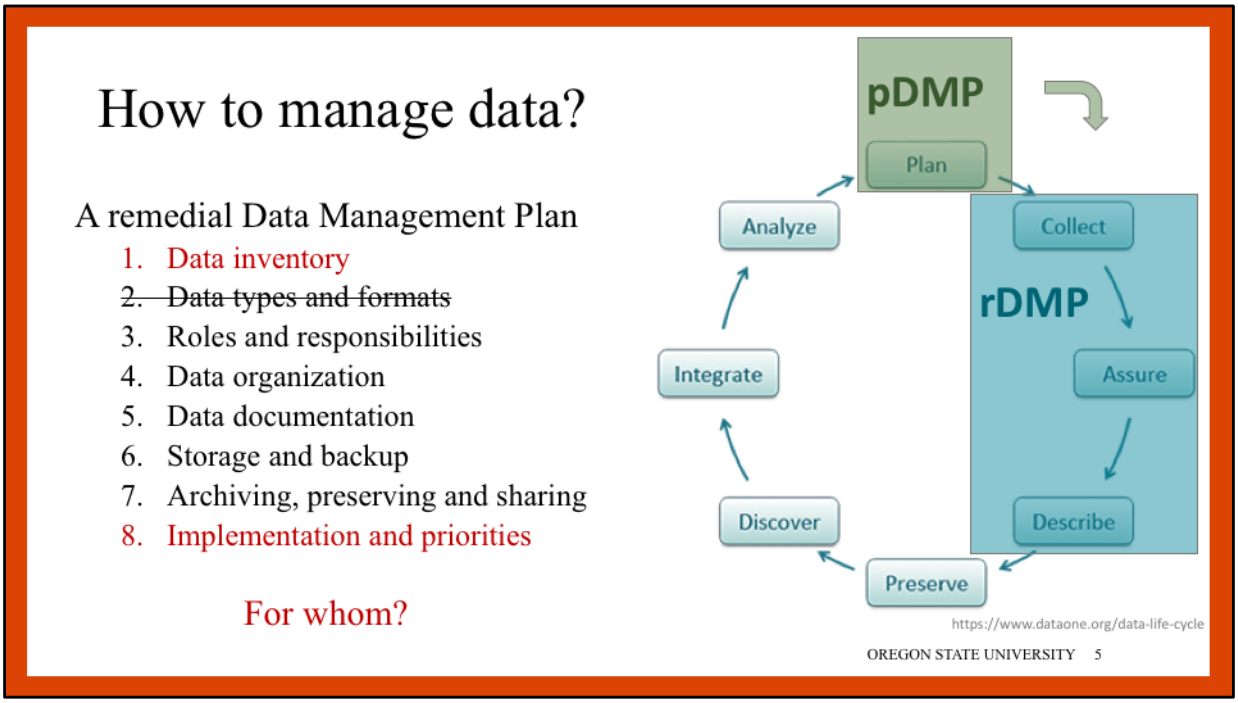


Image credit: Data life cycle from DataONE, https://www.dataone.org/data-life-cycle

What I propose is that the structure of a DMP can be useful for projects that are not in the planning stages. A DMP can be useful for projects that are in the intermediate or final parts of the project. Most or some of the data has already been collected; There may or may not have been quality control; There may or may not have been documentation; There may or may not have been a thorough analysis of the data. Most likely the researchers are in the synthesis phases of the project, and are starting to realize that bad data management practices are a problem.

If we write a remedial data management plan we need to change the structure of a DMP a little bit: add a data inventory section that will substitute the data types and formats. Add an implementation and priorities section. And figure out who we are writing a DMP for.

pDMP vs rDMP

Differences pDMP rDMP

1. Audience and goal
2. Data inventory
3. Implementation strategy

Commonalities pDMP rDMP

1. Data organization
2. Data documentation
3. Sharing and preservation

Red background: case study.
Blue background: general considerations about rDMP.

2. Diff pDMP rDMP: Data inventory

Managing strategies | Datasets/ Data Collections | Disk analyzer

1. Relational databases: 3 relational databases managed by Data Manager
   - Climate
   - Hydrology
   - Temperature
   - Nutrients
   - Dissolved oxygen
   - Sediment
2. Trask tabular data: use of LTER templates
3. Fish database: managed by Researcher acting as Data Manager
4. Other digital data
   - Macroinvertebrates
   - Geospatial
5. Physical samples

