Introduction to Data Management Plans

S. Venkataraman, Training Officer, OpenAIRE

s.venkataraman@openaire.eu

1st August 2023, ICTP, Trieste

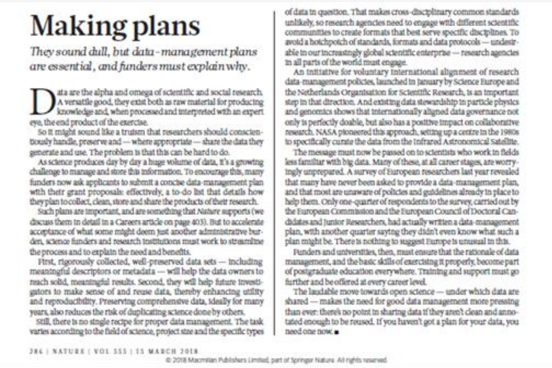
cc BY

or

Go to
www.menti.com
and enter code
83359905

Bringing together what you’ve learnt

- Make informed decisions to anticipate and avoid problems.
- Avoid duplication, data loss and security breaches.
- Develop procedures early on for consistency.
- Ensure data are accurate, complete, reliable and secure.
- Save time and effort to make your life easier!
- Useful both to researchers and institutions



Making plans

They sound dull, but data-management plans are essential, and funders must explain why.

Data are the alpha and omega of scientific and social research. A versatile good, they exist both as raw material for producing knowledge and, when processed and interpreted with an expert eye, the end product of the exercise.

So it might sound like a truism that researchers should conscientiously handle, preserve and — where appropriate — share the data they generate and use. The problem is that this can be hard to do.

As science produces day by day a huge volume of data, it's a growing challenge to manage and store this information. To encourage this, many funders now ask applicants to submit a concise data-management plan with their grant proposals: effectively, a to-do list that details how they plan to collect, clean, store and share the products of their research.

Such plans are important, and are something that *Nature* supports (we discuss them in detail in a Careers article on page 403). But to accelerate acceptance of what some might deem just another administrative burden, science funders and research institutions must work to streamline the process and to explain the need and benefits.

First, rigorously collected, well-preserved data sets — including meaningful descriptors or metadata — will help the data owners to reach solid, meaningful results. Second, they will help future investigators to make sense of and reuse data, thereby enhancing utility and reproducibility. Preserving comprehensive data, ideally for many years, also reduces the risk of duplicating science done by others.

Still, there is no single recipe for proper data management. The task varies according to the field of science, project size and the specific types of data in question. That makes cross-disciplinary common standards unlikely, so research agencies need to engage with different scientific communities to create formats that best serve specific disciplines. To avoid a hotchpotch of standards, formats and data protocols — undesirable in our increasingly global scientific enterprise — research agencies in all parts of the world must engage.

An initiative for voluntary international alignment of research data-management policies, launched in January by Science Europe and the Netherlands Organisation for Scientific Research, is an important step in that direction. And existing data stewardship in particle physics and genomics shows that internationally aligned data governance not only is perfectly doable, but also has a positive impact on collaborative research. NASA pioneered this approach, setting up a centre in the 1980s to specifically curate the data from the Infrared Astronomical Satellite.

The message must now be passed on to scientists who work in fields less familiar with big data. Many of these, at all career stages, are worryingly unprepared. A survey of European researchers last year revealed that many have never been asked to provide a data-management plan, and that most are unaware of policies and guidelines already in place to help them. Only one-quarter of respondents to the survey, carried out by the European Commission and the European Council of Doctoral Candidates and Junior Researchers, had actually written a data-management plan, with another quarter saying they didn't even know what such a plan might be. There is nothing to suggest Europe is unusual in this.

Funders and universities, then, must ensure that the rationale of data management, and the basic skills of exercising it properly, become part of postgraduate education everywhere. Training and support must go further and be offered at every career level.

The laudable move towards open science — under which data are shared — makes the need for good data management more pressing than ever: there's no point in sharing data if they aren't clean and annotated enough to be reused. If you haven't got a plan for your data, you need one now. ■

286 | NATURE | VOL 555 | 15 MARCH 2018

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

CAREERS

PERSONAL ETHICS How a vegetarian biologist balances his beliefs with his work p.405 | BLOG Personal stories and careers counsel http://blogs.nature.com/naturejobs | NATUREJOBS For the latest career listings and advice www.naturejobs.com

ADAPTED FROM GETTY

DATA MANAGEMENT

For the record

Making project data freely available is vital for open science.

BY QUIRIN SCHIERMEIER

When Marjorie Etique learnt that she had to create a data-management plan for her next research project, she was not sure exactly what to do.

The soil chemist, a postdoc at the Swiss Federal Institute of Technology (ETH) in Zurich, studies the interaction of trace elements in sediments and water. While preparing a grant proposal for the Swiss National Science Foundation last October, she learnt of the funder's new data rules. These require applicants to provide a written plan for the organization and long-term storage of their research data, to help minimize the risk of data loss and provide guidance for other scientists on how to use the data in the future.

Etique found the task daunting. "Data management is really not my primary skill," she says. "I had absolutely no idea how to go about it." She was able to get advice from her supervisor and from ETH's digital library service. Other researchers might not be so lucky, and might not even know what a data-management plan is — let alone why they would need one and how to produce it. Here, we answer these questions.

WHAT ARE DATA-MANAGEMENT PLANS?

A data-management plan explains how researchers will handle their data during and after a project, and encompasses creating, sharing and preserving research data of any type, including text, spreadsheets, images, recordings, models, algorithms and software. It does not matter whether the data are generated by large pieces of research equipment, such as imaging tools or particle accelerators, or from straightforward field observation.

Many funders are asking grant applicants to provide data plans. Requirements vary from one discipline to another. But in general, scientists will need to describe — before they begin any research — what data they will generate; how the data will be documented, described, secured and curated; and who will have access to those data after the research is completed. They must also explain any data sharing and reuse restrictions, such as legal and confidentiality issues. Researchers can consult their funder and their host institute's digital library services for assistance. Colleagues who have previously produced data plans may also be able to help (see 'Keeping stock').

WHO NEEDS THEM?

Data management is one example of the way in which public research sponsors and research institutions are implementing 'open science', the push to make scientific research and data freely accessible. Many funding agencies have made data-management plans mandatory for grant applicants in the past decade or so. All US federal agencies, including the National Science Foundation and the National Institutes of Health, have such policies. Data-management plans must also now be included in grant proposals to the European Research Council and other European Union-funded research programmes. And many national funding agencies in Europe — including the UK research councils and the London-based Wellcome Trust, world's largest biomedical research charity — also ask for data plans.

Many scientists already practise data management by default. Astronomers, for example, have been doing so for decades when calibrating their observations and archiving huge amounts of telescope-survey data in standardized, machine-readable catalogues for reuse.

