Title (of Document)	An Introduction to Data Management
Creator (author)	Alejandra Sarmiento Soler, Mara Ort, Juliane Steckel
Contributions	Jens Nieschulze
Project	BEFmate, GFBio
Date	22/02/2016
Date of publication	
access Date of cited URLS	12.01.2016
Version	4
Filename	Reader_GFBio_BefMate_20160222
Internal storage location	\PowerFolder\GFBio Training Material\BEFmate_WoSho_2014\Reader
Subject (key words)	Data management, data life cycle
Description (abstract)	Handbook on data management for researchers. Follows ten steps of the Data Life Cycle (propose, collect, assure, describe, submit, preserve, discover, integrate, analyse, publish). Provides information as well as practical tips and further resources. Informs about GFBio tools and services.
Type	Text
Format	MS Word 2010
Resource Identifier	
Language	English
Licence	CC BY-NC-SA 4.0 
	This work is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-sa/4.0/.

An Introduction to Data Management

BEFmate

Content

1. About this Reader

Data are the fundament of science. In the last years, awareness for the management of data increases more and more and data management is actively performed and integrated in the research process. This reader aims at further raising awareness for data management in the research community and introduce activities related to data management. The structure of the reader follows the concept of the Data Life Cycle with these steps: propose, collect, assure, describe, submit, preserve, discover, integrate, analyse, and publish. After briefly describing each step and its role, the corresponding data management activities are presented, including best practice examples, tips and further resources.
The disciplinary focus of this "Introduction to Data Management" lies on biology and environmental sciences and quantitative data. Nevertheless, many aspects apply universally to quantitative and qualitative data, not depending on discipline. This reader is for everybody who wants to deepen his or her knowledge on data management. A primer on this topic consisting of factsheets about each step of the Data Life Cycle can be found on the GFBio Homepage (http://www.gfbio.org/data-life-cycle). In this reader, the topics are discussed with more detail.

The German Federation for the Curation of Biological Data (GFBio)

This reader is published as training material by GFBio. GFBio aims at establishing a sustainable, service oriented, national data infrastructure facilitating data sharing for biological and environmental research. It acts as a single point of contact for collection, archiving, curation, integration and publication of data. GFBio offers advanced tools to support data-driven research, access to data and archiving data. Data can be discovered by a faceted search. The services support and facilitate several activities of data management. Links to tools and services offered by GFBio are presented in the respective chapters.
GFBio provides further training materials on its website. On a regularly basis workshops are offered about data management and Digital Curation. Visit http://www.gfbio.org/ for more information.

Acknowledgement

The content of this reader builds up upon the materials provided by the Digital Curation Centre (http://www.dcc.ac.uk/) and DataONE (https://www.dataone.org/).

2. What is Data Management?

Data management concerns the dealing with data in the scientific context. Often, more importance is given to results, analysis and derived conclusion than to the data themselves. However, data are a product of the science enterprise and are more and more understood as a valuable research output themselves (DataONE 2012b; Ludwig and Enke 2013; Data Service 2012-2015a). Research data are considered all information collected, observed or created for purposes of analysis and validation of original research results. Data can be quantitative or qualitative and comprises also photos, objects or audio files, resulting from as different sources as field experiments, model outputs or satellite data. In the following, the focus lies on the management of quantitative digital data. One reason why data management is important is that the value of research data is sometimes not yet visible nowadays, which can lead to neglecting proper data management:

> *In many sciences experiments or observations cannot be repeated making at least part of the data so valuable that it needs to be stored for a long time. In many cases the value of data can only be realized after many years by new generations (RDA Europe 2014a).*

One factor making data management even more important is the growing amount of digital data available:

> *The amount of data collected is growing exponentially nowadays. New environmental observing systems […] will provide access to data collected by aerial, ground-based and underwater sensor networks encompassing tens of thousands of sensors that, when combined, will generate terabytes to petabytes of data annually (Michener and Jones 2012).*

In some fields and disciplines, data-intensive research is opening up innovative research possibilities (Mantra et al. 2014). If well-managed, these data can be used in order to answer (new) research questions (Corti et al. 2011).
Ideally, data management is accepted as integral part of the research idea and is already considered early in the research proposal. It includes the collection phase, the processing and analysis of data, the documentation and preservation. Different data management activities are associated with each step, ensuring a reliable and accessible data fundament for the researchers work as well as facilitating sharing and publishing of data. Well-managed data will further facilitate 1) re-use by oneself or others over time, 2) to replicate or validate research results (think of the good scientific practice obligation) 3) processing of so-called wide databases (integration of many small files of varying syntax) and so-called deep data bases (handling of BIG data).

Importance and benefits of data management

An example from Mantra Online Course (Mantra et al. 2014) illustrates what role data management can play and how it may support your research:

> *You have completed your postgraduate study with flying colours and published a couple of papers to disseminate your research results. Your papers have been*

cited widely in the research literature by others who have built upon your findings. However, three years later a researcher has accused you of having falsified the data.

- Do you think you would be able to prove that you had done the work as described? If so, how?
- What would you need to prove that you have not falsified the data?

The documentation of data analysis and transformation as well as the storing of data and research results are integral part of data management and could help you to prove your work. Without data management, not only a solid basis ensuring replicability for your research results may be missing, but your data can also be subject to data loss more easily. This may happen due to technical problems (hardware failure), due to software obsolescence, due to missing information (data cannot be understood in the future) or due to not storing data in an appropriate way (data will never be found again). In Figure 1, the loss of information content of data is related over time to the career of a researcher. It illustrates that often much information is tied to specific persons. If they leave the project or retire, their knowledge is not available anymore. And people do of course also forget details over time. So it doesn't have to be such a drastic case as in the opening example to clarify that data management may help and facilitate research.

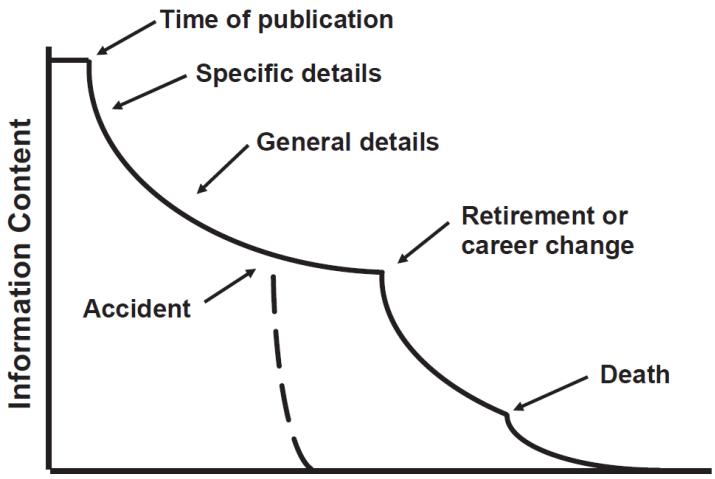


Figure 1: Loss of meta information over time. Michener 2006

Figure 2 displays some benefits of research data management planning. A Data Management Plan provides guidelines and procedures for data management encouraging systematic documentation and description. Data management planning enhances the **security** of data. It safeguards against data loss as storage, backups and archiving are planned. **Compliance** to funder or publisher requirements of the collected data is ensured. DFG proposals require for example a specification about the research data generated in a project (DFG 2014). **Quality** of research in general is enhanced as data management ensures that research data and records are accurate, consistent, complete, authentic and reliable. It also allows for reproducibility of results. As a side effect data management planning streamlines data handling and can

thus create **efficiency** gains for the whole research project. Data management also facilitates the handling of big amounts of data. **Access** and restrictions of use can be documented in Data Management Plans and metadata. Access to data is possible when data are shared and made available. This enables collaboration, prevents duplication and can increase citations for the data creator (DataONE Community 2014; DCC 2008a; UK Data Service 2014a; University of Western Australia 2015).

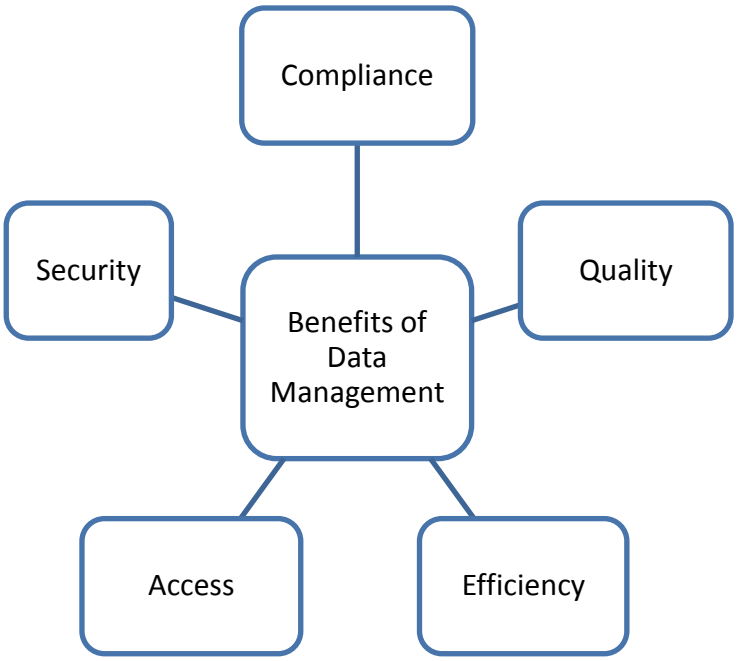


Figure 2: Benefits of Research Data Management Planning. Own design after University of Western Australia 2015

Incorporating data management as a routine part of the research process can save time and resources in the long run. In the beginning, some time is needed to prepare a Data Management Plan and to get used to new practices and activities. This is rewarded by extra funding for your data management, increased citations, and less work organising and understanding data later on (DataONE 2012a).
The costs of data management can be either calculated by total costs of all activities related to the Data Life Cycle (introduced in Chapter 3). As it is often hard to cost data management practices, as many activities are part of standard research activities and data analysis, the costs of data management can also be calculated by focusing on expenses which are additional to standard research procedures (Corti et al. 2011). Some costs and benefits of data management can be measured quantitatively, in terms of people's time or costs of physical resources like hardware or software (see Figure 3). Others have qualitative character or are impossible to measure at all in advance (e.g. possible new scientific findings; Houghton 2011).

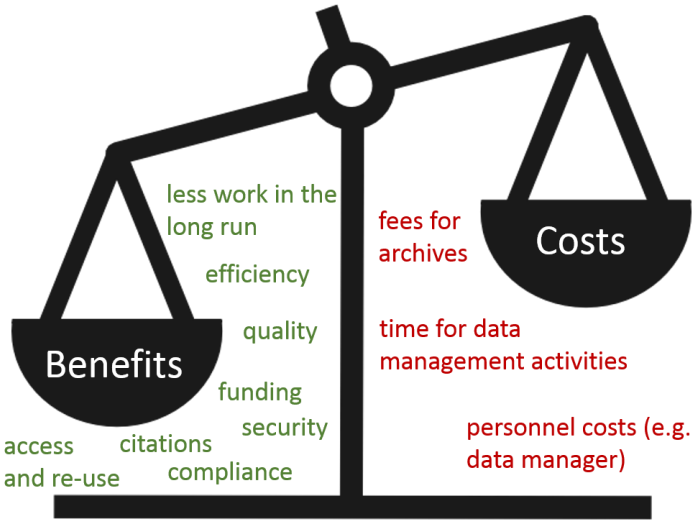


Figure 3: Costs and benefits of data management.

Sharing

The growing awareness for the importance of data results in the conclusion that research data should be made accessible. The DFG states in its "Guidelines on the Handling of Research Data in Biodiversity Research" (2015) that data management should assure a re-use of data also for purposes other than those they were collected for. Furthermore, they emphasize access to data:

> *Enabling free public access to data deriving from DFG-funded research should be the norm.*

Sharing data can bring advantages for individual researchers as well as for the scientific community in general. There are three dimensions of sharing data. First, data can be shared among researchers within the project team. Therefore data is submitted to a shared drive. Second, data can be shared with researchers outside the core research team, e.g. when there is cross-institutional cooperation. In this case, data is submitted to collaboratively used drive, e.g. BExIS. Third, data can be made publicly available. This is referred to as publishing data. Data centres like GFBio preserve data and make them discoverable. A study showed that the publication of data may increase citations (Piwowar and Vision 2013).

Sharing data facilitates the collaboration within and outside research projects and establishes links to the next generation of researchers because data are discoverable and understandable. It also allows to approach research questions which were not thought of when the research started. Another advantage is the prevention of unnecessary duplication of data collection. Furthermore, a key factor for science is replicability, so researchers can collect data and analyse them in order to produce similar results or assess previous work in the light of new approaches (e.g. voice recording of bat signals and the determination of species) (UK Data Service 2012-2015b). However, if that information is not available or poorly documented and difficult to understand, re-use or replication is difficult (Heidorn 2008). Besides the

voluntary sharing of data, many journals already request the submission of data underpinning a paper.

