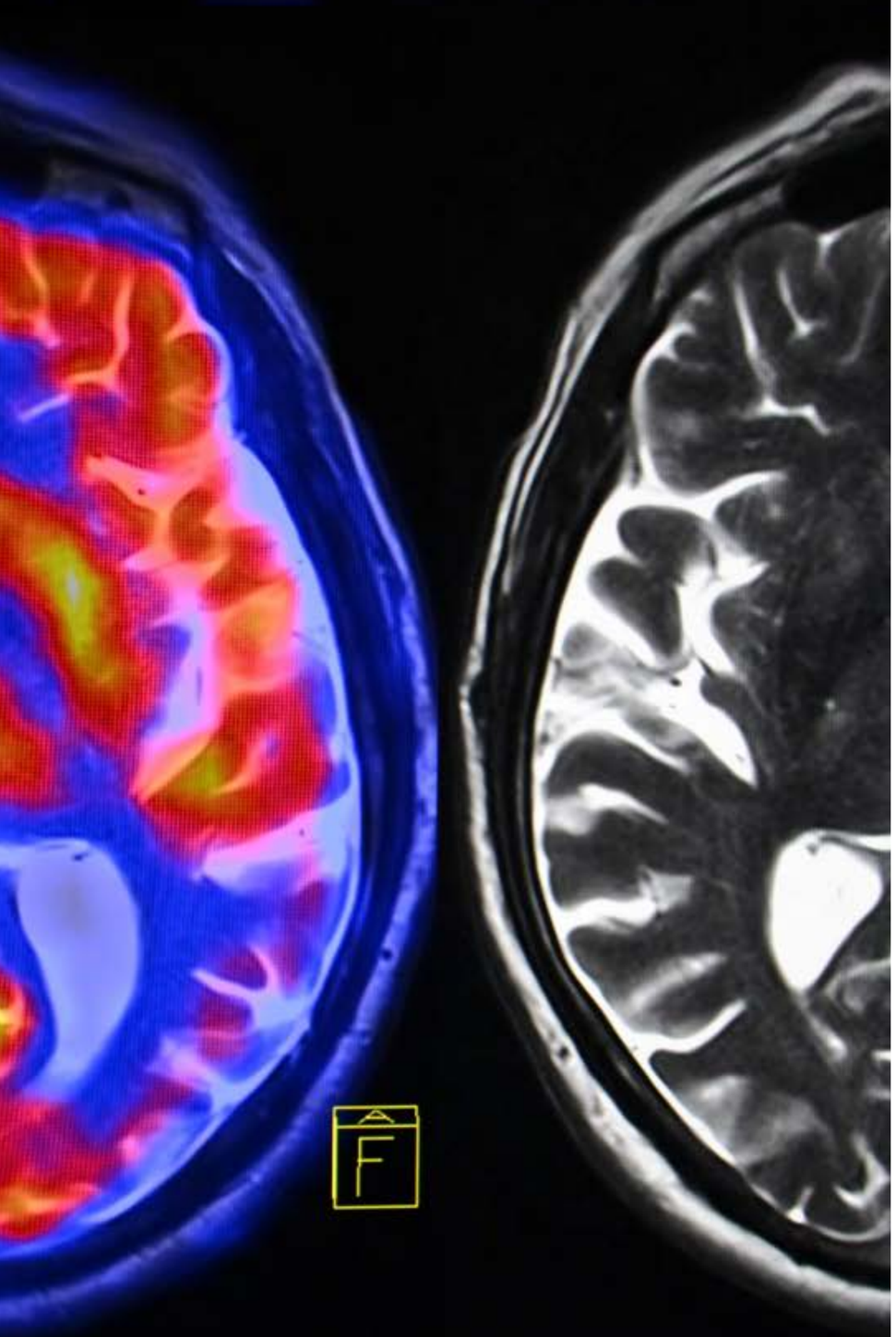
Data Management

Elian E. Jentoft
Stipendiat
Social Work and
Social Policy

03.09.2021

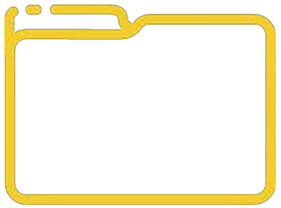


About Me...

...or rather, where I used to work.

