Universität Zürich^UZH

Hauptbibliothek

Data Management Plans (DMPs)

GEO 802 Fall 2020, Data Information Literacy

Anna C. Véron, Dr. sc. nat.

Lesson 10: Data Management Planning

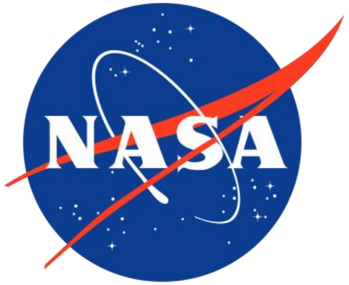
→ **Introduction: Why and what for?**

→ **DMP step by step**

→ **Your turn!**

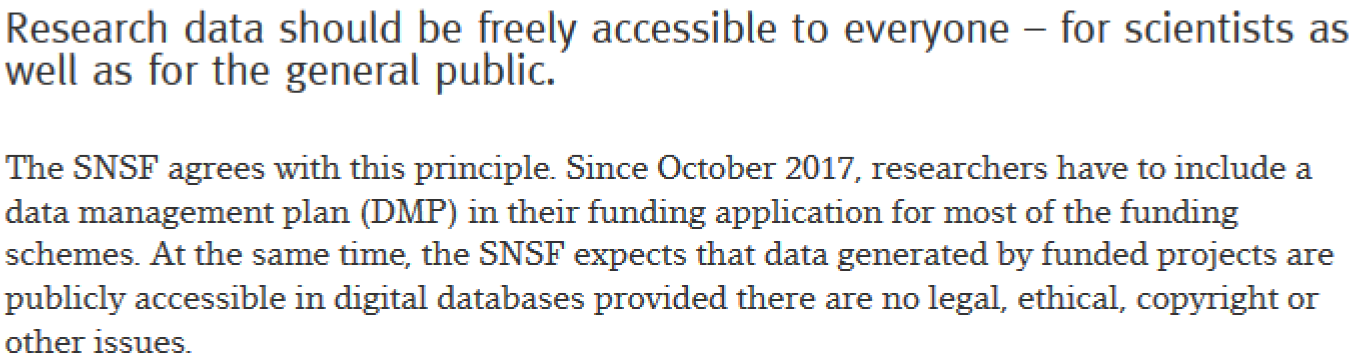
No DMP, no money!

FNSNF
FONDS NATIONAL SUISSE
SCHWEIZERISCHER NATIONALFONDS
FONDO NAZIONALE SVIZZERO
SWISS NATIONAL SCIENCE FOUNDATION



NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION
NOAA
U.S. DEPARTMENT OF COMMERCE

SNSF states…



Research data should be freely accessible to everyone – for scientists as well as for the general public.

The SNSF agrees with this principle. Since October 2017, researchers have to include a data management plan (DMP) in their funding application for most of the funding schemes. At the same time, the SNSF expects that data generated by funded projects are publicly accessible in digital databases provided there are no legal, ethical, copyright or other issues.

http://www.snf.ch/en/theSNSF/research-policies/open_research_data/Pages/default.aspx

Why do we need a DMP?

Why planning ahead is so important.
DID YOU BRING THE SUNBLOCK???
CAPITAL STRATEGIC SOLUTIONS

Save time
Less reorganization later

Research data management is only really efficient if you plan for it already before you start generating data!

Features of a DMP

- Formal document
- Outlines what you will do with your data **during** & **after** you complete your research
- Ensures your data is safe for the **present** & the **future**

