FOSTER

Research Data Management
Data Management Plans

Gwen Franck (EIFL)
Helene Brinken (UGOE)
KTU, Kaunas, Lithuania
April 3rd 2019

Programme

Introduction 9-9.30

Data collection, documentation and metadata 9.30 - 9.50

Storage and backup 9.50 - 10.10

Ethics and legal issues 10.10-10.30

Break 10.30 - 10.45

Selection and preservation 10.45 - 11.05

Data sharing 11.05 - 11.25

Responsibilities and resources 11.25 - 11.45
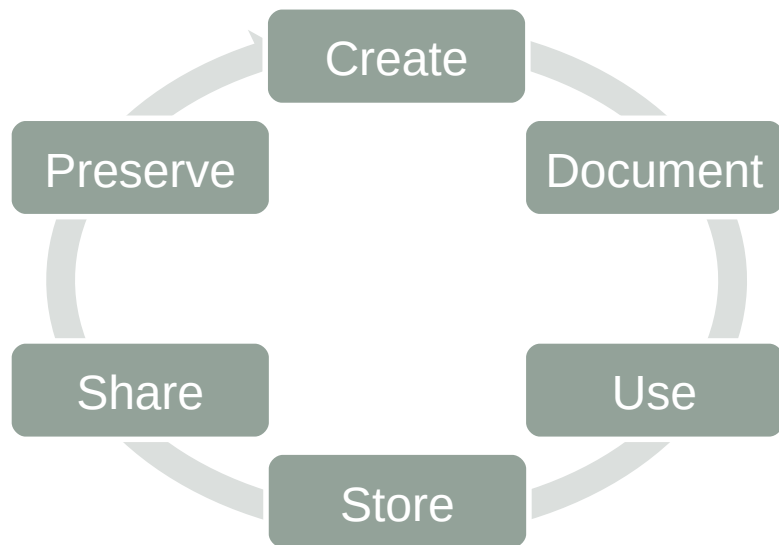
Closing exercise 11.45 - 12.00

Warming up exercises

- Age
- Have you already written a DMP
- Have you ever 'lost' data?
- Did you ever NOT have access to a dataset you needed?
- My data is ready to re-use

INTRODUCTION

(Gwen)

What is Research Data Management?

Create

Document

Use

Store

Share

Preserve

"the active management and appraisal of data over the lifecycle of scholarly and scientific interest"

Data management is part of good research

FOSTER

Go to www.menti.com and use the code 77 08 85

Why manage data?

Pause scroll

Why manage data?

(Not for the research funder, but for life we make data management plans)

- Make your research easier
- Stop yourself drowning in irrelevant stuff
- Save data for later
- Avoid accusations of fraud or bad science
- Write a data paper
- Share your data for re-use
- Get credit for it

NON PECUNIAE INVESTIGATIONIS CURATORE

SED VITAE FACIMUS PROGRAMMAS DATORUM PROCURATIONIS

It's part of good research practice

"It was *never* acceptable to publish papers without making data available."

- Ewan Birney

#OpenData
#OpenScience

Original image via doi:10.1038/461145a. "Research cannot flourish if data are not preserved and made accessible. Data management should be woven into every course in science." - *Nature* 461, 145

FOSTER

Cut down on academic fraud

Validation of results

"It was a mistake in a spreadsheet that could have been easily overlooked: a few rows left out of an equation to average the values in a column.

The spreadsheet was used to draw the conclusion of an influential 2010 economics paper: that public debt of more than 90% of GDP slows down growth. This conclusion was later cited by the International Monetary Fund and the UK Treasury to justify programmes of austerity that

The error that could subvert George Osborne's austerity programme

The theories on which the chancellor based his cuts policies have been shown to be based on an embarrassing mistake

Charles Arthur and **Phillip Inman**
The Guardian, Thursday 18 April 2013 21.10 BST

George Osborne says that Ken Rogoff, the man whose economic error has been uncovered, has strongly influenced his thinking. Photograph: Stefan Wermuth/PA

www.guardian.co.uk/politics/2013/apr/18/uncovered-error-george-osborne-austerity

More scientific breakthroughs

Sharing of Data Leads to Progress on Alzheimer's

By GINA KOLATA
Published: August 12, 2010

In 2003, a group of scientists and executives from the National Institutes of Health, the Food and Drug Administration, the drug and medical-imaging industries, universities and nonprofit groups joined in a project that experts say had no precedent: a collaborative effort to find the biological markers that show the progression of Alzheimer's disease in the human brain.

🔍 Enlarge This Image
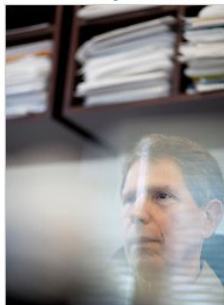
Now, the effort is bearing fruit with a wealth of recent scientific papers on the early diagnosis of Alzheimer's using methods like PET scans and tests of spinal fluid. More than 100 studies are under way to test drugs that might slow or stop the disease.

And the collaboration is already serving as a model for similar efforts against Parkinson's disease. A $40 million project to look for biomarkers for Parkinson's, sponsored by the Michael J. Fox Foundation, plans to enroll 600 study subjects in the United States and Europe.

"It was unbelievable. Its not science the way most of us have practiced in our careers. But we all realised that we would never get biomarkers unless all of us parked our egos and intellectual property noses outside the door and agreed that all of our data would be public immediately."

www.nytimes.com/2010/08/13/health/research/13alzheimer.html?pagewanted=all&_r=0 *John Trojanowski University of Pennsylvania*

FOSTER

A citation advantage

A study that analysed the citation counts of 10,555 papers on gene expression studies that created microarray data, showed:

"studies that made data available in a public repository received 9% more citations than similar studies for which the data was not made available"

Data reuse and the open data citation advantage, Piwowar, H. & Vision, T.   https://peerj.com/articles/175

FOSTER

Increased use and economic benefit

The case of NASA Landsat satellite imagery of the Earth's surface:

Up to 2008

- Sold through the US Geological Survey for US$600 per scene

- Sales of 19,000 scenes per year

- Annual revenue of $11.4 million

Since 2009

- Freely available over the internet

- Google Earth now uses the images

- Transmission of 2,100,000 scenes per year.

- Estimated to have created value for the environmental management industry of $935 million, with direct benefit of more than $100 million per year to the US economy

- Has stimulated the development of applications from a large number of companies worldwide

http://earthobservatory.nasa.gov/IOTD/view.php?id=83394&src=ve

BE PART OF THE NEW ERA OF OPEN SCIENCE
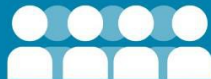
reach more people, have greater impact

avoid duplication of efforts

preserve data for future researchers

simplify final Horizon 2020 reporting thanks to an up-to-date DMP

European Commission

Open Data doesn't just happen - data management planning helps!

- What data will be created (format, types, volume…)

- Standards and methodologies to be used, documentation

- How ethics and Intellectual Property will be addressed

- Plans for storage and back-up

- Plans for data sharing and access
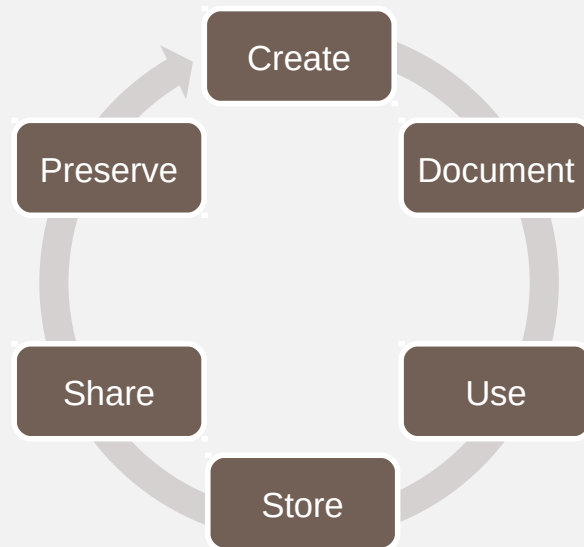
What should NOT be a part of a Data Management Plan?

0 — Information about storage and archiving

0 — Detailed description of each data set

0 — Information about file formats

0 — Ethical and intellectual property considerations

0 — Access and reuse information

eifl KNOWLEDGE WITHOUT BOUNDARIES

Data Management Plans

- Short
- Usable
- 1st version at the start of the project
- Updated during the project
- Internal and/or external use
- Covers all types of data created but not all datasets!

Exercise

- Use the Lithuanian Research Council Template before you
- Keep your own project in mind, or if not applicable, a project you've worked on in the past
- As we discuss each section, fill in (briefly) the part of the DMP template that corresponds with it
- You may not always be able to fill the section in, that's ok!

Sections	Questions	Questions to consider	Answers
A. Data Collection	1. What data will you collect or create?	*1.1. Are there any existing data that you can re-use?* *1.2. What type, format and volume of data?*	
B. Storage and Backup	2. How will the data be stored and backed up during the project?	*2.1. Where will the data be stored?* *2.2. How will the data be recovered in the event of an incident? Will the data be backed up?*	
	3. How will you manage access and security?	*3.1. What are the risks to data security and how will these be managed?* *3.2. How will you ensure that project partners (if applicable) can access your data securely?*	
C. Selection and Preservation	4. Which data are of long-term value and should be retained and preserved?	*4.1. What data must be retained/destroyed for contractual, legal, or regulatory purposes?* *4.2. How long will the data be retained and preserved?*	
D. Data Sharing	5. How will you ensure the availability and sharing of the data?	*5.1. When will you make the data available?* *5.2. How will potential users find out about data?* *5.3. With whom will you share the data, and under what conditions?*	
E. Responsibilities and Resources	6. Who will be responsible for data management?	*6.1. Who is responsible for implementing the DMP, and ensuring it is reviewed and revised?* *6.2. Will data ownership and responsibilities for RDM be part of any consortium agreement or contract agreed between project partners (if applicable)?*	
	7. What human and other resources will you require to draft and deliver your DMP?	*7.1. Is additional specialist expertise required?* *7.2. Do you require hardware or software which is additional or exceptional to existing institutional provision?* *7.3. Have you considered the charges that can be applied by data repositories?*	

FOSTER

Data collection
Documentation and metadata
(Helene)

FOSTER

A. Data Collection

1. What data will you collect or create?

1.1. Are there any existing data that you can re-use?

1.2. What type, format and volume of data?
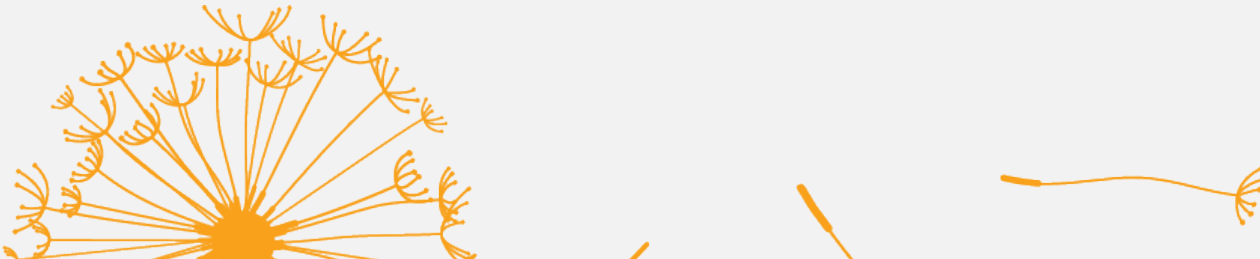
Data Collection

What data will you collect or create?

