Introduction to FAIR and Data Management Plans - Day 1

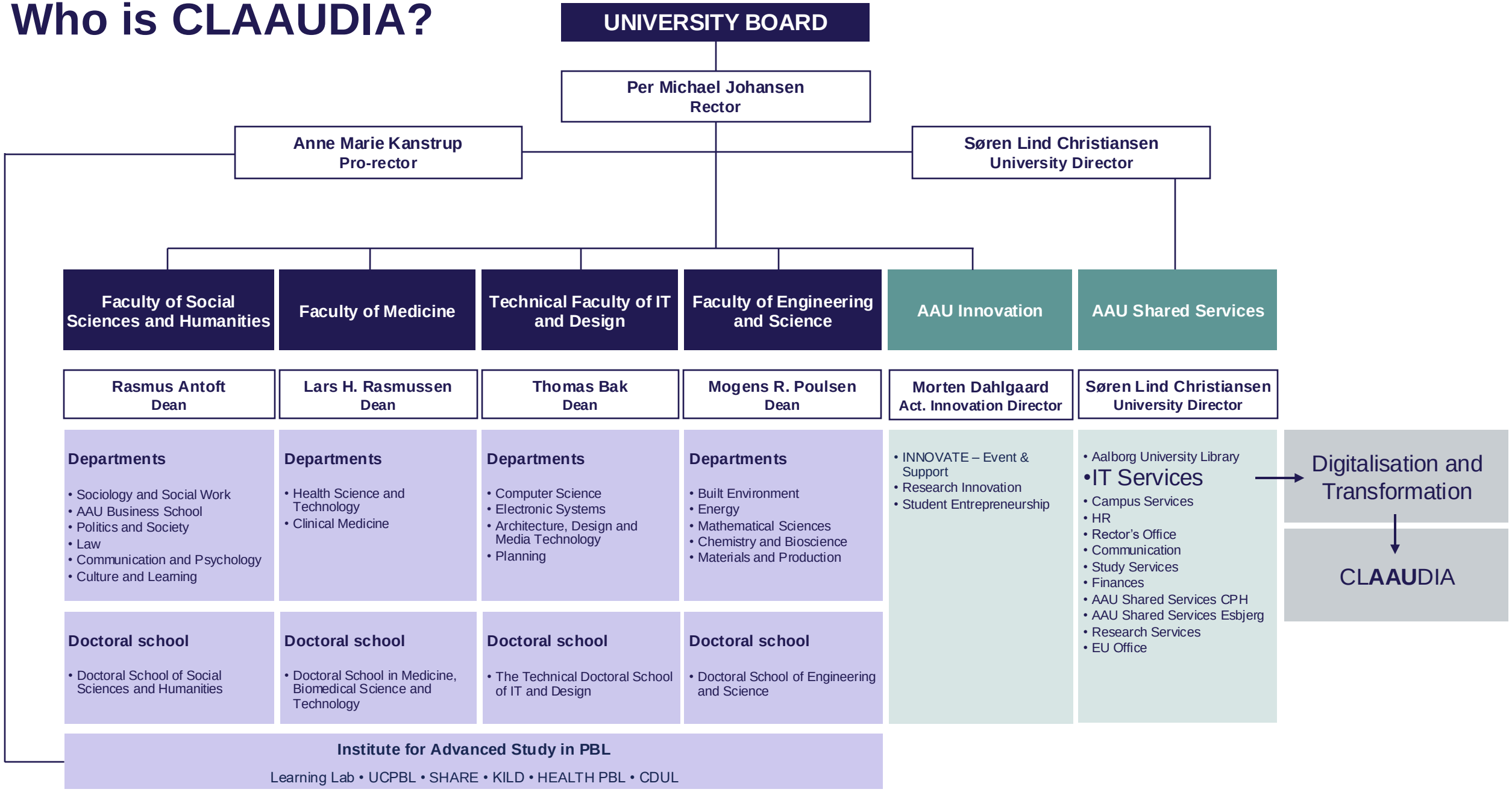
AALBORG UNIVERSITY

CLAAUDIA
RESEARCH DATA SERVICES

Agenda

9:00-9:45	Introduction and presentations Introduction to FAIR and Data Management
9:45-10:15	Exercise with LEGO 1 (spisestue/dining area)
10:15-10:30	Break
10:30-11:45	Introduction to Data Mangement Plans Question 1-4 (back in A.0.03)
11:45-12:15	Lunch
12:15-13:30	Introduction to Data Mangement Plans Question 5-8
13:30-13:45	Break
13:45-14:45	Exercise with LEGO 2
14:45-15:00	Final points and introduction to homework

Who is CLAAUDIA?



CLAAUDIA RESEARCH DATA SERVICES

IT services → Digitalisation and Transformation → CLAAUDIA

Jeanette Dahl-Sørensen
Team Leader

Gergely István Barsi
Data Scientist

Pelle Rosenbeck Gøeg
Data Scientist

Frederik Pertri Svenningsen
Data Scientist

Robert Smith
Data Scientist

Sighvatur Sveinn Davidsson
Data Scientist

Mathias Overgaard Hedegaard
Data Scientist

Nikolaj Andersen
DevOps Engineer

Carina Ollerup Christensen
Data Manager

Thomas Andersen
Data Steward
Primary @ TECH

Freya Vamberg Delfs
Data Steward
Primary @ SSH

Kamilla Hall Kragelund
Data Steward,
Primary @ SUND

Dennis Aagaard Pedersen
Data Steward
Primary @ ENG

Dagmar Knudsen Fallesen
Digital Adoption Consultant

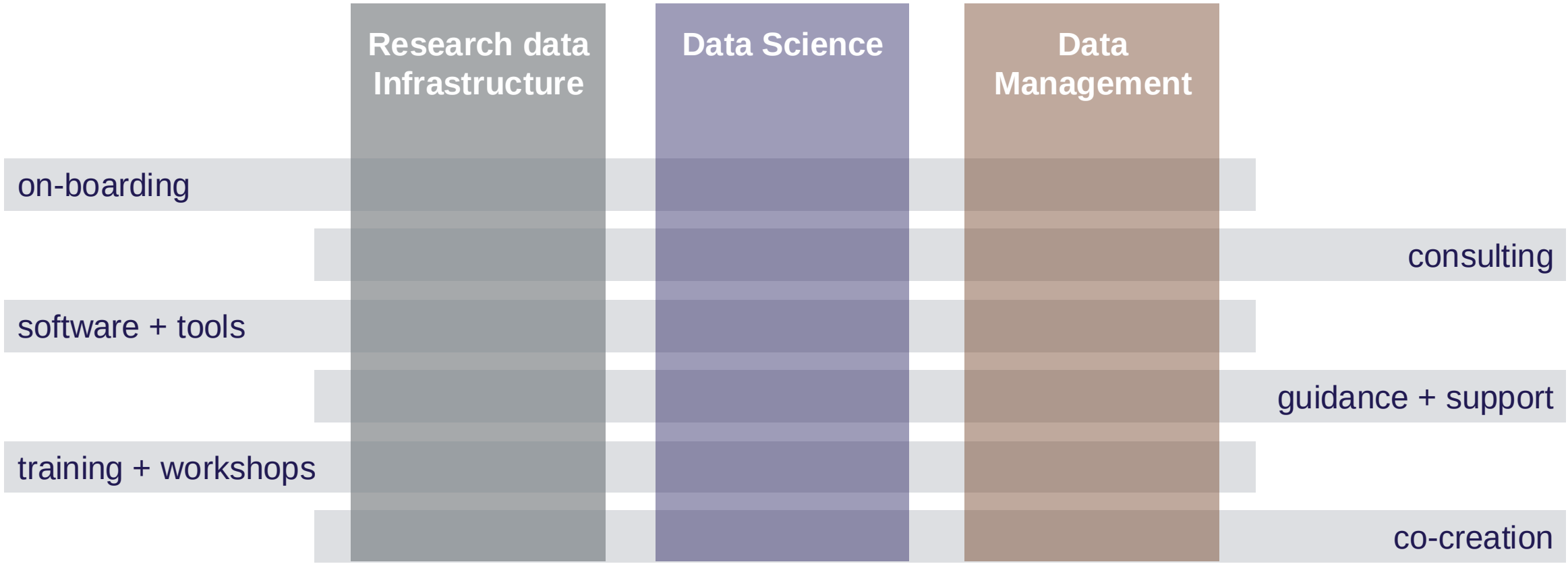
Nina Simone Marstrand Sander Aagaard
Chief Consultant

Rainer Bohm
Enterprise Architect

Rasmus D. Jensen
Digitalization Partner

CLAAUDIA RESEARCH DATA SERVICES

Collaboration across 3 main areas.



CLAAUDIA SERVICES

CONSULTANCY

PLATFORMS, SCIENCE, MANAGEMENT

- Choice of cloud platform
- Access + onboarding to HPC platforms
- Applications for platforms
- Data science methods
- Technical specifications
- Data storage, collaboration and sharing
- FAIR principles
- Data management
- Data management plans
- Funding applications (technical requirements and data management)

SUPPORT AND DEVELOPMENT

PLATFORMS

- Strato
- AI Cloud
- AI Lab
- UCloud
- Computerome
- Sophia
- Genome DK
- LUMI

APPLICATIONS

- DataDeposit
- Deic DMP
- Nvidia Jetson Nano Kit
- REDCap
- Transcriber
- Sundhedsdatanettet

COURSES AND WORKSHOPS

INTRODUCTION TO

- Strato
- AI Cloud
- AI Lab
- Ucloud

INTRODUCTION AND WORKSHOP

- FAIR Data Management and CLAAUDIA
- Python

Ph.D. SCHOOL

- **Data Management and the FAIR principles**
- Scientific software development

Presentation round

- Name
- Research area
- Experience with
  - The FAIR-principles?
  - Data Management Plans?

AALBORG UNIVERSITY

What is Data Management?

It is the practice of how you collect, keep, and use your data.

So, what is data?

Anything collectable that holds information you can use to analyse something or make decisions from.

Examples: Measurements, statistics, transcriptions, code, recorded sound, pictures …

Managing data is about:

- Being aware of the proces your data must go through.
- Being aware of what your data is, how to use it, and potential standards it should meet.
- Being aware if the content or nature of it demands specific attention because of policies or law.
- Making informed choices and documenting them.