FNSNF

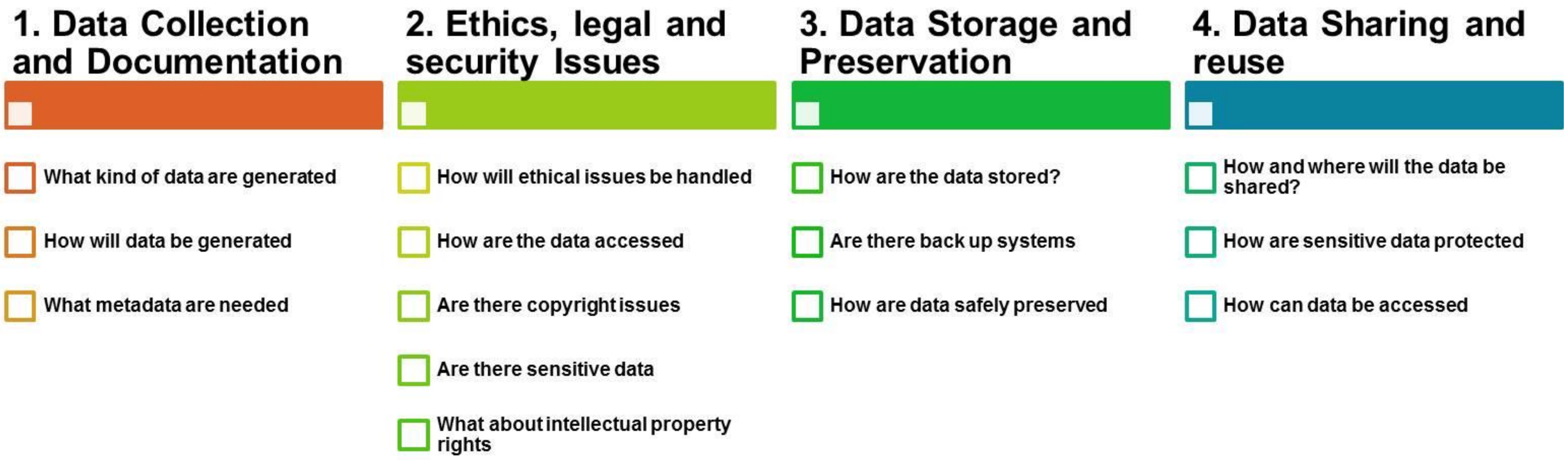
- The DMP is an integral part of the submitted proposal, but…
  - …it is **not part of the scientific evaluation** of a proposal
  - …a **plausible DMP draft** is sufficient at the time of submission

→ Within the project period the **DMP shall be changed and adapted** at any time.

→ At the end of the project **your DMP will be openly shared** on P3 (SNSF public database)

Exercise: DMP and FAIR data

– Here is an overview of the structure of a DMP
  Where do you see parallels to the FAIR principles?



1. Data Collection and Documentation

□ What kind of data are generated

□ How will data be generated

□ What metadata are needed

2. Ethics, legal and security Issues

□ How will ethical issues be handled

□ How are the data accessed

□ Are there copyright issues

□ Are there sensitive data

□ What about intellectual property rights

3. Data Storage and Preservation

□ How are the data stored?

□ Are there back up systems

□ How are data safely preserved

4. Data Sharing and reuse

□ How and where will the data be shared?

□ How are sensitive data protected

□ How can data be accessed

Lesson 10: Data Management Planning

✓ Introduction: Why and what for?

→ DMP step by step

→ Your turn!