- What type, format and volume of data?
- Do your chosen formats and software enable sharing and long-term access to the data?
- Are there any existing data that you can reuse?

Give a brief description of the data, including any **existing data or third-party sources** *that will be used, in each case noting its* **content, type and coverage**. *Outline and justify your choice of format and consider the* **implications of data format and data volumes** *in terms of storage, backup and access.*

Data Collection

How will the data be collected or created?

- What standards or methodologies will you use?
- How will you structure and name your folders and files?
- How will you handle versioning?
- What quality assurance processes will you adopt?

*Outline how the data will be collected/created and which **community data standards** (if any) will be used. Consider how the data will be **organised** during the project, mentioning for example **naming conventions, version control and folder structures**. Explain how the **consistency** and **quality** of data collection will be **controlled** and **documented**. This may include processes such as calibration, repeat samples or measurements, standardised data capture or recording, data entry validation, peer review of data or representation with controlled vocabularies.*

What is research data in your field?

Pause scroll

0

What are research data?

The University of Leeds describes research data as 'any information that has been collected, observed, generated or created to validate original research findings'. Research data can include things like...
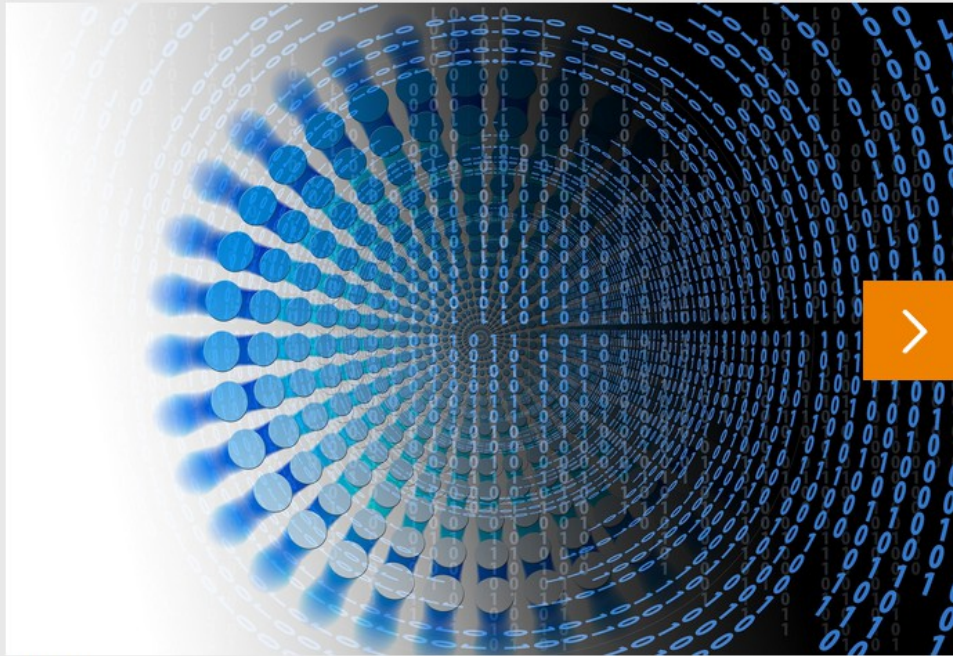
...raw data

Raw data are those which are captured from instruments and sensors such as telescopes, smart phones, and satellites.

Raw data

What are research data?

The University of Leeds describes research data as 'any information that has been collected, observed, generated or created to validate original research findings'. Research data can include things like...
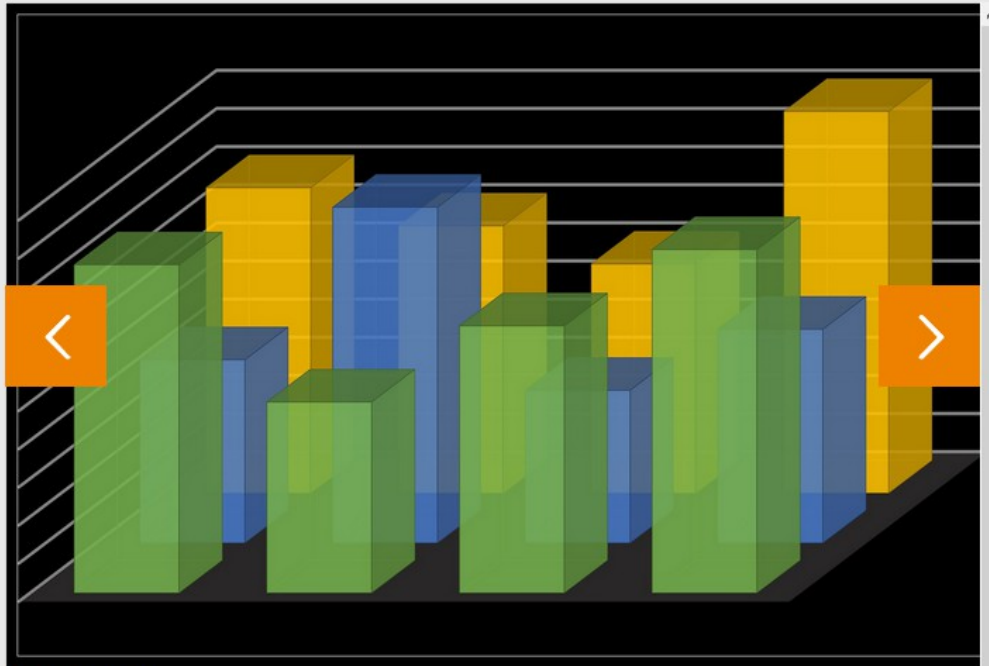
Visualisations

...visualisations, models, and algorithms

Researchers also generate digital resources such as models and algorithms to help them analyse, visualise and present raw data in a meaningful way.

https://www.fosteropenscience.eu/learning/managing-and-sharing-research-data

What are research data?

The University of Leeds describes research data as 'any information that has been collected, observed, generated or created to validate original research findings'. Research data can include things like...

Data can be anything and everything!

...just about anything!

Essentially, research data can be just about anything that researchers produce or work with during the course of their research.

Documentation and Metadata

What documentation and metadata will accompany the data?

- What information is needed for the data to be to be read and interpreted in the future?
- How will you capture / create this documentation and metadata?
- What metadata standards will you use and why?

*Describe the **types of documentation** that will accompany the data to help secondary users to understand and reuse it. This should at least include **basic details** that will help people to find the data, including **who** created or contributed to the data, its **title**, **date** of creation and under what **conditions** it can be accessed. Documentation may also include details on the **methodology** used, analytical and procedural information, **definitions of variables, vocabularies, units of measurement**, any assumptions made, and the **format** and **file type** of the data. Consider how you will capture this information and where it will be recorded. Wherever possible you should identify and use existing community standards.*

YES, MAYBE OR NO?

If you are a capable researcher, you know what the standards in your field are - there's no need to describe
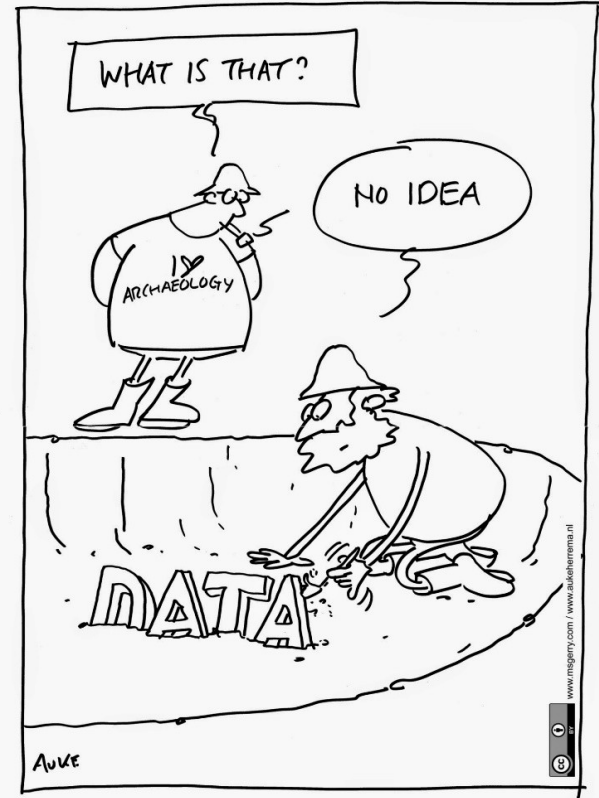
Data organisation

Meaningful file names

Below are tips on meaningful and consistent file names. Read more in '*Choosing a file name*'. (2)

- ◻ Make sure to use consistent file names. When you use a date in the file name, choose a notation (for instance, YYYYMMDD of yymmdd).
- ◻ Do not use strange characters like ?\!@*%{[<> in the file name.
- ◻ Use traceable file names, such as Project_Instrument_locatie_YYYYMMDD.ext.
- ◻ Make sure to only use each file once in the folder structure. If you store a file in more than one place, several versions of the same file can unwillingly be created.
- ◻ See also version management.