- Research Assistant at the Psychology department at University of Oslo
- Human Time Data Project
- Interviewing Researchers on their Data Management Plans (or lack thereof)
- Developing Best Practice Documents

How Does This Relate to Me?

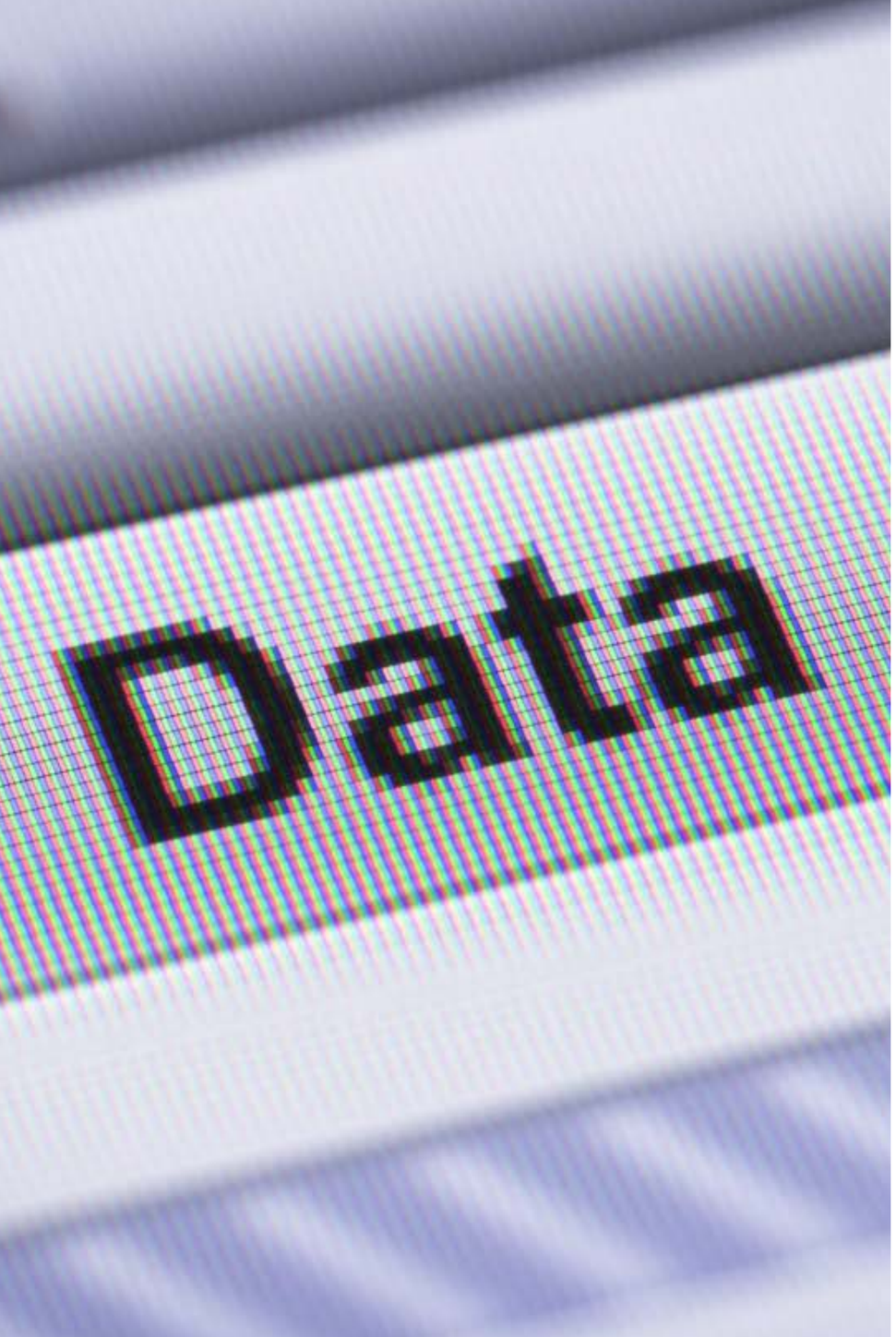


The basic principles of data management are applicable across disciplines.

Regardless of if you're doing qualitative or quantitative research and no matter how large or small your data is, it's important! This is especially true now with the Open Data Movement!

