Data Lifecycle and DMP

Fundamentals of Data Management

Includes material from
DataONE Education Module: Data Management Planning

DataONE. Retrieved Jan, 2014. http://www.dataone.org/sites/all/documents/L03_DataManagementPlanning.pptx

Dr Rob Baxter
Software Development Group Manager, EPCC
r.baxter@epcc.ed.ac.uk
+44 131 651 3579 | +44 7971 437749

Course outline

- What do we mean by “data lifecycle”?
- Why is this a useful concept?
- What is “data management planning”?
- What is it used for?
- How does it fit within the data lifecycle?

- After completing this lesson, you should be able to:
  - Describe the research data lifecycle
  - Understand the importance of preparing a data management plan
  - Identify the key components of a DMP

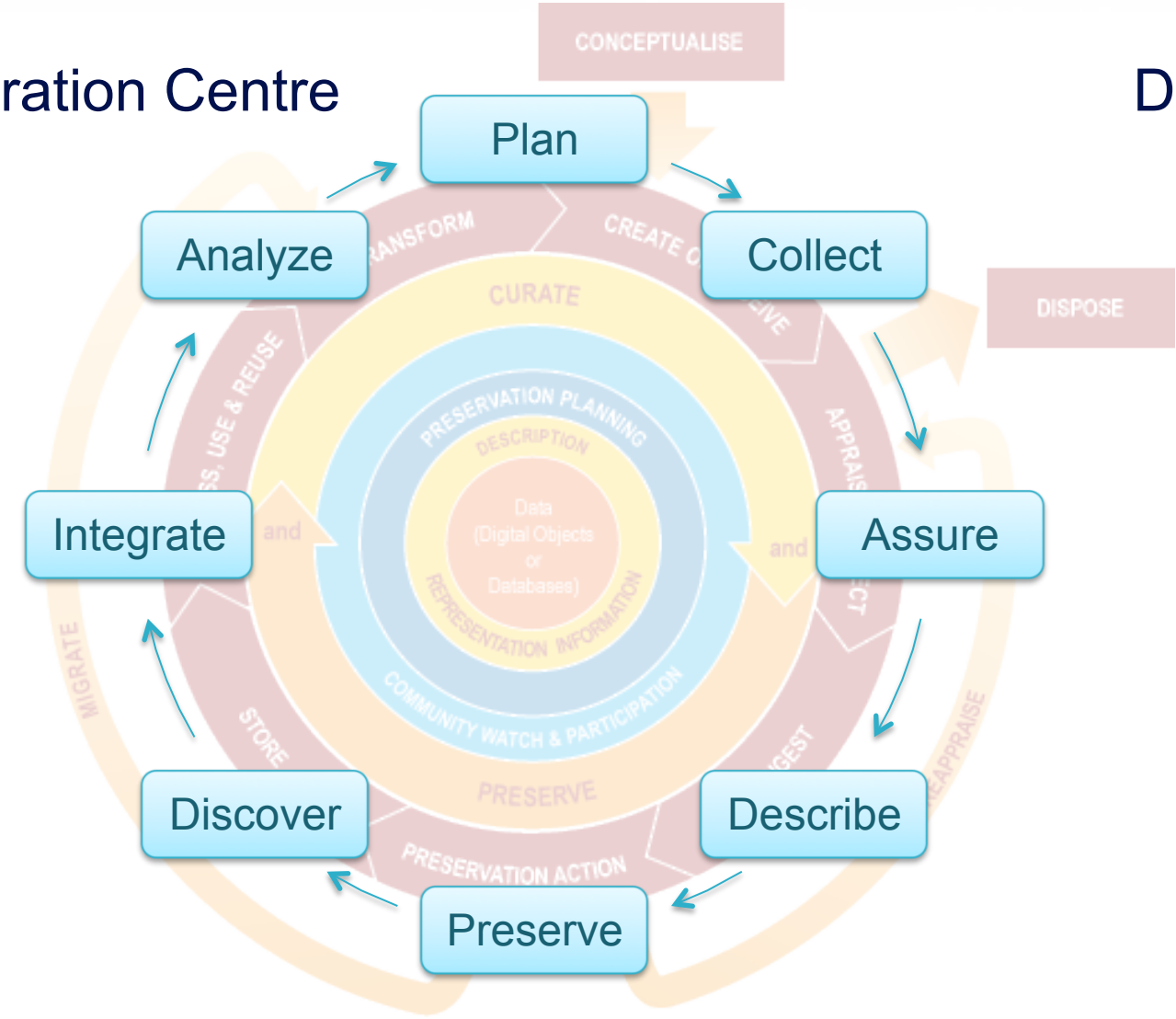
What do we mean by data lifecycle?