Geneticists, too, use special data repositories to archive the vast amounts of DNA and genome-sequencing data (see go.nature.com/2omlrbe). But less data-intensive fields of science and social research also benefit from data management. For example, geochemists analysing soil bacteria and mineral products in different environments can use it to ▶

15 MARCH 2018 | VOL 555 | NATURE | 403

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Schiermeier, Q. “Data management made simple” *Nature* **555**, 403-405 (2018).
https://www.nature.com/articles/d41586-018-03071-1
doi: 10.1038/d41586-018-03071-1

Data Management Plans

A formal document explaining how research data will be handled throughout the data lifecycle.

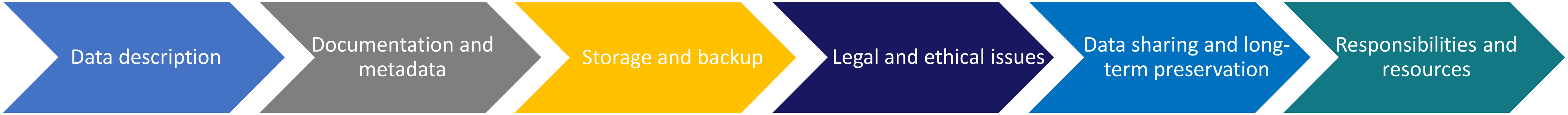
- Mandatory in e.g. European Commission funded projects
- National and institutional mandates
- Increasingly often required as part of PhD requirements in some places

Deliverable and “living” document

Documents processes undertaken throughout data management lifecycle, including costs

What a DMP is not?

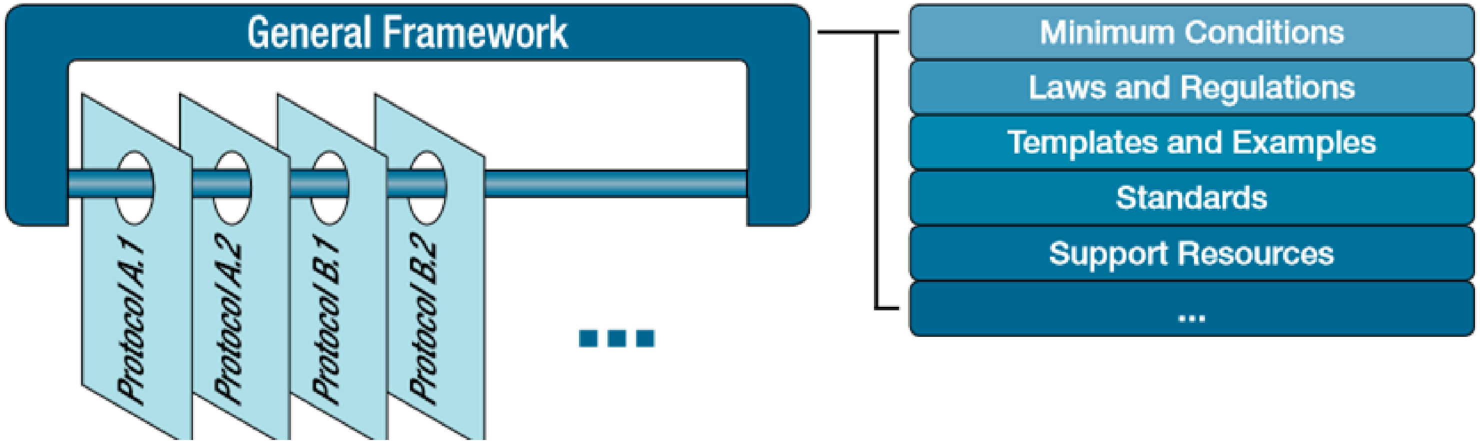
Research assessment method


Data description
Documentation and metadata
Storage and backup
Legal and ethical issues
Data sharing and long-term preservation
Responsibilities and resources

Data Management Plans

- Depends on the funder/institution requirements
- Differences in research communities
  - Formats, standards, documentation etc.

Minimum requirements: Science Europe – **DDPs** (Domain Data Protocols)

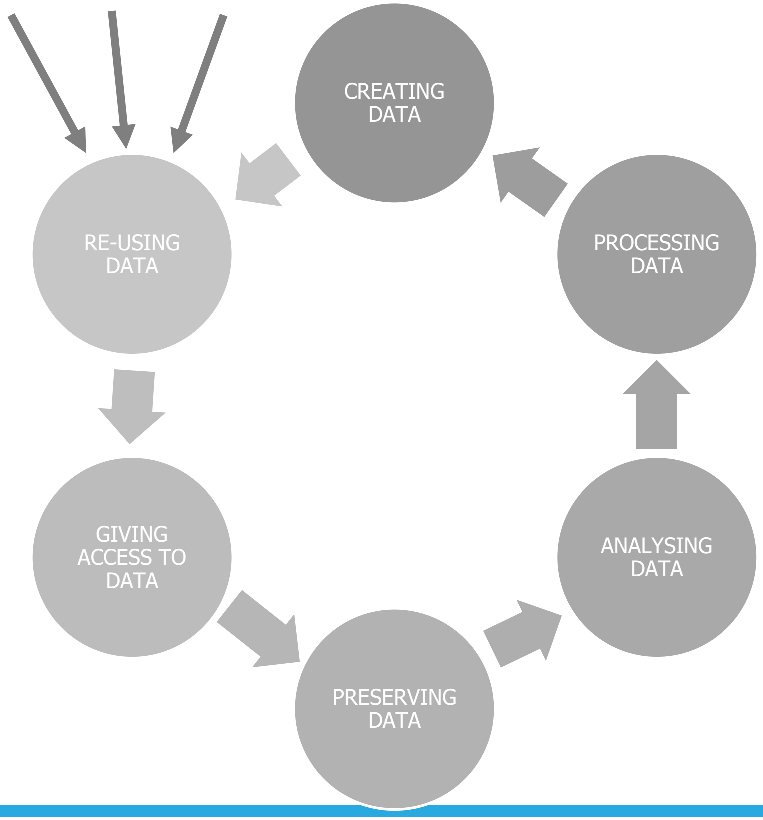


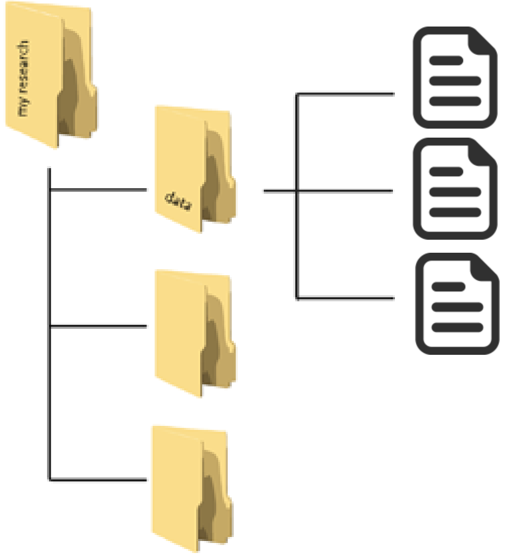
Common themes in DMPs

1. Description of data to be collected / created (i.e. content, type, format, volume...).
2. Standards/methodologies for data collection & management.
3. Ethics and Intellectual Property (highlight any restrictions on data sharing e.g. embargoes, confidentiality).
4. Plans for data sharing and access (i.e. how, when, to whom).
5. Strategy for long-term preservation.