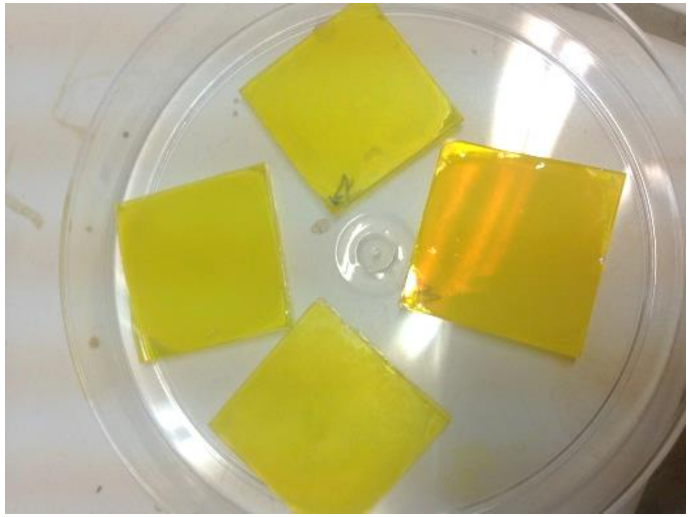
Example Research Project

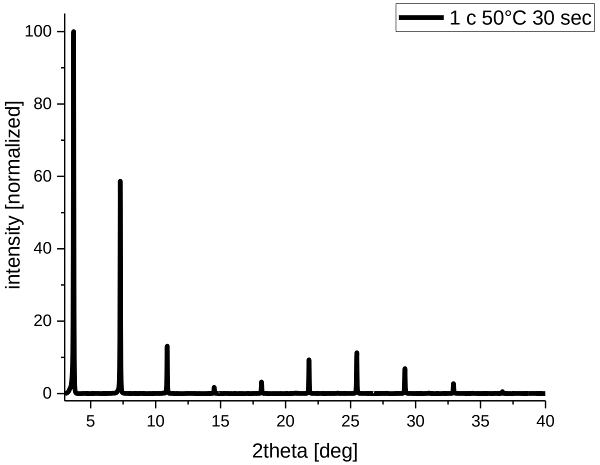
Me in my previous life

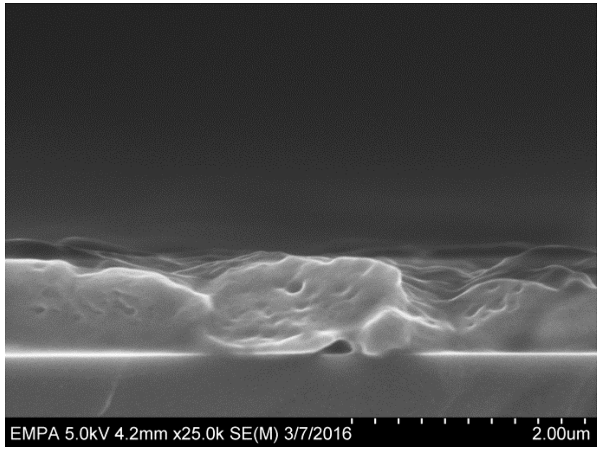
«Perovskite Thin Films»

- preparation of perovskite thin films under different conditions
- characterization by X-ray diffraction

Let's create a DMP for this project!







1. Data Collection and Documentation

1.1 What data will you collect, observe, generate or reuse?

- Type of datasets (e.g. measurement, experiment, observation, survey, etc.)
- Origin of data (if you are reusing: cite the reference)
- Format of data:
  - Raw format (e.g. as produced by instrument)
  - Processed Format – Open standard formats wherever possible!
    → Lists with recommended file formats:

    ETH: https://documentation.library.ethz.ch/display/DD/File+formats+for+archiving

    EPFL: https://researchdata.epfl.ch/files/content/sites/researchdata/files/doc/EPFL_recommended_file_formats.pdf
- volume of data (MB / GB / TB range)

1.1 What data will you collect, observe, generate or reuse?

Estimated amount of samples to be prepared during the project: 100
Each sample will be labelled with a numerical ID from 001 to 100.

		Format	Description / Origin	Volume
1a	Description of sample preparation	Handwritten notes	Description, experiment conditions and observations during sample preparation. Created while experiment is carried out.	-
1b	Description of sample preparation	.pdf	**1a** are scanned once a week.	10-50 MB
1c	Description of sample preparation	.txt	**1b** are converted into text using the handwriting recognition feature of Microsoft OneNote. Additional metadata are added to the file as described in section 1.3 for files **1c.**	100-150 KB
2a	X-ray diffraction data	.xrdml	Raw file created by PANanalytical X-ray diffractometer (includes the instrument type and measurement settings).	1-2 MB
2b	X-ray diffraction data	.asc	**2a** are background corrected and converted to .asc files using Malvern Panalytical Highscore Plus (version 4.6a).	3-5 MB
2c	X-ray diffraction data	.opj	**2b** are imported in Origin2018b and different plots are created for further analysis.	50-150 MB
2d	X-ray diffraction data	.svg	Plots created from **2c** are exported from Origin2018b as vector files.	100-200 MB
2e	X-ray diffraction data	.cif	**2a** raw files are converted to crystallographic information files (CIF) using Malvern Panalytical Highscore Plus (version 4.6a).	10-50 MB

Recommendation

- Create a **list of all your dataset types**
- Label your dataset types (e.g. here: 1a, 1b, etc.)
- It will be easier to refer to these data types later in the DMP
- Work with tables or bullet lists rather than writing full sentences and paragraphs