Public Library of Science (PLOS) (http://journals.plos.org/plosone/s/data-availability)

> *PLOS journals require authors to make all data underlying the findings described in their manuscript fully available without restriction, with rare exception.*

Nature (http://www.nature.com/authors/policies/availability.html)

> *An inherent principle of publication is that others should be able to replicate and build upon the authors' published claims. Therefore, a condition of publication in a Nature journal is that authors are required to make materials, data and associated protocols promptly available to readers without undue qualifications.*

It has to be acknowledged that there are of course also a number of reasons why researchers do not wish or are not able to share research data (see also 3.10 Publish). Not all of these reasons may be overcome (Mantra et al. 2014). These barriers include finances, confidentiality of data and ownership issues. Corti et al. (2011) discuss some of these barriers and show possible solutions (see Chapter 3.10 Publish).

3. Data Life Cycle

The different activities concerning data management can be structured in the so called Data Life Cycle (Figure 4). The Data Life Cycle is a conceptual tool which helps to understand 10 different steps that data management follows from data generation to knowledge creation. The Data Life Cycle incorporates planning and collection of data, quality assurance, metadata creation, submission, preservation, discovery, integration, analysis and publication. Many steps of the Data Life Cycle are not only performed once, but multiple times or continually over the life cycle. The order of the Data Life Cycle is also adapted to the needs of a research project. It can be approached from different perspectives, such as data producer and data re-user. For example, a data re-user does not collect data, and not every data producer integrates data from other researchers.
There are practices and steps within the ideal Data Life Cycle where research has yet to discover its potential, and which are worth adopting as routines (which ones have you thought of?). Some practices and steps are normally carried out by specialists called (digital) curators working at data repositories. Curation is managing digital items in a storage to ensure long-term preservation. One step further is to make them discoverable and accessible as soon as it is possible.
This reader provides best practices for every step of the Data Life Cycle. These practices and activities are a suggestion. Every researcher should check what is suitable and makes sense for her or his specific project and adapt the practices accordingly.

For a short overview over the different steps of the Data Life Cycle, the fact sheets on the GFBio Homepage are recommended (http://www.gfbio.org/data-life-cycle).

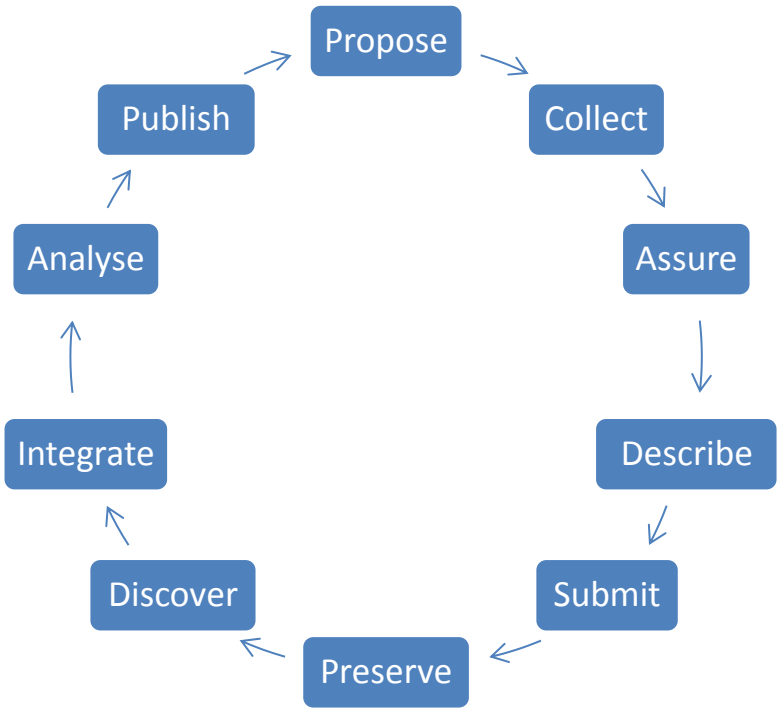


Figure 4: Data Life Cycle after GFBio

The Data Life Cycle can be understood as a part of the Research Life Cycle. The **Research Life Cycle** (Figure 5) is a model for the steps followed in order to create scientific knowledge. The Research Life Cycle, depicted in orange, starts with the research idea and comprises the establishment of cooperation with research partners, the composition of a research proposal, the granting of funding, the Data Life Cycle (except the step "propose", which is already included in the research proposal) and finally the publication of research results. The generated scientific knowledge and information serves as starting point for new research problems.

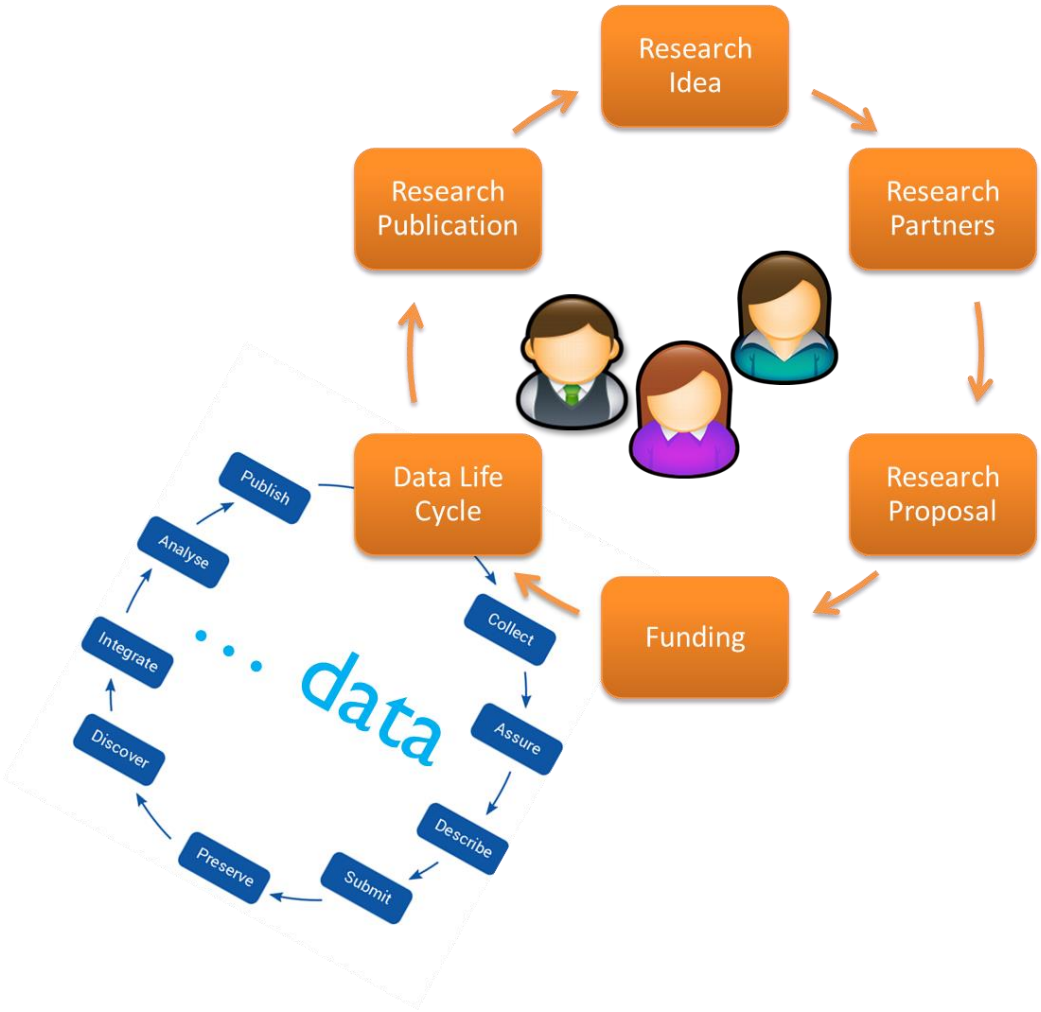


Figure 5: Research Life Cycle after GFBio

3.1 Propose

Data management ideally begins at the planning and proposal phase of the research project. This is the best moment to establish a Data Management Plan to provide a framework that supports researchers and their data throughout the course of research and to provide guidelines for everyone to work with (Mantra et al. 2014). Some funders require a Data Management Plan. And it's also possible to get funding (e.g. from the DFG for archiving costs) for data management activities.

Table 1: Examples for components of a Data Management Plan. Michener and Jones 2012, supplemented

Component	**Description and examples**
Information about data and data format	Types of data that will be produced (e.g. experimental, observational, raw or derived, physical collections, models, images, etc.)
	Volume of data
	When, where and how the data will be acquired (e.g. methods, instruments)
	How the data will be processed (e.g. software, algorithms and workflows)
	File formats (e.g. csv, tab-delimited or naming conventions)
	Quality assurance and control procedures used
	Other sources of data (e.g. origins, relationship to one's data and data integration plans)
	Approaches for managing data in the near-term (e.g. version control, backing up, security and protection, and responsible party)
Metadata content and format	Metadata that are needed
	How metadata will be created or captured (e.g. lab notebooks, auto-generated by instruments, or manually created)
	Format or standard that will be used for the metadata
Policies for access, sharing and re-use	Requirements for sharing (e.g. by research sponsor or host institution)
	Details of data sharing (e.g. when and how one can gain access to the data)
	Ethical and privacy issues associated with data sharing (e.g. human subject confidentiality or endangered species locations)
	Intellectual property and copyright issues
	Intended future uses for data
	Recommendations for how the data can be cited
Long-term storage and technical data management	Identification of data that will be preserved
	Repository or data centre where the data will be preserved
	Data transformations and formats needed (e.g. data centre requirements and community standards)
	Identification of responsible parties
Budget	Anticipated costs (e.g. data preparation and documentation, hardware and software costs, personnel costs and archive costs)
	How costs will be paid (e.g. institutional support or budget line items)
Responsibilities	Who is responsible for what?

Establishing a Data Management Plan at proposal stage in the Data and Research Life Cycle facilitates a structured work with data and saves time later on. A Data Management Plan means to plan all activities and things that should be considered concerning the data foundation of the research project. It clarifies resources needed in terms of money for long term preservation, or skills and software as well as responsibilities and roles of stakeholders, project members and lab staff (e.g. will the computing centre be involved?). A Data Management Plan is a living document that is

to be maintained and kept up-to-date, e.g. if staff changes. It is important to base the plan on available resources and support to ensure that implementation is feasible.
Table 1 gives a first insight in what can be included in a Data Management Plan. A very detailed list elaborated by the WissGrid-Project is available online. From such lists, the suitable aspects can be chosen accordingly to the needs and characteristics of the project.

Resources

WissGrid: Detailed checklist for creating Data Management Plans (in German).
http://www.wissgrid.de/publikationen/deliverables/wp3/WissGrid-oeffentlicher-Entwurf-Checkliste-Forschungsdaten-Management.pdf

Online tools for creating Data Management Plans:
https://dmptool.org/
https://dmponline.dcc.ac.uk

Digital Curation Centre: Advices on DMP writing.
http://www.youtube.com/watch?v=7OJtiA53-Fk

Digital Curation Centre: How to include costs in Data Management Plans.
https://www.youtube.com/watch?v=nKeVPpupsYI&feature=c4-overview&list=UULTOHF6qQrYhEvQzbu03tTg%00

UK Data Service: Data Management Costing Tool. http://www.data-archive.ac.uk/media/247429/costingtool.pdf

3.2 Collect

Ecological data can be collected in many different ways. Collection includes various procedures such as manual recordings of observations in the laboratory or field on hand-written data sheets as well as automated collection by data loggers, satellites or airborne platforms (Michener and Jones 2012). Data created in digital form is "born digital", manually collected data are digitised later on. When collecting data, it is helpful to think of subsequent steps of the Data Life Cycle: what is going to happen with the data? In this way, data collection can be organised in a way that supports following activities and saves time later on. Here are some tips for data collection which are particularly important when several people are involved with data collection and entry:

- Decide what data will be created and how - this should be communicated to the whole research team.
- Be clear about methods.
- Use collecting protocols.
- Develop procedures for consistency and data quality.

There are many activities directly related to data collection. Apart from data entry and file naming (which are discussed below), the choice of an appropriate software format for the collected data has to be taken into consideration (see also 3.5 Submit). Many different programmes are used for data collection, ranging from spreadsheets and statistical software to relational database management systems and geographic information systems (Michener and Jones 2012). Every software and format has advantages and disadvantages, depending on what kind of analyses are planned, software availability and costs, the hardware which is used to capture the data and discipline specific standards and customs (UK Data Service 2014b). Also, the file format for working with the data can differ from the formats used for storing and long-term preservation.

Data entry

Spreadsheet software (like Excel) is very common for working with quantitative data, especially during data collection and entry. Its advantages are that many people already know working with it and first steps are quite simple. However, special care should be taken when working with spreadsheets as they can be quite error-prone. Excel can be a good choice for data entry, but use a syntax that allows information to be stored without loss in csv-files (so that they can be easily accessed with other programs e.g. for analysis). If spreadsheets are used, systematic and accurate work from the beginning on facilitates the process of data exchange to other programmes like R. Especially for analysis and data transformation, it is recommended to use scripted environments like R. But you wouldn't like to enter data in R.