- It's a way of thinking about the different stages through which a digital object (file etc.) passes from creation to storage (or deletion)
- At each stage the DO can be regarded slightly differently, or used in a slightly different way
  - this makes you think about how to manage it accordingly
- The data lifecycle maps well onto a typical research experiment timeline
  - create or measure something
  - analyse and study it
  - file it away somewhere ready for next time

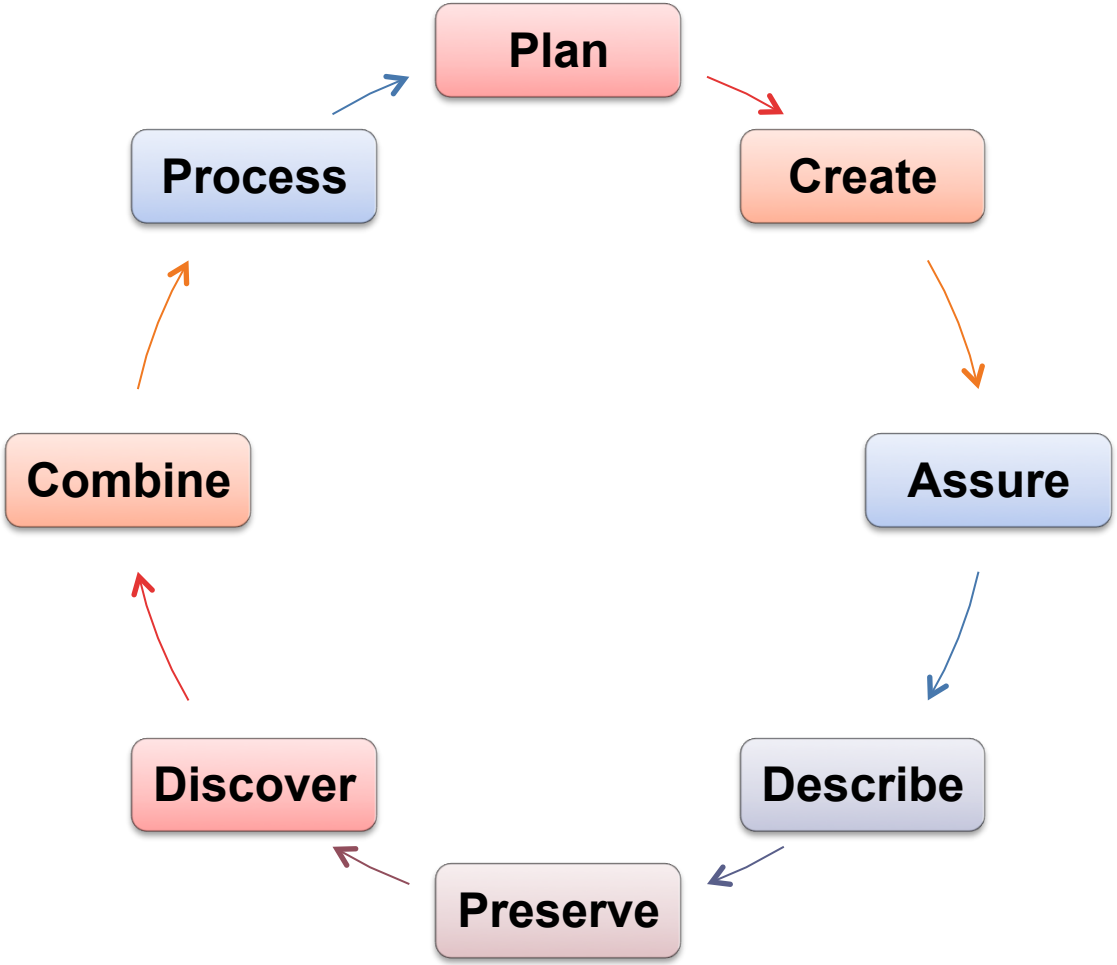
Data lifecycle models

Digital Curation Centre

DataONE



A slight hybrid



Data lifecycle components

- Create
  - Observe, measure, generate by simulation
  - This is your raw data, part of your “laboratory notebook”
  - Organise it from the start – choose standard formats (see later)
- Assure
  - Validate, calibrate
  - Checking the correctness of the methods used to create data
    - In simulation terms, testing your code properly!
  - Perhaps recording the calibration methods used
    - If raw data need to be corrected for instrument bias before they can be interpreted correctly, **you need to record this**!

Data lifecycle components

- Describe
  - "SAM1 = the level of expression of gene…"
  - Use meaningful variable names (**not** SAM1, SAM2…)
  - Record **units** (metres, millimetres, parsecs?)
  - Record information needed to **interpret** the data in 1, 10, 100 years
  - Use metadata standards! (q.v.)
- Preserve (store long-term)
  - If data get this far, they are becoming part of the scientific record
  - Store them carefully. Think about
    - Backup and replication
    - Accessibility
    - Keeping data and metadata together