1. Data Collection and Documentation

1.2 How will the data be collected, observed or generated?

- **Standards, methodologies, quality assurance processes**
  - *e.g. calibration processes, repeated measurements, double blind study, data validation, data peer review, any other internal procedures...*
- **File organization and versioning**
  - provide naming conventions, folder structures, version control
    e.g. *Project-Experiment-Scientist-YYYYMMDD-HHmm-Version.format*
  - For code: use Git
  - If possible, use a data management system, such as an Electronic Lab Journal / Lab Information System (ELN /LIMS)

1.2 How will the data be collected, observed or generated?

Folder hierarchy: Project-name → Sub-Project-name → SampleID
General naming convention: ProjectID-SampleID-DataType-ScientistID-YYYYMMDD-Version.format

		Naming	Quality Assurance
1b	Description of sample preparation	Pero-001-Prep-averon-20181101-v1.pdf	**Describe or link to standards, methodologies or quality assurance protocols.**
1c	Description of sample preparation	Pero-001-Prep-averon-20181101-v1.txt	
2a	X-ray diffraction data	Pero-001-XRD-averon-20181101-v1.xrdml	
2b	X-ray diffraction data	Pero-001-XRD-averon-20181101-v1.asc	
2c	X-ray diffraction data	Pero-001-XRD-averon-20181101-v1.opj	
2d	X-ray diffraction data	Pero-001-XRD-averon-20181101-v1.svg	
2e	X-ray diffraction data	Pero-001-XRD-averon-20181101-v1.cif	Validation with CheckCIF software: http://checkcif.iucr.org/

Recommendation

- Define a file naming system and folder hierarchy that makes sense and is intuitive to use for your research group
- Refer to the data types established in 1.1.

1. Data Collection and Documentation

1.3 What documentation and metadata will you provide with the data?

- How will you ensure that any user will be able to read and interpret the data in the future?
- Are there any metadata community standards that you can use? (Maybe you are using them already? check here: http://rd-alliance.github.io/metadata-directory/standards/)

1.3 What documentation and metadata will you provide with the data?

1.3.1 Metadata added to describe folder contents

Folders	Documentation and Metadata
Project Name : Perovskite-Thin-Films	A file “**Perovskite-Thin-Films_Documentation.tx**t” is placed in the project folder containing: - Grant number, Timeline - Description of project - list of all involved scientists (incl. internal scientist ID, full name, affiliation during project duration, and ORCID) - list of all sub-project folder names (continuously updated)
Sub-Project names: - Concentration-Variation, - Temperature-Variation, - Dipping-Time, - etc.	A File “**Project-name_Sub-project-name_Documentation.txt**” is placed in each sub-project folder containing: - description and scope of the sub-project
Sample ID: - 001 - 002 - 003 - Etc.	Documentation and metadata for each sample are contained in the files **1c** as described under 1.3.2.

Recommendation

Ask yourself: Which metadata are necessary to understand the contents of your folders?

1.3 What documentation and metadata will you provide with the data?

1.3.2 Metadata contained within data files

	Folders	Documentation and Metadata	Community Standards
1c	Description of sample preparation	The following metadata related to the experiment are included in the description files: - *researcher(s) responsible for the creation of the respective sample /data file* - *date / time of experiment* - *detailed experimental procedure and laboratory conditions: temperature, humidity, concentration, dipping time, etc. (additional parameters will be added as required).*	For the description of sample preparation, vocabulary contained in the ScienceWISE ontology is used, whenever possible: http://sciencewise.info/ontology/
2e	X-ray diffraction data	CIF files contain metadata as described by the CIF documentation: https://www.iucr.org/resources/cif/spec/version1.1	CIF (Crystallographic Information Framework) by the International Union of Crystallography https://www.iucr.org/resources/cif The vocabulary and abbreviations used are described in the CIF dictionaries: https://www.iucr.org/resources/cif/dictionaries

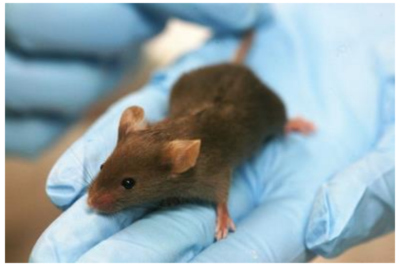
Tipp

Many file types already contain certain metadata (especially raw data created by some measurement instruments).