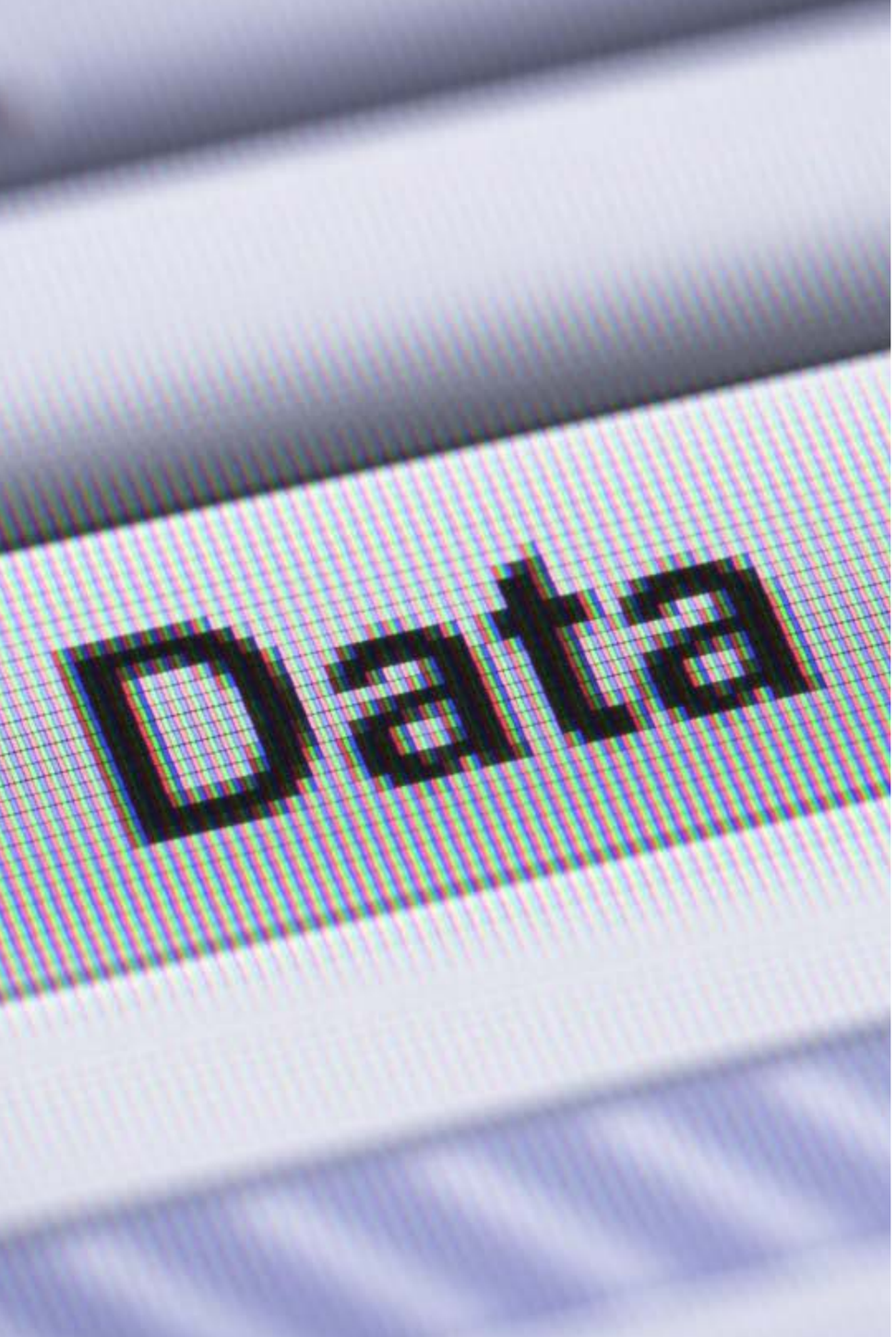
Developing good data management habits early in your career will make things easier!

In the long run, what seems tedious now will save you time and set you up for success (and less stress) later in your career.



Why is this so important?

- Help you and other researchers who will use your dataset in the future (or just future you) by supplying adequate metadata to understand it independently.
- Save time by requiring less support for a dataset's use from the source.
- Prepare data for sharing.
- Save time if deployed from the start of a project, by ensuring data is organized in a similar version to its end state, thus eliminating the need to organize it later.



Why is this so important?

- Improve overall efficiency.
- Support easier reuse of data for multiple publications.
- Provide more control over human errors that occur throughout the project.
- Help meet requirements of ethical review boards which often require at least a basic data management plan to be submitted with clearance applications.