Planning trick 1: think backwards

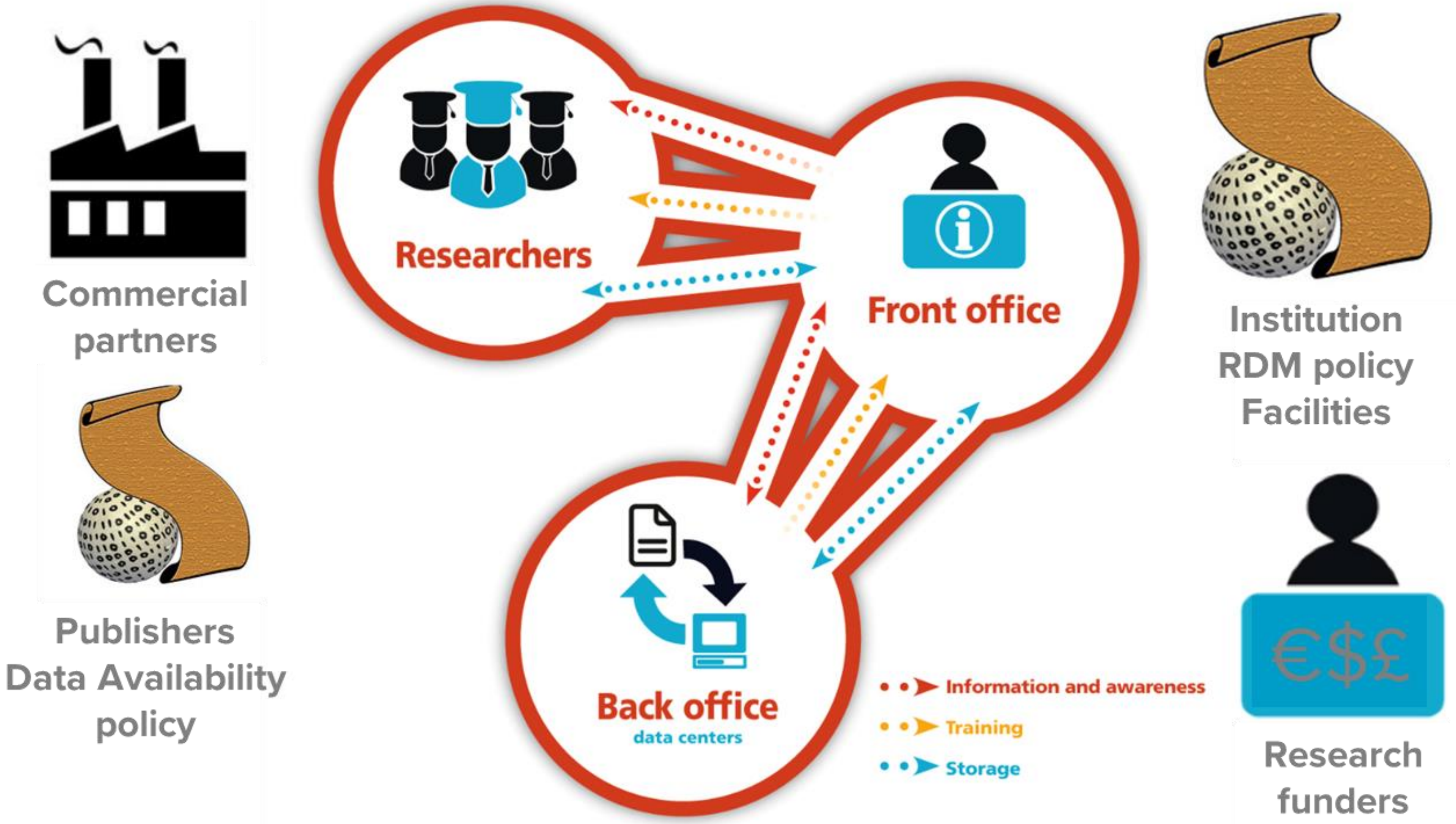
What data organisation would a re-user like?





Design how you will organise data in the project (folder structure, file naming convention, …)

Planning trick 2: include RDM stakeholders



www.openaire.eu/briefpaper-rdm-infonoads

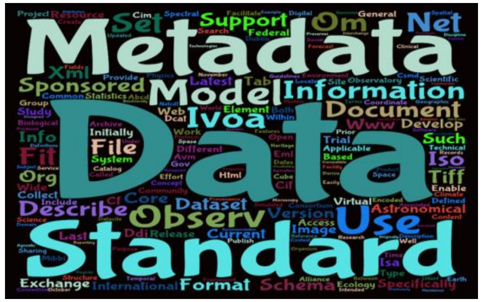
Planning trick 3: ground your plan in reality

Data Schools

Base plans on available skills, support and good practice for the field – show it's feasible to implement.

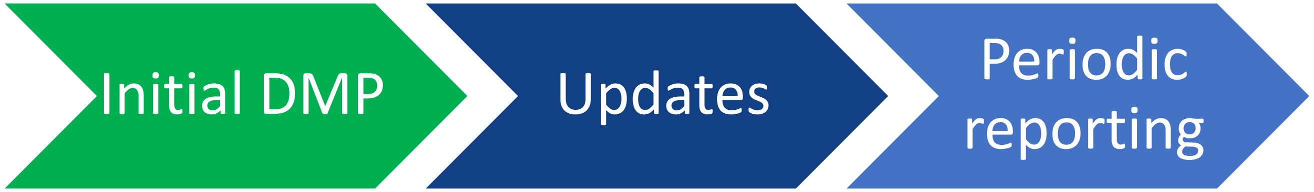
re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES

cc 0 ZERO

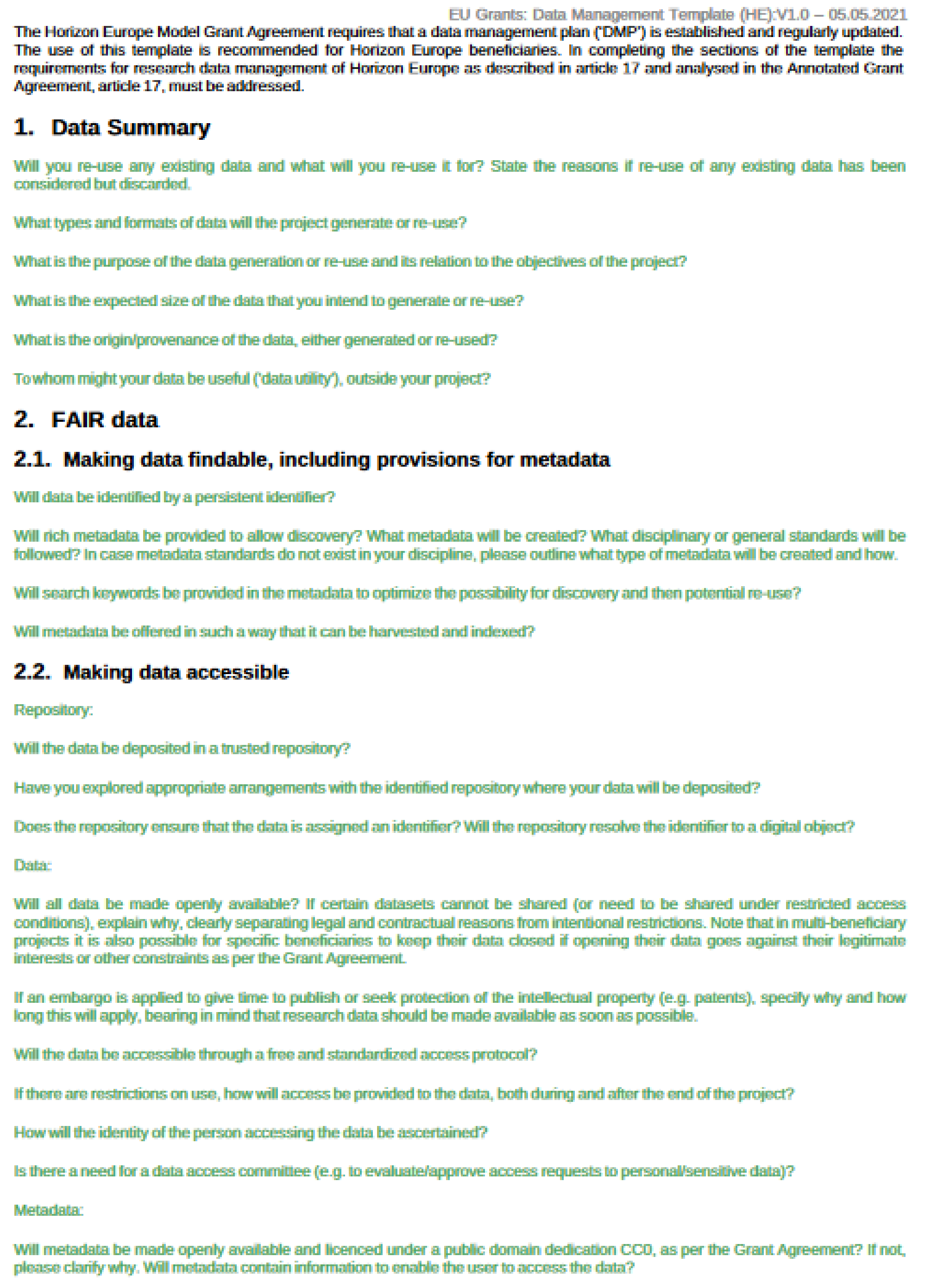


Horizon Europe DMP Template

- Guidance: issues to be covered
- Structured approach
- Living document



https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm
https://enspire.science/wp-content/uploads/2021/09/Horizon-Europe-Data-Management-Plan-Template.pdf


EU Grants: Data Management Template (HE):V1.0 – 05.05.2021

The Horizon Europe Model Grant Agreement requires that a data management plan ('DMP') is established and regularly updated. The use of this template is recommended for Horizon Europe beneficiaries. In completing the sections of the template the requirements for research data management of Horizon Europe as described in article 17 and analysed in the Annotated Grant Agreement, article 17, must be addressed.

1. Data Summary

Will you re-use any existing data and what will you re-use it for? State the reasons if re-use of any existing data has been considered but discarded.

What types and formats of data will the project generate or re-use?

What is the purpose of the data generation or re-use and its relation to the objectives of the project?

What is the expected size of the data that you intend to generate or re-use?

What is the origin/provenance of the data, either generated or re-used?

To whom might your data be useful ('data utility'), outside your project?

2. FAIR data

2.1. Making data findable, including provisions for metadata

Will data be identified by a persistent identifier?

Will rich metadata be provided to allow discovery? What metadata will be created? What disciplinary or general standards will be followed? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.

Will search keywords be provided in the metadata to optimize the possibility for discovery and then potential re-use?

Will metadata be offered in such a way that it can be harvested and indexed?

2.2. Making data accessible

Repository:

Will the data be deposited in a trusted repository?

Have you explored appropriate arrangements with the identified repository where your data will be deposited?

Does the repository ensure that the data is assigned an identifier? Will the repository resolve the identifier to a digital object?

Data:

Will all data be made openly available? If certain datasets cannot be shared (or need to be shared under restricted access conditions), explain why, clearly separating legal and contractual reasons from intentional restrictions. Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if opening their data goes against their legitimate interests or other constraints as per the Grant Agreement.

If an embargo is applied to give time to publish or seek protection of the intellectual property (e.g. patents), specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.

Will the data be accessible through a free and standardized access protocol?

If there are restrictions on use, how will access be provided to the data, both during and after the end of the project?

How will the identity of the person accessing the data be ascertained?

Is there a need for a data access committee (e.g. to evaluate/approve access requests to personal/sensitive data)?

Metadata:

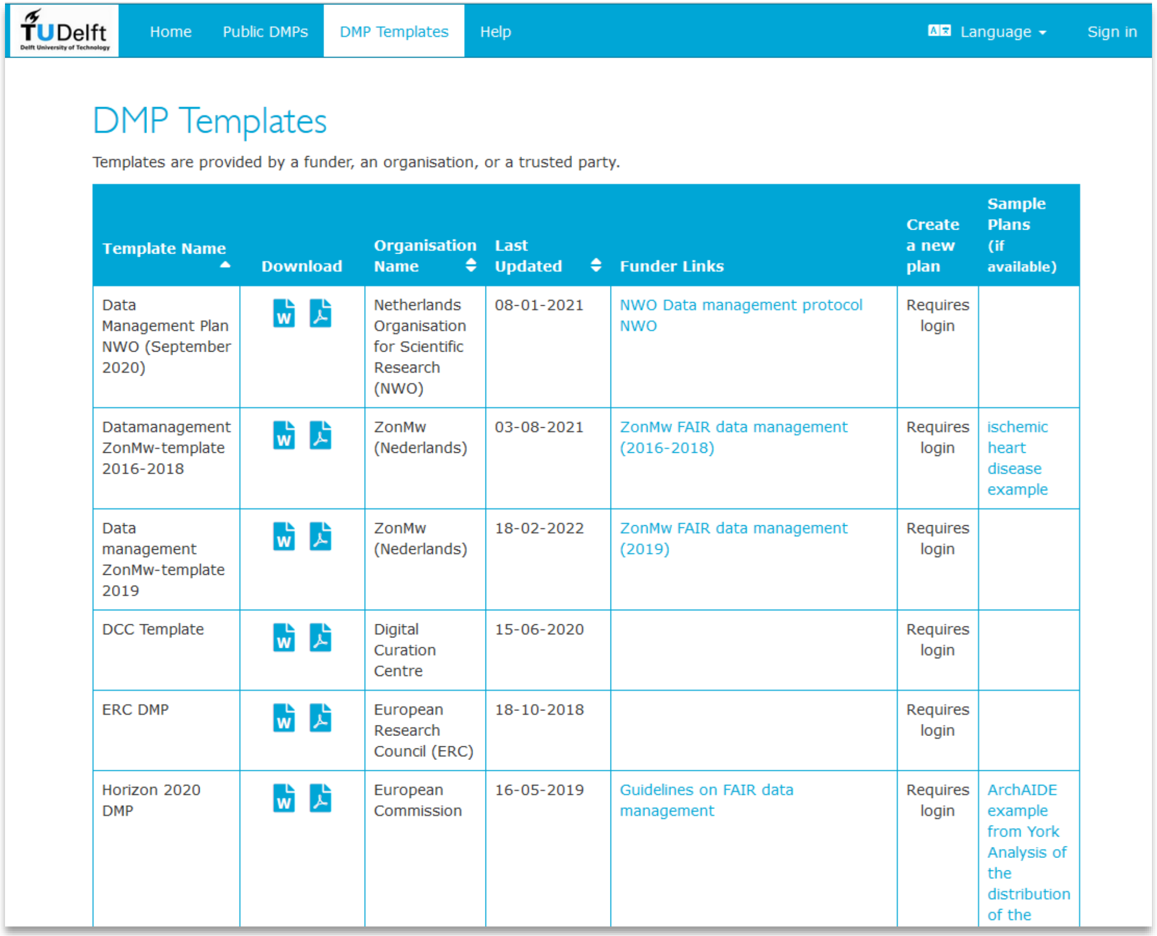
Will metadata be made openly available and licenced under a public domain dedication CC0, as per the Grant Agreement? If not, please clarify why. Will metadata contain information to enable the user to access the data?