For entering data, it is recommended to use codes like ASCII, UTF-8 or ISO 8859-1 (Latin1). These codes contain characters which can be read by most programmes without any problems (in contrast to codes using e.g. umlauts). If a file is opened using a wrong encoding, something like this can happen:

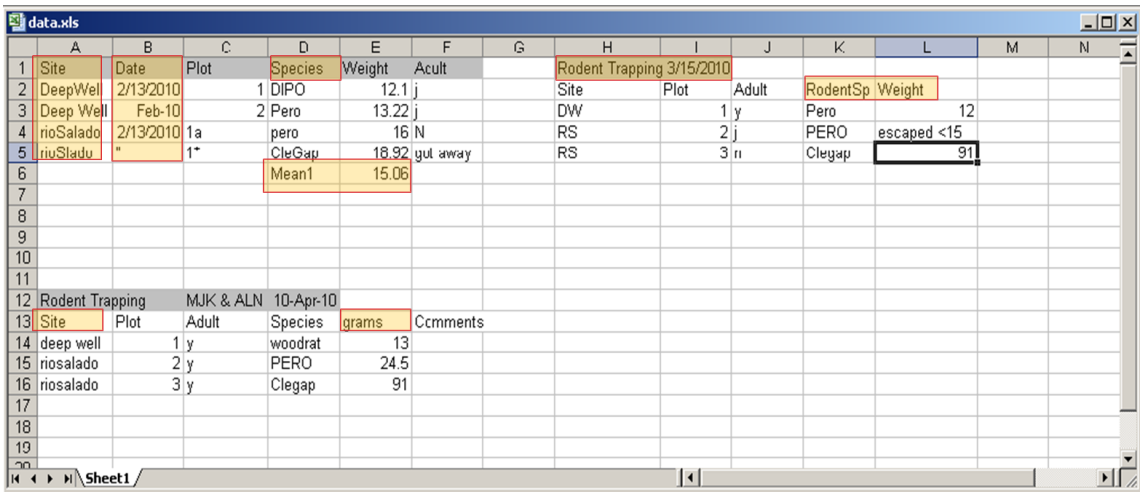
Bärbel Bürgenßen (2015): Encoding Problems in the City of Mörgäl.
ï»¿BÃ¤rbel BÃ¼rgenÃŸen (2015): Encoding Problems in the City of MÃ¶rgÃ¤l.

When you receive data where umlauts are incorrectly displayed, use e.g. the simple notepad editor to change the encoding.

Spreadsheet structure

Some basics for spreadsheet file structure:

- Variable names without spacing, name variables consistently
- Use codes (ASCII, UTF-8), avoid umlauts
- Only one type of information in each cell (atomize values)
- Record full dates, standardize formats (recommended YYYY-MM-DD, allows sorting)



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Site	Date	Plot	Species	Weight	Acult		Rodent Trapping 3/15/2010						
2	DeepWel	2/13/2010	1	DIPO	12.1	j		Site	Plot	Adult	RodentSp	Weight		
3	Deep Well	Feb-10	2	Pero	13.22	j		DW	1	y	Pero	12		
4	rioSalado	2/13/2010	1a	pero	16	N		RS	2	j	PERO	escaped <15		
5	rioSlado	"	1"	CleGap	18.92	gut away		RS	3	n	Clegap	91		
6				Mean1	15.06									
7														
8														
9														
10														
11														
12	Rodent Trapping		MJK & ALN	10-Apr-10										
13	Site	Plot	Adult	Species	grams	Ccmments								
14	deep well	1	y	woodrat	13									
15	riosalado	2	y	PERO	24.5									
16	riosalado	3	y	Clegap	91									
17														
18														
19														

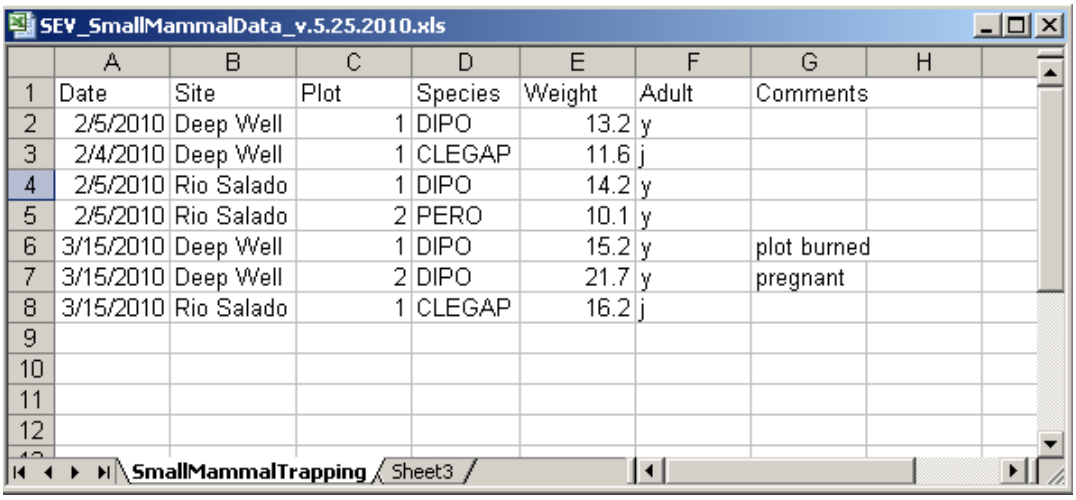
Figure 6: Example for poor data entry. DataONE 2012d

Figure 6 shows an example for poor data entry in Excel. Three different trapping periods of a project were entered in one table. There are many inconsistencies between data collection events:

- Location of date information, inconsistent date format
- Column names
- Order of columns
- Different site spellings, capitalization, spaces in site names
- Mean1 value is in weight column
- Text and numbers in same column

A table structured in that way is hard to filter and to analyse. Data should be structured as consistent as possible. Even if there are any errors, they can be fixed much easier (via scripting) than if the data entry was extremely unstructured.

In Figure 7, a corrected version is shown. The entries are consistent now: only numbers or dates or text was entered. Consistent names, codes and formats (date) are used in each column. And data are all in one table, which is much easier for a statistical programme to work with than multiple small tables which each require human intervention. This record also underlines the importance of additional information about the data (metadata). It is not apparent from the table what measurement unit is used for weight, or what the species abbreviations mean. This information can be given on a separate sheet or in the metadata documentation of the dataset (see 3.4 Describe).

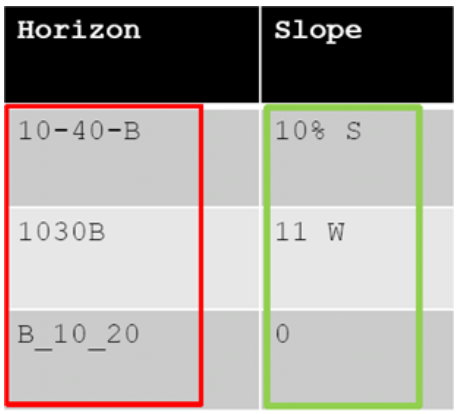

SEV_SmallMammalData_v.5.25.2010.xls

	A	B	C	D	E	F	G	H
1	Date	Site	Plot	Species	Weight	Adult	Comments	
2	2/5/2010	Deep Well	1	DIPO	13.2	y		
3	2/4/2010	Deep Well	1	CLEGAP	11.6	j		
4	2/5/2010	Rio Salado	1	DIPO	14.2	y		
5	2/5/2010	Rio Salado	2	PERO	10.1	y		
6	3/15/2010	Deep Well	1	DIPO	15.2	y	plot burned	
7	3/15/2010	Deep Well	2	DIPO	21.7	y	pregnant	
8	3/15/2010	Rio Salado	1	CLEGAP	16.2	j		
9								
10								
11								
12								

SmallMammalTrapping | Sheet3

Figure 7: Example for good data entry. DataONE 2012d

Figure 8 shows more problematic data records. In the left column (red), the data structure is inconsistent. If you use underscores or hyphens, stick to one mark and be consistent in using it. In the right column (green) there are several values in one cell as well as several units per variable were used. Only data of one unit should be entered, additional information like the aspect can be entered in a separate column.



Horizon	Slope
10-40-B	10% S
1030B	11 W
B_10_20	0

Figure 8: Example for problematic data records.

Practical tips for data entry in spreadsheets

When entering data manually in spreadsheets, the data validation feature of Excel can help to prevent data entry errors and detect erroneous values. The data validation feature can be used to define what can be entered in a cell (e.g. characters, positive values) and warns the user if any other content is entered. Also a valid range between a minimum and a maximum value can be defined. The feature is accessed via Data > Data Tools > Data Validation. Double data entry or controlling the data entry by another person is another way to reduce data entry errors.

A challenge for exporting data or further analysing it is when there is more than one dataset on one sheet. This is sometimes helpful for data entry, but makes subsequent work difficult. Not only in this case it makes sense to separate data entry from the dataset. This can be done by preparing a data entry mask e.g. in MS-Excel (Figure 9). The data which is entered will be automatically saved in a data file.

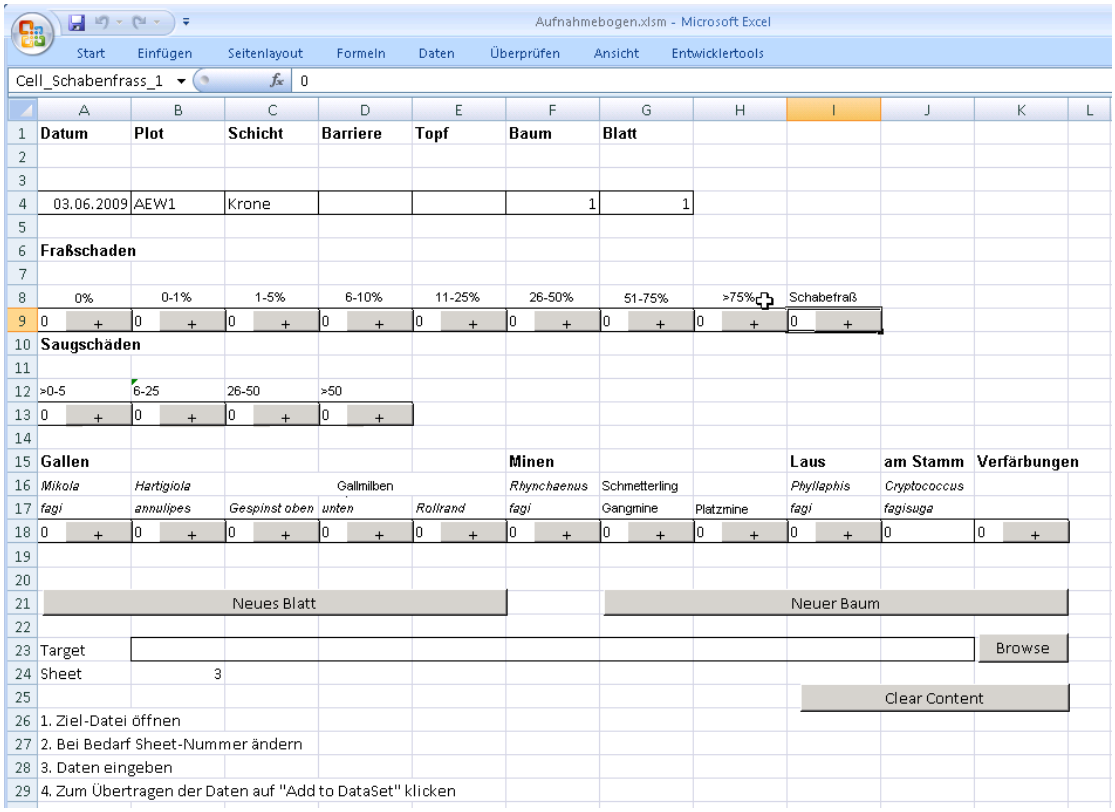


Figure 9: Example for data entry mask in Excel. Dennis Heimann.

File Naming

As well as the file formats, also the file names matter. Naming files is important for organising data on a lab's network drive or personal hard disk and for identifying it later. In data repositories the corresponding information is available in the metadata. File names should be unique and use ASCII characters and avoid spaces. The latter ensures that the file can be read by different operation systems and programs.

The amount and type of information in a file title varies, depending on the type and amount of data and the projects requirements. The content of the file should be reflected in the title. In Table 2 example file namings are displayed. The first name has very little information, whereas the third title is already very long. Much of that information could be documented in the metadata (project, place, time of collection, time of processing, subject).

Table 2: Example file naming.

Title	Information content
Water samples	???
Rhine_water_samples_20140901_V1.0	Rhine (where) water_samples (what) 20140901 (when) V1.0 (version status)
Ecoproject_2011_2014_Water_quality_ Rhine_water_samples_Cologne_ 20140901_V1.0	Additional information can be documented in metadata

Including a version number in the file name is a good idea to identify the most recent file, be able to return to older versions and indicate that changes and transformations have been executed on the data. In Table 3, an example for a versioning system used by BExIS is given. Changes in metadata are indicated by the third digit, smaller changes to the dataset by the second digit and major alterations by the first digit.

