

Research Data Management

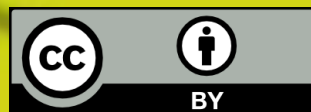
Marta Hoffman-Sommer

Open Science Platform, ICM, University of Warsaw

Jachranka, 25.05.2017



UNIWERSYTET WARSZAWSKI
Interdyscyplinarne Centrum Modelowania
Matematycznego i Komputerowego
www.icm.edu.pl



What is research data?



UNIWERSYTET WARSZAWSKI
Interdyscyplinarne Centrum Modelowania
Matematycznego i Komputerowego
www.icm.edu.pl



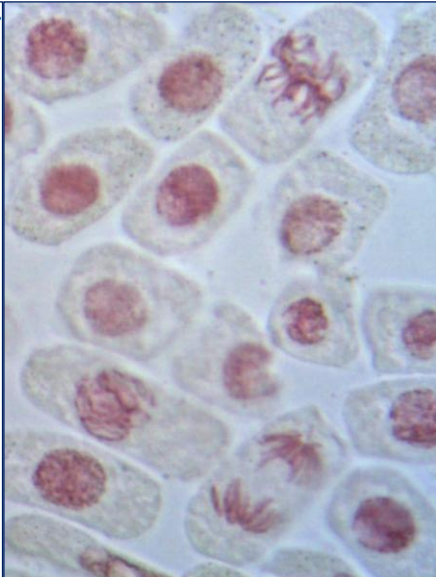
Scientific information

KTH Biblioteket, CC-BY-SA
<https://www.flickr.com/photos/kthbiblioteket/4472640423/>



Articles and books

Channel	Raw Int.	Intensity	Avg.
11481	61,73	69	186
42142	181,65	447	232
37539	151,37	403	248
26707	127,18	302	210
33831	145,82	329	232
30312	135,32	310	224
20118	83,82	125	240
16894	83,22	140	203
16143	82,36	115	196
19950	95	159	210
24331	98,11	174	248
21530	106,06	222	203
11831	67,99	77	174
46601	194,17	428	240
52345	180,5	468	290
43917	177,08	428	248
43813	208,63	478	210
39835	177,83	422	224
20207	103,1	170	196
17899	91,32	136	196
15462	88,86	136	174
18585	94,82	155	196
21416	109,27	197	196
26097	112,49	212	232
11463	63,68	73	180
36909	144,18	277	256
40585	145,47	293	279
32514	140,15	256	232
38101	127	283	300
29338	104,78	203	280
26193	93,88	144	279



```
<TEI version="5.0" xmlns="http://
<teiHeader>
<fileDesc>
<titleStmt>
<title>TEI中文指引</title>
</titleStmt>
<publicationStmt>
<p>將與TEI 中文在地化計劃等文件一
</publicationStmt>
<sourceDesc>
<p>譯自TEI P5 英文指引</p>
</sourceDesc>
</fileDesc>
</teiHeader>
<text>
<body>
<p>這是TEI P5的中文指引...</p>
</body>
</text>
</TEI>
```

Research data

Definitions of research data

„...the recorded factual material commonly accepted in the scientific community as necessary to validate research findings.”

„Research data is data that is collected, observed, or created, for purposes of analysis to produce original research results.”

Examples of research data:

Numerical data

Text documents, lab notes

Questionnaires, responses, transcripts

Audiotapes, videotapes

Photographs, films

Models, algorithms, scripts

Simulation results

Methodologies and workflows

Artifacts, specimens, samples

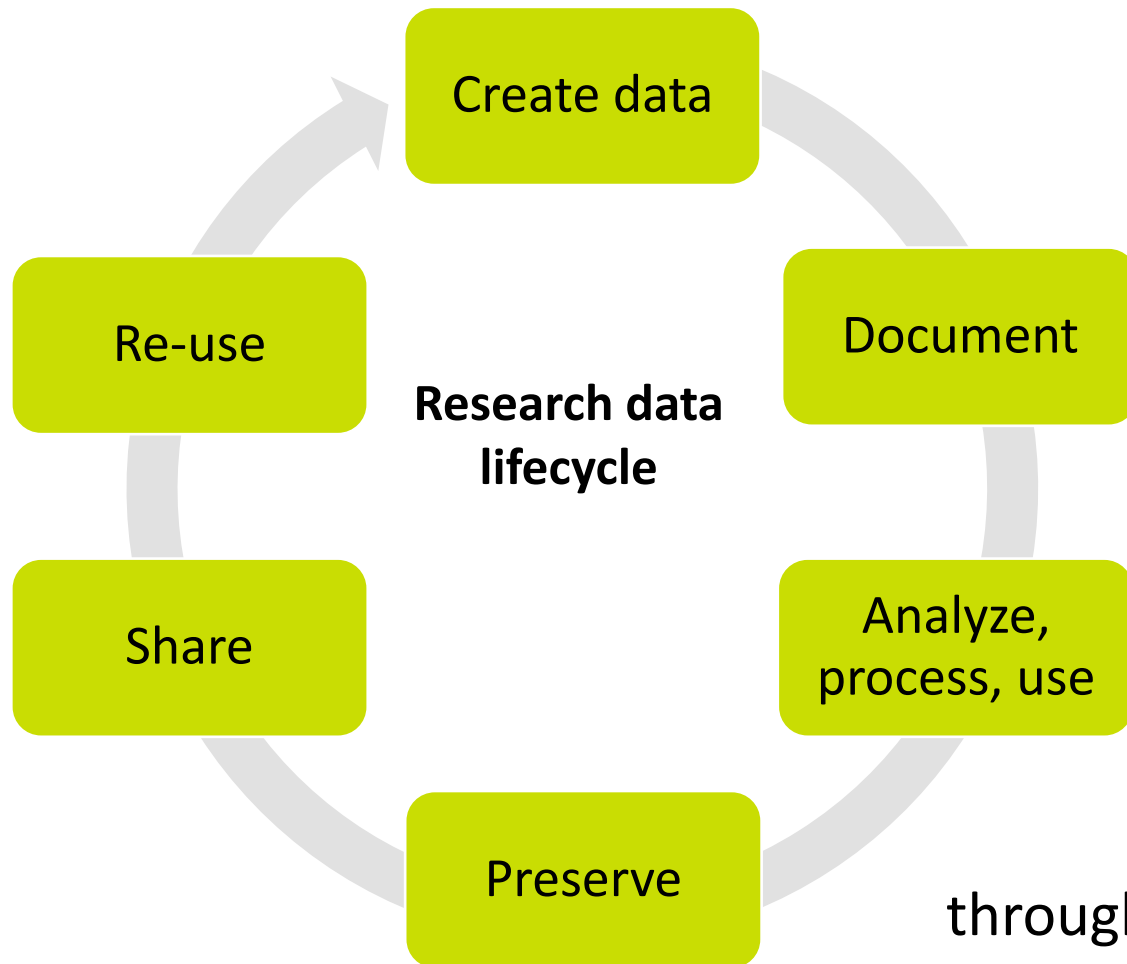
What is research data management and why should we bother?



UNIWERSYTET WARSZAWSKI
Interdyscyplinarne Centrum Modelowania
Matematycznego i Komputerowego
www.icm.edu.pl



Research Data Management is...



...an active approach towards handling data throughout all stages of the research data lifecycle.

Issues to consider in RDM

1. Collecting data: selection of file formats, file naming, metadata, documentation
2. Short- and long-term storage: data selection, safe preservation
3. Access to the data, re-use rules
4. Legal and ethical issues related to data sharing
5. Resources necessary for RDM (funding, skills)

Benefits of good RDM

1. makes your own research easier, now and in the future
2. data is easier to share – more likely to be re-used
3. improves the quality of science (integrity, reproducibility)
4. more cooperation among scientists
5. research moves ahead faster (better communication)
6. less duplication (saves money)

Data sharing requirements

- **Publishing agreements** (journals): require data underlying publication to be made openly available.

Nature, PLoS, Amer. Economic Review...

- **Grant agreements** (funders): require Data Management Plans and/or open availability of research results, including data.

European Commission, Research Councils UK, NSF in USA...



The EU Framework Programme
for Research and Innovation

HORIZON 2020



Guidelines on Open Access
to Scientific Publications and Research Data
in Horizon 2020

Version 1.0
11 December 2013



Open Research Data Pilot in H2020

– since Jan 2017 extended to the program
Open Research Data by Default

„ The use of a detailed data management plan covering individual datasets is required for funded projects participating in the Open Research Data Pilot.”