Assistance: tools

- Wizards
  - Web-based tools
  - Usually free for individual researchers
  - Inbuilt templates
  - Customizable (for institutions)
  - Machine readable DMPs
  - Integration with repositories
- Checklists
  - Swedish National Data Service DMP checklist
  - Customised versions are often found on the websites of academic libraries

DMP Online

- Established by DCC
- Free for individual researchers
- Institutional instances
- Templates
- Public DMPs
- Export: xml, csv, txt, PDF, html



DMP Templates

Templates are provided by a funder, an organisation, or a trusted party.

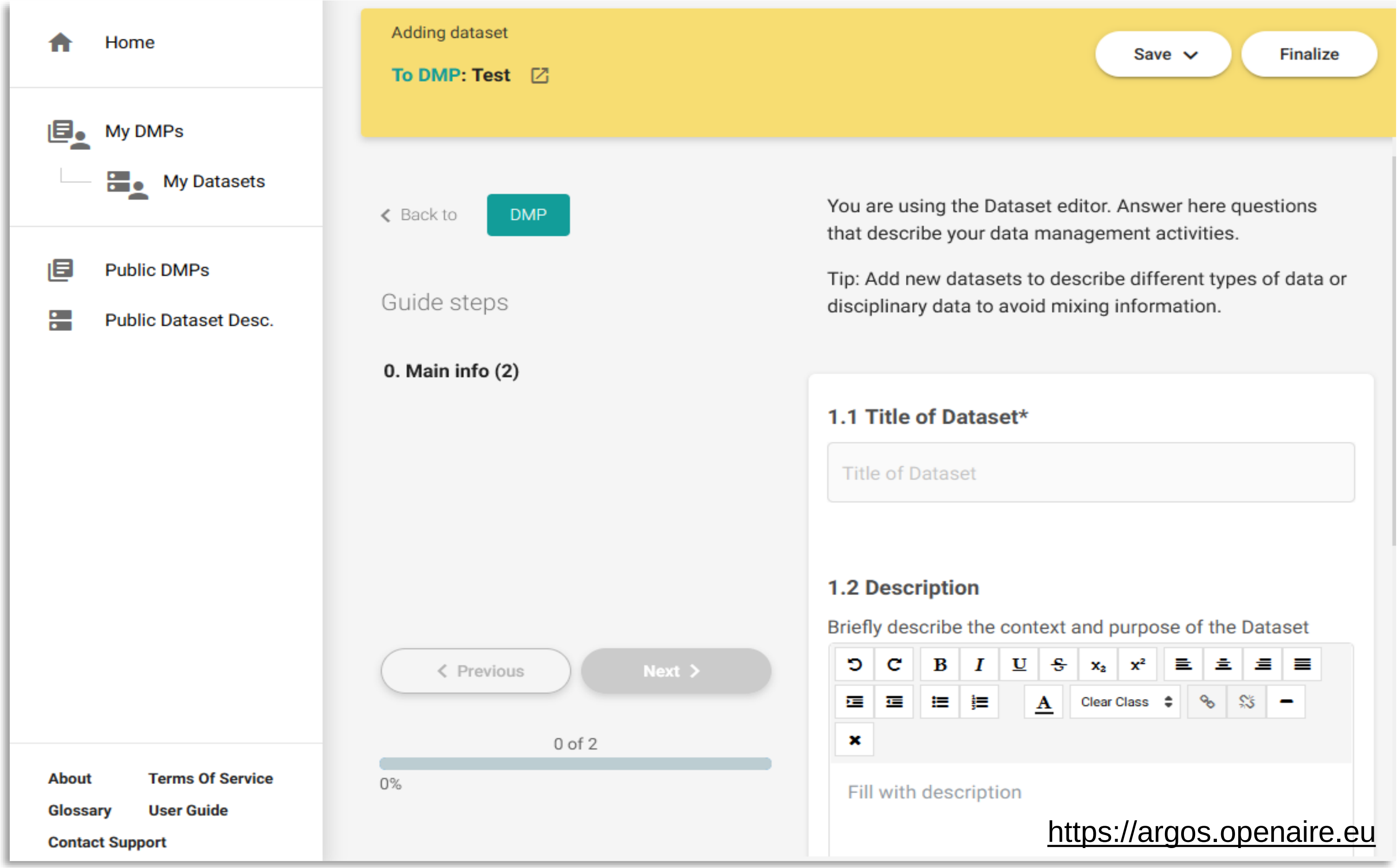
Template Name	Download	Organisation Name	Last Updated	Funder Links	Create a new plan	Sample Plans (if available)
Data Management Plan NWO (September 2020)	W	Netherlands Organisation for Scientific Research (NWO)	08-01-2021	NWO Data management protocol NWO	Requires login	
Datamanagement ZonMw-template 2016-2018	W	ZonMw (Nederlands)	03-08-2021	ZonMw FAIR data management (2016-2018)	Requires login	ischemic heart disease example
Data management ZonMw-template 2019	W	ZonMw (Nederlands)	18-02-2022	ZonMw FAIR data management (2019)	Requires login	
DCC Template	W	Digital Curation Centre	15-06-2020		Requires login	
ERC DMP	W	European Research Council (ERC)	18-10-2018		Requires login	
Horizon 2020 DMP	W	European Commission	16-05-2019	Guidelines on FAIR data management	Requires login	ArchAIDE example from York Analysis of the distribution of the

https://dmponline.dcc.ac.uk/

Argos

- Established by OpenAIRE
- Based on open-source software
- Free for individual researchers
- User interface in many languages
- Templates
- Public DMPs and datasets
- Machine actionable (based on the RDA Common Standard for machine-actionable Data Management Plans)
- Export: JSON, xml
- Publish on Zenodo

- Strong community support – monthly calls: https://www.openaire.eu/argos-community-calls



You are using the Dataset editor. Answer here questions that describe your data management activities.