Action	Version number
I create a new dataset (title does not exist in BExIS).	1.0.0
I upload some data into the dataset.	1.1.0
I make some changes in the metadata (e.g. the address).	1.1.1
I delete some faulty data from the dataset.	1.2.1
The next year, I create a new dataset based upon the dataset I created before.	2.0.0
I upload some data to my newly created dataset.	2.1.0
Etc.	Etc.

Table 3: File versioning in BExIS. BExIS How To: Version numbers in BExIS.

Resources

MANTRA Video: Jeff Haywood talks about the importance of good file management in research. https://www.youtube.com/watch?v=i2jcOJOFUZg

Software Carpentry: Lecture on data management.
https://www.youtube.com/watch?v=3MEJ38BO6Mo&html5=True

New York University Health Sciences Library: How to avoid a Data Management Nightmare.
https://www.youtube.com/watch?v=nNBiCcBlwRA00_SomeDataManagement_1_140902

UK Data Service: Formatting and organising research data.
http://ukdataservice.ac.uk/media/440281/formattingorganising.pdf

3.3 Assure

Assure refers to quality control and assurance. The latter encompasses all those activities which ensure the reliability of data. High quality data are a key element for research and impact replicability of results. Quality checks should be performed during collection, data entry and analysis and answer the following questions:

- Are the data complete?
- Are the data correct?
- Is the format consistent throughout the data set?
- If it contains errors, which errors?
- Are there missing values?

Quality assurance is already applied prior to data collection by defining standards for formats, codes, units and metadata. Also the assignment of responsibility for data quality is part of quality assurance (Michener and Jones 2012). In the validation process it should be checked whether data are incomplete, unreasonable or inaccurate. This can already be included in the data entry process (see 3.2 Collect). Statistical and graphical summaries (e.g. max/min, average, range) help to check for impossible values and outliers. After validation, the dataset is cleaned. This means to check outliers, correct and fix errors.

Basic quality control

Quality control practices are specific to the type of data being collected, but some generalities exist (DataONE n.y. Best Practices: Ensure basic quality control):

Data collected by instruments: Values recorded by instruments should be checked to ensure they are within the sensible range of the instrument and the property being measured. Example: Concentrations cannot be < 0, and wind speed cannot exceed the maximum speed that the anemometer can record.

Analytical results: Values measured in the laboratory should be checked to ensure that they are within the detection limit of the analytical method and are valid for what is being measured. If values are below the detection limit, they should be properly coded and qualified. Any complementary data used to assess data quality should be described and stored. Example: data used to compare instrument readings against known standards.

Observations (such as bird counts or plant cover): Range checks and comparisons with historic maxima will help identify anomalous values that require further investigation. Comparing current and past measurements help identify highly unlikely events. For example, it is unlikely that the girth of a tree will decrease from one year to the next.

Dates and times: Ensure that dates and times are valid. Time zones should be clearly indicated (UTC or local).

Data Types: Values should be consistent with the data type (integer, character, date, time) of the column in which they are entered. Example: 12-20-2000A should not be entered in a column of dates. Use consistent data types in your data files. A database, for instance, will prevent entry of a string into a column identified as holding integer data.

Geographic coordinates: Map coordinates to detect errors.

(DataONE n.y. Best Practices: Ensure basic quality control)

Outliers

Outliers are unexpected values. They may not be the result of actual observations, but rather the result of errors in data collection, data recording, or other parts of the Data Life Cycle. To identify outliers, statistical tests can be used (Dixon's test, Grubbs test, Tietjen-Moore test). Another possibility is the visualization of data (Box plots, scatter plots when there is an expected pattern, such as a daily cycle). A third way for detecting outliers is the comparison to related observations. Difference plots for co-located data streams can show unreasonable variation between data sources (Example: Difference plots from weather stations in close proximity or from redundant sensors can be constructed). Comparisons of two parameters that should covary can indicate data contamination (Example: Declining soil moisture and increasing temperature are likely to result in decreasing evapotranspiration) (DataONE n.y. Best Practices: Identify outliers).
No outliers should be removed without careful consideration and verification that they are not representing true phenomena. Although outliers may be valid observations it is important to identify and examine their validity. Outliers may represent data contamination, a violation of the assumptions of the study, or failure of the instrumentation (DataONE n.y. Best Practices: Identify outliers).

Documentation of quality

Document in the metadata what quality checks were performed and what was their result. The data quality can be communicated by flagging values, coding or labelling of the dataset and in metadata documentation. For example, entire data sets can be labelled 0 for unexamined, -1 for potential problems and 1 for "good data". Some research communities have developed their own standard protocols (DataONE n.y. Best Practices: Mark data with quality control flags). In Table 4, an example for quality codes used by the University of Oregon for data about solar radiation is shown. Similar to codes is the practice of file versioning. Version numbers indicate that a data set was subject to any kind of transformation (see 3.2 Collect).

Table 4: Example for quality codes. http://solardat.uoregon.edu/QualityControlFlags.html

First digit	**Second digit**	
1 - Observed data		
	1	Raw data
	2	Processed data
	3	Possible problems in data
9 – Missing or bad data		
	9	Missing or bad data

Quality assurance also detects missing values. Missing values are common in environmental data. They can occur if they were not collected, not analysed or got lost. Impossible values are values which lie outside the range for a parameter (e.g. a negative pH). The codes in Table 5 can be used to indicate missing values in a data field. Especially "none" and "unknown" should be used meaningfully. "None" can be a proper value (e.g. a "0" for precipitation) whereas "unknown" means that a value is missing. The chosen code should be used consistently throughout the data set and be documented in the metadata.

Table 5: Codes for missing values. DataONE (n.y.) Best Practices: Identify Missing Values

Code	**Use for**
-999.99	Use only if this is an impossible value within the data set
Not applicable Unknown None	Character fields
Pending assignment	When additional information is compiled to complement value

Resources

Open Refine (tool to work with messy data). http://openrefine.org/

R (software environment for statistic computing and graphics). http://www.r-project.org/

Matlab (high-level language and interactive environment). http://uk.mathworks.com/products/matlab/

Finding the Antarctic Ozone Hole: This is an illustrative example how outliers or values rated to be erroneous prevented NASA detecting the Ozone Hole. http://www.statsci.org/data/general/ozonehol.html

Data Validation in Excel. https://support.office.com/en-za/article/Apply-data-validation-to-cells-c743a24a-bc48-41f1-bd92-95b6aeeb73c9

3.4 Describe

Additional information about data is called metadata. Metadata describe all aspects of data (e.g. who, why, what, when and where) that would allow one to understand the physical format, content and context of the data, as well as possibly how to acquire, use and cite the data (Michener and Jones 2012). Sometimes it can be unclear if a value is considered as metadata or as a record. For instance, for one research approach the locational information of an observation or experiment is metadata, whereas in another approach this is "primary" data directly underpinning research results.

Metadata can describe the **data**

- Title of data set
- how, where, by whom and when the data were collected: methods, dates, persons, places
- data content and format
- the data itself: units, parameters, abbreviations and acronyms
- quality of data, version of data set
- information about the project

Metadata can describe the **digital object**

- Technical information about the file: is special software needed? What format do the files have?
- Administrative information: restriction of re-use, how to cite

Metadata accompanying a data set should be written for somebody re-using them 20 years later: what information is necessary to use the data? Data and documentation should be prepared in a way that is understandable for someone unfamiliar with your project and methods. Descriptive and clear writing helps to understand metadata easily. Do not assume that acronyms or abbreviations are eternally comprehensible. Jargon, technical terms, acronyms and abbreviations should be explained. When data are made available for sharing, metadata should give the user enough information to use the dataset without contacting the author. Figure 10 illustrates the importance of metadata: what is meant by RW or HW? What do the values indicate?

1	Plot	RW	HW
2	20	4382600	5674600
3	32	4383921	5675121
4	68	4386710	5672800
5	274	4382900	5673600
6	1620	4386200	5676270
7	87	4387600	5675690

Figure 10: Variable names are hard to understand without metadata.

Metadata is ideally created as soon as possible to minimize the loss of information. Metadata can be provided in simple text documents or spreadsheets where a key/value approach is recommended. In spreadsheets two columns can be used, in a text document, the information can be separated by a colon (e.g. author: John Doe;

taxonomic reference list: WoRMS). This schema is human and machine readable. Especially when it is intended to submit data to a repository, it is helpful and often necessary to use a metadata standard. Data without associated description are impossible to archive, because not enough is known about the data to ascertain what they are, or whether they are worth curation attention (DCC 2004-2015).

Tips for writing Metadata

- Keywords: It is recommended to use thesauri.
- Where: Fully qualify geographic locations. Think of using bounding box coordinates defining an area by two longitudes and two latitudes for multiple locations; name the corresponding coordinate system.
- Methodology: Document what you did (a published article may not give enough information but a reference to it is a good starter).
- Data Quality: Clearly state data limitations and quality (e.g. data set omissions, completeness of data).

Metadata Standards

Metadata can be made available in a simple text form, accompanying the data set. Another possibility is to use a metadata standard. Metadata standards usually structure the information using XML (Extensible Markup Language) and are both human and machine readable. XML can still be viewed with any text editor but allows for easy verification of syntax and partly content. There exist different standards for different purposes, depending on the discipline. It may be tricky to find the right standard. Librarians might be able to give advice. Some institutions also suggest or prescribe metadata standards in their data policies. In Table 6, some example metadata standards are listed.
Remember: a computer will read the metadata. This means you have to be very careful using symbols. Do not use symbols that could be misinterpreted (**! @ # % { }** | - / \ < > ~). Tabs, indents, line feeds or carriage returns could also be misinterpreted. When copying and pasting from other sources, use a text editor to eliminate hidden characters.

Table 6: Examples of metadata standards.

Metadata Standard		**Special focus**	**URL**
Ecological Metadata Language	EML	Ecological data	http://knb.ecoinformatics.org/eml_metadata_guide.html
Darwin Code	DWC	Museum specimen	http://rs.tdwg.org/dwc/index.htm
Dublin Core Element Set	DC	web resources, publications	http://dublincore.org/documents/dces/
ISO 19115 ISO 19139		Geospatial data and services	http://www.fgdc.gov/metadata/geospatial-metadata-standards#fgdcendorsedisostandards

Access to Biological Collection Data	ABCD	Biological collection data	http://www.tdwg.org/standards/115/
Metadata Encoding and Transmission Standard	METS	descriptive, administrative, and structural metadata for objects within a digital library	http://www.loc.gov/standards/mets/
DataCite Metadata Schema		Publication of data	http://schema.datacite.org/ http://schema.datacite.org/meta/kernel-3/example/datacite-example-full-v3.1.xml

Resources

Morpho (Programme to enter metadata, stored in files conform to EML). https://knb.ecoinformatics.org/#tools/morpho

ISA-tab (Tool for creating metadata). http://isatab.sourceforge.net/format.html

Digital Curation Centre: online catalogue of disciplinary metadata standards. http://www.dcc.ac.uk/resources/metadata-standards.

3.5 Submit

Submission is the transfer of data to a curated environment. This is usually an archive, a data centre, a repository, or a collection. Submission to a curated environment ensures safe long term storage and makes data discoverable for other researchers (in the team or outside). Please note: discoverable does not necessarily imply immediately accessible, as many repositories allow for limited (commonly 6-12 months) embargo times. During submission phase, researchers can decide on how the data can be accessed. In any case, if submitted to an archive or repository, the data will be preserved and curated. Access may vary from immediately available for re-use, available with restrictions or an embargo or not accessible for others. See Table 7 for different restriction possibilities on data. If restriction applies make sure that contact details are up to date. The access to some kind of data may be restricted due to its content to protect individuals or interests (confidential information about persons, ethical issues, environmental protection, endangered species, intellectual property, laws, security) (Downs 2013). Contact you library or the targeted repository for possible solutions for anonymisation or secured off-line storage.

Table 7: Restrictions for data. Downs 2013

Access and use	Dissemination and copies
Limit who may access data or how they are used (authorized users)	May not be authorized to re-distribute or copy data
Authorize use only for specific purposes, such as education	Limit distribution to a specific location or service
Limit whether data may be used in new products or services	Limit distribution to a specific time period, possibly in the future
Modifications or derivations of data may be prohibited	May apply to any products created from data

If data are submitted to a data centre, the files receive a persistent identifier. These identifiers are independent of an URL, which might change over time. A common persistent identifier is the Digital Object Identifier (DOI). It consists of numbers and letters. The DOI can be assigned to an internet location, the preferred is http://dx.doi.org/ + DOI = URL (ESRC 2012). Figure 11 shows an example URL containing a DOI. The registration agencies are in charge of administrating DOI registers and resolving DOIs to actual URLs. For research data the agency is "DataCiteConsortium", which is composed of libraries and data centres around the world. TIB German National Library of Science and Technology in Hanover (Germany) is the leading organization for the DataCite Consortium.