It is good practice to note the file naming and its meaning in a readme.txt.

white_data_20140708.csv

blue_data_20140708.docx

red_data_20140708.R

red_data_20140708_v02.R

File naming and version management

Even if a researcher is well underway with his project consistent file naming is still an option by using a *bulk file rename utility*. (3) It is important, however, to check if this bulk renamer delivers on its promises.

http://datasupport.researchdata.nl/en/start-de-cursus/iii-onderzoeksfase/organising-data

cessda eric

About ▾   Consortium ▾   Projects ▾   Research Infrastructure ▾   Contact

TIP: Batch renaming of automatically generated files

Batch renaming is organising research data files and folders in a consistent and automated way with software tools (also known as mass file renaming, bulk renaming).

Batch renaming software exists for most operating systems. See the accordion for examples.

  + Batch renaming tools

It may be useful to rename files in a batch when:

- Images from digital cameras are automatically assigned base filenames consisting of sequential numbers;
- Proprietary software or instrumentation generate crude, default or multiple filenames;
- Files are transferred from a system that supports spaces and/or non-English characters in filenames to one that doesn't (or vice versa). Batch renaming software can be used to substitute such characters with acceptable ones.

https://www.cessda.eu/Research-Infrastructure/Training/Expert-Tour-Guide-on-Data-Management/2.-Organise-Document/File-naming-and-folder-structure

How to ... use Bulk Rename Utility

cessda eric

About ▾   Consortium ▾   Projects ▾   Research Infrastructure ▾   Contact

How to ... use Bulk Rename Utility

Follow the steps in the video to use Bulk Rename Utility to batch rename your files.

Expert Tour Guide on Data Management

1. Plan

2. Organise & Document

Designing a data file structure

Organisation of variables

File naming and folder structure

Documentation and metadata

Adapt your DMP: part 2

Sources and further reading

3. Process

4. Store

5. Protect

6. Archive & Publish
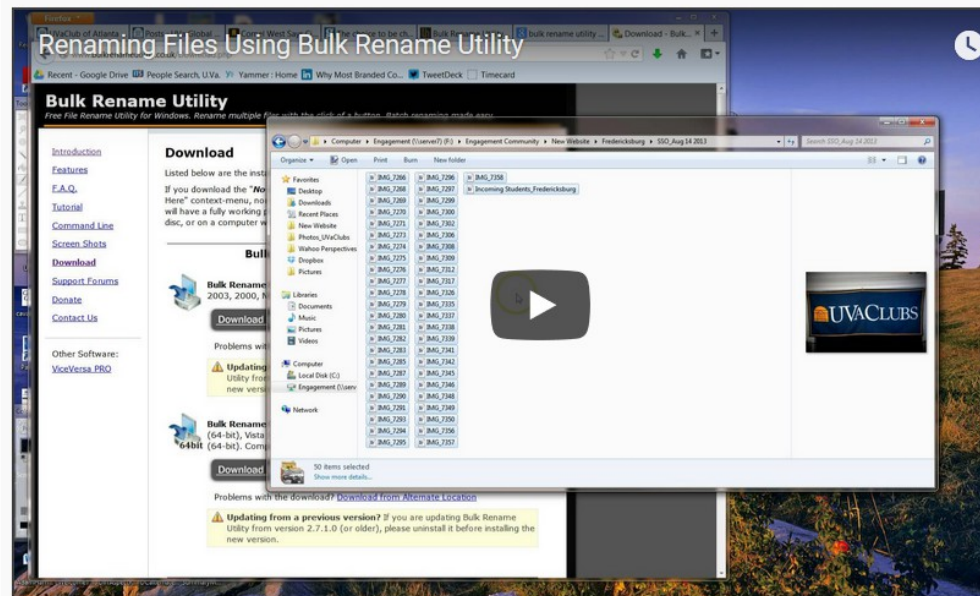
Renaming Files Using Bulk Rename Utility

https://www.cessda.eu/Research-Infrastructure/Training/Expert-Tour-Guide-on-Data-Management/2.-Organise-Document/File-naming-and-folder-structure

« Previous   Next »

www.dcc.ac.uk/resources/subject-areas/social-science-humani   120%

Search

Contact us

D|C|C

because good research needs good data

Search

| Home | Digital curation | About us | News | Events | Resources | Training | Projects | Community | Tailored support |

Home > Resources > Subject Areas > Social Science Humanities

http://www.dcc.ac.uk/resources/subject-areas/social-science-humanities

Social Science & Humanities

Social Policy Rural and Regional) Multi-disciplinary **Statistics** Heritage Studies Creative art and design **Historical and Philosophical Studies** Building **Conservation** Anthropology **Archaeology** Music **Economics** Planning (Urban **Health Policy** History **History by Area** **Sociology** Planning (Urban, Rural and Regional) **Politics** Law **Human and Social Geography** General **Demography** Architecture

Metadata Standards

DDI - Data Documentation Initiative
An international standard for describing data from the social, behavioral, and economic sciences. Expressed in XML, the DDI metadata specification supports the entire research data life cycle.

Encoded Archival Description (EAD)
A standard for encoding archival finding aids using XML in archival and manuscript repositories, implementing the recommendations of the International Council on Archives ISAD(G): General International Standard Archival Description ⊕.

MIDAS-Heritage
A British cultural heritage standard for recording information on buildings, archaeological sites, shipwrecks, parks and gardens, battlefields, areas of interest and artefacts.

OAI-ORE - Open Archives Initiative Object Reuse and Exchange
Defines standards for the description and exchange of aggregations of Web resources.

QuDEx - Qualitative Data Exchange Format

Metadata Standards

DDI - Data Documentation Initiative

An international standard for describing data from the social, behavioral, and economic sciences. Expressed in XML, the DDI metadata specification supports the entire research data life cycle.

SDMX - Statistical Data and Metadata Exchange

A set of common technical and statistical standards and guidelines to be used for the efficient exchange and sharing of statistical data and metadata.

Extensions

CESSDA MLI - Council of European Social Science Data Archives Minimum Level of Information

A common base profile of DDI for use by the member archives of CESSDA.

GESMES/TS - GEneric Statistical MESsage for Time Series

An extension of SDMX used to exchange statistical data and metadata.

Tools

DDI on Rails

Server-side software for building a data portal, with a particular focus on survey datasets. It uses DDI to provide access to the data at the level of concepts and variables. For an example of it in use, see the SOEPinfo data portal.

DDI Tools

The Data Documentaion Initiative website's list of tools to implement the DDI standard.

Istat SDMX Framework Project

A suite of tools for managing data and metadata in SDMX.

SDMX Editor

A simple tool for managing and accessing statistical metadata, using the SDMX framework.

SDMX Mapping Assistant

A tool to facilitate the mapping between the structural metadata provided by an SDMX-ML Data Structure Definition and those that reside in a database of a dissemination environment.

SDMX Tool Repository

A list of software tools supporting the SDMX standard.

Contact us

D|C|C because good research needs good data

[Search]   **Search**

| Home | Digital curation | About us | News | Events | Resources | Training | Projects | Community | Tailored support |

Home > Resources > Metadata Standards > Disciplinary > Materials Science

http://www.dcc.ac.uk/resources/metadata-standards/disciplinary/materials-science

Materials Science

Metadata Standards

CIF - Crystallographic Information Framework
An extensible standard file format and set of protocols for the exchange of crystallographic and related structured data.

CSMD-CCLRC Core Scientific Metadata Model
A study-data oriented model that captures high-level information about scientific studies and the data that they produce, primarily tailored for the physical sciences.

NeXus
NeXus is an international standard for the storage and exchange of neutron, x-ray, and muon experiment data.

http://rd-alliance.github.io/metadata-directory/subjects/

Metadata

RDA | Metadata Directory

Edit this page

View the standards

View the extensions

View the tools

View the use cases

Browse by subject areas

Arts and Humanities ✔ Edit

- Archaeology ✔ Edit
- Creative art and design ✔ Edit
- Heritage Studies ✔ Edit
- Historical and Philosophical Studies ✔ Edit
- History by Area ✔ Edit
- History ✔ Edit
- Law ✔ Edit
- Music ✔ Edit

Engineering ✔ Edit

- Architecture ✔ Edit
- Building Conservation ✔ Edit

Materials Science

Metadata

RDA | Metadata Directory

Edit this page

View the standards

View the extensions

View the tools

View the use cases

Browse by subject areas

Contribute

Add standards

Add extensions

Add tools

Standards ⚙ Add

CIF (Crystallographic Information Framework) ⚙ Edit

A well-established standard file structure for the archiving and distribution of crystallographic information, CIF is in regular use for reporting crystal structure determinations to Acta Crystallographica and other journals.

Sponsored by the International Union of Crystallography, the current standard dates from 1997. As of July 2011, a new version of the CIF standard is under consideration.

CSMD (Core Scientific Metadata Model) ⚙ Edit

A study-data oriented model, primarily in support of the ICAT data managment infrastructure software. The CSMD is designed to support data collected within a large-scale facility's scientific workflow; however the model is also designed to be generic across scientific disciplines.

Sponsored by the Science and Technologies Facilities Council, the latest full specification available is v 4.0, from 2013.

NeXus ⚙ Edit

NeXus is an international standard for the storage and exchange of neutron, x-ray, and muon experiment data. The structure of NeXus files is extremely flexible, allowing the storage of both simple data sets, such as a single data array and its axes, and highly complex data and their associated metadata, such as measurements on a multi-component instrument or numerical simulations. NeXus is built on top of the container format HDF5, and adds domain-specific rules for organizing data within HDF5 files in addition to a dictionary of well-defined domain-specific field names.