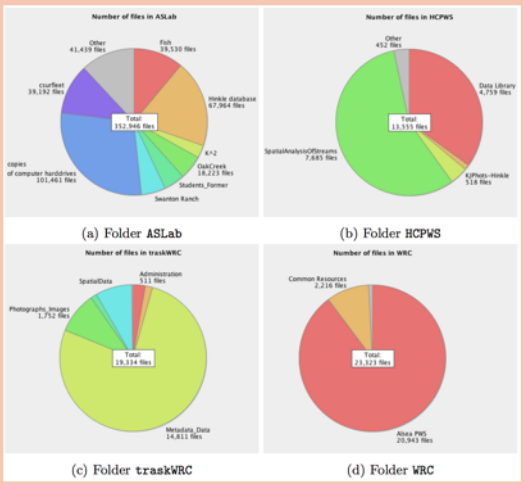
2. Diff pDMP rDMP: Data inventory

Total of 409158 files and 344 GB of data

- 50000 excel files (73GB)
- 52000 images (79 GB)
- 334 databases (17 GB)
- 9000 documents pdf ppt (30 GB)
- 17000 geospatial (22GB)

Mix of

- Data
- Admin
- Personal folders
- Publications
- Backups



(a) Folder ASLab (b) Folder HCPWS

(c) Folder traskWRC (d) Folder WRC

2. Diff pDMP rDMP: Data inventory

- Collect metadata about data
  - Subject
  - Location
  - Responsibilities
  - Manager
  - Versions
  - Formats
  - Documentation
  - Sensitivity
  - Sharing status

- Detail according to the size of the project

Small projects

- File to file?

Medium/large projects

- Management level categories
- Data collection categories
- File analysis (automated).

3. Diff pDMP rDMP: implementation

- **Priority 1**: Clean, document, and preserve in ScholarsArchive@OSU quality controlled datasets.
  - Use article publications as triggers to do the work.
- **Priority 2**: Clean, document and preserve data associated to past publications.
- **Priority 3**: Triage data in shared drive folders.

Challenge: HOW? HOW MUCH? WHO?

rDMP

Something needs to change!

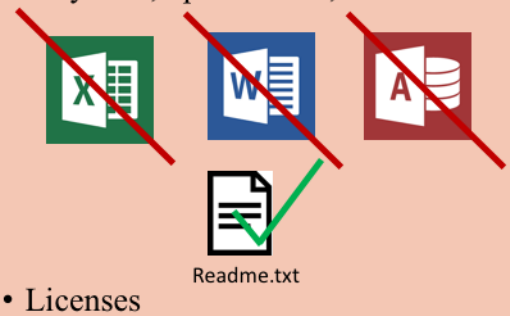
- Priorities
- Resources
- Individuals
- Motivate researchers: carrots or sticks?

Commonalities: Data organization; Documentation

- Levels of data quality:
  - Level zero (L0): raw data downloaded directly from instrument or model.
  - Level one (L1): Raw data in a format that is understandable by the researcher.
  - Level two (L2): Verified data that have undergone quality control
  - Level three (L3): L2 data that have been analyzed to answer specific research questions. Typically used for figures in a publication.
- Organization in datasets
- Folder structure
- File naming strategy
- Metadata standard: Ecological Metadata Language
- Define scenarios (minimum & ideal) for data documentation.

Commonalities: Preservation strategy

- Preservation in OSU's institutional repository ScholarsArchive@OSU.
- Tidy data, open format, documented



Readme.txt

- Licenses

Conclusion

DMP useful structure

Some differences that need to be taken into account.

Also useful for other stages in the data life cycle? Maybe an archival Data Management Plan?

Clara Llebot Lorente
Data Management Specialist
clara.llebot@oregonstate.edu

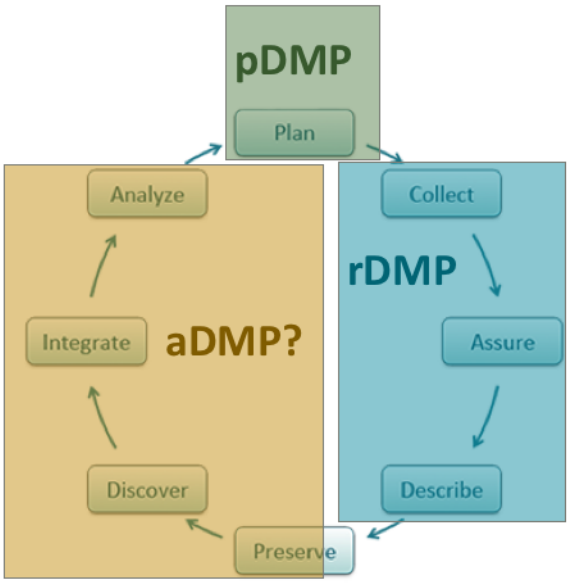


Image credit: Data life cycle from DataONE, https://www.dataone.org/data-life-cycle

Clara Llebot Lorente | Data Management Specialist
clara.llebot@oregonstate.edu

ResearchDataServices@oregonstate.edu
http://bit.ly/OSUData