How you do this, will determine the quality, efficiency, and impact of your data.

What is Data Management?

It sounds important, but is it worth all the work?

Yes it is:

1. Notable benefits to your research proces and outcome quality because you make **deliberate and informed choices** about your data.

2. Data management is **the essential practical part** of working with the FAIR-principles that are made to get more use and value from your data.

3. Data management based on the FAIR-principles is already part of international and national strategies for research data and funding. →

Research data strategies

FAIR-principles

Not new ideas, but the unifying work began in *2014.*

The FAIR Guiding Principles for scientific data management and stewardship was published in 2016.

EU

Public Sector Information Directive (PSI 2018):

All publicly funded research data should be openly available and compatible with the FAIR principles.

National

Strategy for National Coorporation on Digital Research Infrastructure in 2018.

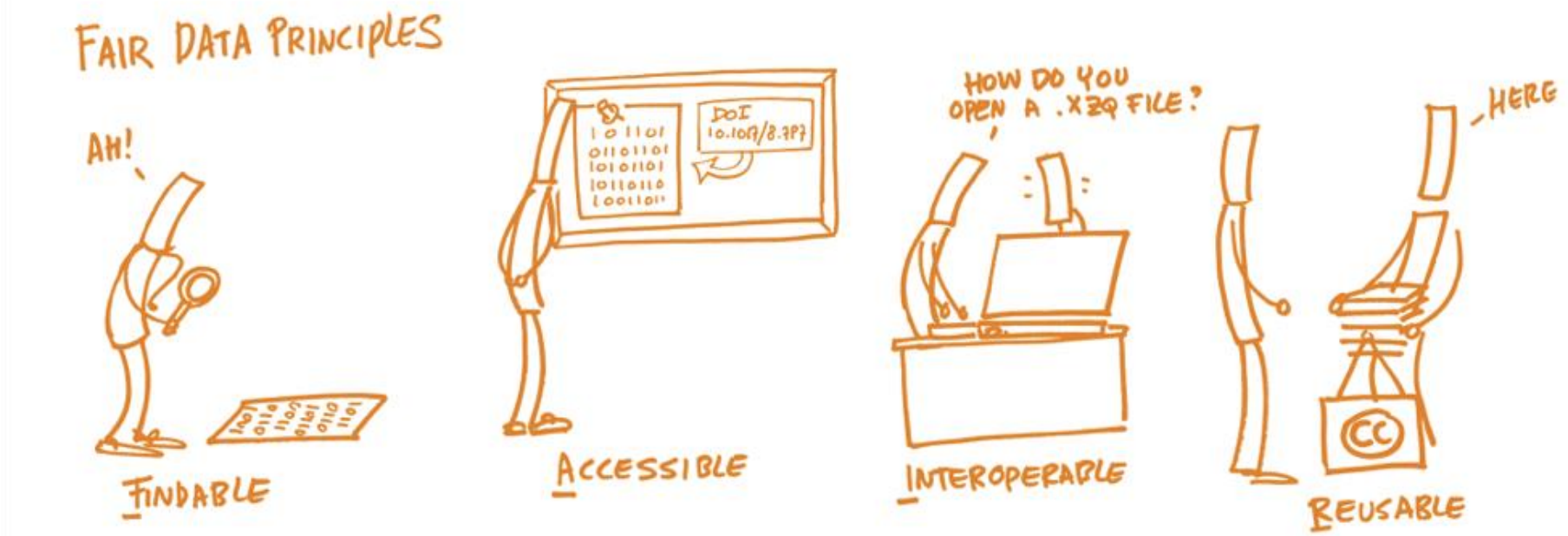
Strategy for Data Management based on the FAIR principles by DeiC in 2021.

AAU

Policy for research data:

4.3. The management of research data takes place in accordance with the intentions of the FAIR principles … FAIR thinking applies to the entire research process, including choice of scientific method, data formats, documentation etc.

The FAIR principles



- Each letter holds a range of principles.
- The principles are guidelines for good data management practice.
- All open data must be FAIR, but not all FAIR data must be open. The FAIR saying: *"As open as possible and as closed as necessary!"*
- Not either-or. FAIR is a continuum, and the goal is be as FAIR as you can.

Dette billede efter Ukendt forfatter er licenseret under CC BY

Findable

Humans and machines can find out the data exits.

Because you:

- Publish searchable metadata.
- Assign a unique persistent identifier.

The FAIR Guiding Principles for scientific data management and stewardship says:

F1: Data, and/or metadata, assigned a globally unique and persistent identifier

F2: Data are described with rich metadata

F3: Metadata clearly and explicitly include the identifier of the data they describe

F4: Data, and/or metadata, are registered or indexed in a searchable resource

Dette billede efter Ukendt forfatter er licenseret under CC BY

Accessible

Humans and machines can find out how to get access to the data.

Because you:

- Upload to public data repository.
- Define access conditions for data and metadata.

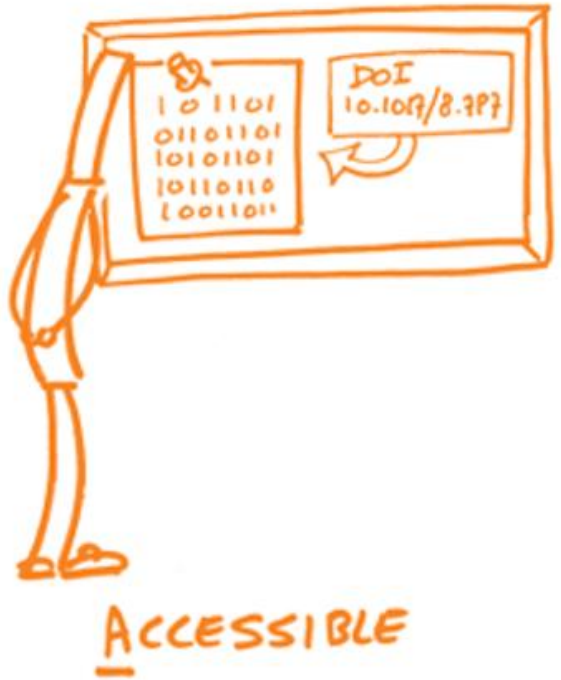
The FAIR Guiding Principles for scientific data management and stewardship says:

A1. Data, and/or metadata, are retrievable by their identifier using a standardised communications protocol

A1.1 The protocol is open, free, and universally implementable

A1.2 The protocol allows for an authentication and authorization procedure, where necessary

A2. Metadata are accessible, even when the data are no longer available



Dette billede efter Ukendt forfatter er licenseret under CC BY

Interoperable

Humans (and machines) can open the data and work with it.

Because you:

- Use open file formats.
- Use community standards, keywords and ontologies.

The FAIR Guiding Principles for scientific data management and stewardship says:

I1. Data, and/or metadata, use a formal, accessible, shared, and broadly applicable language for knowledge representation.

I2. Data, and/or metadata, use vocabularies that follow FAIR principles

I3. Data, and/or metadata, include qualified references to other Data, and/or metadata

Dette billede efter Ukendt forfatter er licenseret under CC BY

Reusable

Humans and machines understand how the data was created and how to reuse it.

Because you:

- Attach sufficient documentation.
- Add a usage license.

The FAIR Guiding Principles for scientific data management and stewardship says:

R1. Data, and/or metadata, are richly described with a plurality of accurate and relevant attributes

R1.1. Data, and/or metadata, are released with a clear and accessible data usage license

R1.2. Data, and/or metadata, are associated with detailed provenance

R1.3. Data, and/or metadata, meet domain-relevant community standards

Dette billede efter Ukendt forfatter er licenseret under CC BY

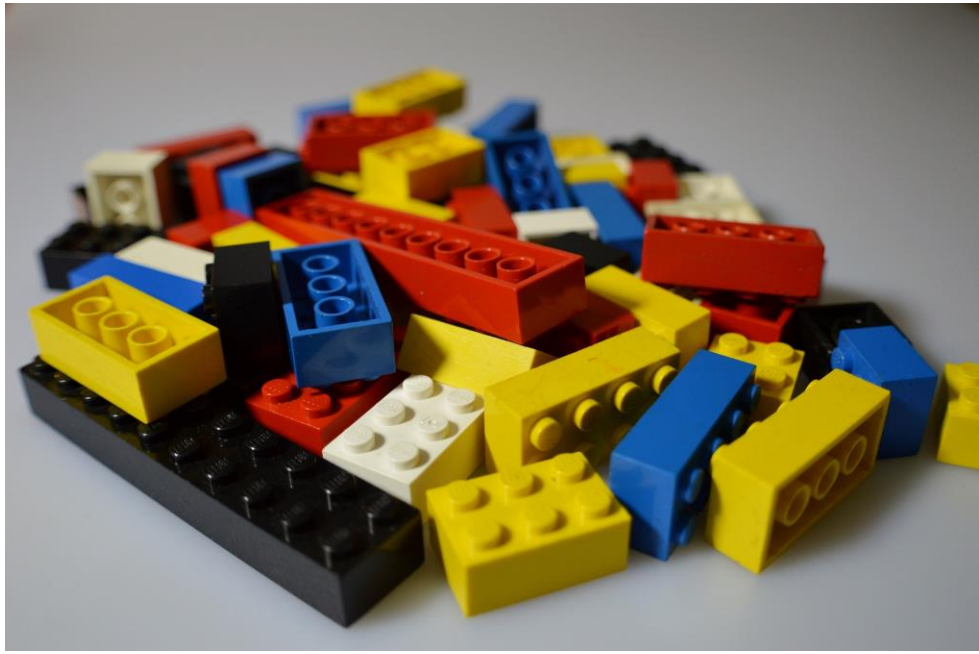
To sum up

FAIR: Framework for research data

CLAAUDIA is here to help you!