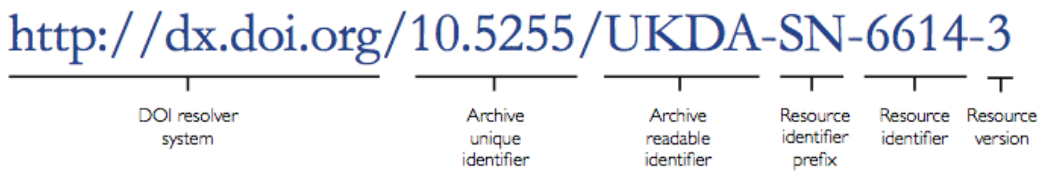


Figure 11: Elements of a Digital Object Identifier (DOI). ESRC 2012

How to submit data

- Use appropriate repositories and data catalogues (see links below).
- License the data so it is clear how they can be re-used (see 3.10 Publish).
- Make sure it's clear how to cite the data (see link below).

What to keep?

Often, it's not necessary or possible to keep everything. Decision criteria should be documented.

- Which data set or digital resources (e.g. the script/source code employed in the analysis) do you want to keep?
- Which elements or characteristics of those data sets or digital resources do you want to keep?
- What do you need to keep to support you research findings?
- What has to be kept (e.g. data underlying publications)?
- What can't be recreated (e.g. environmental recordings)?
- What is potentially useful to others, to whom, what should they be able to do with the data?
- What has scientific, cultural or historical value?
- What must be destroyed due to legislation?

(DCC 2010, 2014)

File formats

It is worth thinking about software and file formats when preparing data for long term storage. Proprietary software may become obsolete, or newer versions won't open older files anymore. Additionally, open formats and standards facilitate exchange and can often be handled with the programme of choice. Also, ASCII formats should usually be preferred over binary formats. For instance, R changed its internal binary storage format and it is likely that reading/importing such data will incur additional work at best and render re-use impossible at worst. In Table 8, recommended file formats are listed for different types of data. Some formats are suitable or even necessary to work with during collection or analysis phase, but are not recommended for storing and archiving data (e.g. Excel). Recommended formats either do not need special software or are fully and openly described. In any case, you should keep the raw data in a format that can be easily processed.

Table 8: Recommended file formats for storing and sharing data. UK Data Service 2012-2015c

Type	Recommended	Avoid for data sharing
Tabular data	CSV (Comma separated values), TSV (Tab separated values)	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

Resources

Repositories:
PANGAEA (cooperation with GFBio): http://www.pangaea.de/
Dryad: http://datadryad.org/
Figshare: http://figshare.com/
Symbiota (biology): http://symbiota.org/docs/
VertNet (biodiversity): http://www.vertnet.org
Morphbank (biological images): http://www.morphbank.net/
iDigBio (Integrated Digitized Biocollections, specimen): https://www.idigbio.org

Re3data: Registry of research data repositories. http://www.re3data.org/

Five steps to decide what data to keep (DCC): http://www.dcc.ac.uk/resources/how-guides/five-steps-decide-what-data-keep

How to cite Data sets (DCC):http://www.dcc.ac.uk/resources/how-guides/cite-datasets

Data Deposition Decision Tree (University of Oxford):
http://researchdata.ox.ac.uk/files/2014/01/Data-deposit-decision-tree.pdf

Nature (http://www.nature.com/authors/policies/availability.html):

> *Unstructured repositories like figshare and Dryad are suitable alternatives if no structured public repositories exist.*
>
> *When repositories do not exist for a particular data type, authors can deposit and share data via figshare or Dryad, two general-purpose scientific data repositories.*

PLOS (http://journals.plos.org/plosone/s/data-availability):

> *PLOS requires that authors comply with field-specific standards for preparation and recording of data and select repositories appropriate to their field, for example deposition of microarray data in ArrayExpress or GEO; deposition of gene sequences in GenBank, EMBL or DDBJ; and deposition of ecological data in Dryad. Authors are encouraged to select repositories that meet accepted criteria as trustworthy digital repositories.*

3.6 Preserve

Digital data are fragile. Hardware fails, software becomes obsolete, files are subject to bit rot which produces bit errors. Digital preservation as one part of digital curation encompasses a set of actions which ensure long-term usability by maintaining the accessibility, integrity and longevity of data:

- Longevity: extend the lifespan of data for current and future user requirements
- Integrity: ensure the authenticity and reliability of data (no undocumented manipulations)
- Accessibility: store data in formats which ensure their future use

Preservation protects against data loss and obsolescence. Migration and emulation are principal preservation actions (see below). These actions are carried out by specialized curators. Researchers and project staff can also carry out some measures by themselves. At an early stage, researchers should ask themselves what they want others to be able to do with their data as this will impact many aspects of preservation (DCC 2009). Data can be stored or migrated to a suitable format, avoiding proprietary file formats (see 3.5 Submit). Also, a suitable medium should be chosen with care. The lifespan of digital storage media have to be considered. A burned CD may only last some years and USB-sticks get lost easily. Backups and storage as short-term preservation is the duty of the research staff (see below). Also responsibilities have to be taken into account. Who looks after stored data, who is responsible for transferring data to a curated environment? The creation of metadata and documentation of preservation actions helps to understand and use data later on. Significant properties of data recorded in the metadata help the people carrying out preservation actions. The actual archiving and long-term preservation actions are performed by specialists like professional curators (DCC 2008d).
Be critical when reviewing “best practices”. They might works for specific scenarios but not for you. For support, consult and work with experts in your field. Data centres are there to offer solutions, help and professional curation of data.

Backup

Backup of data is not the same as preservation. Backups are short-term recovery solutions whereas preservation includes measures taken for long term storage and archiving. For backups, there exist ideally at least 3 copies of a file, on at least 2 different media, with at least 1 offsite. Managed services like university drives are always a better choice than external hard drives or USB-sticks, but make sure you know the conditions as e.g. normal network drives may keep backup copies of a file for only 6 months and if overwritten with messy data and looked at after 7 months you may run into trouble. The IT-Team of your institution might give support and advice.

Digital preservation methods in data curation

Preservation actions aim to ensure that data are in the best possible shape to be stored or archived by a repository. The proceeding steps of the Data Life Cycle (3.3 Assure, 3.4 Describe) are crucial for this.

Migration

Migration refers to the transformation of data or other digital material from one format or technology to another (software or hardware). The ability to retrieve, display and use the contents is maintained by this action (DCC 2008b). Files can be migrated within one software product to a newer version or to other file formats when obsolescence occurs. Every migration changes data a little bit. If migrations are carried out multiple times, data can be subject to major changes. Chris Rusbridge describes a commonly- experienced migration result:

> *So if I migrate from those PowerPoint 4 files to today's PowerPoint, and then from today's to tomorrow's PowerPoint, and then from tomorrow's to the next great thing, I will introduce cumulative errors whose impact I will only be able to assess at some horribly cringe-making moment, like in the middle of a presentation using a host's machine. So the best way to do migration is to start from the original file and migrate to today's version (DCC 2008b).*

Because functionality is lost and integrity comprised as a result of migration, care must be taken to strict quality checking procedures, for example, to compare the original bit stream and the migrated bit stream. To check data integrity and detect bit errors, checksums can be used. A checksum algorithm calculates a value which can be compared to the value of the original file. A widely used algorithm is MD5.

Refreshment

Media refreshment refers to copying data to a new medium after a fixed time (e.g. 5 years) to preserve the bit stream and consider the life span of digital storage media.

Emulation

Emulation refers to the development of software that can mimic (obsolete) systems on current and future generation of computers. By emulating applications, operating systems and hardware architecture, older files and programmes can still be used and software can be kept alive (e.g. using a 16 bit application like Harvard Graphics on a DOS System) (DCC 2008c).

Resources

University of Cambridge: Preservation.
http://www.lib.cam.ac.uk/dataman/pages/preservation.html

UK Data Service: Store your data. http://ukdataservice.ac.uk/manage-data/store.aspx

3.7 Discover

Discovering data means to search and find data collected by other researchers. This data can be used for different purposes like long-time analysis, modelling or comparative studies. Data and metadata are made accessible through submission on any kind of shared environment. The pre-requisite for discovering data is that the authors are willing to share their data with the research community. Michener and Jones (2012) name the problem:

> *Many valuable and relevant data products are not readily available as they are stored on laptops and computers in the offices of individual scientists, projects and institutions.*

So, only submitted and published data can be discovered. Specialized data archives, centres and repositories offer search functions to discover relevant data. Another way to discover data is via journals, when data are appended to articles. Often it is necessary to register at a repository or archive to get access to the data. A persistent identifier is important to locate the data, e.g. if the URL changes. The persistent identifier is also an important element for citing data which were discovered and used. GFBio offers a faceted search tool that searches the associated data centres (Figure 12). In the field at the top, any word can be entered. When the results of the search are displayed, a search facet appears on the left. Here, you can specify your search by choosing authors, publication years, geographic regions and data centres. The search results are displayed geographically by clicking on the trolley symbol.

Figure 12: GFBio search screenshot. http://www.gfbio.org/archives/-data-centers

Checklist

- Are data accessible, are there any use restrictions?
- Do appropriate metadata exist for citing, understanding and evaluating the data?
- Is the provenance of the data visible?

If you want to make your own data available and discoverable, it is indispensable to create metadata to communicate significant properties of the data. Consider if there should be any access or use restrictions (DCC 2008d).

Resources

GFBio search function: http://www.gfbio.org/data-portal

PANGAEA (cooperation with GFBio): http://www.pangaea.de/

knb (Knowledge Network for Biocomplexity): https://knb.ecoinformatics.org/#about

Thredds Data Server (data server which provides metadata and data access for scientific data sets): http://www.unidata.ucar.edu/software/thredds/current/tds/

ModisLand (land cover satellite data): http://modis-land.gsfc.nasa.gov

Dryad (especially natural and environmental sciences): http://datadryad.org/discover?query=&submit=Search#

Global Spatial Data Infrastructure (Links to geographic data portals): http://www.gsdi.org/ElectronicGateways (international)

NOAA US National Oceanic and Atmospheric Administration Data Center: http://www.ngdc.noaa.gov/ngdcinfo/onlineaccess.html

NASA Global Change Master Directory: http://gcmd.nasa.gov/index.html

Symbiota (biology): http://symbiota.org/docs/

VertNet (biodiversity): http://www.vertnet.org

Morphbank (biological images): http://www.morphbank.net/

iDigBio (Integrated Digitized Biocollections, specimen): https://www.idigbio.org

BISE Biodiversity Information System for Europe: http://biodiversity.europa.eu/data

GBIF Global Biodiversity Information Facility: http://www.gbif.org/

Re3data: Registry of research data repositories. http://www.re3data.org/

3.8 Integrate

Integration is the merging of multiple datasets from different sources, like your recently collected data with former data from other owners, resulting in a new, bigger dataset. You might want to integrate a dataset that you discovered and that fits to your own data in order to verify your results, as a starting point for an integrative study or just to test a new hypothesis for a follow-up study. Large-scale ecological studies require the integration of data from different studies and disciplines (e.g. population studies, hydrology and meteorology; Michener and Jones 2012). Integrating data from different sources is

> *labor intensive and time consuming, because it requires understanding methodological differences, transforming data into a common representation, and manually converting and recoding data to compatible semantics before analysis can begin. Data integration for crosscutting studies is generally a manual process and can consume the majority of time involved in conducting collaborative research […] (Michener and Jones 2012).*

For an effective integration it is important that in the early steps of the Data Life Cycle, good data management practices were applied. That means for example that metadata about the dataset were created, like information about the quality of the data, measurement methods or abbreviations. This allows to assess fitness of data for re-use. When other authors' data are re-used, it is fundamental to provide credit to the data creators through a robust data citation mechanism.

Data can be integrated manually or automatically. It is important to document the integration workflow, all the steps used for cleaning, analysing and visualizing the data and new data products so that the output is comprehensible.
There are common problems when integrating data that complicate database compilation. These include for example the lack of a consistent coordinate reference system, taxonomic naming inconsistencies, and the semantics of variable names (Rüegg et al. 2014). If integrating datasets, ensure that formats and parameters are compatible (datum, resolution, metric units) and assess the quality of the data.

Document the integration of multiple datasets

Ideally, mechanisms to systematically capture the integration process are adopted, e.g. in an executable form such as a script or workflow. Another possibility is to document the process, scripts, or queries in the metadata or any other accompanying document. Also describe the relationship between datasets from different sources.
When you use data from other sources than your own collection, it is important to indicate this. Citing allows for tracing the chain of use of datasets and data elements, credit the creators of the data as well as the possibility that if errors or new information about the original datasets or data elements comes to light, that any impact on your new datasets and interpretation of such could be traced (DataONE n.y. Best Practices: Document the integration of multiple data sets).