Data lifecycle components

- Discover
  - If you can't find data, they may as well not exist
  - How should you make your data discoverable by others?
  - How can you find other researchers' data that might be useful?
  - Description and accessibility are key
- Combine
  - Combining, integrating, merging data to create new insights
  - Good metadata are essential, as are good tools
  - And an appreciation of licensing conditions
- Process
  - Applying computer software to create "new data from old"
  - Analysis of digital sensor data; simulation input; re-analysis of integrated third-party data

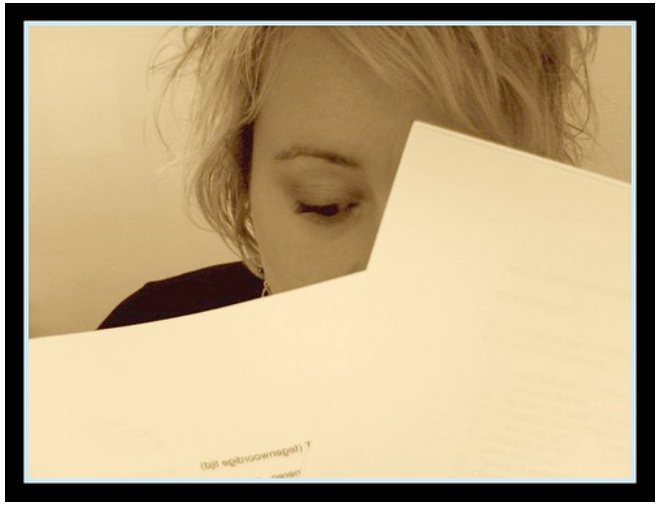
Data management planning

Plan

- First part of the lifecycle
- Lays down plans for all the rest
  - What data will I create?
  - How will I describe them?
  - How will I store them?
  - Will I publish and share them? If not, why not?
  - How will others find them?
- What is a Data Management Plan?
  - A formal document that captures the above
  - Outlines what you will do with your data **during** and **after** you complete your research
  - Ensures your data are safe for the **present** and the **future**

Why prepare a DMP? (1)

- Save time
  - Less reorganization later
- Increase research efficiency
  - Ensures you and others will be able to understand and use data in future
- And, increasingly, you have to!
  - Research funding agencies now ask for DMPs for most proposals



Why prepare a DMP? (2)

- Easier to preserve your data
- Prevents duplication of effort
- Can lead to new, unanticipated discoveries
- Increases visibility of research
- Makes research and data more relevant

- And did we mention it's a funding agency requirement?