Tip: Add new datasets to describe different types of data or disciplinary data to avoid mixing information.

1.1 Title of Dataset*

Title of Dataset

1.2 Description

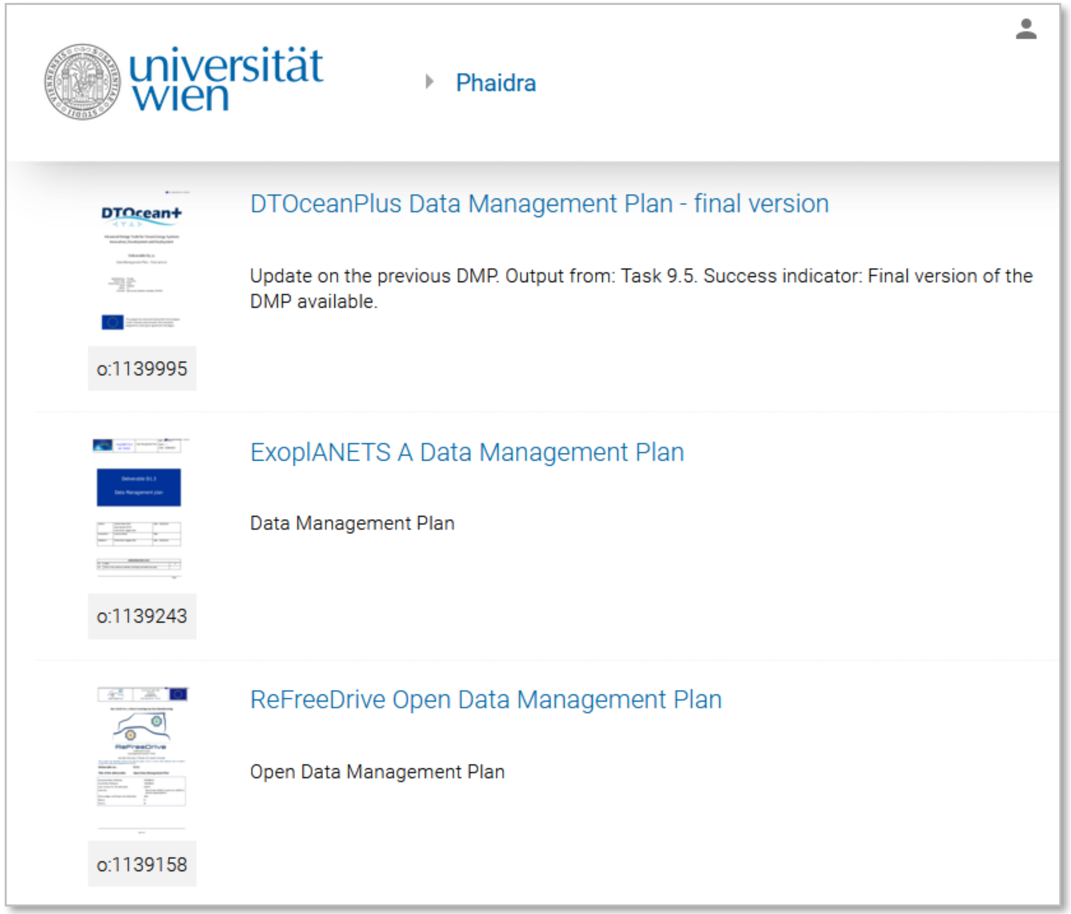
Briefly describe the context and purpose of the Dataset

Fill with description

https://argos.openaire.eu

Best practice: DMP examples and use cases

- DMP Use Case Project (OpenAIRE Austria): a collection of public DMPs of EC-funded projects: https://hdl.handle.net/11353/10.1140797
- Public DMPs in Argos, DMP Online etc.
- Search in Zenodo



DTOceanPlus Data Management Plan - final version

Update on the previous DMP. Output from: Task 9.5. Success indicator: Final version of the DMP available.

o:1139995

ExoplANETS A Data Management Plan

Data Management Plan

o:1139243

ReFreeDrive Open Data Management Plan

Open Data Management Plan

o:1139158

DCC Checklist for a DMP

- The DCC assessed existing funder requirements, DMP templates and other best practice to see what should be included in plans. This was synthesised down into common themes and questions.
- 13 questions on what’s asked across the board.
- Prompts/pointers to help researchers get started.
- Guidance on how to answer.

DCC Checklist for a Data Management Plan

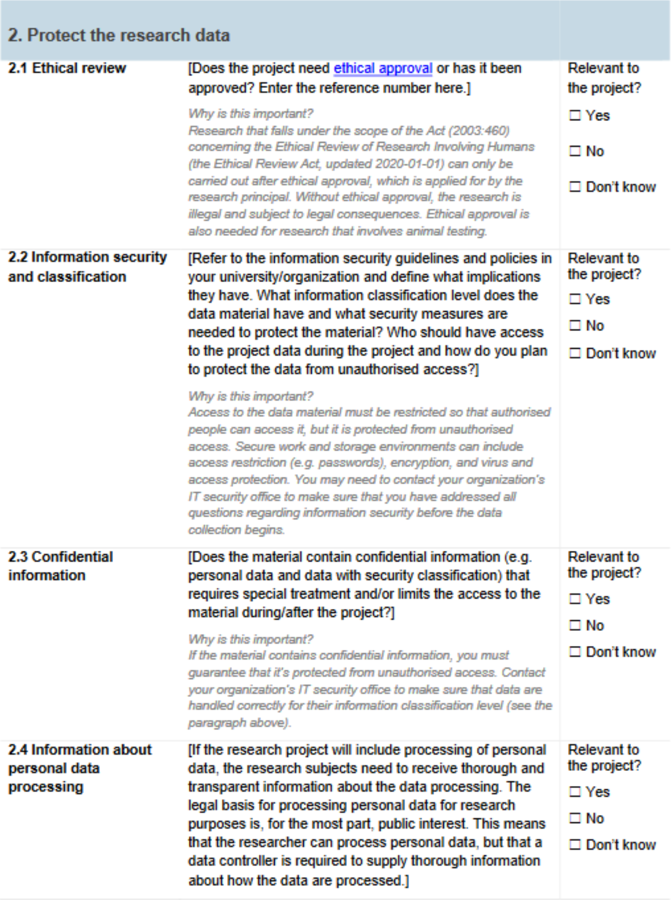
DCC

www.dcc.ac.uk/sites/default/files/documents/resource/DMP_Checklist_2013.pdf

Swedish National Data Service DMP checklist

https://snd.gu.se/en/manage-data/guides/dmp-checklist

- "The SND DMP checklist is designed so that it can be used as a complete and comprehensive plan for an entire research project. It can be used for different research areas, data types, phases of the research process, requirements from funding bodies, as well as various legal requirements."
- Very useful to those working with sensitive data