Data Citation

Data have to be cited like bibliographic resources have to be cited in publications. But there are some differences between data and publications (Mayernik 2013):

- Data sets are extremely variable in structure, representation schemes, and access mechanisms, unlike traditional scholarly publications.
- Data sets can be dynamic (changing or growing) whereas scholarly articles do not change once published.
- Data sets are often hierarchical composites of data and metadata files, software, and other related documents.
- Authorship is a problematic notion in relation to data sets, particularly in distributed and collaborative work.

However, data which are integrated from other sources can be cited. Figure 13 displays an example citation with all properties which should be included in the citation of a data set (ESRC 2012; Mayernik 2013). If available, also more information can be included, like the type of resource or a version number.

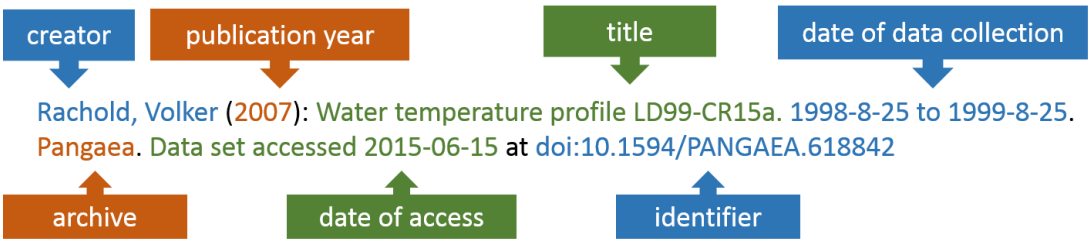


Figure 13: Example citation

When you are a data user, it is useful to provide feedback about the datasets you have used. By providing feedback it is possible to reduce repetition of mistakes, improve data quality and provide chances for further publications (Duerr 2011).

Resources

Digital Curation Centre (DCC): How to Cite Datasets and Link to Publications.
http://www.dcc.ac.uk/resources/how-guides/cite-datasets

Digital Curation Centre (DCC): Data Citation and linking.
www.dcc.ac.uk/resources/briefing-papers/introduction-curation/data-citation-and-linking

Data Citation what you need to know.
www.esds.ac.uk/news/publications/data_citation_online.pdf

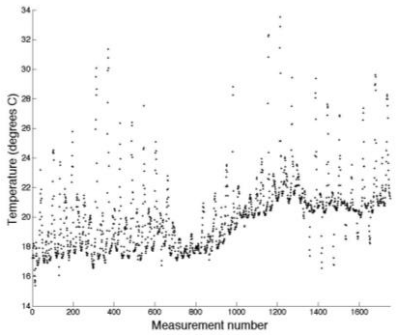
3.9 Analyse

Analysis comprises the actions used to derive and understand information from data. The types of analyses depend on the discipline and on the research questions to be answered. Furthermore, the software and hardware used to analyse data also vary.

Statistics and visualisation

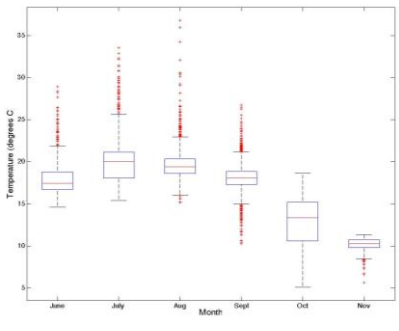
Statistics are one of the most common types of analyses used for quantitative data. They comprise for example (multivariate) analyses of variance or regressions. Conventional statistics tend to rely on assumptions such as random error and homogeneous variance. Descriptive statistics are traditionally applied to observational data. Descriptive data might include the distribution of organisms in space or community-habitat relationships. Statistics used to understand these types of data include diversity indices, cluster analysis, and principle components analysis, among many others. Statistical analyses might also include temporal or spatial analyses, and nonparametric approaches which do not rely on specific assumptions about the data's distribution. Other kinds of data analyses are simulations, model runs, parameter estimations and visualisations (also compromising plots and graphs; DataONE 2012b). Visual representation allows to detect patterns within the data. Plots can be used to assure the quality of data as well. They can quickly show you potential data errors such as impossible values. In Figure 14, on the left is a scatter plot of temperatures for the month of August. The general pattern is easily discernible, and particularly warm measurements are readily apparent. On the right is a box and whisker plot of monthly temperatures. The boxes indicate averages, and measurements far from the averages are visible as red dots outside of the error bars.

Scatter plot of August temperatures



Strasser, unpub. data

Box and whisker plot of temperature by month



Strasser, unpub. data

Figure 14: Examples of visualisations. DataONE 2012b

Before starting to analyse the data, it can be necessary to **process** it to be more usable. Processing can include for example selecting a subset, merging multiple datasets or reduce the amount of data. It can also be necessary to transform data prior to analysis. That includes the conversion of values to common units or the adjustment of data

collected by multiple people. Analysing very large datasets requires special considerations. Data of interest have to be discovered and identified. Often high-performance computing systems are necessary to process big amounts of data (DataONE 2012b).

Documentation

Data analysis often consist of several steps. Analyses are run, hypotheses modified, data transformed and complemented, and again analysed. This complex process can be hard to reproduce if it is not properly documented. The documentation of data cleaning, transformation and analysis is called process metadata. Related to process metadata is the concept of data provenance, which explains the origins of the data at hand. In order to make data provenance clear, process metadata allow others to follow the data through their life cycle and understand all steps used to create results and output. In other words, the scientific procedure is described by the process metadata. Good provenance allows for the ability to replicate analyses and reproduce results (DataONE 2012b).
Process metadata can be documented in informal or formal workflows. A workflow is essentially higher-level metadata. It is a conceptualized series of data ingestion, transformation, and analytical steps. Informal workflows consist of diagrams or flowcharts, guiding you through data processing. In Figure 15, a flowchart is displayed. The green boxes in the middle indicate the transformation rules (what you are doing to your data), and the blue boxes on the left and right represent the actual data. First, temperature and salinity data are imported in R. The output is the data in R format. These data are then controlled and cleaned, resulting in "clean" data. These clean data are the input for the analysis step. The output, the summary statistics, is used to produce a graph.

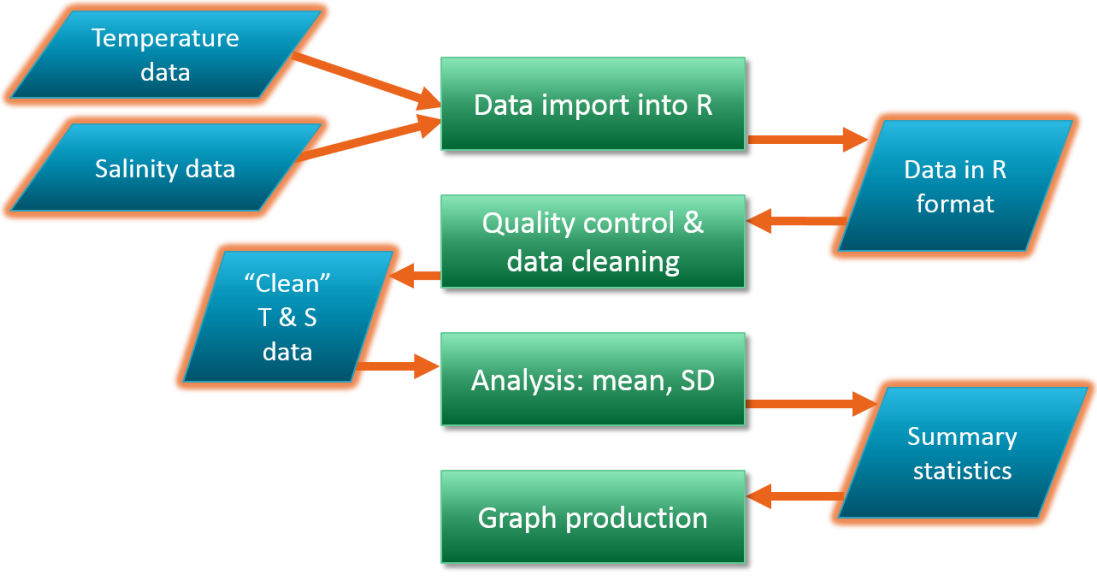


Figure 15: Informal workflow in form of a flowchart. DataONE 2012b

Another type of an informal flowchart would be commented scripts. These can be applied when working with scripted languages, e.g. R. A commented script is a text file containing the used commands and additional information, like descriptions of the

commands. The advantage of using R or Matlab for analysis instead of Excel is that you can produce a script and therefore know exactly what transformations the data have gone through. Michener and Jones (2012) state:

> *Ecologists have traditionally used tools such as Excel to manipulate and convert data manually for integration; however, this process is error-prone and is not reproducible because of the lack of provenance regarding these operations. Scripted analysis environments, such as R and Matlab, improve this by providing a record of data manipulations, but are still largely a record of procedural manipulations of data.*

Besides informal workflows, also formal ones exist. Formal workflows are produced using a specialized software like Kepler or VisTrails. They allow to integrate different software systems in one workflow. In the process of performing analyses, each step and its parameters and requirements are formally recorded (DataONE 2012b). This allows both the individual steps and the overall workflow to be re-used, either by the original scientist or someone else. The open-source and free cross-platform programme Kepler uses a drag-and-drop interface for scientists to construct their workflow. There exist pre-programmed components e.g. for R for easy application (Kepler 2008). Additionally, Kepler provides access to data repositories, computing resources and workflow libraries (Kepler 2008).

Resources

Data Pub articles about workflows (Blog by California Digital Library):
http://datapub.cdlib.org/2012/06/05/workflows-part-i-informal/
http://datapub.cdlib.org/2012/06/12/workflows-part-ii-formal/

Kepler: https://kepler-project.org

VisTrails: http://www.vistrails.org/index.php/Main_Page

3.10 Publish

The last step of the Data Life Cycle deals with the publication of data, especially with datasets linked with the publication of related academic papers. It is based on the fact that datasets are a fundamental part of the research process, as important as discussions and conclusions derived from them (AGU 2012):

> *The scientific community should recognize the professional value of such [data] activities by endorsing the concept of publication of data, to be credited and cited like the products of any other scientific activity, and encouraging peer-review of such publications.*

Furthermore, publication of datasets promotes transparency in the research cycle and facilitates verification and reproducibility. Therefore, it is important to establish mechanisms which allow sharing datasets while at the same time promoting rewards mechanisms for authors, as journals do for academic publications.
More and more academic journals (e.g. Nature, PLOS) include the publication of the datasets within their publication requirements. Publishing research data also brings advantages to the researcher. The work becomes more visible and the citation rate may increase.

Best practices for publishing and sharing data

- Work with journal publishers and data repositories to archive data during the publication process.
- Information about how to access your data set can be published with your article. Data repositories typically offer to embargo archived data for a pre-determined time period after publication.
- Archived data sets should get a persistent identifier (DOI) through the holding data centre so that they can be uniquely cited.
- Encourage other data authors to cite data and to make their own data available for re-use.
- Provide full citation information for data whenever you publish work that makes use of another author's data.
- Archive your own data in a repository that supports data discovery and re-use.
- Update your archived data sets when newer versions are available.

(DataONE 2012c)

Where to publish your data

There are many different possibilities where to publish your data. There exist a lot of data centres, often specialised on one discipline or a field of research (e.g. GFBio for biological end environmental data). See the resource list in chapter 3.7 Discover for data centre suggestions. Specialized data journals offer the possibility to publish short papers cross-linked to and describing a data set which is deposited in a data centre. If a dataset is published in relation to a paper, the journal might prescribe a data centre where to archive the data.

Data publishing process

The following steps are a general overview about how the deposition of data in the UK Data Archive works (UK Data Archive 2002-2015):

1. Data transfer: depositor transfers data to the archive
2. Data processing: curators (in cooperation with depositors) check data against the documentation supplied, prepare archival and dissemination versions and document processing
3. Metadata creation: a catalogue record based on specific standards is created
4. Additional user information: additional information is provided e.g. by a "read me"-file
5. Publishing data: after data and documentation are complete, material is transferred to preservation systems. The catalogue record is published and the data becomes available to users.

Licences

Published data is ideally licenced. By this, it is clear for others how they can use the data and what they are allowed to do with them. Two common licences are described below.

Creative Commons Licences: http://creativecommons.org/
As data do not underlie the copyright, the Creative Commons Licence Zero CC0 applies. It means that no rights are reserved and data may be re-used and enhanced. Nonetheless, appropriate citation can be requested. Data repositories like PANGAEA or Dryad publish data under the CC0 licence.
For data other than raw data like figures, media, posters and papers, the CC-BY licence can be used. This licence states that the work has an author. It allows to use and enhance as long as the author is credited.

Open Data Licences: http://opendatacommons.org/licenses/
Open Data Commons offers three different licences:
Public Domain Dedication and License (PDDL): You are free to copy, distribute and use the database; to produce works from the database and to modify, transform and build upon the database without restriction.
Attribution License (ODC-By): You are free to copy, distribute and use the database; to produce works from the database and to modify, transform and build upon the database as long as you attribute as specified in the licence.
Open Database License (ODC-ODbL): You are free to copy, distribute and use the database; to produce works from the database and to modify, transform and build upon the database as long as you attribute, share alike and keep the data open (no technological measures to restrict the work).