But also...

- Help you meet requirements of funders, many of whom now demand that data management plans accompany grant applications. Some also demand data sharing as a condition for funding.
- Some major journals also require data sharing upon publication.
- Expand your future employment possibilities. Working with data requires the skills to manage it, and this is unlikely to change any time soon!

Think ahead!

What will happen to your data when you're done?

Plan for the end of your study as early as possible.

Plan to share your data if possible

If you are able to share your data you will be encouraged to do so.

Know where it will be stored.

Data curation and storage entities exist for different types of data and they all have their own data management conventions.

Understanding the data lifecycle

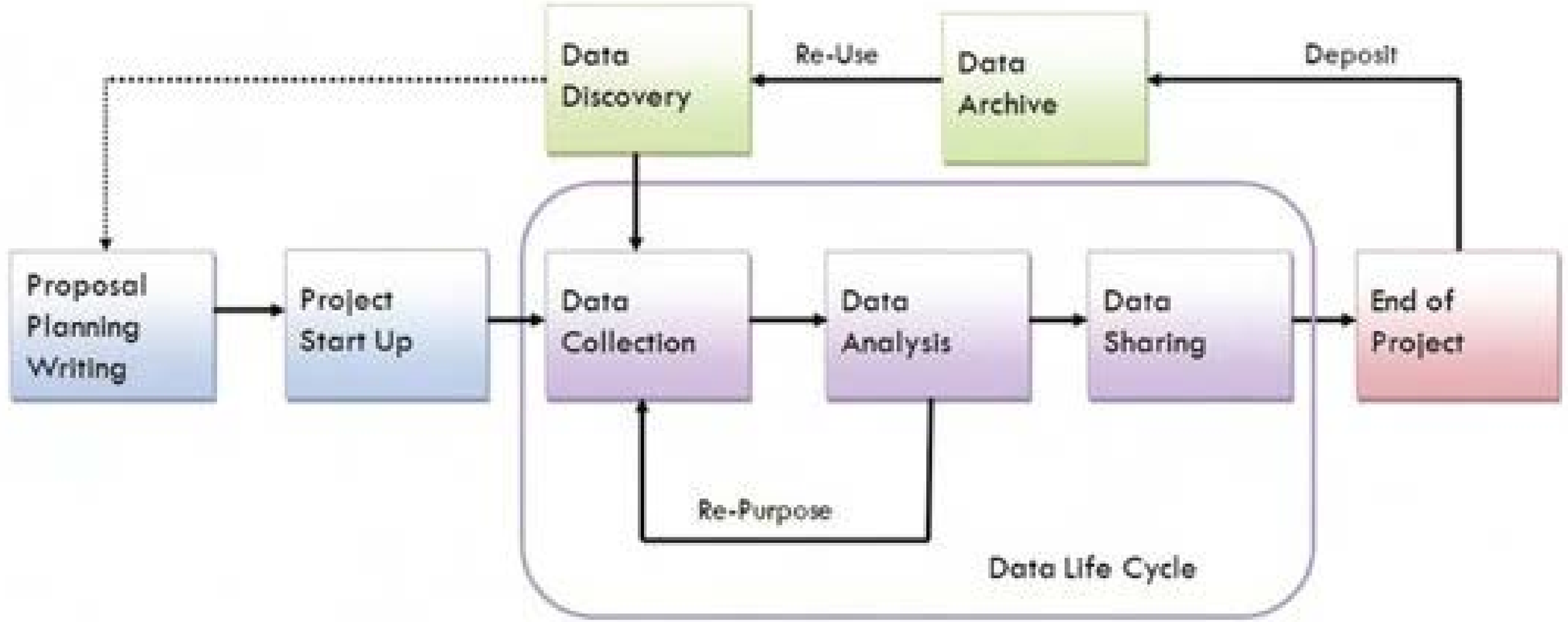
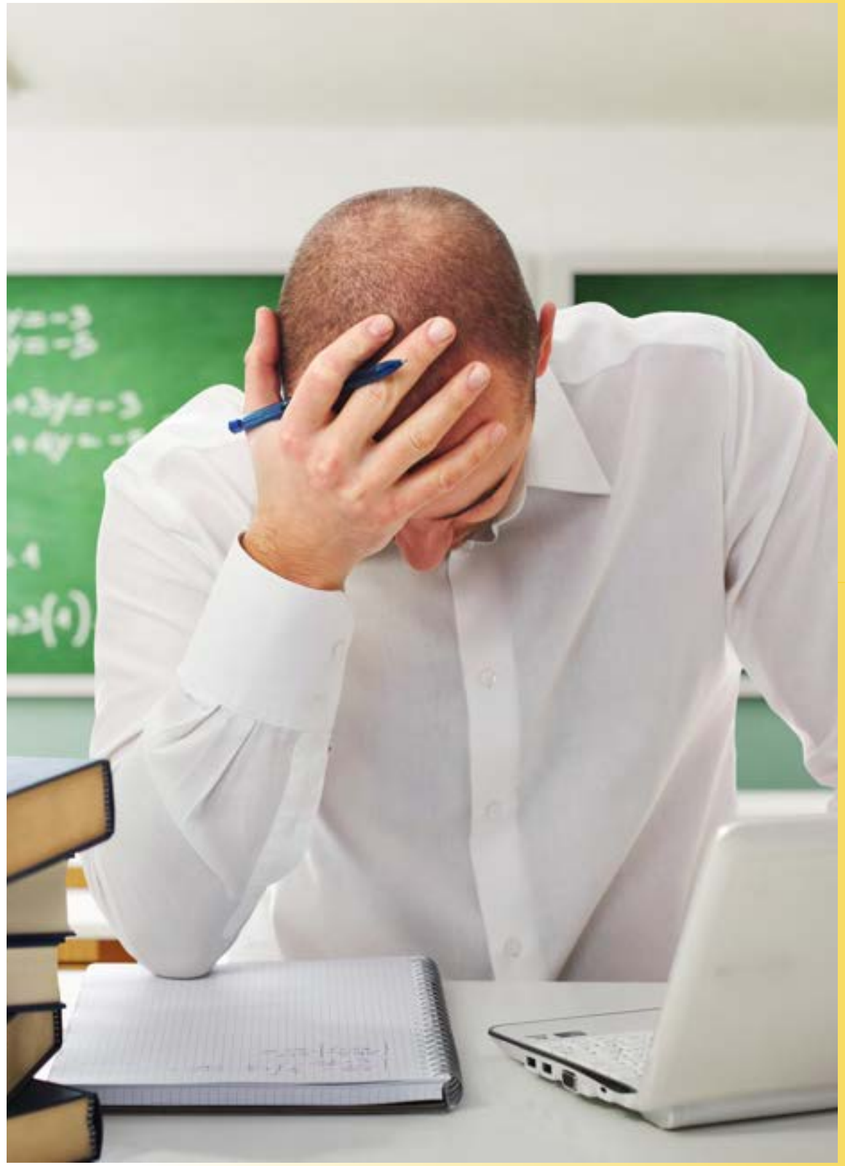