Exercise

Sections	Questions	Questions to consider	Answers
A. Data Collection	1. What data will you collect or create?	*1.1. Are there any existing data that you can re-use?* *1.2. What type, format and volume of data?*	
B. Storage and Backup	2. How will the data be stored and backed up during the project?	*2.1. Where will the data be stored?* *2.2. How will the data be recovered in the event of an incident? Will the data be backed up?*	
	3. How will you manage access and security?	*3.1. What are the risks to data security and how will these be managed?* *3.2. How will you ensure that project partners (if applicable) can access your data securely?*	
C. Selection and Preservation	4. Which data are of long-term value and should be retained and preserved?	*4.1. What data must be retained/destroyed for contractual, legal, or regulatory purposes?* *4.2. How long will the data be retained and preserved?*	
D. Data Sharing	5. How will you ensure the availability and sharing of the data?	*5.1. When will you make the data available?* *5.2. How will potential users find out about data?* *5.3. With whom will you share the data, and under what conditions?*	
E. Responsibilities and Resources	6. Who will be responsible for data management?	*6.1. Who is responsible for implementing the DMP, and ensuring it is reviewed and revised?* *6.2. Will data ownership and responsibilities for RDM be part of any consortium agreement or contract agreed between project partners (if applicable)?*	
	7. What human and other resources will you require to draft and deliver your DMP?	*7.1. Is additional specialist expertise required?* *7.2. Do you require hardware or software which is additional or exceptional to existing institutional provision?* *7.3. Have you considered the charges that can be applied by data repositories?*	

FOSTER

Storage and backup (Gwen)

B. Storage and Backup
2. How will the data be stored and backed up during the project?
2.1. Where will the data be stored?
2.2. How will the data be recovered in the event of an incident? Will the data be backed up?

FOSTER
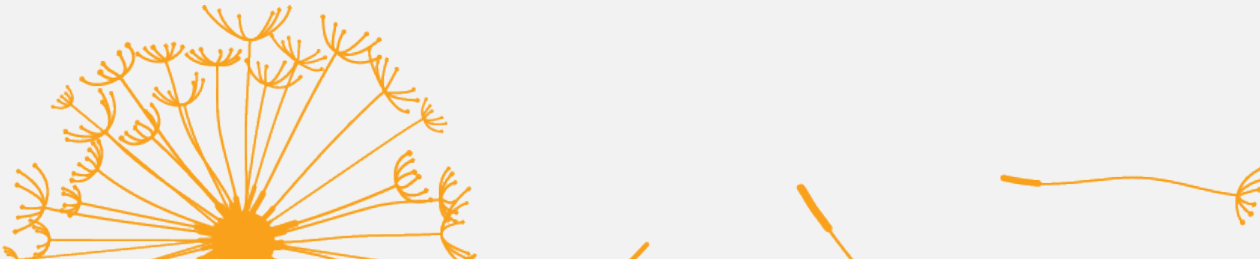
B. Storage and Backup

3. How will you manage access and security?

3.1. What are the risks to data security and how will these be managed?

3.2. How will you ensure that project partners (if applicable) can access your data securely?

YES, MAYBE OR NO?

If you want to keep data 'safe' before sharing it (for example, during your research), it's best to keep it offline on a USB stick or a local server.

Storage and Backup

How will the data be stored and backed up during the research?

- Do you have sufficient storage or will you need to include charges for additional services?
- How will the data be backed up?
- Who will be responsible for backup and recovery?
- How will the data be recovered in the event of an incident?

State how often the data will be backed up and to which locations. How many copies are being made? Storing data on laptops, computer hard drives or external storage devices alone is very risky. The use of robust, managed storage provided by university IT teams is preferable. Similarly, it is normally better to use automatic backup services provided by IT Services than rely on manual processes. If you choose to use a third-party service, you should ensure that this does not conflict with any funder, institutional, departmental or group policies, for example in terms of the legal jurisdiction in which data are held or the protection of sensitive data.
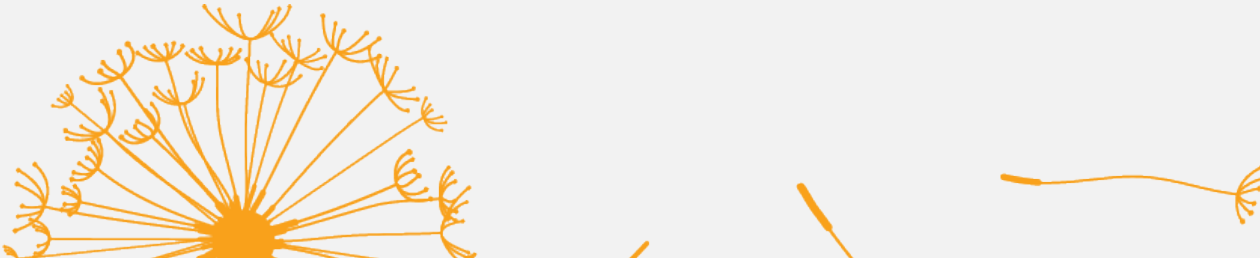
Storage and Backup

How will you manage access and security?

- What are the risks to data security and how will these be managed?
- How will you control access to keep the data secure?
- How will you ensure that collaborators can access your data securely?
- If creating or collecting data in the field how will you ensure its safe transfer into your main secured systems?

If your data is confidential (e.g. personal data not already in the public domain, confidential information or trade secrets), you should outline any appropriate security measures and note any formal standards that you will comply with e.g. ISO 27001."

Storing data: what is meant?

Storing and backing up files while research is active

Archiving or preserving data in the long-term
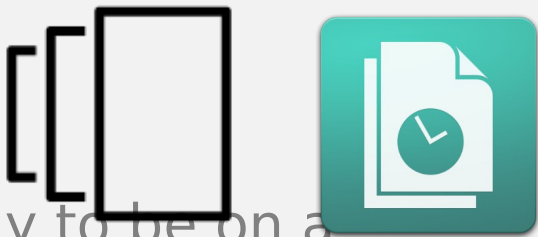
Likely to be on a networked filestore or hard drive

Easy to change or delete
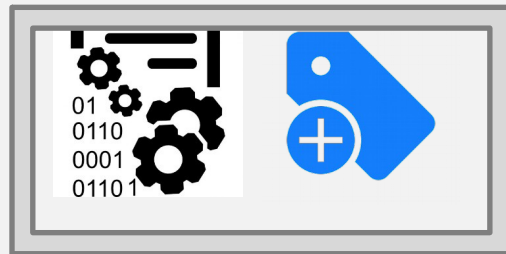
Likely to be deposited in a digital repository

safeguarded and preserved

Archiving, repositories, ehm?

Select a data repository that will preserve your data, metadata and possibly tools in the long term.

It is advisable to contact the repository of your choice when writing the first version of your DMP.

Repositories may offer guidelines for sustainable data formats and metadata standards, as well as support for dealing with sensitive data and licensing.

Where to find a repository?
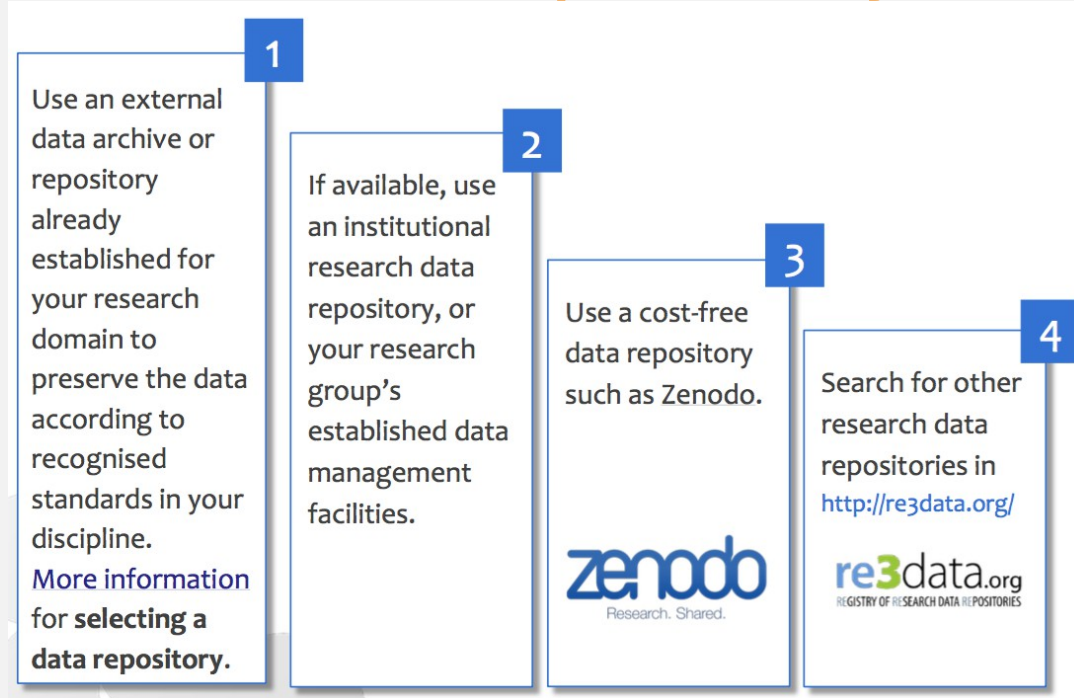
More information: https://www.openaire.eu/opendatapilot-repository
Zenodo: http://www.zenodo.org
Re3data.org: http://www.re3data.org

How to select a repository?

Main criteria for choosing a data repository:

Certification as a 'Trustworthy Digital Repository', with an explicit ambition to keep the data available in the long term.

•Three common certification standards for TDRs:

Data Seal of Approval: http://datasealofapproval.org/en
nestor seal:
http://www.langzeitarchivierung.de/Subsites/nestor/EN/nestor-Siegel/siegel_node.html
ISO 16363: http://www.iso16363.org

How to select a repository? (2)

•Matches your particular data needs: e.g. formats accepted; mixture of Open and Restricted Access.

•Provides guidance on how to cite the data that has been deposited.