Sharing

Sharing and publication of data have already been discussed previously in this reader (Chapters 2, 3.5). Its advantages were explained and possible barriers discussed, some of them which are hard to solve. But there are also some less serious opinions against

sharing and publication which can be resolved. See Figure 16 for some arguments from people reluctant to data sharing. But there are answers and solutions at hand, many of them discussed in the proceeding chapters. So, go for it, share your data and enjoy the benefits (Figure 17)!

We will get too many enquiries.	Data centres offer solutions for publishing data.
The data has poor quality.	Data labels allow to document data quality. Maybe it's not useful to you, but to somebody else with another approach?
We might want to use it in a paper.	Choose an embargo time. And the journal might even ask for data publication.
It's too big.	Data centres are specialised in dealing with big amounts of data
Lawyers want a custom licence.	There are many possibilities to license your data.
People may misinterpret the data.	Metadata explain characteristics of the data.
It's too complicated.	Data managers, librarians and curators can help you.
Data protection	Data centres can give special advice on how to publish data with content that has to be protected.

Figure 16: Solutions to doubts about data sharing. Own design, after Corti et al. 2011 and Open Data Bingo

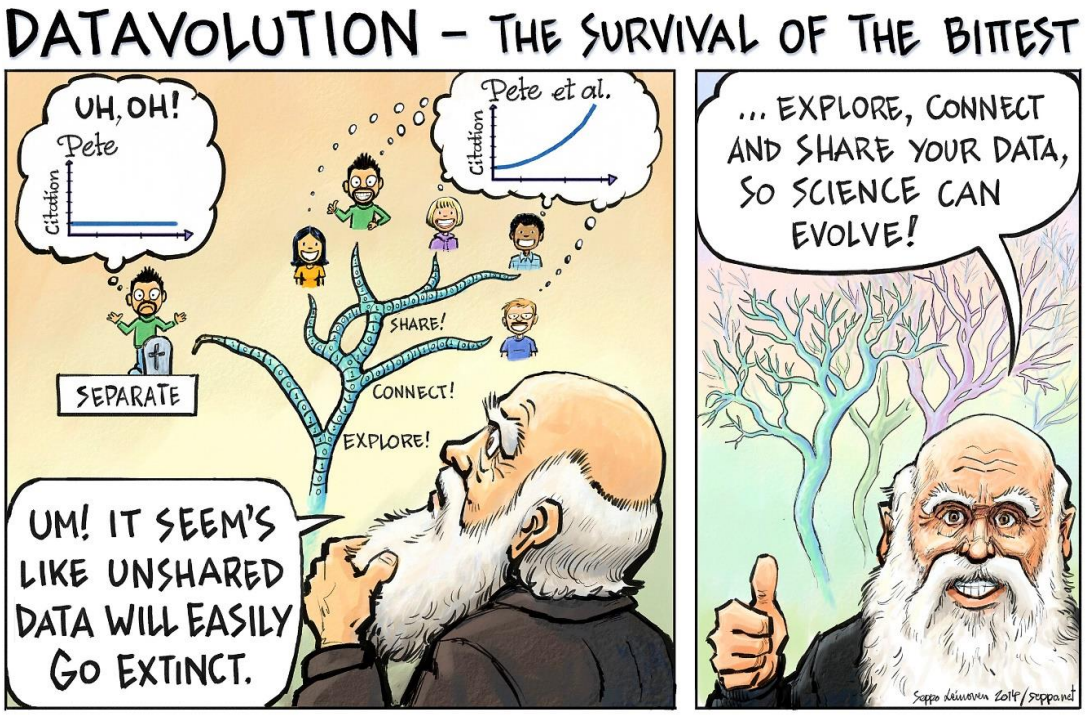


Figure 17: Benefits of sharing data. GFBio-Postcard realised by Seppo

Resources

Korn, N., Oppenheim, C. (2011). Licensing Open Data: A Practical Guide.
http://discovery.ac.uk/files/pdf/Licensing_Open_Data_A_Practical_Guide.pdf

Digital Curation Centre: How to license research data.
www.dcc.ac.uk/resources/how-guides/license-research-data

Research Data Netherlands: Addressing a researcher's data sharing concerns.
https://www.youtube.com/watch?v=8qRLgQa1wT4

Data Pub article about sharing excuses (Blog by California Digital Library):
http://datapub.cdlib.org/2013/04/24/closed-data-excuses-excuses/

Data Journals:
Nature Scientific Data: http://www.nature.com/sdata/
Biodiversity Data Journal: http://biodiversitydatajournal.com/
UP meta journals (Journal for meta papers): http://metajnl.com/
Geoscience data journal:
http://onlinelibrary.wiley.com/journal/10.1002/%28ISSN%292049-6060

Re3data: Registriy of research data repositories. http://www.re3data.org/

4. Data Management with BExIS, version 1

BExIS is a platform for data storage and information exchange in scientific projects, especially aimed at joint research projects, collaborative research centres or research training groups. It provides central storage, access and exchange facilities for data, with a focus on tabular data.
In this chapter, some functionality of BExIS is described. More detailed descriptions and How-to-guides are distributed with BExIS-instances.

Metadata

The first action in the process of uploading data in BExIS is to create metadata. This ensures that no data can be uploaded without corresponding metadata. The metadata created by BExIS is compatible to EML. Nine steps, each on a separate page, guide you through metadata creation:

1. general information: title of the dataset and information about the owner
2. research objects
3. methodology: objectives of the project, research methods
4. description: acronyms and keywords
5. time: set format of time and date used in the data set, start and end of data collection
6. data: all information to understand the data set, data quality, data structure, variables (name, type, units, description)
7. references: links to additional documentation, publications
8. comment
9. validation

Especially the description of the data in step 6 is important. The structure of the data set defined here will determine what data can be uploaded later on. If the data is structured data, BExIS reads the data while uploading them and compares them to the predefined structure. This function is part of quality assurance as it helps to ensure that data are consistent and detects erroneous values (if they have another type). If for example a variable is defined as integer number, the upload of a dataset containing characters in a field of that variable will fail.

Upload structured and unstructured data

Once a metadata entry is generated, data can be uploaded. In the metadata entry, the file type to be uploaded is already determined. Structured data allow for filtering or exporting selected parts. The data are stored in spreadsheets, csv or tsv. The file type unstructured data is intended for the upload of documents, models, images and other data with no tabular arrangement.

Edit structured data

Single records in structured data sets can be directly edited in BExIS using the online interface. Extensive transformations are better carried out by downloading the data, manipulating it with a programme of choice and then uploading it again. BExIS automatically creates a new version number if any changes have been performed on the data set.

File Versioning

BExIS generates file version numbers automatically. The versioning system is explained in Figure 18.

Action	Version number
I create a new dataset (title does not exist in BExIS).	1.0.0
I upload some data into the dataset.	1.1.0
I make some changes in the metadata (e.g. the address).	1.1.1
I delete some faulty data from the dataset.	1.2.1
The next year, I create a new dataset based upon the dataset I created before.	2.0.0
I upload some data to my newly created dataset.	2.1.0
Etc.	Etc.

Figure 18: BExIS versioning system. BExIS How To: Version numbers in BExIS.

Access data and right management

BExIS is designed for an easy exchange of data. Metadata uploaded in BExIS is visible to everybody. BExIS is conservative with access rights. Data is by default only visible to the owner and available to others on request. The owner can decide to make data available to a certain group or the entire project by asking the administrator to grant access rights. The data owner gets an overview about who downloaded the data and when.

Download data

Download of data is only possible when you have appropriate access rights. If you download any data you will receive an automatically generated email once a month if there have been any changes or new versions of the data set. If somebody downloaded your data, you will get a notification E-Mail. Under Upload > Owner Info you can see, who downloaded the data.

5. Data Management at a glance: Summary

Data management is about work with research data in a structured, efficient, and conscious way, all throughout the 10-step Data Life Cycle:

1. Propose: To get the most benefit out of data management, start early. Establish a Data Management Plan at proposal stage.
2. Collect: Core activity for generating data. Take special care with data entry in spreadsheets. Organize files in a structured way, optimized for analysis. Avoid redundancies.
3. Assure: Assure data quality prior to analysis. Document quality using data labels and value flags. Also transformations should be documented so people can judge if data are fit for use.
4. Describe: Metadata makes data understandable (values, parameters, methods…) and findable. It also provides information about the creator, which is necessary for correct citation.
5. Submit: Submit data to a curated environment, especially for long-term storage and for sharing and publication.
6. Preserve: Back-ups and short-term storage within the project to secure data. Long-term storage and special preservation actions are carried out by repositories.
7. Discover: Search, find and include already existing data to enrich your research.
8. Integrate: Merging of data sets, data may be created by others. Special focus in this phase lies on compatibility and citing.
9. Analyse: Statistical, visual and other analyses to derive information from data.
10. Publish: Publish a data set individually or underpinning a publication.

The way we work with data is currently undergoing a period of change. Access to and re-use of data is a growing concern of funding agencies the world over. The digital transformation causes a deluge of data as well as collaboration capabilities and changes the way research is conducted. What proper data management should look like in the future will very much depend on the discipline and research approach.
In this handbook we tried to cover basic concepts of proper data management, including practical advices, following the Data Life Cycle. Success will hinge on how well introduced concepts can be integrated into day-to-day research. It will depend on both an understanding of researchers about data management and data management providers about research processes and needs. Tangible benefits of data management are already evident in form of security, compliance, access, quality and efficiency and we hope that interest in and adoption of data management will further grow among the scientific community.

6. More Data Management: Recommended further reading

German Federation for the Curation of Biological Data (GFBio):
http://www.gfbio.org

Provides access to data, tools for visualisation, integration and management of data, as well as training resources, like fact-sheets for each step in the Data Life Cycle (http://www.gfbio.org/fact-sheets).

Digital Curation Centre (DCC): http://www.dcc.ac.uk/

Centre of expertise in digital curation, with a focus on building capacity, capability and skills for research data management across the UK's higher education research community. Provides a lot of detailed information material, including checklists, briefing papers, case studies and workshop materials for professional curators.

UK Data Service: http://ukdataservice.ac.uk/

Guides, advice and training on data management, with focus on social data.

UK Data Archive: http://www.data-archive.ac.uk/

Collection of digital research data in social sciences and humanities, with information about curation und data management.

Handbook on Data Management:

Corti L, Van der Eynden V, Bishop L, Morgan-Brett B (2011) Managing and sharing data: training resources. UK Data Archive (2014 edition published at SAGE)
http://repository.essex.ac.uk/2398/1/TrainingResourcesPack.pdf

Van der Eynden, V (2014): Managing and sharing your research data. UK Data Archive http://ukdataservice.ac.uk/media/455282/dm_stirling_2014.pdf (extensive presentation corresponding to the handbook)

Data Observation Network for Earth (DataONE): https://www.dataone.org/

A US based cyberinfrastructure or the environmental science, huge best practice collection and education modules prepared in presentations and fact sheets.

DataONE Best Practice Primer:
https://www.dataone.org/sites/all/documents/DataONE_BP_Primer_020212.pdf

Primer which gives a first overview about best practices in data management.

Mantra: http://datalib.edina.ac.uk/mantra/

Research Data Management Training online course by the University of Edinburgh. Includes material for students, researchers and information professionals as well as data handling tutorials in R and ArcGIS.

7. Glossary

Access Possibility to search and retrieve data. Covers also rights to download or re-use data

Archiving To place or store data in a data centre or otherwise managed environment; typically done for ensuring long-term preservation of the data and to promote discovery and use of the data (Strasser et al. 2012). Also refers to a curation activity which ensures that data is properly selected, stored, can be accessed and that its logical and physical integrity is maintained over time (RDA Europe 2014b).

Big data Data sets encompassing a big amount of records, normally collected automatically by satellites, remote sensing, instruments or sensor networks. Often personal computers cannot handle such data, grid computing or clouds are used. The term Big Data is not defined by a specific quantity but "is often used when speaking about petabytes and exabytes of data" (NIST 2015:10 after Dutcher 2014).

Bit stream Sequence of bits that is identified as a unit, constituting a digital object.

Citable data Citable data is a type of referable data that has undergone quality assessment and can be referred to as citation in publications (RDA Europe 2014b).

Curation Work of specialised curators to ensure good management of data, especially ensuring preservation and promoting re-use.

Data Information used for research.

Database An organized collection of data. A database can be classified by the type of content included in it (e.g. bibliographic, statistical, document-text) or by its application area (e.g. biological, geological; Strasser et al. 2012).

Data catalogue Curated collection of metadata about datasets (RDA Europe 2014b).

Data centre (see also repository) Facility specialized in the professional work with data (e.g. acquiring data from providers and make it available to re-users), can be a repository, archive or other kind of centre. Data centres also offer e.g. user help desk support and training, support data processing activities and other services.

Data citation A reference to data created by others, used analogical to bibliographic references like citation of journal articles.

Data documentation Metadata or information about data that enables one to understand and use the data. Comprises also the documentation of data transformations and analyses (process metadata; Strasser et al. 2012).