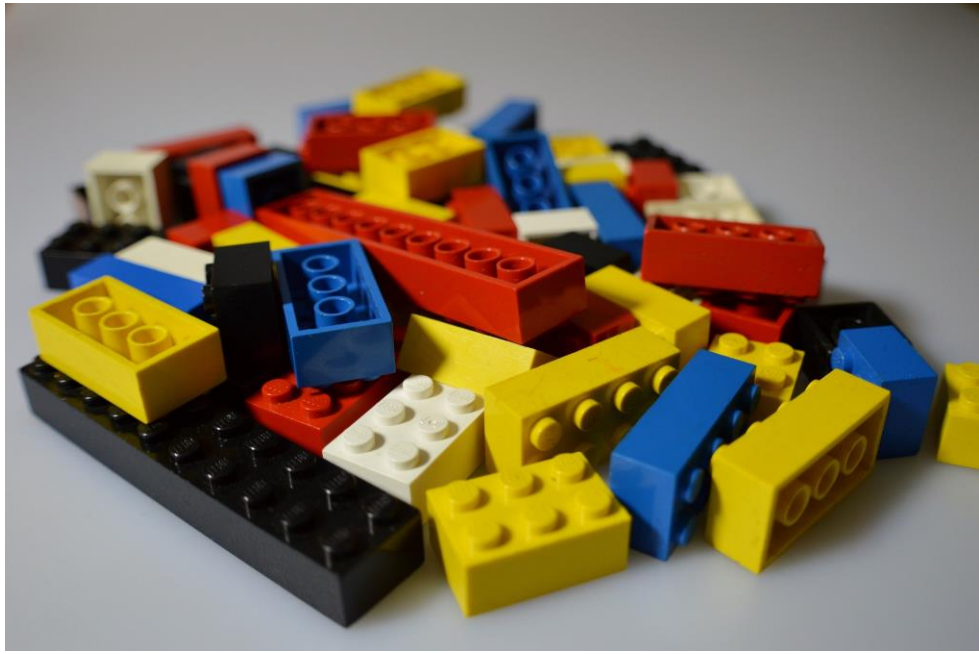
LEGO exercise (part 1)

- Pick 3-5 LEGO bricks.
- Describe your bricks in writing so that others would be able to identify them.
- Build a figure with your chosen bricks.
- Describe the figure in writing so that others would be able to build it.
- Disassemble your figure and place the bricks back into the plastic bag. (keep the plastic bag with Lego bricks)



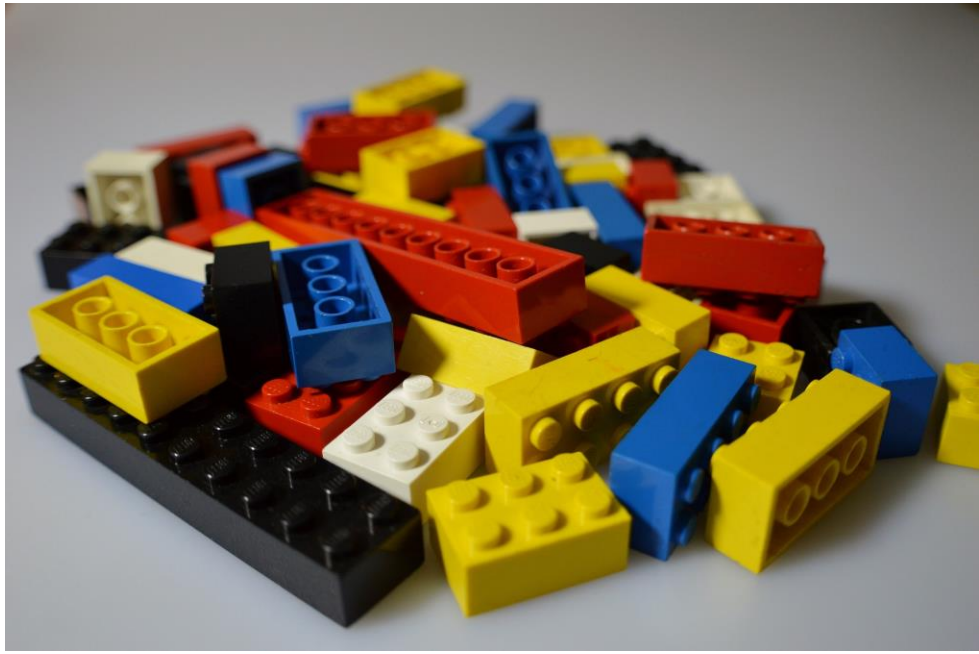
LEGO exercise (part 2)

- Swap descriptions with another group (but keep your bag with Lego bricks).
- Read the description of bricks and try to find them.
- Read the description of the figure and try to build it.
- Present the figure to its original creators.



LEGO exercise (part 3)

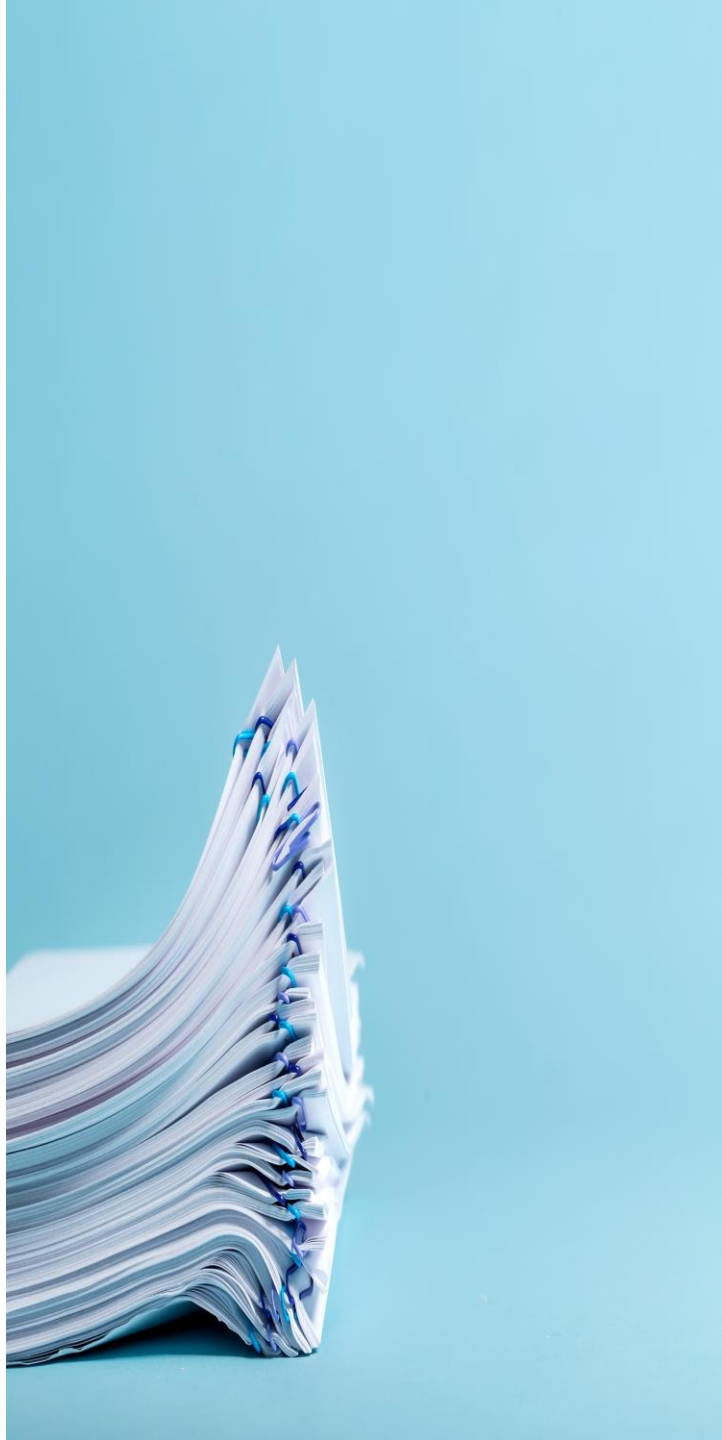
- Does the figure look like the original?
- Did you find the description understandable?
- Did you experience any challenges?

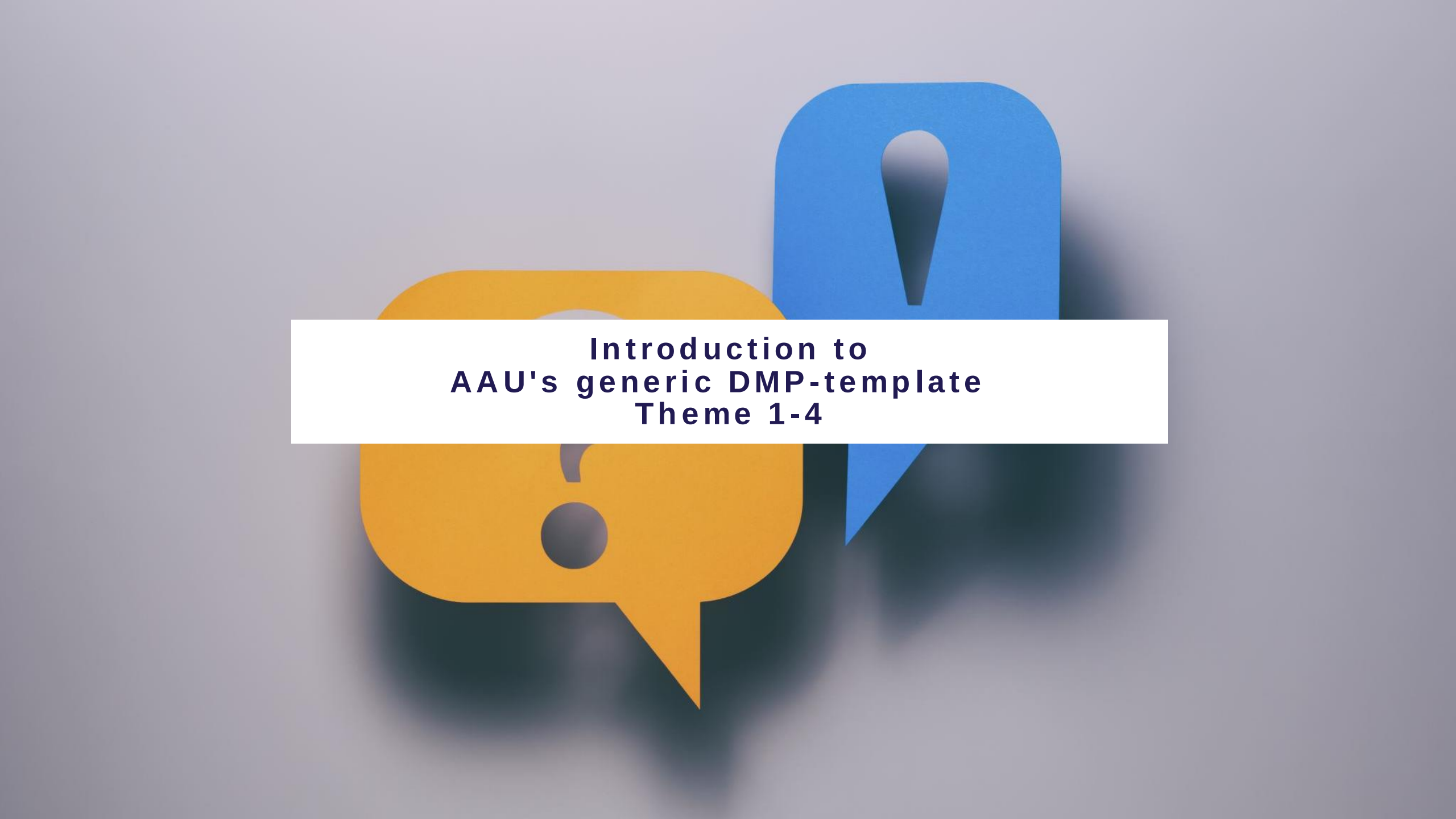


Introduction to Data Management Plans

What is a DMP?