„ The Open Research Data Pilot applies to two types of data:

1) the **data (...)** needed to validate the results presented in scientific publications as soon as possible;

2) **other data (...)** as specified and within the deadlines laid down in the data management plan.

(...) Participating projects are required to deposit the research data described above, preferably into a research data repository.”

„ As far as possible, projects must then take measures to enable for third parties to **access, mine, exploit, reproduce** and **disseminate (free of charge** for any user) this research data.

One straightforward and effective way of doing this is to attach a Creative Commons Licence (CC-BY or CC0 tool) to the data deposited.”

➤ **FAIR data**

Findable - good metadata, visibility

Accessible - in a repository or archive

Interoperable - easy to combine with other datasets (format, structure)

Reusable - legal and ethical issues, as well as technical (documentation)

➤ ***As open as possible, as closed as necessary***

– legal and/or ethical restrictions, or openness not compatible with the main goal of the project (e.g. commercialization)

Exemptions – reasons for opting out

- If results are expected to be commercially or industrially exploited
- If participation is incompatible with the need for confidentiality in connection with security issues
- If incompatible with existing rules on the protection of personal data
- Would jeopardise the achievement of the main aim of the action
- If the project will not generate / collect any research data
- If there are other legitimate reasons to not take part in the Pilot

Can opt out at proposal stage OR during lifetime of project.

Should describe issues in the project Data Management Plan.



Managing data



UNIWERSYTET WARSZAWSKI
Interdyscyplinarne Centrum Modelowania
Matematycznego i Komputerowego
www.icm.edu.pl



1. Identifying your data

Where does your data come from?

How often do you get new data?

How much data do you generate?

What format(s) are your data in?



Workbook for Writing a Data Management Plan,
<http://www.dcc.ac.uk/training/digital-curation-101/dmp-workshop-uct>



2. Managing data during the project

How do you structure and name your folders and files?

What additional information is required to understand each data file?

Where do you store your data?

How is your data backed up?



Workbook for Writing a Data Management Plan,
<http://www.dcc.ac.uk/training/digital-curation-101/dmp-workshop-uct>



More links on everyday data management:

<http://libraries.mit.edu/data-management/services/workshops/>

How to organize data:

http://libraries.mit.edu/data-management/files/2014/05/FileOrg_20160121.pdf

and more:

<https://www.ukdataservice.ac.uk/manage-data/format/organising>

<http://datalib.edina.ac.uk/mantra/organisingdata/>

<https://library.stanford.edu/research/data-management-services/data-best-practices/best-practices-file-naming>

3. Selecting data for archiving: what to keep, what to discard

What should be archived beyond the end of the project?

Where will you archive your material?

For how long should it be stored?

Who should have access and under what conditions?



Workbook for Writing a Data Management Plan,
<http://www.dcc.ac.uk/training/digital-curation-101/dmp-workshop-uct>



Data Selection – guidelines

1. **Legal requirements** to retain the data beyond its immediate use.
2. **Scientific or Historical Value**: this involves inferring anticipated future use.
3. **Uniqueness**: does it duplicate existing datasets?
4. **Non-Replicability**: would it be feasible to replicate the data? (high costs, one-time events)
5. **Potential for Redistribution**: the reliability, integrity, and usability of the data files (do formats meet technical criteria? are IPRs addressed?)
6. **Economic Case**: costs for managing and preserving the data are justifiable when assessed against evidence of potential future benefits.
7. **Full documentation**: documentation is comprehensive and correct.



Whyte, A. & Wilson, A. (2010). "How to Appraise and Select Research Data for Curation". DCC How-to Guides. Edinburgh: Digital Curation Centre. Available online: <http://www.dcc.ac.uk/resources/how-guides/appraise-select-data>



Data which you will not keep

...should nevertheless be documented:
what, when and why has been discarded.

Archiving: long-term storage

1. Data safety – trusted repository/archive
2. Visibility – known among the target research community and/or visible in search engines
3. Identifiable – persistent identifier (e.g. DOI – *digital object identifier*)

Digital research data repositories

- specialized (one type of data, e.g. GenBank)
- broad (disciplinary, e.g. Dryad for biology)
- general (catch-all, for all data, e.g. Zenodo)



Institutional repositories:

Purdue University Research Repository





PANGAEA.