Figure 3. "The Research Life Cycle" model developed by the University of Virginia Library's Scientific Data Consulting Group.

But don't worry

Those of you nearing the end of your PhD:

Everyone has to start somewhere! Even if you're at the end of your PhD, you're at the beginning of your academic career and that's something!



What are the FAIR Principles?

Funders are going to want you to know this!

The FAIR Principles

Four foundational principles toward Open Science

Metadata

It's data about data!

- Who created the data?
- Who worked with the data?
- What year was it created?
- Which institution/country?
- File naming conventions and what they mean.
- File types.
- Units of measurement.
- Conditions for usage.
- Population data.
- Experiment parameters, description
- ETC

Findable

What does it mean for data to be findable?

- Data should be findable by both humans and computers.
- The dataset should have a unique, persistent identifier such as a DOI.
- The dataset should contain strong metadata that makes working with the dataset easier.
- The metadata should be registered within a searchable database or other resource.

Accessible

What does it mean for data to be accessible?

- The data is able to be retrieved via its identifier (but authorization procedures can be in place as needed).
- Protocols should be open and able to be implemented by others.
- The metadata should remain available even when the actual data is not.

Interoperable

What does it mean for data to be interoperable?

- Metadata should be written in a common, accessible language with a broad range of applications, for example .txt, .json
- Propriatory file types should be avoided when possible.

Reusable

What does it mean for data to be reusable?

- Metadata should be well-described
- The data usage license should be clearly stated.
- The metadata should conform to standards applied in the relevant community of researchers.