Components of a general DMP

1. Information about data & data formats
2. Metadata content and format
3. Policies for access, sharing and re-use
4. Long-term storage and data management
5. Budget

1. Information about data & data format

1.1 Description of data to be produced

- Experimental
- Observational
- Raw or derived
- Physical collections
- Models and their outputs
- Simulation outputs
- Curriculum materials
- Software
- Images
- Etc…

1. Information about data & data formats

1.2 How data will be created or acquired

- When?
- Where?

1.3 How data will be processed

- Software used
- Algorithms
- Workflows

CC image by Ryan Sandridge on Flickr

1. Information about data & data formats

1.4 File formats

- Justification
- Naming conventions

1.5 Quality assurance & control during sample collection, analysis, and processing

1. Information about data & data formats

1.6 Existing data

- If existing data are used, what are their origins?
- Will your data be combined with existing data?
- What is the relationship between your data and existing data?

1.7 How data will be managed in short-term

- Version control
- Backing up
- Security & protection
- Who will be responsible?

2. Metadata content & format

A quick definition of metadata:

- “Data about data”
- Documentation and reporting of data
- Contextual details: Critical information about the dataset
- Information important for using the data
- Descriptions of temporal and spatial details, instruments, parameters, units, files, etc.

2. Metadata content & format

2.1 What metadata are needed

- Any details that make data meaningful

2.2 How metadata will be created and/or captured

- Lab notebooks?
- Simulation parameters?
- Auto-saved on instrument?

2.3 What format will be used for the metadata

- Standards for community
- Justification for format chosen

- See later lectures for more detail on metadata standards

3. Policies for access, sharing, reuse

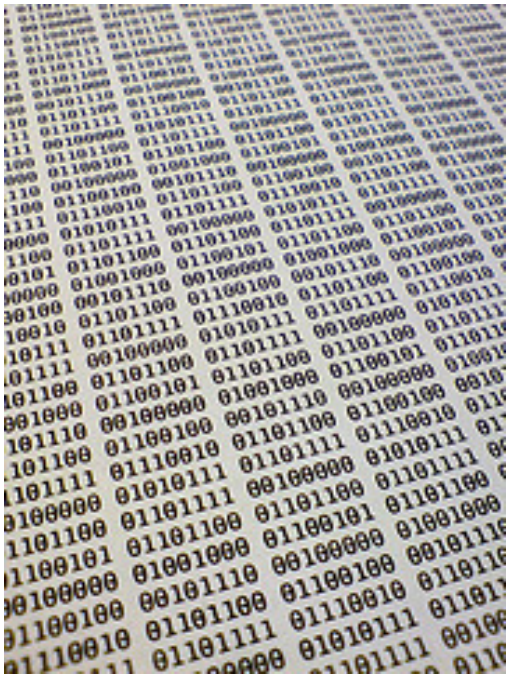
3.1 Obligations for sharing

- Funding agency
- Institution
- Other organization
- Legal

3.2 Details of data sharing

- How long?
- When?
- How access can be gained?
- Data creator/collector rights

3.2 Ethical/privacy issues with data sharing



3. Policies for access, sharing, reuse

3.4 Intellectual property & copyright issues

- Who owns the copyright?
- Institutional policies
- Funding agency policies
- Embargos for political/commercial reasons

3.5 Intended future uses/users for data

3.6 Citation

- How should data be cited when used?
- Persistent citation?