•Gives your submitted dataset a persistent and globally unique identifier: for sustainable citations – both for data and publications – and to link back to particular researchers and grants.

www.re3data.org/browse/by-subject/

Browse by subject

Graphical | **Text**

A. Humanities and Social Sciences
   a. Humanities
      I. Ancient Cultures
         1. Prehistory
         2. Classical Philology
         3. Ancient History
         4. Classical Archaeology
         5. Egyptology and Ancient Near Eastern Studies
      II. History
         1. Medieval History
         2. Early Modern History
         3. Modern and Current History
         4. History of Science
      III. Fine Arts, Music, Theatre and Media Studies
         1. Art History
         2. Musicology
         3. Theatre and Media Studies
      IV. Linguistics
         1. General and Applied Linguistics
         2. Individual Linguistics
         3. Typology, Non-European Languages, Historical Linguistics
      V. Literary Studies
         1. Medieval German Literature
         2. Modern German Literature
         3. European and American Literature
         4. General and Comparative Literature and Cultural Studies

re3data.org

Search   Browse ▾   Suggest   Resources ▾   Contact   DataCite
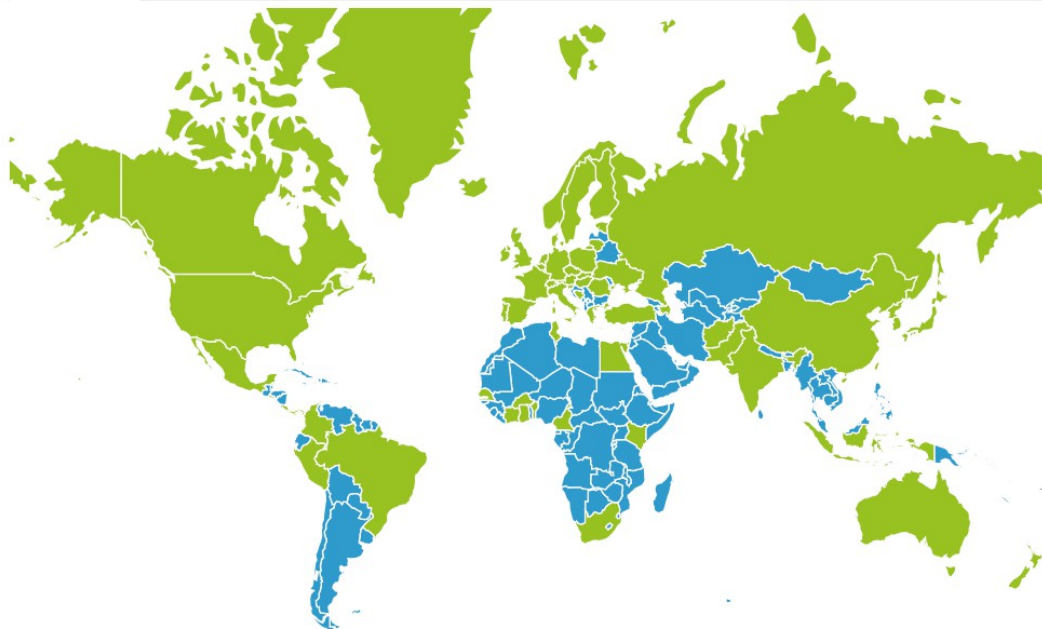
Found 5 result(s)

MIDAS
Nacionalinis atviros prieigos mokslinių tyrimų duomenų archyvas

i	🔓	©	doi	◉	§

Subject(s)   Humanities and Social Sciences   Life Sciences   Natural Sciences   Engineering Sciences

Content type(s)   Databases   Images   Structured graphics   Audiovisual data   Scientific and statistical data formats
                    Raw data   Plain text   Structured text   Archived data   Software applications   Source code
                    other

Country   Lithuania

MIDAS is national research data archive. The aim of the MIDAS is to collect, process, store and analyse scientific research data and other relevant information in all fields of knowledge, enabling free, easy and convenient access to it via the Internet. MIDAS provides services for registered and not-registered users: students, listeners, academics, researchers, scientific workers, research data evaluation and quality assurance experts, other participants in a science and studies system as well as individuals interested in research data. MIDAS consists of 2 parts: MIDAS portal (all users) and user account (internal portal for registered users).The Vilnius University is controller and main processor of MIDAS system.

CLARIN-LT
CLARIN-LT Repository

i	🔓	©	hdl	◉	§

Exercise

Sections	Questions	Questions to consider	Answers
A. Data Collection	1. What data will you collect or create?	*1.1. Are there any existing data that you can re-use?* *1.2. What type, format and volume of data?*	
B. Storage and Backup	2. How will the data be stored and backed up during the project?	*2.1. Where will the data be stored?* *2.2. How will the data be recovered in the event of an incident? Will the data be backed up?*	
	3. How will you manage access and security?	*3.1. What are the risks to data security and how will these be managed?* *3.2. How will you ensure that project partners (if applicable) can access your data securely?*	
C. Selection and Preservation	4. Which data are of long-term value and should be retained and preserved?	*4.1. What data must be retained/destroyed for contractual, legal, or regulatory purposes?* *4.2. How long will the data be retained and preserved?*	
D. Data Sharing	5. How will you ensure the availability and sharing of the data?	*5.1. When will you make the data available?* *5.2. How will potential users find out about data?* *5.3. With whom will you share the data, and under what conditions?*	
E. Responsibilities and Resources	6. Who will be responsible for data management?	*6.1. Who is responsible for implementing the DMP, and ensuring it is reviewed and revised?* *6.2. Will data ownership and responsibilities for RDM be part of any consortium agreement or contract agreed between project partners (if applicable)?*	
	7. What human and other resources will you require to draft and deliver your DMP?	*7.1. Is additional specialist expertise required?* *7.2. Do you require hardware or software which is additional or exceptional to existing institutional provision?* *7.3. Have you considered the charges that can be applied by data repositories?*	

Ethics and Legal Compliance

How will you manage any ethical issues?

- Have you gained consent for data preservation and sharing?
- How will you protect the identity of participants if required? e.g. via anonymisation
- How will sensitive data be handled to ensure it is stored and transferred securely?

*Ethical issues affect how you store data, who can see/use it and how long it is kept. Managing ethical concerns may include: **anonymisation of data; referral to departmental or institutional ethics committees; and formal consent agreements**. You should show that you are aware of any issues and have planned accordingly. If you are carrying out research involving human participants, you must also ensure that consent is requested to allow data to be shared and reused.*

Ethics and Legal Compliance

How will you manage copyright and Intellectual Property Rights (IPR) issues?

- Who owns the data?
- How will the data be licensed for reuse?
- Are there any restrictions on the reuse of third-party data?
- Will data sharing be postponed / restricted e.g. to publish or seek patents?

*State **who** will own the **copyright and IPR** of any data that you will collect or create, along with the **licence**(s) for its use and reuse. For multi-partner projects, IPR ownership may be worth covering in a consortium **agreement**. Consider any relevant **funder, institutional, departmental or group policies** on copyright or IPR. Also consider **permissions** to reuse third-party data and any **restrictions** needed on data sharing.*

YES, MAYBE OR NO?

If I share my data (early), I'll be scooped!
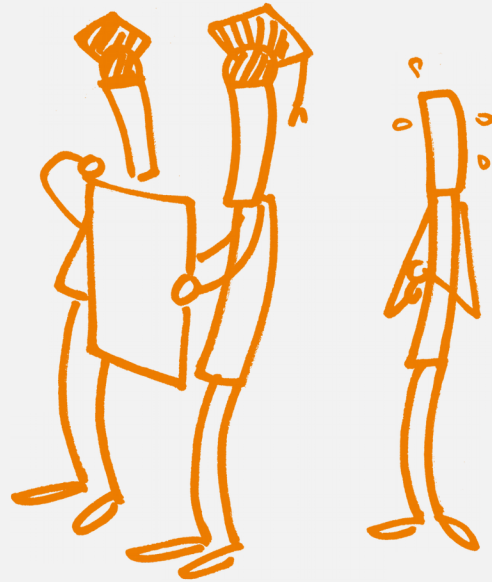
Pre-registration timestamps your work

Register Your Project

Open Science Framework

A registration on OSF creates a frozen, time-stamped version of a project that cannot be edited or deleted. The *original project* can still be edited, while the registered version cannot. You might create a registration to capture a snapshot of your project at certain points in time - such as right before data collection begins, when you submit a manuscript for peer review, or upon completion of a project.

Registrations can be made public immediately or embargoed for up to 4 years. Registrations cannot be deleted, but they can be withdrawn. Withdrawing a registration removes the content of the registration but leaves behind basic metadata, like registration title, contributors, and a reason for the withdrawal (not required).

http://help.osf.io/m/registrations/l/524205-register-your-project

Tips – share preprints too

- Early feedback on methods and initial findings
- Time to correct and mistakes before publishing
- Recognition for your ideas by peers

Licensing research data

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

D|C|C

JISClegal information

How to License Research Data

Alex Ball (DCC)
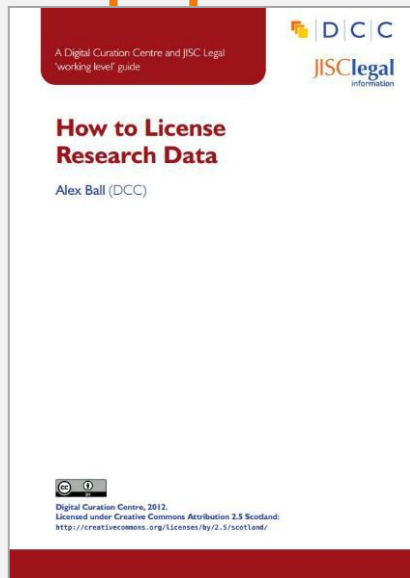
www.dcc.ac.uk/resources/how-guides/license-research-data

This DCC guide outlines the pros and cons of each approach and gives practical advice on how to implement your licence.

Horizon 2020 Open Access guidelines point to:

cc 0 ZERO

or

cc BY

CREATIVE COMMONS LIMITATIONS

$ Commercial

NC   Non-Commercial

What counts as commercial?

= Derivatives

ND   No Derivatives

Severely restricts use

CC-0 or CC-BY as a straightforward and effective way to make it **possible for others to mine, exploit & reproduce** data.

These clauses are not open licenses

OpenAIRE guides on legal issues

OpenAIRE

Guides for Researchers

How do I license my research data?