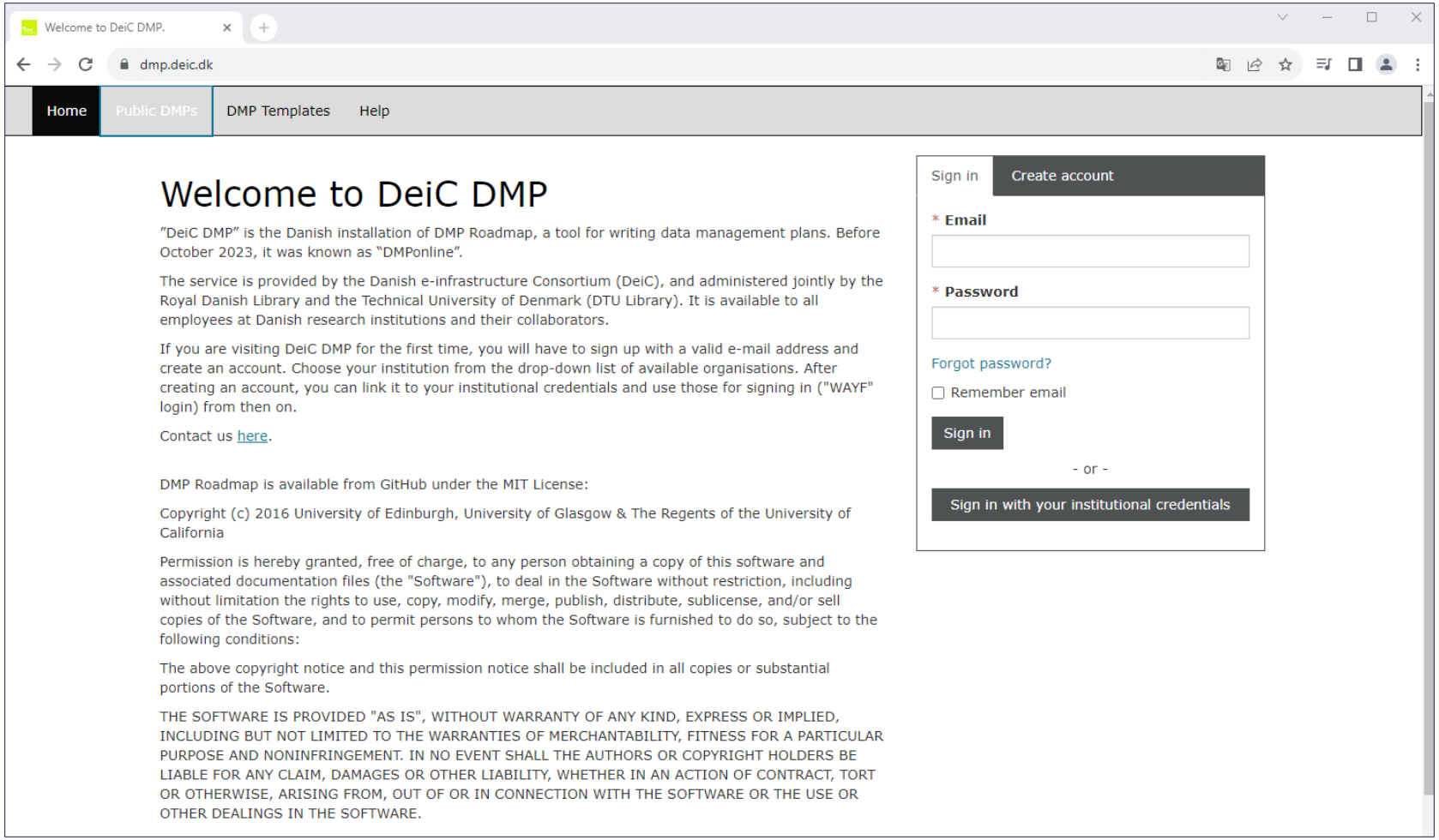
- A data management plan (DMP) is a document about **how** you handle your data.
- A DMP will help you make **informed choices** to prevent future obstacles. It is a reflection tool.
- A living document – your DMP **will change and grow** with your knowledge and progress.
- A way to **structure** data management with relevant topics, alignment and responsibilities that can increase the quality of your research and be timesaving in the long run.





Introduction to AAU's generic DMP-template Theme 1-4

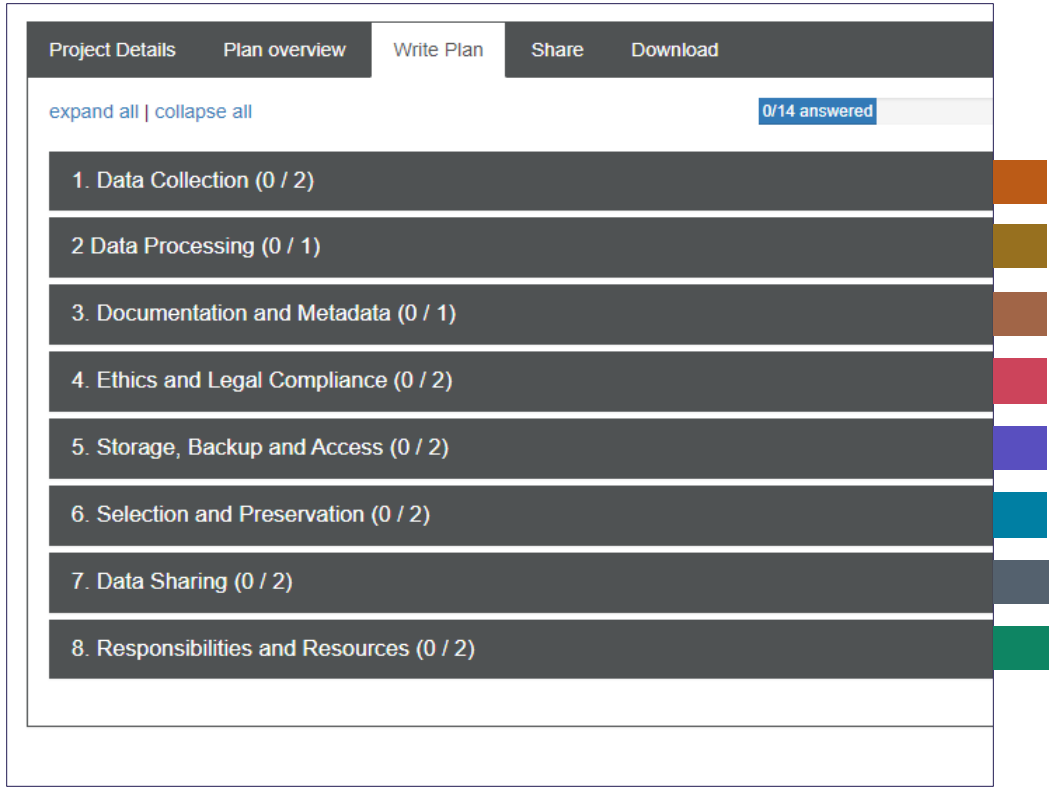
Find the AAU DMP-template at www.dmponline.deic.dk



AAU's generic DMP-template

Content of the generic AAU template:

- The template has 8 themes.
- Each theme consists of 1-2 questions, and they are supported by a guiding text and a range of questions to consider and/or to be inspired by.
- Again, the idea is to let your plan grow with your progress.



1. Data Collection

- Describe the main characteristics of the data you work with or the data you imagine working with.
- The rest of the 7 themes in the DMP is about how you handle these data throughout the project.
- You don’t have to start with question 1.
- Research data can be very different depending on what research area you are working with. We work with this description:

> Research data refer to all the information, documentation and materials that forms the basis of the research output.
>
> Research data include all materials created in the course of academic work. It could be old records, measurements, surveys, interviews, code, and much more.

Dette billede efter Ukendt forfatter er licenseret under CC BY-SA-NC

1. Data Collection

- What are the main characteristics of your data?

- Data type: What type of data are you working with?
- Data format: What format(s) will your data be in and why?
- Data volume: What is the expected volume of data (in MB/GB or TB)?

Data type and format:

- Text documents (doc, txt, odf, etc.)
- Audio (mp3, wav, aiff, etc.)
- Video (mp4, avi, wmv, etc.)
- Graphics/images (jpeg, png, gif, etc.)
- Spreadsheets (spss, stata, csv, xls, etc.)
- Structured text (html, json, xml, etc.)
- Databases (MySQL, Oracle, etc)
- Software code (JavaScript, Java, etc.)
- Physical samples (prototype, etc.)
- Experiments
- Simulations
- Sensor data

1. Data Collection

- How will you collect or reuse existing data?

What data are you working with and where did it come from?

- New collected data:
  - How is the data collected? Is special software required?
  - How does the new data interact with existing data?

- Existing data:
  - Can existing data be used? Where does the data come from, and how was it collected?
  - How is the existing data used in the research project?