Making a Data Management Plan (DMP)

What do I need to think about?

What does your funder require?

Sponsors of research are beginning to place more and more emphasis on data management and data sharing. *Forskningsrådet* offers guides on DMP requirements and encourages data sharing where appropriate. If you don't submit a DMP with your grant application, you may need to justify it.

What does the ethics board require?

Norwegian ethics boards like REK and NSD will require the submission of a DMP which indicates how data will be stored securely during and after the project, especially if you are collecting sensitive data.

Ethics Boards

Where will consent forms be stored during the project and what will happen to them once the project ends?

How will you anonymize your data?

Where will it be stored during and after the study?

Making a Data Management Plan (DMP)

What do I need to think about?

What does your data repository require?

The repository should be listed in DMPs submitted to funders and ethical review boards.

This is important because:

- You will need to budget for costs involved with curation, storage, maintanence and sharing of your data.
- You will also need to obtain informed consent for long-term storage and sharing of your data from participants.
- Certain repositories have specific data structuring and curation rules you may want to follow from the start.

Making a Data Management Plan (DMP)

What do I need to think about?

What kind of data will you collect?

What is the nature of your data? Many projects cross modalities (video and transcripts, etc.). Some projects will reuse data from other projects. Include file formats that will be generated and if any of them are proprietary. You will also want to indicate the volume of data that will be collected.

Who is responsible for data collection?

Who will be involved in data collection and what role and tasks are they each responsible for? What training will they receive to ensure they are prepared to collect the data for your project?

Making a Data Management Plan (DMP)

What do I need to think about?

Where will your data be stored during the study?

What servers will the data be stored on? Are international servers involved? What is the security status of the server? Will hard copies exist?

How is data transported?

If you are working across institutions, extracting data from a lab or doing field work, you need to consider how data gets securely from point A to point B.

Making a Data Management Plan (DMP)

What do I need to think about?

How will your data be organized?

Having a shared plan in place for exactly what folder and file-naming conventions will be used will spare grief when others join your project.

- What programs will you use in data capture and analysis and what file types do the programs generate?
- How the folders and files will be structured is also an important consideration. Decide before data collection begins if you will use a preexisting file structure accepted in your field or create your own.

Making a Data Management Plan (DMP)

What do I need to think about?

Who will have access?

You may designate access permissions in your ethics applications as part of ensuring data security. Access permissions may also need to be granted by the IT department.

What are your plans for analysis?

What steps will be taken in the analysis process? Include not only the methods, but the software used that will be used and the files that each step of the process will produce.

Making a Data Management Plan (DMP)

What do I need to think about?

How will you document your data?

Without proper documentation of metadata, others who attempt to use your data, even inter-departmentally, may have difficulty understanding it, or replicating your results. Essentially a record of all aspects of your dataset, changes to the data including pre-processing, and steps taken during analysis should be clearly documented.

Making a Data Management Plan (DMP)

What do I need to think about?

How will you ensure quality of the data?

- What tasks will be performed during the quality control process and who will perform them?
- Will programs be used to find duplicate files or formatting consistencies in data fields (for example, when one data entry field reads "34 cm" and another is entered as "48,0 centimeters")?
- Ensuring data quality also involves version control, or measures to indicate when changes were made to files and by whom. This is especially important when several people will work with the data.

Making a Data Management Plan (DMP)

What do I need to think about?

Who owns the data?

You will need to know who the true owner of the data is to determine who makes access and storage decisions. This also determines what can be done after the project is complete and by whom.

Plans should also be put in place for what happens to the data if you leave the institution. Who will be the contact person for matters pertaining to data sharing?

Making a Data Management Plan (DMP)

Last but not least!

What licenses will be applied for re-use?

A variety of licenses that can be applied to the sharing and use of your data exist. These state who may use the data, for what purpose, and how they must obtain access.

Examples:

- Creative Commons Public Domain Dedication (CC-0)
- Open Data Commons Public Domain Dedication and License (PDDL)
- Creative Commons Attribution-NoDerivatives 4.0 International (CC BY-ND)

Best Practices

A primer on the bare minimum!