Learn more about licenses for research data and how to apply it

https://www.openaire.eu/how-do-i-license-my-research-data

EUDAT licensing tool

Answer questions to determine which licence(s) are appropriate to use

Do you own copyright and similar rights in your dataset and all its constitutive parts?

[Yes] [No]

Do you allow others to make commercial use of you data?

[Yes] [No]

Creative Commons Attribution (CC-BY)
This is the standard creative commons license that gives others maximum freedom to do what they want with your work.

Public Domain Dedication (CC Zero)
CC Zero enables scientists, educators, artists and other creators and owners of copyright- or database-protected content to waive those interests in their works and thereby place them as completely as possible in the public domain, so that others may freely build upon, enhance and reuse the works for any purposes without restriction under copyright or database law.

http://ufal.github.io/public-license-selector

Data Protection and Ethics

This course helps you to get to grips with data protection and the ethics around responsible data sharing.

https://www.fosteropenscience.eu/learning/data-protection-and-ethics
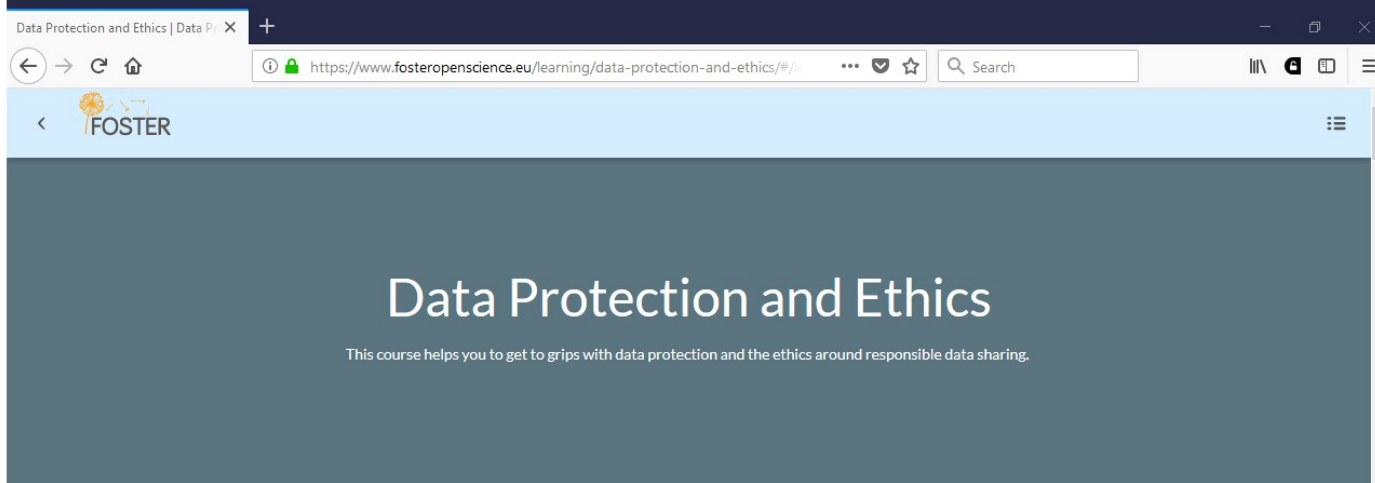
Introduction

This course covers data protection in particular and ethics more generally. It will help you understand the basic principles of data protection and introduces techniques for implementing data protection in your research processes. Upon completing this course, you will know:

- what personal data are and how you can protect them
- what to consider when developing consent forms
- how to store your data securely
- how to anonymise your data

Data protection and ethics

FOSTER

Open Licensing

Open Licensing course

https://www.fosteropenscience.eu/learning/open-licensing
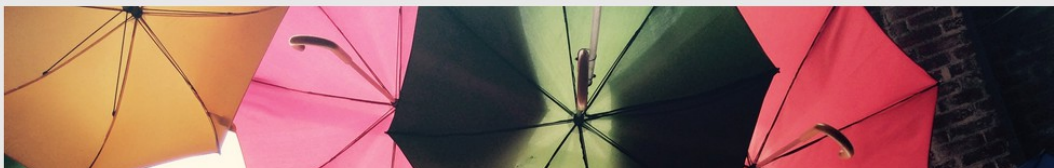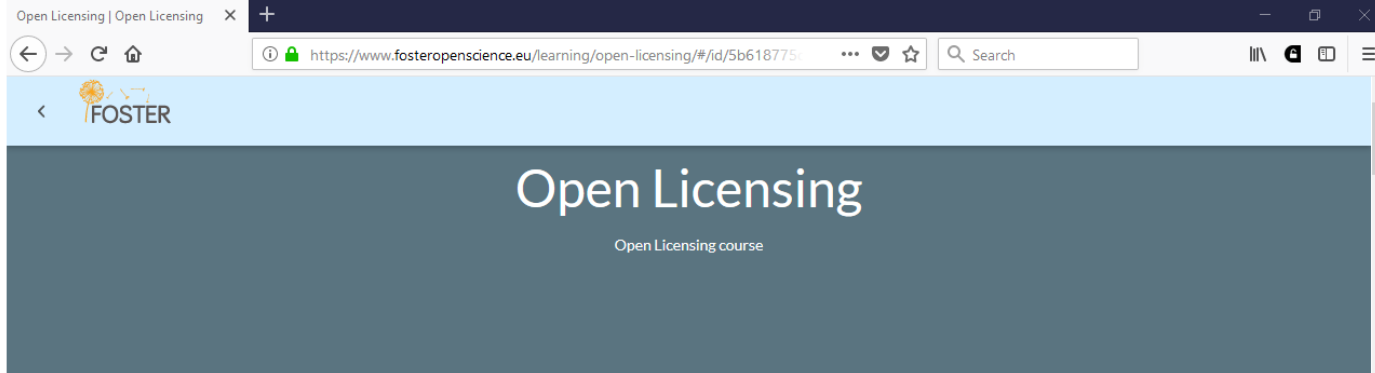
Introduction

Licensing your research outputs is an important part of practicing Open Science. In this course, you will:

- know what licenses are, how they work, and how to apply them
- understand how different types of licenses can affect research output reuse
- know how to select the appropriate license for your research

Why do you need apply a license?

Licensing is an important aspect of practising Open Science. By applying licenses to your outputs, you remove any ambiguity over what others can - and can't - do with your work.

An open license, Creative Commons or any other open license, consists of different elements that can be combined. Each element consists of a condition that needs to be followed by the re-user. The different combinations allow for great variation in the type of open license you apply: some being very open, others being very restrictive.

Open licenses

CC

Guides for Researchers

How to deal with sensitive data

Learn how to preserve your sensitive data safely

https://www.openaire.eu/sensitive-data-guide

WHAT IS
SENSITIVE
DATA

HOW TO
PREPARE
SENSITIVE
DATA FOR
STORAGE AND
SHARING?

STORING
SENSITIVE
DATA

COSTS

MORE
INFORMATION

What is Sensitive data

What is the best way of managing access to sensitive data? This is not a straightforward question as it involves ethical, legal and technical issues to be tackled. This guide will help you on your way to preserve your sensitive data safely. It explains the different types of sensitive data, how to prepare them for storage and the possible cost involved in the process.

What is Sensitive data

Sensitive data is data that must be protected against unwanted disclosure. Access to sensitive data should be safeguarded. Protection of sensitive data may be required for legal or ethical reasons, for issues pertaining to personal privacy, or for proprietary considerations.

The EU has strong regulations regarding personal (e.g. General Data Protection Regulation - GDPR) and sensitive non-personal data. In Horizon 2020 an Ethical review is required when applying for a grant.

Exercise

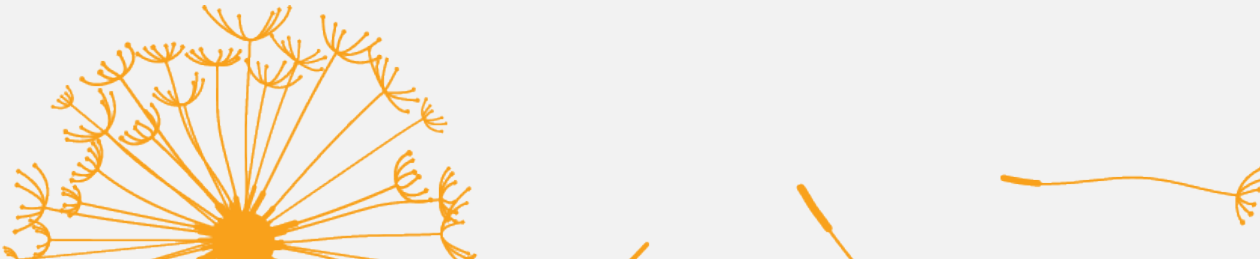| Sections	Questions	Questions to consider	Answers
A. Data Collection	1. What data will you collect or create?	*1.1. Are there any existing data that you can re-use?* *1.2. What type, format and volume of data?*	
B. Storage and Backup	2. How will the data be stored and backed up during the project?	*2.1. Where will the data be stored?* *2.2. How will the data be recovered in the event of an incident? Will the data be backed up?*	
	3. How will you manage access and security?	*3.1. What are the risks to data security and how will these be managed?* *3.2. How will you ensure that project partners (if applicable) can access your data securely?*	
C. Selection and Preservation	4. Which data are of long-term value and should be retained and preserved?	*4.1. What data must be retained/destroyed for contractual, legal, or regulatory purposes?* *4.2. How long will the data be retained and preserved?*	
D. Data Sharing	5. How will you ensure the availability and sharing of the data?	*5.1. When will you make the data available?* *5.2. How will potential users find out about data?* *5.3. With whom will you share the data, and under what conditions?*	
E. Responsibilities and Resources	6. Who will be responsible for data management?	*6.1. Who is responsible for implementing the DMP, and ensuring it is reviewed and revised?* *6.2. Will data ownership and responsibilities for RDM be part of any consortium agreement or contract agreed between project partners (if applicable)?*	
	7. What human and other resources will you require to draft and deliver your DMP?	*7.1. Is additional specialist expertise required?* *7.2. Do you require hardware or software which is additional or exceptional to existing institutional provision?* *7.3. Have you considered the charges that can be applied by data repositories?*	

FOSTER

C. Selection and Preservation
4. Which data are of long-term value and should be retained and preserved?
4.1. What data must be retained/destroyed for contractual, legal, or regulatory purposes?
4.2. How long will the data be retained and preserved?