1. Data Collection

- How will you collect or reuse existing data?

Potential limitations when working with existing data:

- **Usage rights:** You must comply with the terms of the usage rights given by the data license such as Creative Commons.
- **Contract substance:** Some companies want to keep data confidential. So, you need to be aware of how the data must be handled. This is typically described in the contracts you enter.
- **Personal data:** Datasets that include personal data are likely to have restrictions on how results from the dataset are reported and what happens to the data when the project is complete.

There could also be other things to be aware of. It all depends on the type of data that you are working with and the possible contracts or agreements you have made to get the data.

1. Data Collection

- Example answer – data description

The most relevant data and research outputs expected in the project include:

- *Experimental data, test results (csv file), (WP1).*
- *Models e.g., SW files, plots, animations and drawing, test results (csv file), reports (Latex code or Word file), (WP2).*
- *Experimental data, computational data, images, text, numerical data from simulations (WP3).*
- *Plant drawings, schemes, block diagrams (WP4).*
- *Stakeholders attitude towards hydrogen survey/interview data, Technology and Market intelligence elaborated based on a mix of public data and conducted interviews (WP5).*

1. Data Collection

- Example answer- data volume

The expected volume of the data is presently unknown, but it is foreseen that it will stay small enough to be stored in the standard available servers used in the project.

There is no indication of a need to budget for repository storage outside of the scope of what is accepted by repositories (in the case of this project; Zenodo) within their free limit quotas.

1. Data Collection

- Example answer- data overview

A preliminary overview on the data types expected to be used in the project is provided in table 1.

WP partner	Type of data	Short description	How created/source	Format	Label
WP 1	Algorithms for model sampling and formulation of data-driven models	Algorithm sheets and respective source code of data-driven modelling applications (sampling on rigorous models, training, and testing).	Existing data Source: www.xxxxxx.com	Executable code; .py Graphical schemes; .pdf	O
WP 2	CFD simulation results of the lab-scale reactor, System level modelling of ammonia synthesis process	Visible output data from CFD simulation of the lab-scale NH3 synthesis reactor, including pictures and animations, Visible data from 1D models which include pictures, and data generated from simulations.	Done by WP2 in system; XXX	Simulations: .txt, .CSV, Mat and .xlsx file formats. Pictures: JPEG/PNG, animations in MPEG	O
WP 5	Demonstrator installation	Documents related to the installation of the demonstrator and connection to the media. Plot plans, demonstrator drawings, BFD, PFD, P&ID.	Done by WP5	Preferred file format; .pdf	C
Wp 5	Lab scale reactor testing for ammonia synthesis	Output data from reactor testing for ammonia synthesis with novel catalysts and sorbent materials.	Existing data from company; XXX	Tabular data in CSV/XLSX files	L

1. Data Collection

- Questions

- What type of data was collected?
- What is the volume of data?
- What file format are the data and does it facilitate easy reuse?

2. Data Processing

Describe how data is handled in the analytical phase.

How will you organize your data?

- Naming and structuring of files and folders.
- Version control.

What are your software and hardware requirements?

- Generic or specialized?

How will you ensure high quality data?

- Methods for quality control?

2. Data Processing

What are the benefits of planning and describing the way you process your data?

Clear structures and folder systems give you

- a reliable way to navigate your data.
- clarity on where to keep specific parts or categories.

Consistent naming conventions will make it

- possible to identify the content of files and folders.
- easier to collaborate and enter new people.

Version control helps you

- manage changes and who made them.
- not to work on old or wrong documents.

Documented software/hardware requirements can help

- to align collaboration.
- making necessary acquirements visible.

A useable method for quality control is

- how you proof of the consistency and quality of your data.
- a way to align the quality of multiple sources.

2. Data Processing

How to organise and create naming conventions and version control.

What makes sense is very subjective so think of common denominators that are relevant and easy to understand.

Consider how you can

- design a structure that is intuitive and easy to navigate.
- name files and folders consistently.
  - with description of work, location, or analysis.
  - and short but descriptive.
- use capitals and underscores instead of periods or spaces or slashes.
- include a version number.
- if relevant use a date format such as YYYYMMDD.
- write it down so it can be understood and followed by everyone in the project.

2. Data Processing

Examples

Naming convention:

YYYYMMDD_Image_Modification

Filenames:

- 20130421_tina_original.tiff
- 20130421_tina_cropped.jpeg
- 20130421_tina_mustache.jpeg

Naming convention:

LocationAnalysisVersion

Filenames:

- CarnegieLakeWordCloudV1
- CarnegieLakeMapV1
- CarnegieLakeMapV2

No naming convention:

- Manuscript1
- Manuscript1finalv1
- Manuscript1FINAL
- Manuscript1FINALv2
- Manuscript1_Really_FINAL

No naming convention:

- 01042022interview
- 01042022interview1
- 20220104interview2
- 20220104interv_v2
- 01042022interview output

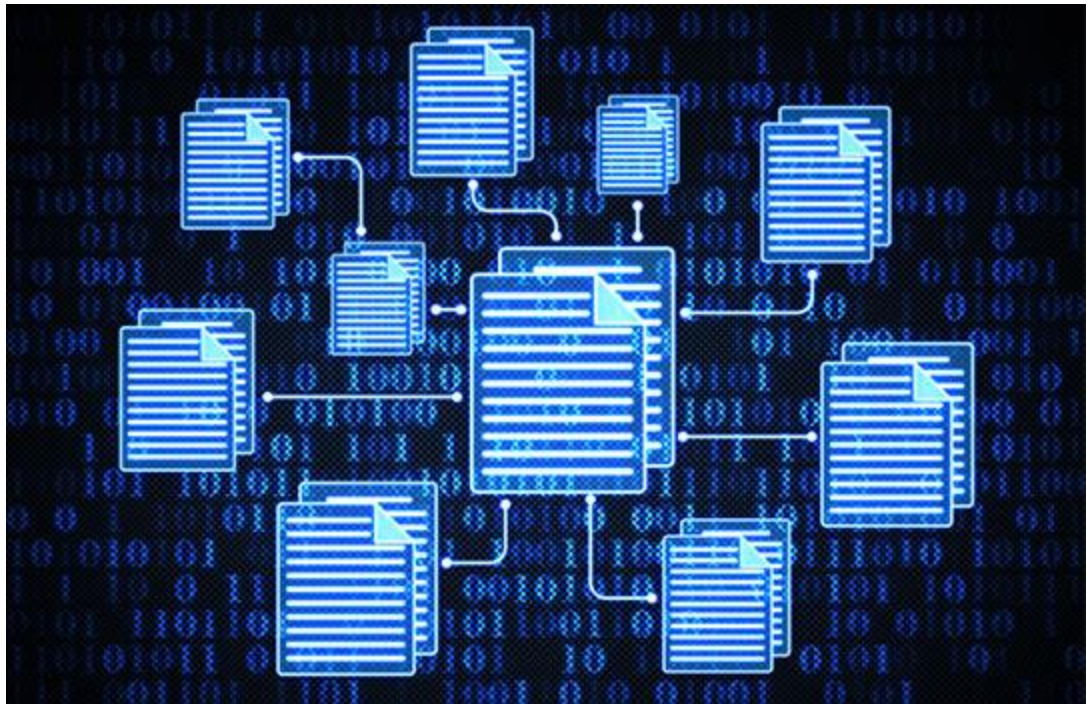
2. Data Processing

Software and hardware requirements

- Is there a need to align the use of certain software?
- Do you have a need for specialized software or hardware to make your analysis?
- Specifying even generic software may save you from having to convert files later.

Methods for data quality control

- Use of vocabulary specific for your field.
- Repeat samples or measurements.
- Data entry validation.
- Peer review of data.



2. Data Processing

- Example answers

The Embedded Computing Systems Group has an in-house GitLab platform to handle proper planning, management, versioning, monitoring of and collaboration on different R&D efforts. We will use this platform, where for each work package, we will create separate repositories.

OSF provides an automated version control, logging by whom and when changes were made, and storing the previous versions. Therefore, files with the same type of content should not include version numbers, dates, user initials, words as final, first , draft , etc. Instead, they should be named exactly the same as the previous version of the file.

The structure of the OSF storage is as following: Project -> Component -> Folder -> File.

The respective work package leader will handle the structure and versioning of the research data. For the whole project a standardized folder structure and naming conventions will be in place.

File naming is done according to the following standard: [focus group]_[location]_[YYYYMMDD].mp4.

2. Data Processing

- Example answers

Data quality checks will be done, e.g. checks of consistency of labels, logical errors in the data, data curation, and version control.

Source code review and refactoring will be done.

A document describing the dataset will be generated to supper the interpretation and use. Standard vocabulary will be used for all data types present in the dataset to allow inter-disciplinary interoperability. In addition, the documentation will include a general glossary used to share information about the vocabulary and general methodologies employed for the generation of the dataset.

A far as possible, we will use controlled vocabularies for our data to allow interdisciplinary interoperability and machine-actionability.

2. Data Processing

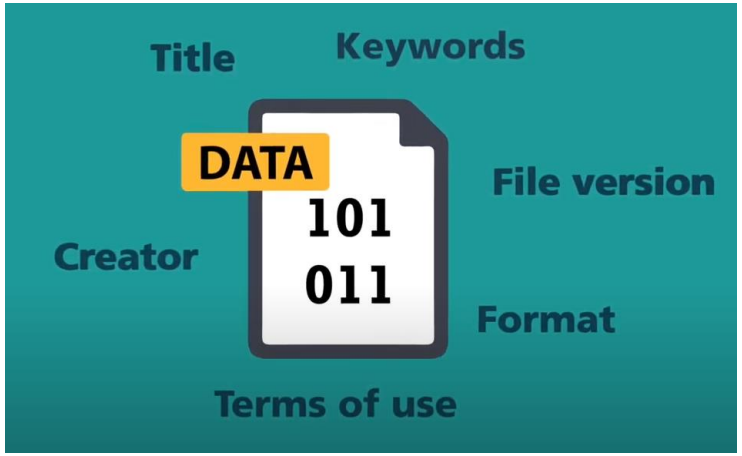
- Questions

- Quality control
  - Are you able to find the questions that were asked in the survey and interview?
- Version control
  - Is it possible to identify their use of version control?