List them all while referring to the file types introduced in section 1.1.

2. Ethics, legal and security issues

2.1 How will ethical issues be addressed and handled?



- **List any ethical isssues involved in the research project**
  - Human participants
  - Privacy issues (confidential or sensitive data)
  - Animal experiments
  - …
- **Describe how these ethical issues will be managed.**
  - Consent from ethics committee, anonymization of personal data, sensitive data is not stored on cloud services, etc.

→ Delegate for Data Protection of UZH: https://www.dsd.uzh.ch/en/contact.html

→ UZH Ethics Commission: https://www.ethik.uzh.ch/en/ethikkommission.html

→ Zurich Cantonal Ethics Commission (in German): https://kek.zh.ch/internet/gesundheitsdirektion/kek/de/home.html

2.1 How will ethical issues be addressed and handled?

There are no ethical issues to be addressed in this research project. The principal investigator has the permission to obtain, process, preserve and share all data mentioned in 1.1.

2. Ethics, legal and security issues

2.2 How will data access and security be managed?

- **Describe any security concerns**
  - Refer to the sensitive data described in 2.1

- **Describe what measures are in place to handle security risks**
  - Who will have access to the data and how are these access rights / permissions regulated?
  - How can you ensure safe storage of personal or sensitive data?

→ **Responsibilities to manage access: check with your institutional IT services!**
Who is responsible to grant and / or manage access to the data?

This section is only important for those who create sensitive data!

2.2 How will data access and security be managed?

No sensitive data are collected and thus no specific measures to manage data access and security are necessary.

Access to the data is managed through an identity management system and is possible off-site via a VPN connection. User access is authorized by the project PI *(**name of PI**)* and managed by the institutional IT services of ***name of institute** (**name of responsible person**).*

Recommendation

Make sure that sections 2.2 (data access) and 3.1 (data storage) do not contradict each other.

→ data storage also relates to access (e.g. in case you store data on personal drives, where access is not really controlled!)

2. Ethics, legal and security issues

2.3 How will you handle copyright and Intellectual Property Rights issues?

- **Which licenses will be applied to the data?**
  - Refer to the data you will be sharing (section 4.1).
  - Recommendation: choose a Creative Commons license, e.g.:

    CC0 = no copyright reserved
    CC-BY = enables the data set to get cited
    CC-BY-NC = allows you to keep the exclusive right to commercial use

- **Who will be the owner of the data?**
  - Declare who will be the owners (“authors”) of the data and **how you would like your data to be cited**.
  - Comply **with institutional policies –** Currently, UZH does not have such a policy (yet)!

- **What restrictions apply to the reuse of third-party data?**
  - If you declared the reuse of third-party data in 1.1, indicate the licenses and required permissions that allow their reuse.

2.3 How will you handle copyright and Intellectual Property Rights issues?

All data described in 4.1 will be available under a Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/). The owners of the data are all researchers who scientifically contributed to the creation of the data. They will be listed in the metadata of the shared dataset and a recommendation on how to cite will be provided.

Recommendation

- Refer to the data in section 4.1 (sharing of data).
- A Creative Commons Attribution License («CC-BY») is recommended whenever possible, since it is an open license that allows you to get cited.
- Make sure you use the newest license version (currently 4.0).

3. Data storage and preservation

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

- **What are your storage capacities and where will the data be stored?**
  - List all locations where the data are stored.
    → Storage through IT services (e.g. network drives) is safer than hard drives or laptops!

- **What are the back-up procedures?**
  - Who is responsible?
  - Automatic or manual processes?

→ **Check with your institutional IT services!**

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

All digital data described in section 1 will be stored on a network drive (500 GB) maintained and backed-up by the institutional IT services of ***name of institute** (**name of responsible person**).*

Ask IT-services to provide a description of their back-up procedures to paste here.