YES, MAYBE OR NO?

It is important that you permanently archive every dataset you create.
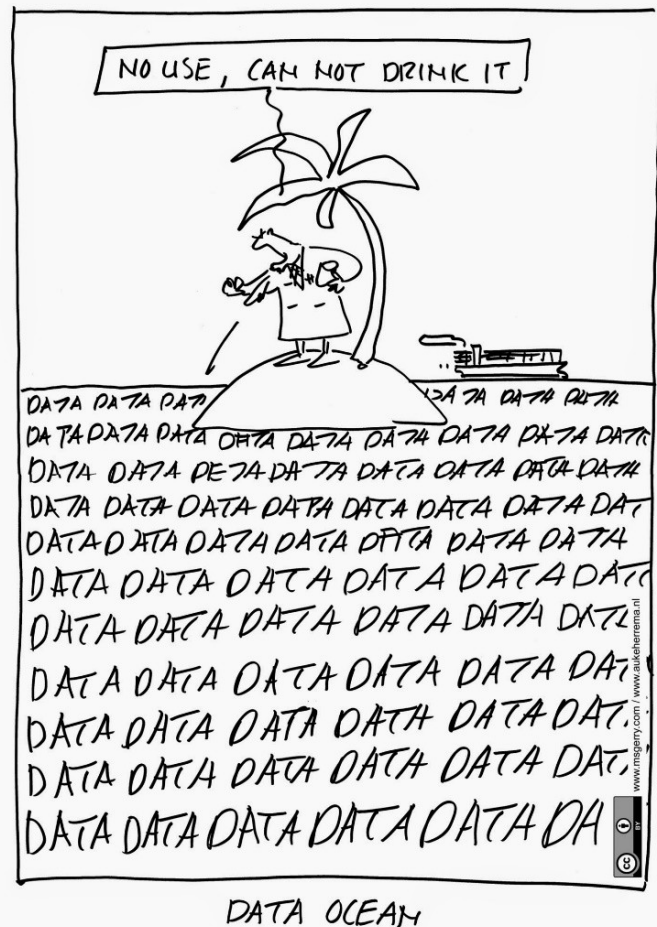
Selection and Preservation

Which data are of long-term value and should be retained, shared, and/or preserved?

- What data must be retained/destroyed for contractual, legal, or regulatory purposes?
- How will you decide what other data to keep?
- What are the foreseeable research uses for the data?
- How long will the data be retained and preserved?

Consider how the data may be reused e.g. to validate your research findings, conduct new studies, or for teaching. Decide which data to keep and for how long. This could be based on any obligations to retain certain data, the potential reuse value, what is economically viable to keep, and any additional effort required to prepare the data for data sharing and preservation. Remember to consider any additional effort required to prepare the data for sharing and preservation, such as changing file formats.

Selection and Preservation

What is the long-term preservation plan for the dataset?

- Where e.g. in which repository or archive will the data be held?
- What costs if any will your selected data repository or archive charge?
- Have you costed in time and effort to prepare the data for sharing / preservation?

Consider how datasets that have long-term value will be preserved and curated beyond the lifetime of the grant. Also outline the plans for preparing and documenting data for sharing and archiving. If you do not propose to use an established repository, the data management plan should demonstrate that resources and systems will be in place to enable the data to be curated effectively beyond the lifetime of the grant.

YES OR NO?

Ultimately, as the 'creator' of the research data, you are the one who

Deciding which data need to be kept after the project ends

Five steps to follow

1. **Could** this data be re-used

2.   **Must** it be kept as evidence or for legal reasons or must it be removed?

3.   **Should** it be kept for its potential value

4. **Consider costs** – do benefits outweigh cost?

5. **Evaluate criteria** to decide what to keep

5 steps to decide what data to keep

www.dcc.ac.uk/resources/how-guides/five-steps-decide-what-data-keep

What should be preserved and shared?

- The data needed to validate results in scientific publications (minimally!).
- The associated metadata: the dataset's creator, title, year of publication, repository, identifier etc.
  - Follow a metadata standard in your line of work, or a generic standard, e.g. Dublin Core or DataCite, and be FAIR.
  - A repository will assign a persistent ID to the dataset: important for discovering and citing the data.

What should be preserved and shared? (2)

- **Documentation**: code books, lab journals, informed consent forms – domain-dependent, and important for understanding the data and combining them with other data sources.

- Software, hardware, tools, syntax queries, machine configurations – domain-dependent, and important for using the data. (Alternative: information about the software etc.)

Basically, everything that is needed to replicate a study should be available. Plus everything that is potentially useful for others.

Exercise

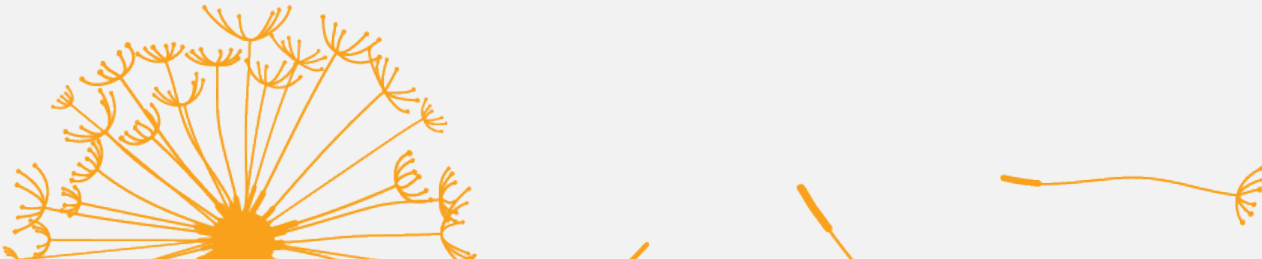| Sections	Questions	Questions to consider	Answers
A. Data Collection	1. What data will you collect or create?	*1.1. Are there any existing data that you can re-use?* *1.2. What type, format and volume of data?*	
B. Storage and Backup	2. How will the data be stored and backed up during the project?	*2.1. Where will the data be stored?* *2.2. How will the data be recovered in the event of an incident? Will the data be backed up?*	
	3. How will you manage access and security?	*3.1. What are the risks to data security and how will these be managed?* *3.2. How will you ensure that project partners (if applicable) can access your data securely?*	
C. Selection and Preservation	4. Which data are of long-term value and should be retained and preserved?	*4.1. What data must be retained/destroyed for contractual, legal, or regulatory purposes?* *4.2. How long will the data be retained and preserved?*	
D. Data Sharing	5. How will you ensure the availability and sharing of the data?	*5.1. When will you make the data available?* *5.2. How will potential users find out about data?* *5.3. With whom will you share the data, and under what conditions?*	
E. Responsibilities and Resources	6. Who will be responsible for data management?	*6.1. Who is responsible for implementing the DMP, and ensuring it is reviewed and revised?* *6.2. Will data ownership and responsibilities for RDM be part of any consortium agreement or contract agreed between project partners (if applicable)?*	
	7. What human and other resources will you require to draft and deliver your DMP?	*7.1. Is additional specialist expertise required?* *7.2. Do you require hardware or software which is additional or exceptional to existing institutional provision?* *7.3. Have you considered the charges that can be applied by data repositories?*	

FOSTER

D. Data Sharing

5. How will you ensure the availability and sharing of the data?

5.1. When will you make the data available?

5.2. How will potential users find out about data?

5.3. With whom will you share the data, and under what conditions?

Data Sharing

How will you share the data?

How will potential users find out about your data?

●With whom will you share the data, and under what conditions?

●Will you share data via a repository, handle requests directly or use another mechanism?

●When will you make the data available?

●Will you pursue getting a persistent identifier for your data?

*Consider **where, how, and to whom data** with acknowledged long-term value should be made **available**. The methods used to share data will be dependent on a number of factors such as the **type, size, complexity and sensitivity of data**. If possible, mention earlier examples to show a track record of effective data sharing. Consider how people might acknowledge the reuse of your data.*

Data Sharing

Are any restrictions on data sharing required?

● What action will you take to overcome or minimise restrictions?

● For how long do you need exclusive use of the data and why?

● Will a data sharing agreement (or equivalent) be required?

*Outline any **expected difficulties** in sharing data with acknowledged long-term value, along with **causes and possible measures** to overcome these. Restrictions may be due to confidentiality, lack of consent agreements or IPR, for example. Consider whether a non-disclosure agreement would give sufficient protection for confidential data.*

YES, MAYBE OR NO?

I don't need to decide now if I want to share. I can wait and see what I want to do at the end of my project.

Sharing data: what is meant?

With collaborators while research is active
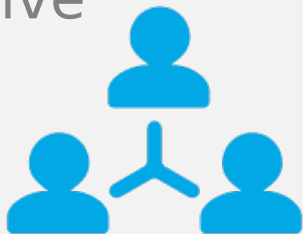
(Open) data sharing

Data are mutable

Data are stable, searchable, citable, clearly licensed

Definition of Open Data

Open Data are online, free of cost, accessible data that can be used, reused and distributed provided that the data source is attributed.

How to make data open?

https://okfn.org

1. ## Choose your dataset(s)
   - What can you open? You may need to revisit this step if you encounter problems later.

2. ## Apply an open license
   - Determine what IP exists. Apply a suitable licence e.g. CC-BY

3. ## Make the data available
   - Provide the data in a suitable format. Use repositories.

4. ## Make it discoverable
   - Post on the web, register in catalogues…

FOSTER

Making data FAIR

Findable – Assign persistent IDs, provide rich metadata, register in a searchable resource,...

Accessible - Retrievable by their ID using a standard protocol, metadata remain accessible even if data aren't...

Interoperable - Use formal, broadly applicable languages, use standard vocabularies, qualified references...