3. Documentation and metadata

Metadata = data about data

- A way to provide and structure information that describes, explains and locates data.
- Essential to identify, discover, and reuse data = one of the main tools to make FAIR data.



3. Documentation and metadata

- What documentation and metadata will accompany the data?

Metadata standards

- Some research disciplines have a common agreed upon way to describe metadata.
- An easy way to provide clear and relevant metadata.
- This will increase the data's interoperability and reusability.

Find metadata standards here:

- The Digital Curation Centre (DCC)
- www.re3data.org (filter: metadata standards).

3. Documentation and metadata

- What documentation and metadata will accompany the data?

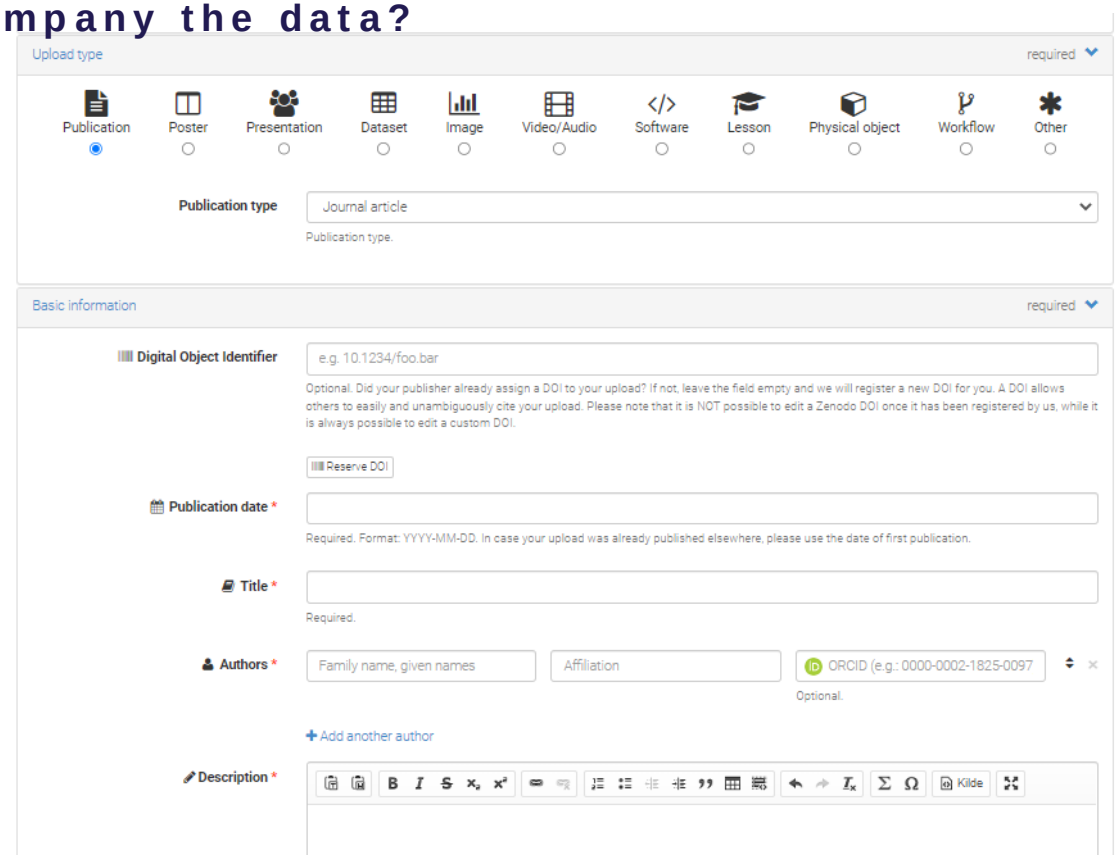
Is there a metadata standard within your research area?

If yes:

- Find a repository that supports the metadata standard.

If no:

- Use a generic repository and provide as much information as possible when uploading data.



Extract from Zenodo's metadata schema. Zenodo is a generic repository.

3. Documentation and metadata

- Example answers

There are no domain specific metadata standards applicable. We will provide a README file with an explanation of all values and terms used next to a file with data. Additionally, we will provide common metadata such as title, description, or keywords when publishing data in open access repositories. In such a case, we will follow the default template provided by the repository, such as Data Cite Metadata or Dublin Core.

We will include README-files at file-level, dataset-level, and project-level. The files will contain data inventories and important contextual information such as the software used to collect/process the data and any assumptions made during analysis. Keywords will be assigned according to the subject-specific research areas.

Each dataset has metadata associated to it, for descriptive, structural and administrative issues. The embedded metadata in data will be part of the data collection and documentation process, which are elaborated before the publication of the metadata, both in terms of sanitizing and for adding metadata. Some metadata will be created during the publication procedure to a given repository and follow the (metadata) standard(s), for the specific repository. As for the specific format of metadata, Zenodo utilizes DataCite which is a de facto standard for describing datasets.

3. Documentation and metadata

- Questions

1. Is there any documentation of data?
2. Number of respondents
3. How were the respondents in the survey selected?
4. In what geographical area where the data collected?

4. Ethics and Legal Compliance

When working with data, always consider if there are any special requirements or circumstances in relation to the data that you are going to work with.

In the DMP you describe if there are any requirements linked to data collection/use and lawful processing - and how you comply with these.

Remember to:

- Register your research project when working with personal data: Grants and contracts
- Self-screen if your research entails ethical aspects: The Research Ethics Committee

4. Ethics and Legal Compliance

- How will you manage ethical issues and GDPR?

In the DMP you consider whether ethical/GDPR issues can affect:

- How data are stored.
- How data are transferred.
- Who can see or use data.
- How long data are kept.

And which measures you take.

For clarification and help contact: AAU Ethics

Examples of ethical and GDPR issues:

- Studies or experiments that involve people (surveys, tests, interviews, focus groups etc.).
- Sequrity-related research, where there is a potential for misuse of research results.
- Materials, knowledge and technology with dual use potential.
- Environmental research that can pose a risk to people themselves and the environment.
- Research involving human samples (e.g., cells), laboratory animals or genetically modified organisms etc..

4. Ethics and Legal Compliance

- How will you manage copyright and intellectual Property rights (IPR)?

Things to consider

Multiparter projects

- Who has the rights to data?
- Who stores data and upload it to a repository?
- Who has the rights to access and grant access to data?

Use of existing data

- Usage rights?
- Special requirements about storage and sharing?

Access to data

- Any legal or ethical requirements about data access and sharing?
- Embargo periods?

Potential legal issues:
- Personal data (GDPR)
- Ethics approvals
- Data ownership (IPR)
- Publication restrictions
- Embargo periods
- Etc..

Please note: *Your answers to the legal issues will be very different depending on your project and your requirements.*

4. Ethics and Legal Compliance

- Example answers

Only project staff who sign a non-disclosure agreement will be granted access to the raw data. Before publication, the data will be anonymized to protect respondents' privacy according to the GDPR and XY national law.

Datasets referring to certain elements at risk, such as people and critical infrastructures, are not open by default as their publication may pose privacy, ethical or security risks.

Datasets listed as Limited or Closed are restricted since making those openly available is flagged against the beneficiaries' legitimate interests, as industrial partners expect commercial exploitation from these data in the project.

4. Ethics and Legal Compliance

- Example answers

Project partner(s) xxx will be the owner(s) of the data generated and have the rights to control access. Further details will be covered in the consortium agreement.

All project partners will jointly own the data generated and have the rights to control access. Further details will be covered in the consortium agreement.

Within our research group, we have agreed to the following IPR and licensing conditions:

- *Data owner: [name]*
- *Licenses: license CC-By 4.0 for datasets and license XY for software*
- *Restriction: for datasets AB and CD until [date]*
- *Embargos: for dataset XY until [date]*

4. Ethics and Legal Compliance

- Questions

1. Are there any ethical aspects to consider if you were to use the Malawi dataset?

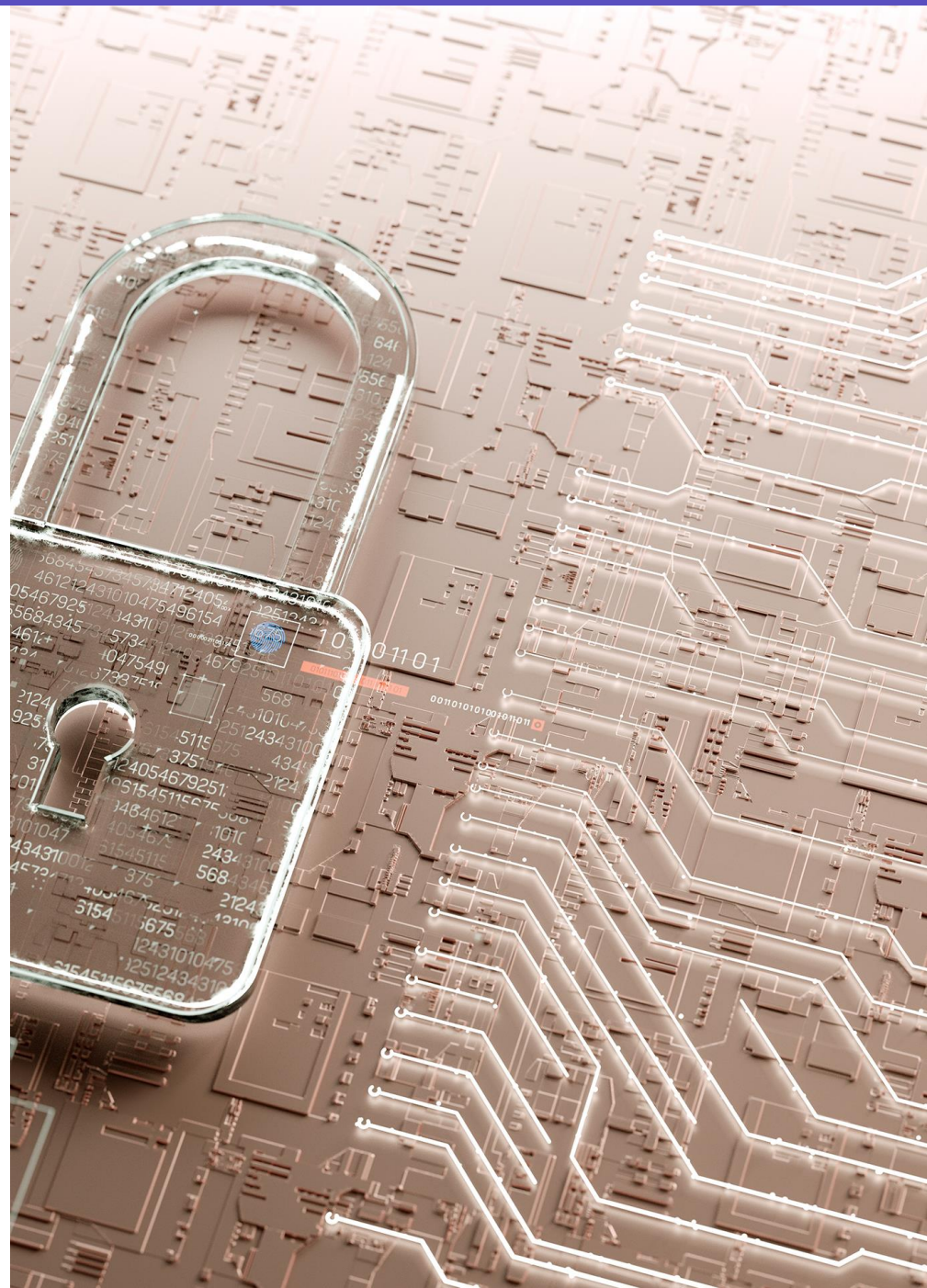
Introduction to Data Management Plans Theme 5-8