CC image by buddawiggi on Flickr

4. Long-term storage & data management

4.1 What data will be preserved

4.2 Where will it be archived

- Most appropriate archive for data
- Community standards

3.6 Data transformations/formats needed

- Consider archive policies

4.4 Who will be responsible

- Contact person for archive

5. Budget

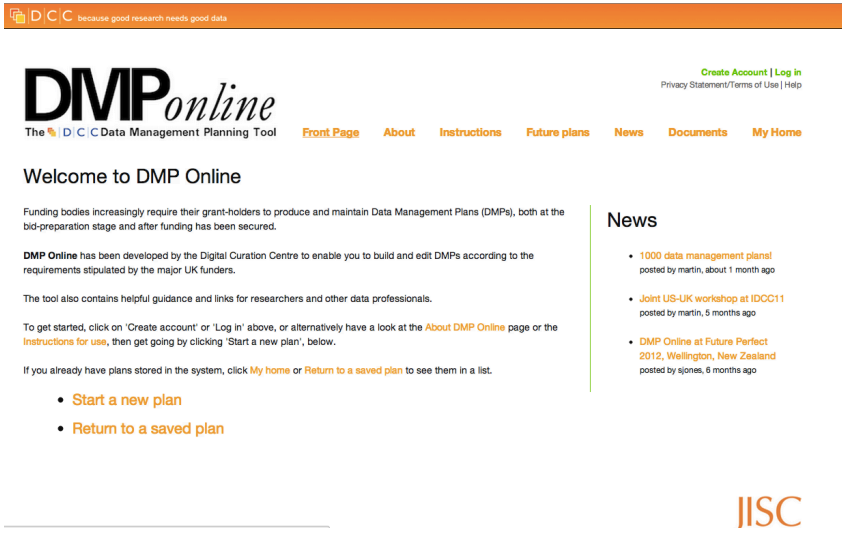
5.1 Anticipated costs

- Time for data preparation & documentation
- Hardware/software for data preparation & documentation
- Personnel
- Archive costs

5.2 How costs will be paid

- Up front?
- Over time?

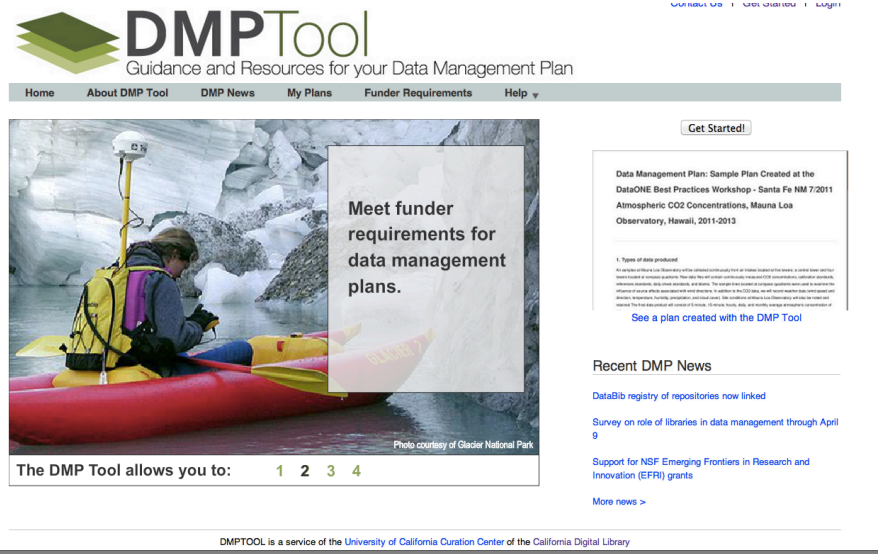
Tools for Creating Data Management Plans