Data Publisher for Earth & Environmental Science

Not logged in  

[SEARCH](#) [SUBMIT](#) [ABOUT](#) [CONTACT](#)

Submit Data




Submit Data to PANGAEA

Welcome to the PANGAEA data submission system. Any data from earth and life sciences are accepted. We highly appreciate you to archive and publish your data with PANGAEA.

When you start the data submission process, you will be redirected to the PANGAEA issue tracker that will assist you in providing metadata and uploading data files. Any communication with our editors will go through this issue tracker.

Before submitting your data you need to log in or sign up:

 [Log in](#)

 [Sign up as new user](#)

<https://www.pangaea.de/>

Get started

Search over **5553** data sets, services and maps, ...

anyPlaceholder



Browse by **INSPIRE themes** topics



Coordinate reference systems

3



Elevation

57



Land cover

84



Orthoimagery

2



Geology

474



Soil

111



Land use

31



Human health and safety

8



Geographical grid systems

13



Environmental monitoring f...

138



Atmospheric conditions

47



Meteorological geographica...

55

Browse resources



Dataset

5179



Series

315



Service

45



Non geographic dataset

14

Data Portal

Here you will find data, services and maps and more.

Data journals



- Articles describing data (*data descriptors*)
- Data deposited in repositories
- Some journals also allow attaching data as Supplementary Material
 - Complementary to repositories, not an alternative solution!

Browse

re3data.org supports the following browsing options:

- [Browse by subject](#)
- [Browse by content type](#)
- [Browse by country](#)

Exercise: Identifying and selecting data

1. Identify all data in a selected project
2. Which data do you want to archive for the long-term (after the project ends)?

4. Preparing data for archiving

Preparing files

Metadata and additional documentation

Preferred formats for archiving

Preferred formats:

- Without compression
- Not requiring proprietary software
- Open, documented
- Standard coding (ASCII, Unicode)

Type	Recommended	Non-preferred
Tabular data	CSV, TSV, SPSS portable	Excel
Text	Plain text, HTML, RTF PDF/A only if layout matters	Word
Media	Container: MP4, Ogg Codec: Theora, Dirac, FLAC	Quicktime H264
Images	TIFF, JPEG2000, PNG	GIF, JPG
Structured data	XML, RDF	RDBMS

Based on: UK Data Archive (social sciences and humanities)

<http://www.data-archive.ac.uk/create-manage/format/formats>



Selecting file formats for archiving

- In everyday work most convenient formats should be used – before archiving the files should be converted to open, non-proprietary formats.
- Some repositories suggest depositing two formats:
 - (1) most suitable for long-term preservation,
 - (2) and most used in the relevant research community.

Metadata and documentation

Catalogue metadata: basic information about the dataset (author, title, date of creation, license, etc.)

Other documentation: context, methodology, other files necessary to understand the data (including scripts), standard vocabularies, etc.

Metadata standards
(Digital Curation Centre)

Search by Discipline



Biology



Earth Science



General Research Data



Physical Science



Social Science & Humanities

Where to put the documentation?

→ Included in the data files

e.g. units of measurement in the headers of columns for tabular data

→ Attached to the dataset as separate files

ReadMe.txt file with additional information,
attached questionnaires, scripts, etc.

Guidelines to writing „readme” style metadata

Author: Wendy Kozlowski,
Research Data Management Service Group,
Cornell University Libraries
CC-BY

http://data.research.cornell.edu/sites/default/files/SciMD_ReadMe_Guidelines_v4_1_0.pdf



1. Introductory information

- a. For each filename, a short description of what data it contains
- b. Format of the file if not obvious from the file name
- c. If the data set includes multiple files that relate to each other, the relationship between the files or a description of the file structure that holds them (possible terminology might include “dataset” or “study” or “data package”)
- d. Name/institution/address/email information for
 - i. Principle investigator (or person responsible for collecting the data)
 - ii. Associate or co-investigators
 - iii. Contact person for questions
- e. Date of data collection (can be a single date, or a range)
- f. Information about geographic location of data collection
- g. Date that the file was created
- h. Date(s) that the file(s) was updated and the nature of the update(s), if applicable
- i. Keywords used to describe the data topic
- j. Language information