5. Storage, Backup and Access

Consider the types of data that will be created or used in the project. Will the project…

- Generate large amounts of data?
- Require platforms for collaboration with external/internal partners?
- Use data that has licensing agreements or other restrictions on its use?
- Involve human or animal subjects?

Or other things that you need to be aware of in relation to storage, back-up and access.



5. Storage, Backup and Access

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

AAU Data Classification model

- Find out what type your data is.
- Determine the security level.
- Check if the system or service is authorized for use with your data type.

Level 0: Public information	
Study descriptions	News articles
Books	AAU's website
Open data	Research reports
Person names	Job title
Work phone number	AAU email addresses

Level 1: Internal information	
AAU username	Nationality
Course descriptions	Birthdays
Salary information	Time of illness and absence
Pension information	Department budgets
Purchasing agreements	Educational material

Level 2: Confidential information	
CPR number	Private address
Private phone number	Private email address
Personality tests	Alcohol and narcotics tests
Registered exam cheating	Grades
Family relations	Drivers license photo

Level 3: Sensitive information	
Inventions valued over DKK 5 million	Research valued over DKK 5 million
Race and ethnicity	Political beliefs
Religion	Health information
Sexual relations	Sexual orientation
Philosophical beliefs	Union

	Dataklassifikation			
System / service	Niveau 0 Offentlig	Niveau 1 Intern	Niveau 2 Fortrolig	Niveau 3 Følsom
Fildrev – Fællesdrev Fileshare	Ja	Ja	Ja	Ja
Workzone	Ja	Ja	Ja	Ja
UCloud	Ja	Ja	Ja	Ja
Email (Outlook eller mail.aau.dk)	Ja	Ja	Ja	Ja
REDcap (kun administreret af ITS)	Ja	Ja	Ja	Ja
Filesender (DeiC Service)	Ja	Ja	Nej	Nej
Git-svn (kun administreret af ITS)	Ja	Ja	Nej	Nej
Microsoft Office 365	Ja	Ja	Nej	Nej
Sciencedata.dk	Ja	Nej	Nej	Nej
Andre cloud-services (f.eks. Google Drive)	Ja	Nej	Nej	Nej

5. Storage, Backup and Access

-How will you manage access and security?

Map your dataflow and consider if there is anything you need to be aware of in relation to access and security.

- What type of data do you work with?
- What is the data's provenance?
- Where do you store the data?
- Do you share the data with anyone? Internally at the university or with external partners? How do you share the data, which systems do you use?
- Are you sending the data anywhere or to someone? How and which systems do you use?
- Who has access to the data?

5. Storage, Backup and Access

-Example answers

During the project, the data is stored on a Fileshare, a central and redundant network drive, protected by password with daily backups and regular snapshots provided by AAU. Only authorized staff members and project partners have access.

Each partner is required to follow the data security standards from their own institution, including guidelines for information security. This will also include local guidelines for backup and restore procedures, as well as for ensuring, e.g., the proper user management for confidentiality, integrity and accessibility of data. Partners will transfer data using, either secure encryption transportation protocols, or by trusted encryption techniques, including proper transfer of encryption key(s) – as necessary.

The computational modelling activities carried MATLAB/Python are carried out in UCloud and stored in a closed folder only the project partners have access to.

5. Storage, Backup and Access

- Questions

1. Where would you place the Malawi data in the AAU data classification model?

6. Selection and Preservation

Think long-term

- What data do you intend to keep after the research project?
- How can others benefit from your research?
- Where will your data go at the end of the project?

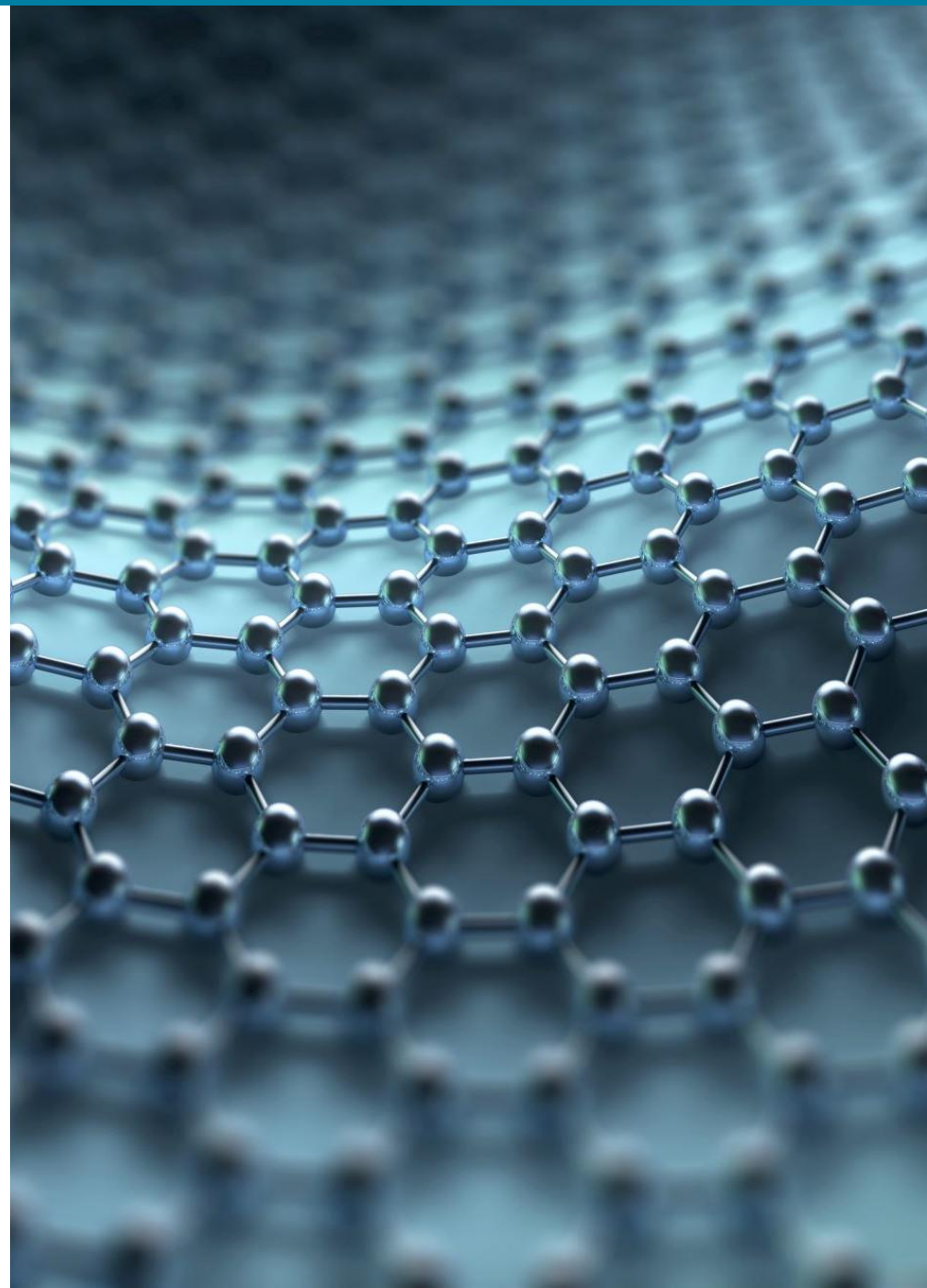
6. Selection and Preservation

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

Rules and regulations when preserving data:

- Danish Code of Conduct for Research Integrity says: "Data should in general be kept for a period of at least five years from the date of publication".
- Health data: Be aware, that some providers of health data have special rules about when to delete the data for instance after 3, 5 or 10 years.
- Company/industry data: Some companies require that data is deleted after the end of the project. This will be stated in the contractual obligations.

If the data you are working with do not have any special requirements for preservation and deletion of data, you need to follow the code of conduct, that says to keep the data for a minimum of 5 years.



6. Selection and Preservation

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

Considerations on what data to preserve:

- It does not make sense to preserve all collected and produced data in all versions.
- If you are using existing data, it is sufficient to quate the source, but you do not have to save existing data as they are already saved somewhere else.
- Data associated with publications need to be preserved long-term due to requirement of reproduction of results.
- Data that are not possible to replicate have great value and need to be preserved.
- Data with great value for reuse need to be preserved.

There might be data, that need to be destroyed for legal or ethical reasons.

6. Selection and Preservation

-What is the long-term preservation plan for your data?

Things to remember and consider in relation to long-term preservation of your data:

- Deposit your data in a data repository and assign your data with a persistent identifier such as DOI.
- Remember to use commonly used file formats that do not require special programs or licenses.
- Provide your data with descriptions of standard metadata including search keywords.
- Remember to link to related research outputs and resources such as instruments.
- Attach sufficient documentation about your data, so others can understand and use the data without you.
- Clarify who is responsible for depositing the data. What happens if the researcher leaves the university?
- Check whether there are any funder or journal requirements.