3. Data storage and preservation

3.2 What is your data preservation plan?

- **Describe the selection criteria that will be applied to select data to be preserved**
  - e.g. reusability or data, value / quality of data, ethical considerations, stakeholders requirements, costs Which data will be kept, which ones deleted?
  - Who is the responsible person during the selection process and after the end of the project?
- **Which file formats will be used for data preservation?**
  - File formats suitable for long-term preservation: https://documentation.library.ethz.ch/display/DD/File+formats+for+archiving
  - Format migration – refer to file types listed in section 1.1

→ This section of the DMP is particularly difficult to write before the research project has started. **Remember that only a draft is required at this stage.**

→ Think about this topic while your project is running and **create the final version towards the end of the project!**

3.2 What is your data preservation plan?

All digital data described in section 1 will be stored on a network drive (500 GB) maintained and backed-up by the institutional IT services of **name of institute** (**name of responsible person**).

	File format	Retention period	Storage location / Comments	Responsible Person
1b	PDF-A encoded as UTF-8	10 years or more	Institutional server	**name & contact**
1c	.txt encoded as UTF-8	20 years or more	Zenodo and institutional server	**name & contact**
2a	.xrdml	20 years or more	Zenodo and institutional server	**name & contact**
2b	.asc encoded as UTF-8	20 years or more	Zenodo and institutional server	**name & contact**
2c	.opj (proprietary format, not suitable for archiving)	10 years or more	Institutional server	**name & contact**
2d	.svg	20 years or more	Zenodo and institutional server	**name & contact**
2e	.cif	20 years or more	Zenodo and institutional server	**name & contact**

Tipps

- Describe selection criteria: Which data are worth archiving and why? (resuability)
- Zenodo claims to archive data for at least 20 years
- Primary data on your institutional servers should be archived for at least 10 years

4. Data sharing and reuse

4.1 How and where will the data be shared?

- **Choose a repository to share your data**
  - Check **re3data.org** to find a subject-specific repository for your field
  - If no suitable subject repositories are available: Choose **Zenodo**
    Zenodo is CERN’s repository which fulfills SNSF requirements and it accepts research data across all disciplines.

→ Remember: Only the data related to a publication are mandatory to be published (at the time of the publication).

→ All other data are optional and can have an embargo period of your choice.

4.1 How and where will the data be shared?

Data of the types **1c, 2b, 2d, 2e, 2f** will be shared on the repository Zenodo under a Creative Commons Attribution 4.0 International License (see 2.3).

Datasets will be created grouping all data related to each sub-project (see **1.3.1**) and will be made available at the time the data are cited in a publication.

Any data which are not related to a publication will be made available with an embargo period of 3 years after the end of the project (under a Creative Commons Attribution 4.0 International License).

- Remember:
  Data need to be open at the time of publication
- Any data not (yet) mentioned in a publication are optional and can thus have an embargo period.

4. Data sharing and reuse

4.2 Are there any necessary limitations to protect sensitive data?

- **Sharing your sensitive data under controlled access is possible**
    - Refer to any sensitive data listed in section 2.1. Under which conditions will the data be made available?
    - Explain any legal, ethical, copyright, confidentiality or other clauses which prevent you from sharing the data openly.
    - Consider whether a non-disclosure agreement would be sufficient for your confidential data.

This section is only important for those who create sensitive data!

4.2 Are there any necessary limitations to protect sensitive data?

No sensitive data are collected and thus no limitations are necessary.

4. Data sharing and reuse

4.3 I will choose digital repositories that are conform to the FAIR Data Principles.

☑

Needs to be checked. If your data cannot be shared (as you have stated in section 4.2), this is a statement of principles.

4.4 I will choose digital repositories maintained by a non-profit organization.

YES: SNSF will contribute up to 10’000 CHF to the costs of data sharing.

NO: SNSF will only contribute to data management, but not to the costs of uploading data to the (commercial) repository.

Summary of Lesson 10

A DMP is essential for successful data management.

The four main topics of a DMP are:

1. **Data collection and documentation**
2. **Ethics, legal and security issues**
3. **Data storage and preservation**
4. **Data sharing and reuse**

The DMP is a “work in progress”, i.e. it can and *should* be updated during the research project.

Use a tabular format to create your DMP. Make it clear and concise and avoid long “flowery” sentences.

Lesson 10: Data Management Planning

✓ Introduction: Why and what for?

→ DMP step by step

→ Your turn!