Data element A logical, identifiable unit of data that forms the basic organizational component in a database. Usually a combination of characters or bytes referring to one separate piece of information. A data element may combine with one or more other data elements or digital objects to form a digital record**.**

Data entropy Normal degradation in information content associated with data and metadata over time.

Data Life Cycle Steps data undergo in their life cycle concerning data management. The Data Life Cycle encompasses all facets of data generation to knowledge creation, including the proposal, collection and quality assurance, metadata creation, data submission and preservation, discovery, integration, analysis and data publication.

Deposit The act of submitting data, as to a repository (Strasser et al. 2012).

Derived data set New dataset created by using multiple existing datasets and their data elements as sources (Strasser et al. 2012).

Digital curation: see curation

Digital data Refers to a structured sequence of bits/bytes that represents information content. In many contexts digital data and data are used interchangeably implying both the bits and the content (RDA Europe 2014b).

Digital object Consists of a structured sequence of bits/bytes. As an object it is named. This bit sequence can be identified and accessed by a unique and persistent identifier or by use of referencing attributes describing its properties.

DOI (**Digital Object Identifier)** A DOI is a string of letters and numbers that can be used to make resources directly available to anyone over the internet. Used for data products so that they can be uniquely identified and cited (Michener and Jones 2012).

Faceted search Enables users to discover specific data products by filtering a set of available descriptors (such as author, repository where the data are stored, sensors used to collect the data, geographic location; Michener and Jones 2012).

Header Meaningful name for referencing the content contained in a row or column, as in a spreadsheet (Strasser et al. 2012).

Integrity Accuracy, validity and consistency of data. Integrity means that data has not been changed or manipulated undocumented.

Meta-analysis An analysis that combines the results of multiple studies.

Metadata Provides information about and documentation of data content, context, quality and structure. Data without metadata cannot be used.

Metadata editing tool Software tool to input, edit, and view metadata using a metadata standard.

Metadata standard Machine readable metadata using extensible markup language (XML) and following a prescribed structure, depending on the chosen standard. Ensures consistency in content and format of metadata.

Non-proprietary Refers to software not protected by patent, copyright, or trademark, often with open source code.

Ontology A framework for interrelated concepts within a domain. An ontology would for example link the terms “water vapour”, “relative humidity”, and “H_2O vapour pressure”, so that a user searching for one, would also see the other related terms and their relationships (Strasser et al. 2012).

Operations Data manipulation

Parameter A variable and measurable factor that determines or characterizes a system.

Persistent identifiers Persistent identifiers are labels for digital objects that remain the same regardless of where the object is located. The use of persistent identifiers is not limited to web material. It is equally essential for linking and citation of primary research with datasets (DCC Glossary).

Provenance Data provenance refers to the ability to track data from creation through all transformations, analyses and interpretations, enabling full understanding of the processes used to create derived scientific products (Michener and Jones 2012).

Quality assurance/quality control (QA/QC) Refers to the mechanisms preventing errors from entering a data set and to monitor and maintain data quality (Michener and Jones 2012). Quality control is testing and other activities designed to identify problems and errors (Strasser et al. 2012).

Quality level flag Indicator within the data file that identifies the level of quality of a particular data point or data set. Flags can be defined within the metadata (Strasser et al. 2012).

Referable data Type of data that is persistently stored and has a persistent identifier (RDA Europe 2014b).

Relational database Collection of tables (often called relations) with defined relationships to one another (Strasser et al. 2012).

Repository Facility specialized on storing data and make it available to re-users. Repositories provide curation and stewardship services, access to data products and search interfaces.

Re-use Use of data by other researchers or for a purpose other than that for which they were collected. Re-use of data can be limited by file formats or missing and incomplete metadata.

Scripted programme Programme (requiring a command line interface or similar) that performs an action or task. Scripts can be saved, modified and re-used as necessary (Strasser et al. 2012).

Scientific workflow Description of the scientific procedure, for example what analyses have been carried out on data. Can be informal (e.g. flowchart) or formal (e.g. workflow systems).

Structured data Tabular data stored in spreadsheets, csv or tsv. Structured data allow for filtering or exporting selected parts.

Unstructured data Data with no tabular arrangement, like documents or models.

8. References

AGU - American Geophysical Union. (2012). Data Position Statement. Retrieved from http://sciencepolicy.agu.org/files/2013/07/AGU-Data-Position-Statement_March-2012.pdf

Corti, L., Van der Eynden, V., Bishop, L., & Morgan-Brett, B. (2011). Managing and sharing data: training resources. UK Data Archive.

DataONE. (n.y.). Best Practices. Retrieved May 28, 2015, from https://www.dataone.org/best-practices

DataONE. (2012a). DataONE Education Module: Data Management. Retrieved May 22, 2015, from http://www.dataone.org/sites/all/documents/L01_DataManagement.pptx

DataONE. (2012b). DataONE Education Module: Analysis and Workflows. Retrieved May 29, 2015, from http://www.dataone.org/sites/all/documents/L10_Analysis Workflows.pptx

DataONE. (2012c). DataONE Education Module: Data Citation. Retrieved April 22, 2012, from http://www.dataone.org/sites/all/documents/L09_DataCitation.pptx

DataONE. (2012d). DataONE Education Module: Data Entry and Manipulation. Retrieved April 29, 2015, from http://www.dataone.org/sites/all/documents/L04_DataEntryManipulation.pptx

DataONE Community. (2014). DataONE Education Module Lesson 1: Data Management Handout. Retrieved April 7, 2015, from https://www.dataone.org/sites/all/documents/L01_DataManagement_Handout_FINAL.pdf

DCC - Digital Curation Centre. (2008a). Digital Curation 101 Incentives for Digital Curation. Retrieved from http://www.dcc.ac.uk/sites/default/files/documents/DC%20101%20Incentives%20for%20Curation.pdf

DCC - Digital Curation Centre. (2008b). Digital Curation 101 Migrate. Retrieved from http://www.dcc.ac.uk/sites/default/files/documents/DC%20101%20Migrate.pdf

DCC - Digital Curation Centre. (2008c). Digital Curation 101 Preservation Methods. Retrieved from http://www.dcc.ac.uk/sites/default/files/documents/DC%20101%20Preservation%20Methods.pdf

DCC - Digital Curation Centre. (2008d). Digital Curation 101 - Access, use and reuse. Retrieved from http://www.dcc.ac.uk/sites/default/files/documents/DC%20101%20Access%20and%20Use.pdf

DCC - Digital Curation Centre. (n.y.). Glossary. Retrieved July 2, 2015, from http://www.dcc.ac.uk/digital-curation/glossary#B

DCC - Digital Curation Centre. (2004-2015). Using Metadata Standards. Retrieved June 2, 2015, from http://www.dcc.ac.uk/resources/briefing-papers/standards-watch-papers/using-metadata-standards

DCC - Digital Curation Centre. (2008-2015). DC 101 materials. Retrieved from http://www.dcc.ac.uk/training/train-the-trainer/dc-101-training-materials

DCC - Digital Curation Centre. (2009). Preservation Action Checklist. Retrieved from http://www.dcc.ac.uk/sites/default/files/documents/Preservation%20Action%20Checklist.pdf

DCC - Digital Curation Centre. (2010). How to Appraise and Select Research Data for Curation. Retrieved May 29, 2015, from http://www.dcc.ac.uk/resources/how-guides/appraise-select-data

DCC - Digital Curation Centre. (2014). Five steps to decide what data to keep. Retrieved May 29, 2015, from http://www.dcc.ac.uk/resources/how-guides/five-steps-decide-what-data-keep

DFG - Deutsche Forschungsgemeinschaft. (2014). Leitfaden für die Antragstellung: Projektanträge. Retrieved from http://www.dfg.de/formulare/54_01/54_01_de.pdf

DFG - Deutsche Forschungsgemeinschaft. (2015). Guidelines on the Handling of Research Data in Biodiversity Research. Working Group on Data Management of the DFG Senate Commission on Biodiversity Research. Information for Researchers No. 36. Retrieved from http://www.dfg.de/download/pdf/foerderung/antragstellung/forschungsdaten/guidelines_biodiversity_research.pdf

Downs, R. R. (2013). Responsible Data Use: Data restrictions. Retrieved from commons.esipfed.org/sites/default/files/RDUDataRestrictionsDowns_final.pptx

Duerr, R. (2011). Responsible data use. Presented at the AGU Data Management 101 for the Earth Scientist. ESIP. Retrieved from http://wiki.esipfed.org/images/7/71/ResponsibleDataUse_FINAL.ppt

ESRC - Economic and Social Research Council. (2012). Data Citation. What you need to know. Retrieved from http://www.esrc.ac.uk/_images/Data_citation_booklet_tcm8-21453.pdf

Heidorn, B. (2008). Shedding Light on the Dark Data in the Long Tail of Science. Library Trends, 57(2). http://doi.org/http://hdl.handle.net/2142/9127

Houghton, J. (2011). Costs and Benefits of Data Provision. Centre for Strategic Economic Studies, Victoria University. Retrieved from http://ands.org.au/resource/houghton-cost-benefit-study.pdf

Kepler. (2008). Getting started with Kepler. Retrieved from https://kepler-project.org/users/documentation

Ludwig, J., & Enke, H. (Eds.). (2013). Leitfaden zum Forschungsdaten-Management: Handreichungen aus dem WissGrid-Projekt. Glückstadt: Hülsbusch. Retrieved from http://www.univerlag.uni-goettingen.de/bitstream/handle/3/isbn-978-3-86488-032-2/leitfaden_DGRID.pdf?sequence=1

Mantra, EDINA, & Data Library, University of Edinburgh. (2014). Research Data MANTRA (online course). Retrieved April 20, 2015, from http://datalib.edina.ac.uk/mantra#sthash.c9HnlX89.dpuf

Mayernik, M. (2013). Responsible Data Use: Citation and Credit. Presented at the Data Management for Scientists Short Course; Federation of Earth Science Information Partners: ESIP Commons. National Center for Atmospheric Research, DOI 10.7269/P3QC01D5. Retrieved from http://live.commons.esipfed.bluedotapps.org/node/1428

Michener, W. K. (2006). Meta-information concepts for ecological data management. Ecological Informatics, 1(1), 3–7. http://doi.org/doi:10.1016/j.ecoinf.2005.08.004

Michener, W. K., & Jones, M. B. (2012). Ecoinformatics: supporting ecology as a data-intensive science. Trends in Ecology & Evolution, 27(2), 85–93. http://doi.org/10.1016/j.tree.2011.11.016

NIST - National Institute of Standards and Technology. (2015). NIST Big Data Interoperability Framework: Volume 1, Definitions. DRAFT. Big Data Public Working Group.

Open Data Bingo. (n.y.). Retrieved May 15, 2015, from http://data.dev8d.org/devbingo/

Piwowar, H. A., & Vision, T. J. (2013). Data reuse and the open data citation advantage. PeerJ, 1(e175). http://doi.org/https://dx.doi.org/10.7717/peerj.175

RDA Europe - Research Data Alliance. (2014a). Report on the RDA-MPG Science Workshop on Data. Retrieved from https://europe.rd-alliance.org/sites/default/files/report/RDA-Europe-Science-Workshop-Report_final_April2014.pdf

RDA Europe - Research Data Alliance. (2014b). DFT Model Overview & Term Definitions - Prepared for Plenary 3. Dublin.

Rüegg, J., Gried, C., Bond-Lamberty, B., Bowen, G. J., Felzer, B. S., McIntyre, N. E., Soranno, P.A., Vanderbilt, K.L., Weathers, K. C. (2014). Completing the data life cycle: using information management in macrosystems ecology research. Frontiers in Ecology and the Environment, 12(1), 24–30. http://doi.org/10.1890/120375

Strasser, C., Cook, R., Michener, W., & Budden, A. (2012). Primer on Data Management: What you always wanted to know. (DataONE, Ed.). Retrieved from www.dataone.org

UK Data Archive. (2002-2015). How we curate data: the process. Retrieved June 1, 2015, from http://www.data-archive.ac.uk/curate/process

UK Data Service. (2012-2015a). Prepare and manage data. Retrieved May 22, 2015, from http://ukdataservice.ac.uk/manage-data.aspx

UK Data Service. (2012-2015b). Why share data? Retrieved May 22, 2015, from http://ukdataservice.ac.uk/manage-data/plan/why-share.aspx

UK Data Service. (2012-2015c). Recommended formats. Retrieved June 1, 2015, from http://ukdataservice.ac.uk/manage-data/format/recommended-formats.aspx

UK Data Service (Ed.). (2014a). Benefits of managing and sharing your data. University of Essex. Retrieved from http://ukdataservice.ac.uk/media/440285/whysharedata.pdf

UK Data Service. (2014b). Formatting and organising research data. Retrieved June 2, 2015, from http://ukdataservice.ac.uk/media/440281/formattingorganising.pdf

University of Western Australia. (2015). Research Data Management Toolkit. Retrieved May 20, 2015, from http://guides.is.uwa.edu.au/RDMtoolkit