Reusable - Rich, accurate metadata, clear licences, provenance, use of community standards

www.force11.org/group/fairgroup/fairprinciples

https://ec.europa.eu/research/press/2016/pdf/opendata-infographic_072016.pdf

Tip – use 5 Star Open Data Model to explain FAIR

★ make your stuff available on the Web (whatever format) under an open license

★★ make it available as structured data (e.g., Excel instead of image scan of a table)

★★★ make it available in a non-proprietary open format (e.g., CSV instead of Excel)

★★★★ use URIs to denote things, so that people can point at your stuff

★★★★★ link your data to other data to provide context

Tim Berners-Lee's proposal for five star open data - http://5stardata.info

Options for open data

- Domain repository (first choice)
- General repository (Figshare, Zenodo)
- Institutional repository
- Data journal
- Journal supplementary material

OPEN ACCESS

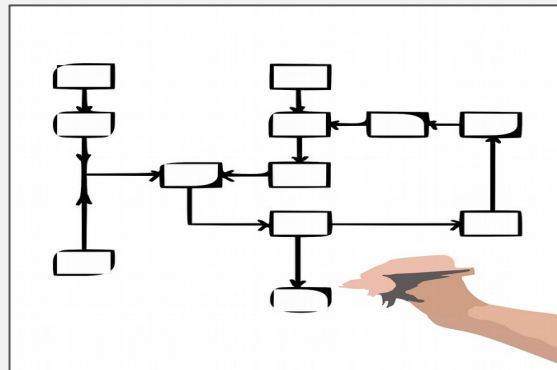# Tip – link data to other outputs for context (reuse)

Open Data

To support validation and facilitate reuse

Open Code

Software created to analyse and/or visualise the data

Open Workflows

What steps were taken and in what order?

5 star open data

Use URIs to denote things so that people can point at your data.

Make your data available on the Web (whatever format) under an open licence.

Use non-proprietary formats (e.g., CSV instead of Excel for tables).

Link your data to other data to provide context

Make them available as structured data (e.g., Excel instead if a scan of a table).

1 star

5 stars

Exercise

Sections	Questions	Questions to consider	Answers
A. Data Collection	1. What data will you collect or create?	*1.1. Are there any existing data that you can re-use?* *1.2. What type, format and volume of data?*	
B. Storage and Backup	2. How will the data be stored and backed up during the project?	*2.1. Where will the data be stored?* *2.2. How will the data be recovered in the event of an incident? Will the data be backed up?*	
	3. How will you manage access and security?	*3.1. What are the risks to data security and how will these be managed?* *3.2. How will you ensure that project partners (if applicable) can access your data securely?*	
C. Selection and Preservation	4. Which data are of long-term value and should be retained and preserved?	*4.1. What data must be retained/destroyed for contractual, legal, or regulatory purposes?* *4.2. How long will the data be retained and preserved?*	
D. Data Sharing	5. How will you ensure the availability and sharing of the data?	*5.1. When will you make the data available?* *5.2. How will potential users find out about data?* *5.3. With whom will you share the data, and under what conditions?*	
E. Responsibilities and Resources	6. Who will be responsible for data management?	*6.1. Who is responsible for implementing the DMP, and ensuring it is reviewed and revised?* *6.2. Will data ownership and responsibilities for RDM be part of any consortium agreement or contract agreed between project partners (if applicable)?*	
	7. What human and other resources will you require to draft and deliver your DMP?	*7.1. Is additional specialist expertise required?* *7.2. Do you require hardware or software which is additional or exceptional to existing institutional provision?* *7.3. Have you considered the charges that can be applied by data repositories?*	

FOSTER

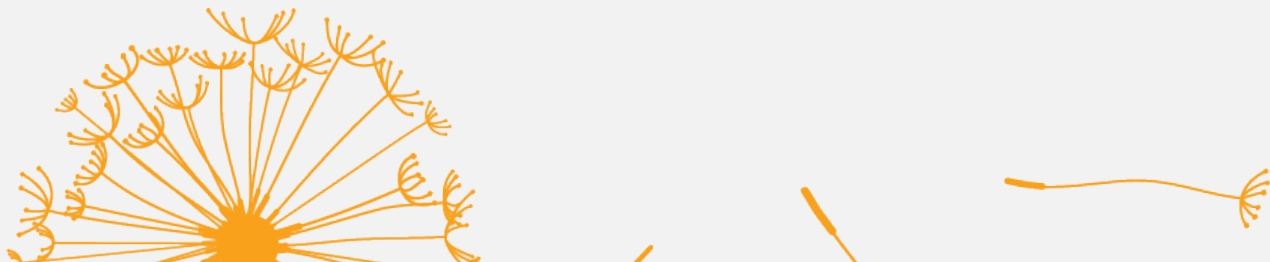## Responsibilities and Resources (Gwen)

FOSTER

E. Responsibilities and Resources

6. Who will be responsible for data management?

6.1. Who is responsible for implementing the DMP, and ensuring it is reviewed and revised?

6.2. Will data ownership and responsibilities for RDM be part of any consortium agreement or contract agreed between project partners (if applicable)?
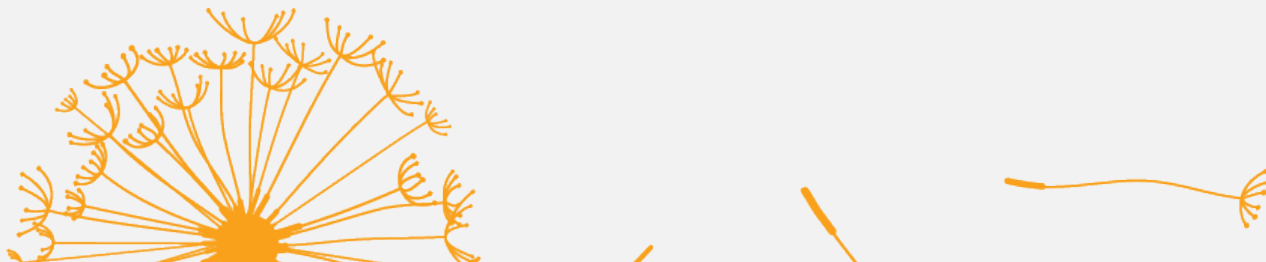
FOSTER

E. Responsibilities and Resources
7. What human and other resources will you require to draft and deliver your DMP?
7.1. Is additional specialist expertise required?
7.2. Do you require hardware or software which is additional or exceptional to existing institutional provision?
7.3. Have you considered the charges that can be applied by data repositories?

Responsibilities and Resources
Who will be responsible for data management?

●Who is responsible for implementing the DMP, and ensuring it is reviewed and revised?

●Who will be responsible for each data management activity?

●How will responsibilities be split across partner sites in collaborative research projects?

●Will data ownership and responsibilities for RDM be part of any consortium agreement or contract agreed between partners?

Outline the roles and responsibilities for all activities e.g. data capture, metadata production, data quality, storage and backup, data archiving & data sharing. Consider who will be responsible for ensuring relevant policies will be respected. Individuals should be named where possible.

Responsibilities and Resources

What resources will you require to deliver your plan?

● Is additional specialist expertise (or training for existing staff) required?

● Do you require hardware or software which is additional or exceptional to existing institutional provision?

● Will charges be applied by data repositories?

Carefully consider any resources needed to deliver the plan, e.g. software, hardware, technical expertise, etc. Where dedicated resources are needed, these should be outlined and justified.

YES, MAYBE OR NO?

Open Data also means that the tools you use should be available at no or very small extra cost.

Consider who else has a say about sharing data

- Collaborators

- Research participants

- Commercial partners

- Data repository

- Publishers

- Institutions, funders

Costs

Guides for Researchers

How to identify and assess Research Data Management (RDM) costs

THE COST OF DATA MANAGEMENT

HOW TO CALCULATE COSTS?

HOW TO USE THIS COSTING TOOL?

ESTIMATING COSTS RDM TOOL

Estimating costs RDM tool

DMP PHASE	ACTIVITY	COMMENTS AND SUGGESTIONS	COSTS
Preparing	Make a Data Management Plan	Make a DMP before you start creating data; make decisions about managing your data. You can find the template for H2020 DMPs here.	2 hrs to 2 days, depending on the complexity of your project

https://www.openaire.eu/how-to-comply-to-h2020-mandates-rdm-costs

Exercise

Sections	Questions	Questions to consider	Answers
A. Data Collection	1. What data will you collect or create?	*1.1. Are there any existing data that you can re-use?* *1.2. What type, format and volume of data?*	
B. Storage and Backup	2. How will the data be stored and backed up during the project?	*2.1. Where will the data be stored?* *2.2. How will the data be recovered in the event of an incident? Will the data be backed up?*	
	3. How will you manage access and security?	*3.1. What are the risks to data security and how will these be managed?* *3.2. How will you ensure that project partners (if applicable) can access your data securely?*	
C. Selection and Preservation	4. Which data are of long-term value and should be retained and preserved?	*4.1. What data must be retained/destroyed for contractual, legal, or regulatory purposes?* *4.2. How long will the data be retained and preserved?*	
D. Data Sharing	5. How will you ensure the availability and sharing of the data?	*5.1. When will you make the data available?* *5.2. How will potential users find out about data?* *5.3. With whom will you share the data, and under what conditions?*	
E. Responsibilities and Resources	6. Who will be responsible for data management?	*6.1. Who is responsible for implementing the DMP, and ensuring it is reviewed and revised?* *6.2. Will data ownership and responsibilities for RDM be part of any consortium agreement or contract agreed between project partners (if applicable)?*	
	7. What human and other resources will you require to draft and deliver your DMP?	*7.1. Is additional specialist expertise required?* *7.2. Do you require hardware or software which is additional or exceptional to existing institutional provision?* *7.3. Have you considered the charges that can be applied by data repositories?*	

Exercise

Discuss your model DMP with your colleagues.
Identify common discussion points
Briefly discuss with the whole group

With thanks to

S. Venkataraman, DCC
Joy Davidson, University of Glasgow
Bianca Kramer & Jeroen Bosman, Utrecht University Library
Marjan Grootveld, DANS
Sarah Jones, DCC
Iryna Kuchma, EIFL
Acknowledgements:
Jonathan Rans, DCC
Thanks to DANS and DCC for reuse of

gwen.franck@eifl.net

@g_fra

brinken@sub.uni-goettingen.de

@HeleneBrinken