2. Methodological information

- a. Method description, links or references to publications or other documentation containing experimental design or protocols used in data collection
- b. Any instrument-specific information needed to understand or interpret the data
- c. Standards and calibration information, if appropriate
- d. Describe any quality-assurance procedures performed on the data
- e. Definitions of codes or symbols used to note or characterize low quality/questionable/outliers that people should be aware of
- f. People involved with sample collection, processing, analysis and/or submission

3. Data-specific information

- a. Full names and definitions (spell out abbreviated words) of column headings for tabular data
- b. Units of measurement
- c. Definitions for codes or symbols used to record missing data
- d. Specialized formats or abbreviations used

4. Sharing/Access information

- a. Licenses or restrictions placed on the data²
- b. Links to publications that cite or use the data
- c. Links to other publicly accessible locations of the data
- d. Recommended citation for the data
- e. Information about funding sources that supported the collection of the data

Legal issues



UNIWERSYTET WARSZAWSKI
Interdyscyplinarne Centrum Modelowania
Matematycznego i Komputerowego
www.icm.edu.pl



- Am I entitled to make the decision about opening my data?
 - Who owns the intellectual property rights (IPRs) to my dataset?
- Are there any additional legal restrictions that I have to take into account?
- What will the data user be allowed to do with the data? Who and how decides about user rights?
 - What is fair use and what is licensing?

Rights associated with research data

- **Copyright** – has to do with **creativity** in a work (facts cannot be copyrighted)
- **Database rights** – have to do with **investment** made in the creation of the collection (EU only)
- **Third party rights** – they have to do with **other people** involved (creators of contents, human subjects tested, ...)

- Different in different countries
- Ownership of rights may lie with your employer (university, research institute) – may be different for students and employees

Step-by-step:

1. Contact all co-authors
2. Check employer regulations
3. Consider third-party rights

Legal status of data

You can deposit your data in a repository:

- without any license: under the conditions of **fair use**
- under a chosen license (e.g. an open Creative Commons license)
- under a rights waiver (e.g. Creative Commons Zero)

Fair use...

...default rules that by law apply to all published works.

„ As far as possible, projects must then take measures to enable for third parties to **access, mine, exploit, reproduce** and **disseminate (free of charge** for any user) this research data.

One straightforward and effective way of doing this is to attach a **Creative Commons Licence (CC-BY or CC0 tool)** to the data deposited.”



What are Creative Commons Licenses?

BY – Attribution

NC – Non-commercial

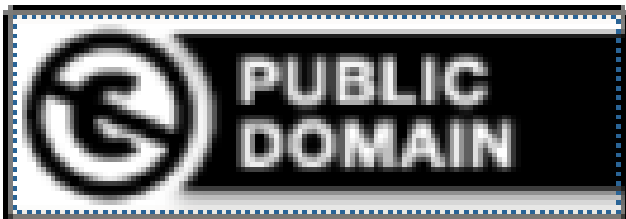
SA – Share Alike

ND – No derivatives



Public Domain

- works not protected by copyright



Public Domain Mark



Use if you know that the work is not protected



Public Domain Dedication



Use if you want to release your work from copyright and database protection

A Digital Curation Centre and JISC Legal
'working level' guide



How to License Research Data

Alex Ball (DCC)



Digital Curation Centre, 2012.
Licensed under Creative Commons Attribution 2.5 Scotland:
<http://creativecommons.org/licenses/by/2.5/scotland/>

www.dcc.ac.uk/resources/how-guides/license-research-data



Should all data be open? No.

Privacy protection (human subjects!)

National security issues

Protection of endangered species, of archaeological sites, etc.