dmponline.dcc.ac.uk

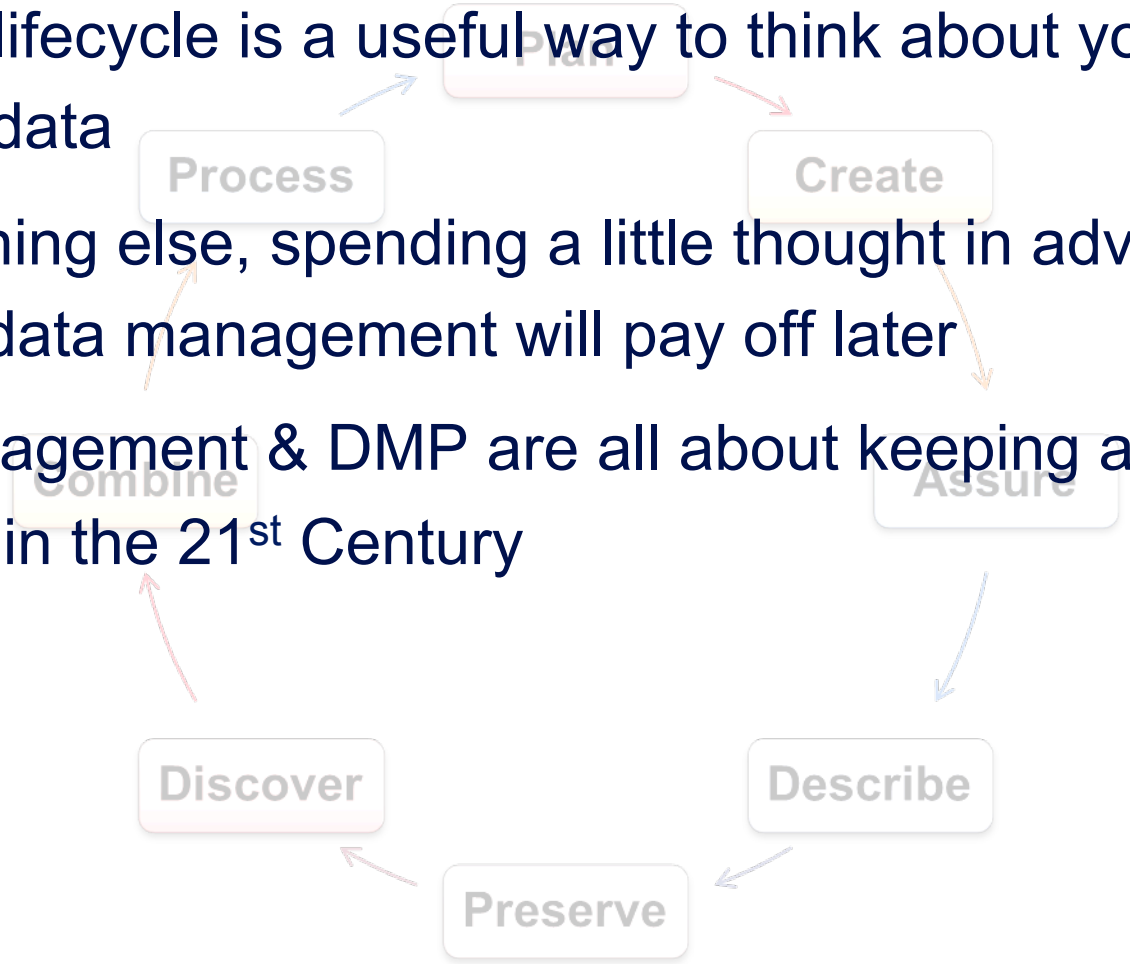
JISC

dmp.cdlib.org



Summary

- The data lifecycle is a useful way to think about your research data
- Like anything else, spending a little thought in advance planning data management will pay off later
- Data management & DMP are all about keeping a tidy lab notebook in the 21st Century



Acknowledgements

- Includes material from
  - DataONE Education Module: Data Management Planning.
    - DataONE. Retrieved Jan, 2014.
    - http://www.dataone.org/sites/all/documents/L03_DataManagementPlanning.pptx

DataONE
Data Observation Network for Earth