2. Protect the research data		
2.1 Ethical review	[Does the project need ethical approval or has it been approved? Enter the reference number here.] *Why is this important?* *Research that falls under the scope of the Act (2003:460) concerning the Ethical Review of Research Involving Humans (the Ethical Review Act, updated 2020-01-01) can only be carried out after ethical approval, which is applied for by the research principal. Without ethical approval, the research is illegal and subject to legal consequences. Ethical approval is also needed for research that involves animal testing.*	Relevant to the project? ☐ Yes ☐ No ☐ Don't know
2.2 Information security and classification	[Refer to the information security guidelines and policies in your university/organization and define what implications they have. What information classification level does the data material have and what security measures are needed to protect the material? Who should have access to the project data during the project and how do you plan to protect the data from unauthorised access?] *Why is this important?* *Access to the data material must be restricted so that authorised people can access it, but it is protected from unauthorised access. Secure work and storage environments can include access restriction (e.g. passwords), encryption, and virus and access protection. You may need to contact your organization's IT security office to make sure that you have addressed all questions regarding information security before the data collection begins.*	Relevant to the project? ☐ Yes ☐ No ☐ Don't know
2.3 Confidential information	[Does the material contain confidential information (e.g. personal data and data with security classification) that requires special treatment and/or limits the access to the material during/after the project?] *Why is this important?* *If the material contains confidential information, you must guarantee that it's protected from unauthorised access. Contact your organization's IT security office to make sure that data are handled correctly for their information classification level (see the paragraph above).*	Relevant to the project? ☐ Yes ☐ No ☐ Don't know
2.4 Information about personal data processing	[If the research project will include processing of personal data, the research subjects need to receive thorough and transparent information about the data processing. The legal basis for processing personal data for research purposes is, for the most part, public interest. This means that the researcher can process personal data, but that a data controller is required to supply thorough information about how the data are processed.]	Relevant to the project? ☐ Yes ☐ No ☐ Don't know

Discipline-specific DMPs

Science Europe Guidance Document

Presenting a Framework for Discipline-specific Research Data Management

January 2018

SCIENCE EUROPE

- Guidance Document Presenting a Framework for Discipline-specific Research Data Management (Science Europe, 2018), https://www.scienceeurope.org/our-resources/guidance-document-presenting-a-framework-for-discipline-specific-research-data-management
- Training materials in the SSH Open Marketplace: https://marketplace.sshopencloud.eu/search?order=score&q=data+management+plan&categories=training-material

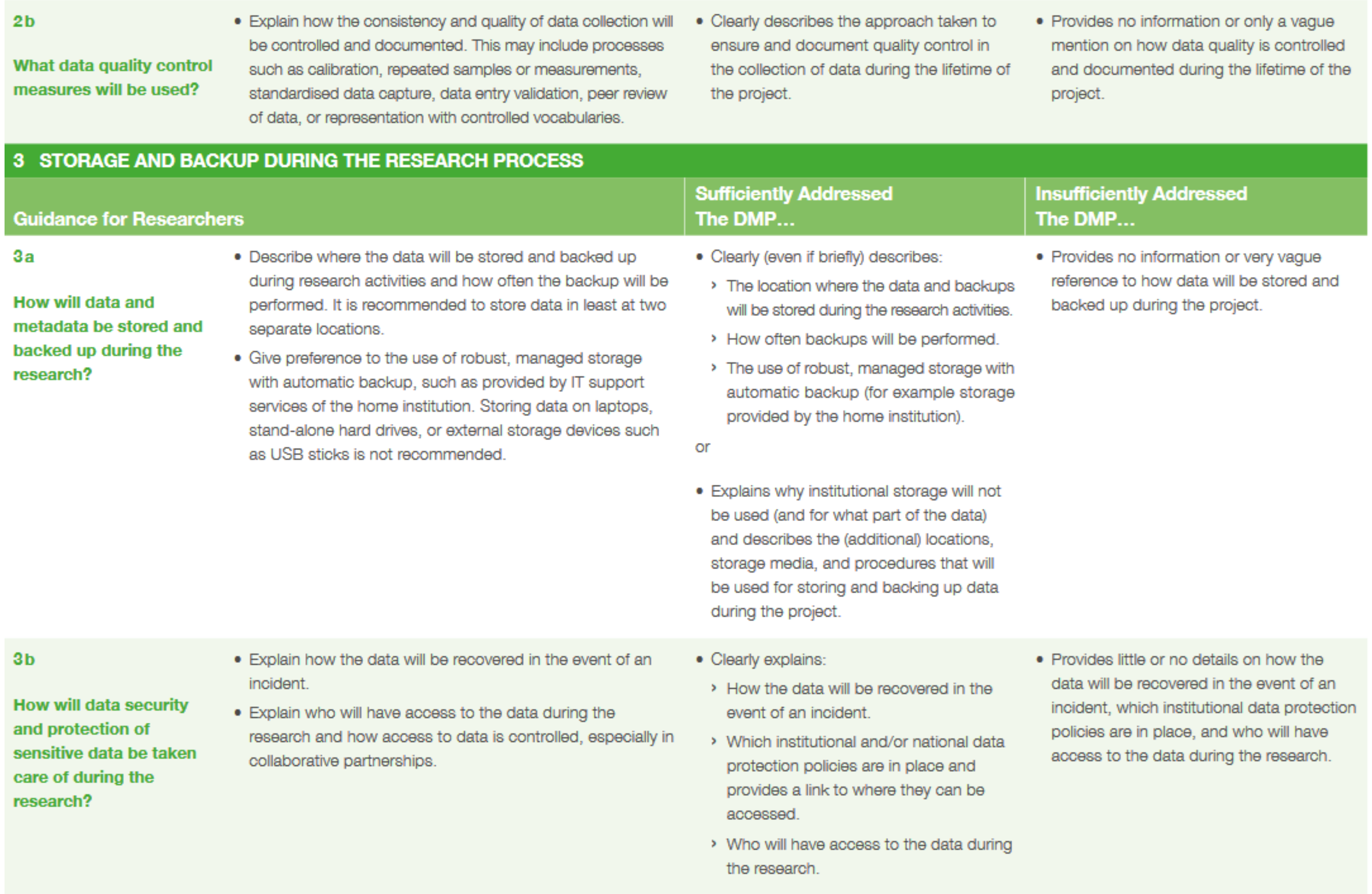
Assessment frameworks for DMPs

- DMP Evaluation Rubric (part of Practical Guide to the International Alignment of Research Data Management)
- “Reverse engineering”
- Structured approach
- Customised versions available on the websites of academic libraries

DMP question	DMP guidance	Performance level	
		Sufficiently addressed	Insufficiently addressed

Science Europe. (2021). Practical Guide to the International Alignment of Research Data Management - Extended Edition. https://doi.org/10.5281/zenodo.4915862

Science Europe. (2021). Practical Guide to the International Alignment of Research Data Management - Extended Edition. https://doi.org/10.5281/zenodo.4915862