Interference with commercialization plans

But data existence should always be open:

- Allows discovery & negotiation on use
- Avoids pointless replication

Licensing

- Use **standard licenses**
- Use **international licenses** – prepared so as to achieve the closest possible effect in many different jurisdictions
- Ranging from restrictive to liberal (open licenses)

Step-by-step:

- Choose an appropriate license
- Make sure you can apply it
- Make sure that the chosen repository will accept it
- Make the license clear at deposit and in documentation

Data Management Plan (DMP)



UNIWERSYTET WARSZAWSKI
Interdyscyplinarne Centrum Modelowania
Matematycznego i Komputerowego
www.icm.edu.pl



Five common themes

1. Description of data to be collected / created
(i.e. how will it be collected, content, type, format, volume...)
2. Documentation & metadata
(standards and formats, structure of file naming, etc.)
2. Ethics and Intellectual Property
(highlight any restrictions on data sharing e.g. privacy, confidentiality)
4. Plans for data sharing and access
(i.e. how, when, to whom)
5. Strategy for long-term preservation
(i.e. where, how long)

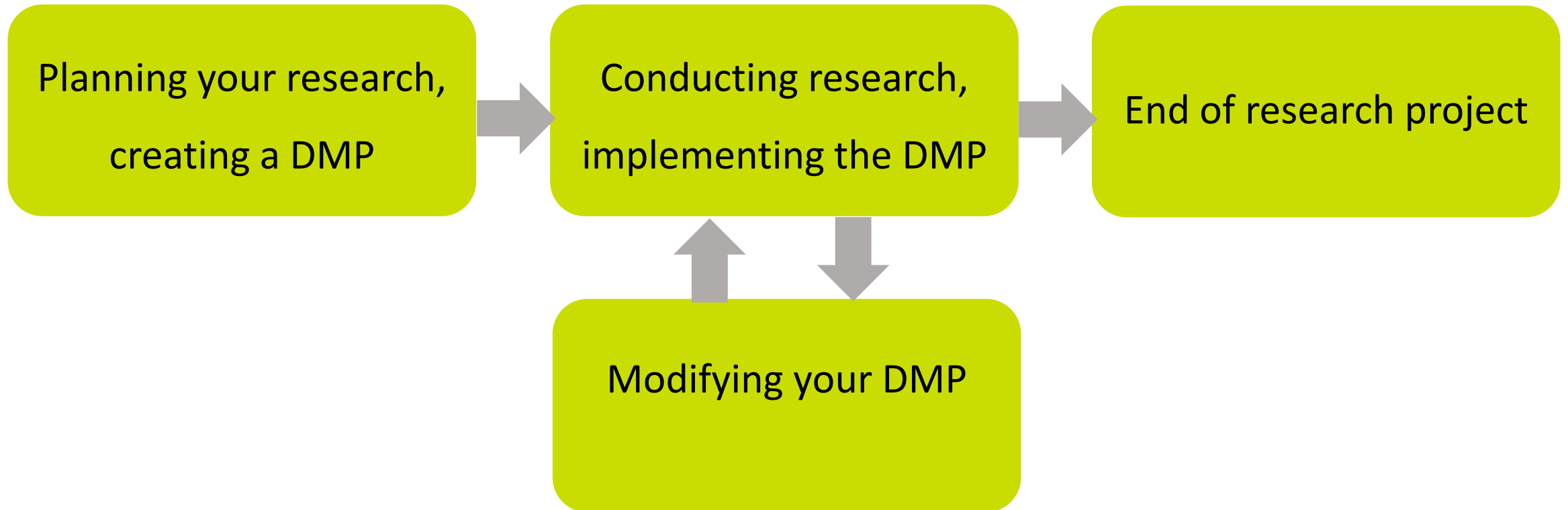


www.dcc.ac.uk/resources/data-management-plans/checklist

Slide adapted from Kevin Ashley, DCC, CC-BY



Data management plan



NSF template (USA):

1. Types of data produced
2. Data and metadata standards
3. Policies for Access and Sharing
4. Policies for Re-use, Distribution
5. Plans for Archiving and Preservation

DataONE www.dataone.org

Data Management Plan
Arthropod responses to grassland nutrient limitation.

1. Types of Data Produced
We will collect insects annually from the 30 experimental plots at each of the eight sites (see body of proposal for sampling details). Samples will be immediately deposited in sealable containers labeled with the date, site code (already existing), block, plot, and subsample. An associated record of any observations or notes will be entered in a field tablet computer and labeled with the same information. We will also record environmental information including temperature and general observations. Labeled samples will be transported back to the laboratory, where they will be sorted and identified using a dissecting microscope. We will identify and count the arthropods to the classification of order, with the exception of members of the order Auchenorrhyncha, which will be identified to species or morphospecies. Identifications will be reviewed by multiple researchers associated with the project and verified with the assistance of Stuart McKamey of the Systematic Entomology Laboratory of the USDA Agricultural Research Service. Representatives of the identified species and morphospecies will be vouchered to the Bell Museum of Natural History at the University of Minnesota (U of M).