Best Practices

File names

Include metadata right in the file names. Example: "Sub01_Interview2_jan2019.txt"

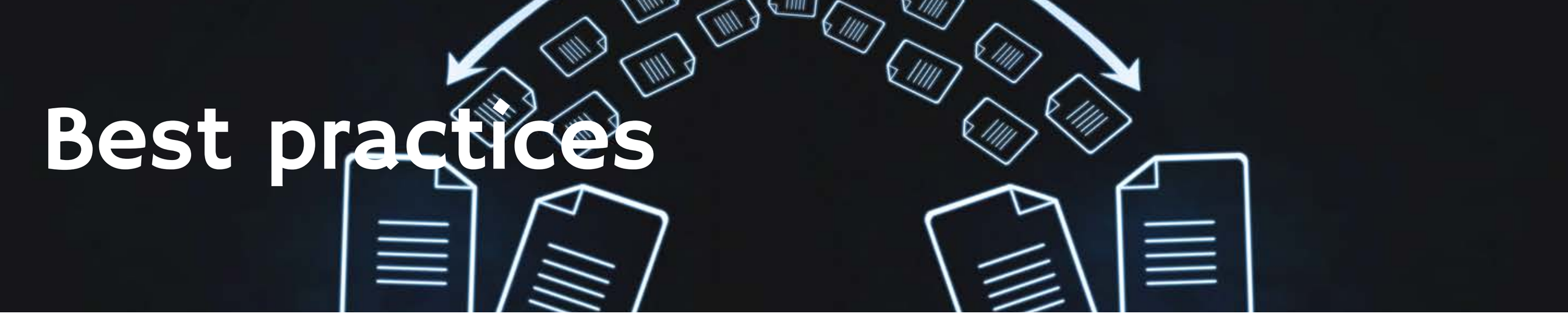
Automate

Automate mass name changes, quality assurance, even analysis if you can! Coding is skill that is in high demand, by the way.

File types

Use non-proprietary file types when possible. This assures longevity of your data.

Best practices



Data degradation

That .zip file you made 20 years ago isn't going to last forever.

File storage

File conventions and storage changes over time. Files and storage may need to be updated over time (especially with longitudinal studies).

Programming Languages

If you are using codes for analysiss or data handling try to use open-source languages like Python, R, etc.

Data Degradation

Avoid 'bit rot' and keep your data healthy over time.

- Avoid reading CDs of older data sets on old optical drives (they have a greater tendency to damage the disc than newer ones).
- Always eject flash drives correctly prior to removing them from the computer.
- Avoid leaving files open for long periods.
- Save new versions of files once another has been written over multiple times and document when it was created.
- Save older data in archive files (.rar, .zip) using non-solid compression (the default). This allows you to right click and check for damage using 'Check Archive' and then repair if necessary.
- Perform backups on two separate forms of media. Copy files to a new storage medium every 2-5 years, as the disc/drive itself can also degrade over time.

Anonymization Basics

1 Make a Data Key

Always keep the data key, which connects the participants to the ID linking to their data, stored separately from the data itself. Only allow access to those who have a true need to access.

2 Date of Collection

Shift dates for participation so that participants cannot be identified based on the date of the session(s).

3 Avoid re-linkage

Generalize data such as age, birthdate, location, and other indirect identifiers when possible. Re-linkage occurs when metadata (name, age, or gender) can be used to identify the participant along with other data you collected.

4 Mind the file headers

Anonymize headers to files that contain sensitive information embedded in them as a default at the source. Some file types (medical files for example) inherently contain sensitive data.

Vulnerable Populations

Anonymization with small groups of participants with distinguishable characteristics proves very difficult. Have a plan in place from the start!

Recommended Resources

BRINEY, KRISTIN. (2015).

Data Management for Researchers: Organize, maintain and share your data for research success. First edition. Exeter: Pelagic Publishing.

NORWEGIAN CENTER FOR RESEARCH DATA. (N.D.)

NSD *Data Management Plan.* https://nsd.no/arkivering/en/data_management_plan.html

NSD

The Norwegian National Research Ethics Committees. The Research Ethics Library. https://www.etikkom.no/en/library/

DATA ONE

DataOne's best practices for the creation of a data management plan https://www.dataone.org/best-practices/plan-data-management-early-your-project

MIT LIBRARIES

Data management and publishing: Organize your files. http://libraries.mit.edu/data-management/store/organize

LANGTVEDT, N.J. (2015)

Protection of privacy. The Norwegian National Research Ethics Committees. https://www.etikkom.no/en/library/topics/data-protection-and-responsibility-concerning-the-individual/protection-of-privacy/