2b **What data quality control measures will be used?**	• Explain how the consistency and quality of data collection will be controlled and documented. This may include processes such as calibration, repeated samples or measurements, standardised data capture, data entry validation, peer review of data, or representation with controlled vocabularies.	• Clearly describes the approach taken to ensure and document quality control in the collection of data during the lifetime of the project.	• Provides no information or only a vague mention on how data quality is controlled and documented during the lifetime of the project.
3 STORAGE AND BACKUP DURING THE RESEARCH PROCESS			
Guidance for Researchers		**Sufficiently Addressed The DMP...**	**Insufficiently Addressed The DMP...**
3a **How will data and metadata be stored and backed up during the research?**	• Describe where the data will be stored and backed up during research activities and how often the backup will be performed. It is recommended to store data in least at two separate locations. • Give preference to the use of robust, managed storage with automatic backup, such as provided by IT support services of the home institution. Storing data on laptops, stand-alone hard drives, or external storage devices such as USB sticks is not recommended.	• Clearly (even if briefly) describes: › The location where the data and backups will be stored during the research activities. › How often backups will be performed. › The use of robust, managed storage with automatic backup (for example storage provided by the home institution). or • Explains why institutional storage will not be used (and for what part of the data) and describes the (additional) locations, storage media, and procedures that will be used for storing and backing up data during the project.	• Provides no information or very vague reference to how data will be stored and backed up during the project.
3b **How will data security and protection of sensitive data be taken care of during the research?**	• Explain how the data will be recovered in the event of an incident. • Explain who will have access to the data during the research and how access to data is controlled, especially in collaborative partnerships.	• Clearly explains: › How the data will be recovered in the event of an incident. › Which institutional and/or national data protection policies are in place and provides a link to where they can be accessed. › Who will have access to the data during the research.	• Provides little or no details on how the data will be recovered in the event of an incident, which institutional data protection policies are in place, and who will have access to the data during the research.

What makes a good DMP?

- Clear, detailed information that is relevant to the science:
  - adopting recognised standards.
  - practices in line with norms for that field.
  - use of support services e.g. university storage, subject repositories…
- Realistic approach that is feasible to implement.
- Evidence of consultation and seeking advice.
- Proper justification of restrictions and costs.
- **Have you taken time to reflect on what to do?**

Is the information specific enough?

“we will use suitable formats to ensure that our data can be preserved and sustained over the long term”

- Which standards? Name them!
- Show that you know which are suitable.
- Does your chosen repository have preferences?

Are decisions justified?

“data will be made available upon request to bona fide medieval historians”

- ○ Why is it restricted?
- ○ Could other communities not reuse the data?
- ○ Will the research team be around to handle access requests in the future?

A better response...

“We will provide MP3 audio files for online dissemination. While this is not an open format, it is well-established and the most widely supported. High-resolution WAV files will be used for the archival master recordings.”

- Be clear, specific and detailed.
- Justify decisions.

Example plans

- Plans from several funders and disciplines via DCC www.dcc.ac.uk/resources/data-management-plans/guidance-examples

- Scientific DMPs submitted to the NSF (USA) provided by DataOne https://www.dataone.org/data-management-planning

- DMPs published in RIO journal http://riojournal.com/browse_user_collection_documents.php?collection_id=3&journal_id=17

- Share yours! - www.dcc.ac.uk/share-DMPs

Data description examples

The final dataset will include self-reported demographic and behavioural data from interviews with the subjects and laboratory data from urine specimens provided.

From NIH data sharing statements

Every two days, we will subsample E. affinis populations growing under our treatment conditions. We will use a microscope to identify the life stage and sex of the subsampled individuals. We will document the information first in a laboratory notebook and then copy the data into an Excel spreadsheet. The Excel spreadsheet will be saved as a comma separated value (.csv) file.

From DataOne – E. affinis DMP example

Metadata examples

Metadata will be tagged in XML using the Data Documentation Initiative (DDI) format. The codebook will contain information on study design, sampling methodology, fieldwork, variable-level detail, and all information necessary for a secondary analyst to use the data accurately and effectively.

From ICPSR Framework for Creating a DMP

We will first document our metadata by taking careful notes in the laboratory notebook that refer to specific data files and describe all columns, units, abbreviations, and missing value identifiers. These notes will be transcribed into a .txt document that will be stored with the data file. After all of the data are collected, we will then use EML (Ecological Metadata Language) to digitize our metadata. EML is one of the accepted formats used in ecology, and works well for the types of data we will be producing. We will create these metadata using Morpho software, available through KNB. The metadata will fully describe the data files and the context of the measurements.

From DataOne – E. affinis DMP example

Data sharing examples

The videos will be made available via the bristol.ac.uk website (both as streaming media and downloads) HD and SD versions will be provided to accommodate those with lower bandwidth. Videos will also be made available via Vimeo, a platform that is already well used by research students at Bristol. Appropriate metadata will also be provided to the existing Vimeo standard.

All video will also be available for download and re-editing by third parties. To facilitate this Creative Commons licenses will be assigned to each item. In order to ensure this usage is possible, the required permissions will be gathered from participants (using a suitable release form) before recording commences.

From University of Bristol Kitchen Cosmology DMP

We will make the data and associated documentation available to users under a data-sharing agreement that provides for: (1) a commitment to using the data only for research purposes and not to identify any individual participant; (2) a commitment to securing the data using appropriate computer technology; and (3) a commitment to destroying or returning the data after analyses are completed.

From NIH data sharing statements

Restrictions examples

Because the STDs being studied are reportable diseases, we will be collecting identifying information. Even though the final dataset will be stripped of identifiers prior to release for sharing, we believe that there remains the possibility of deductive disclosure of subjects with unusual characteristics. Thus, we will make the data and associated documentation available to users only under a data-sharing agreement.

From NIH data sharing statements

1. Share data privately within 1 year.
   Data will be held in Private Repository, but metadata will be public
2. Release data to public within 2 years.
   Encouraged after one year to release data for public access.
3. Request, in writing, data privacy up to 4 years.
   Extensions beyond 3 years will only be granted for compelling cases.
4. Consult with creators of private CZO datasets prior to use.
   Pis required to seek consent before using private data they can access

From Boulder Creek Critical Zone Observatory DMP

Archiving examples

The investigators will work with staff at the UKDA to determine what to archive and how long the deposited data should be retained. Future long-term use of the data will be ensured by placing a copy of the data into the repository.

From ICPSR Framework for Creating a DMP

Data will be provided in file formats considered appropriate for long-term access, as recommended by the UK Data Service. For example, SPSS Portal format and tab-delimited text for qualitative tabular data and RTF and PDF/A for interview transcripts. Appropriate documentation necessary to understand the data will also be provided. Anonymised data will be held for a minimum of 10 years following project completion, in compliance with LSHTM's Records Retention and Disposal Schedule. Biological samples (output 3) will be deposited with the UK BioBank for future use.

From Writing a Wellcome Trust Data Management and Sharing Plan

Data Schools

Thank you!

Questions?

(Please get in touch!)