Abundance for each group will be recorded by hand in a laboratory notebook during sorting. These data will be transcribed into an Excel spreadsheet as each sample is completed. The spreadsheets will be stored on a controlled-access U of M server directory that is backed up offsite nightly. Files will be named according to the format `site_mmd/yyyy_plot.csv` using existing unique site codes. Lind will be responsible for the data during and after data collection until publication.

After identification, Arthropod samples from each experimental plot will be subsampled and sent to the University of St. Thomas Kay lab for stoichiometric analysis. We will receive a spreadsheet of data after processing is complete. This spreadsheet will include the insect identification (including site code, date, year, plot, and arthropod identification) and percent by mass of carbon, phosphorus and nitrogen. These files will be saved as `.csv` files in the previously described server directory.

Our data set will be used in combination with the existing Nutrient Network (nutnet.unm.edu) data on plant responses to nutrient manipulation. The NutNet data is currently stored and managed in a MySQL relational database housed at the Minnesota Supercomputing Institute and accessed through a secure internet connection. We will add our data and metadata to the NutNet relational database. The existing `csv` files will be read into temporary tables in the MySQL database, and then inserted into permanent data tables using insert query statements. The existing database schema links tables of data observations to a "plot" table describing the experimental unit. New tables will be created for each of the arthropod data types (abundance and stoichiometry) containing the unique plot identifier. Multiple tables may be necessary for efficient data storage and management; for example, an "Arthropod" table holding scientific names for use can be used to constrain the labels of abundance records to acceptable possibilities.

2. Data and Metadata Standards
The project will leverage existing metadata standards currently stored in Ecological Metadata Language (EML) format for the NutNet project. We will add additional metadata entries for the arthropod community composition and arthropod stoichiometry; field notes taken during the time of collection will be recorded. Morpho software will be used to generate the metadata file in EML. We chose EML format for our metadata since it allows integration with existing NutNet data housed in the Knowledge Network for Biocomplexity (KNB) data repository.

2 Example DMP - NutNet.
© DataONE 2011

3. Policies for Access and Sharing
After publication of manuscripts based on the data we collect, we will share our data and metadata with the NutNet community via data updates sent annually as `.csv` files from the existing central relational database. Other NutNet users will need to contact Lind for access to the data.

We will also submit both of our datasets (abundance and stoichiometry) to the U of M Digital Conservancy, an archive for digital preservation. Borer has access to this resource as a faculty member. This will occur within a year of publication. The data will be publicly available via the Digital Conservancy, which provides a permanent URL for digital documents.

4. Policies for Re-use, Distribution
Access to databases and associated software tools generated under the project will be available for educational, research and non-profit purposes. Such access will be provided using web-based applications, as appropriate.

Materials generated under the project will be disseminated in accordance with University/Participating institutional and NSF policies. Depending on such policies, materials may be transferred to others under the terms of a material transfer agreement.

Those that use the data (as opposed to any resulting manuscripts) should cite it as follows:
Lind, E, E Borer and A Kay. yyyy. Grassland Arthropod abundance and stoichiometry associated with nutrient manipulation. [URL]; accessed on ddmmyyyy.

This information will be described in the metadata.

Intended and foreseeable users of the data are NutNet collaborators and participants, as well as other scientists interested in arthropod-plant relationships. This data set could be used in combination with similar data sets from other NutNet sites or for meta-analysis.

5. Plans for Archiving and Preservation
We will preserve both arthropod datasets generated during this project (abundance and stoichiometry) for the long term in the Digital Conservancy at the U of M. We will include the `.csv` files, along with the associated metadata files. We will also submit an abstract with the datasets that describe their original context and any potentially relevant project information. Borer will be responsible for preparing data for long-term preservation and for updating contact information for investigators.

Example DMP - NutNet.
© DataONE 2011 3

Abundance for each group will be recorded by hand in a laboratory notebook during sorting. These data will be transcribed into an Excel spreadsheet as each sample is completed. The spreadsheets will be stored on a controlled-access U of M server directory that is backed up offsite nightly. Files will be named according to the format *site_mmddyyyy_plot.csv* using existing unique site codes. Lind will be responsible for the data during and after data collection until publication.

The project will leverage existing metadata standards currently stored in Ecological Metadata Language (EML) format for the NutNet project. We will add additional metadata entries for the arthropod community composition and arthropod stoichiometry; field notes taken during the time of collection will be recorded. Morpho software will be used to generate the metadata file in EML. We

We will also submit both of our datasets (abundance and stoichiometry) to the U of M Digital Conservancy, an archive for digital preservation. Borer has access to this resource as a faculty member. This will occur within a year of publication. The data will be publicly available via the Digital Conservancy, which provides a permanent URL for digital documents.

Access to databases and associated software tools generated under the project will be available for educational, research and non-profit purposes. Such access will be provided using web-based applications, as appropriate.

Those that use the data (as opposed to any resulting manuscripts) should cite it as follows:

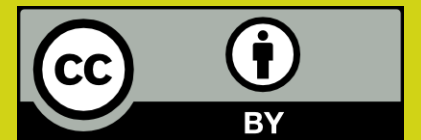
Lind, E, E Borer and A Kay. yyyy. Grassland Arthropod abundance and stoichiometry associated with nutrient manipulation. [URL]; accessed on ddmmyyyy.

This information will be described in the metadata.

Thank you for your attention

Contact:

msommer@icm.edu.pl



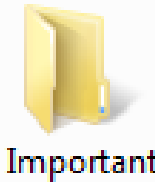
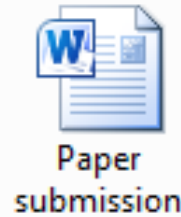
<http://creativecommons.org/licenses/by/3.0/pl/legalcode>



UNIwersytet Warszawski
Interdyscyplinarne Centrum Modelowania
Matematycznego i Komputerowego
www.icm.edu.pl



File naming



- What information do you need in your file name?
(type of data, initials of the researcher, sample number, date, version, etc.)
- Are the names unique?
- How do you want to sort the files?

http://www.data.cam.ac.uk/files/gdl_tilsdocnaming_v1_20090612.pdf

Example:

Document naming for the TILS Division should follow this convention:

GDL_TILSDocNaming_V1_20090612.docx

A prefix shows the document type

The document title describes the content

The version number

The date in the format yyyymmdd

Place an explanation of file names in the documentation!

At least 2 backups, including one off-site:



Monday morning



daily
- automatically



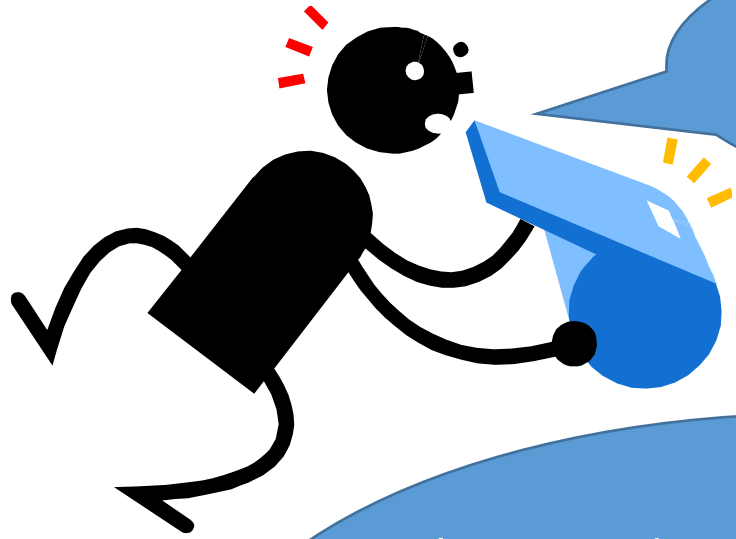
- Regularity
- Automatically

Based on slide:

Y. Creba, V. Philips, C. Sewell, M. Teperek,
University of Cambridge, CC-BY

Free software to manage backups (example):
<http://www.2brightsparks.com/download-syncbackfree.html>

Data quality



I need that data now!!! I don't care how messy it is – I can fix it!

I've wasted too much of my life fixing other's people's bad data. I'm not interested until you've cleaned it up and documented it. Besides, I have other things to think about...