6. Selection and Preservation

-Example answers

All data required to reproduce published research results will be archived in long-term storage solutions and made publicly available as far as previously existing license restrictions permit. Data that is outdated or temporary in nature will be deleted at the end of the project.

Research data on which a publication is based, but also other relevant milestone files of the project are stored safely for at least ten years in system xxx. The data protection officer of the university is involved in this process regarding personal data. The expected total size of the remaining data is about 100 GB.

Public published data will be deposited in Zenodo. The data must remain available and findable for a period of five years after the end of the project. According to Zenodo's website the data will be available for at least 20 years.

6. Selection and Preservation

- Questions

- Is there a PID connected to data?

7. Data Sharing

Data sharing in this part of your DMP is a direct continuation of Selection and Preservation.

Again, it is about your final datasets.

It is about your choice of:

- Type of repository
- Mode of access
- License for use

7. Data Sharing

Three types of repositories:

1. **Discipline or data type specific repositories:**
   Specific and relevant attention but perhaps also limited exposure.
2. **Generic and public repositories:**
   Possible wide exposure and attention. Ex: Figshare, Zenodo.
3. **Institutional repositories:**
   Possible advantages due to security and cost. Limited exposure. Ex: AAU DataDeposit

For inspiration go to **re3data.org**

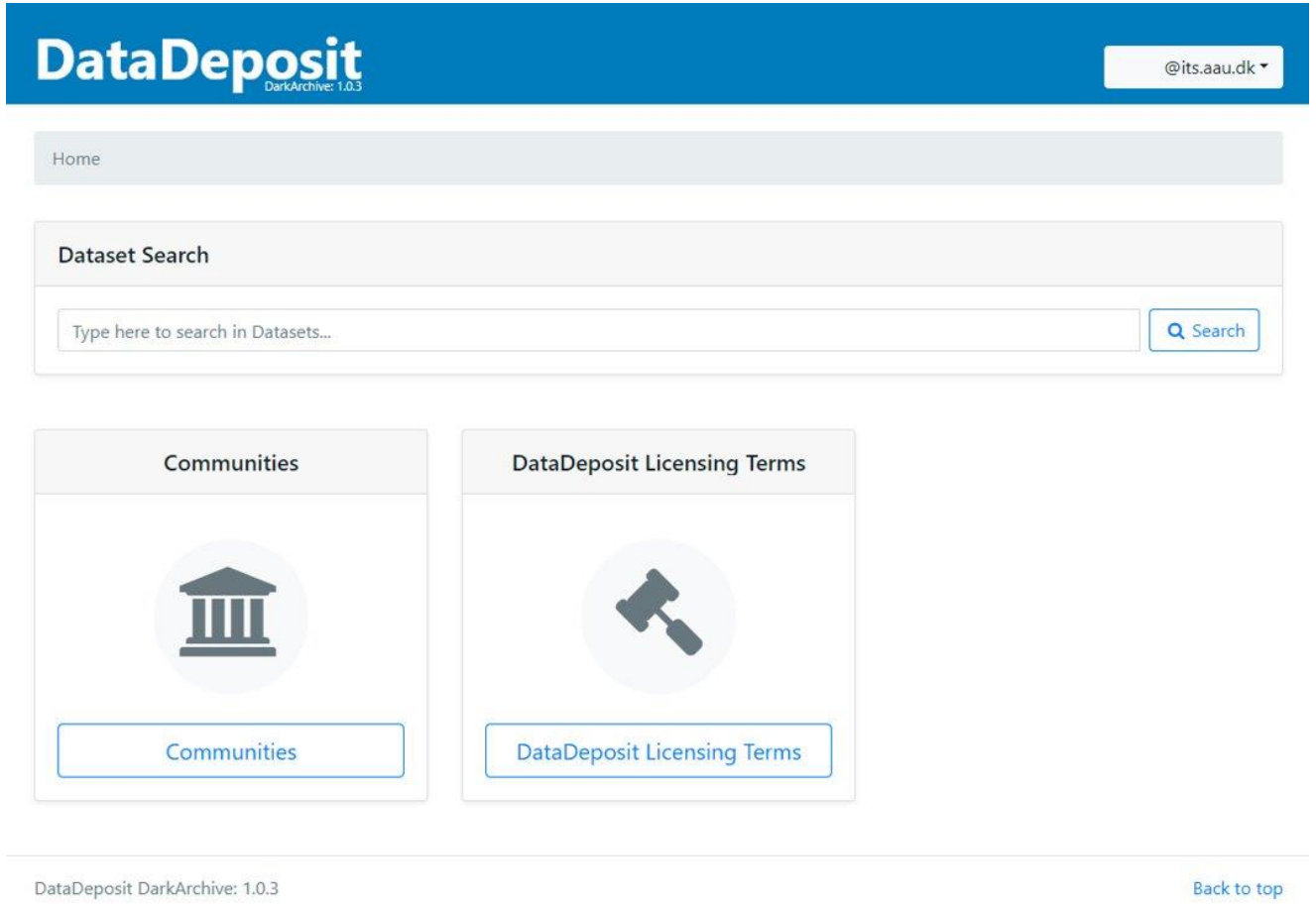
Large registry of repositories and you can browse and search research area, type of content and country.

AAU DataDeposit

AAU's local archiving solution:

- Available to all researchers at AAU.
- Deposit your finalized research data.
- Handles all data types and security levels.
- Public metadata exposed on AAU's VBN/PURE
- Guidance and hands on support available from CLAAUDIA.
- Free for now!

https://datadeposit.claaudia.aau.dk/



7. Data Sharing

Type of access

You must consider the type of access there will be to your datasets.

Also explore what types of access are possible in your chosen repository.

- Open Access
- Embargoed Access
- Restricted Access
- No Access

Data license

If others may use your data, but not completely without any restrictions.

You must determine **how** others can use the data.

You can do this by adding a **data license**.

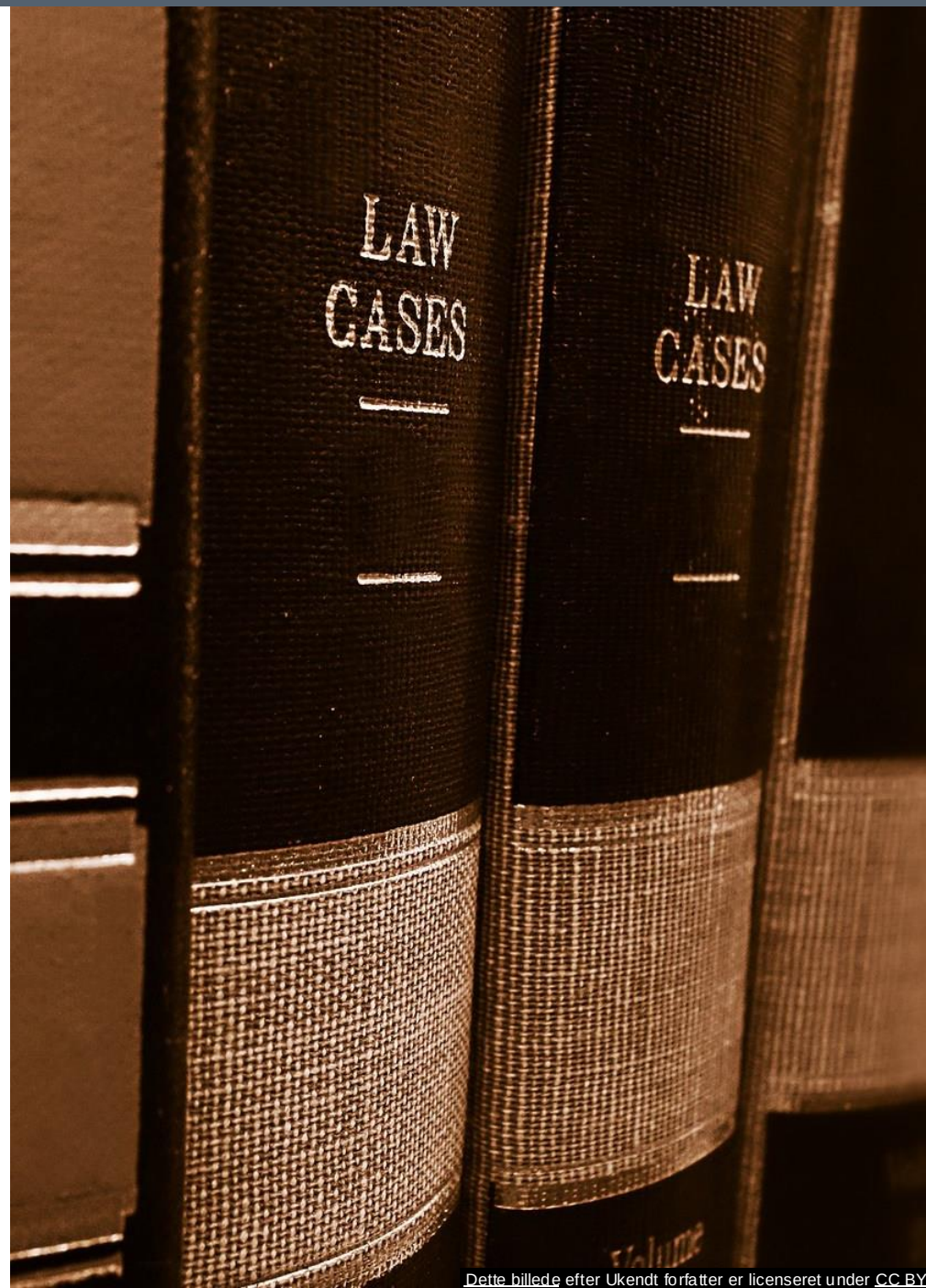
The most used, are the **Creative Commons** - a suite of six licenses, but there are others.

Here is a link to an article on FigShare called **What is the most appropriate license for my research?**

7. Data Sharing

Topics that can have a big influence on your choices:

- Read the repositories' policies to see how long the data are preserved.
- You cannot deposit sensitive data in a public repository. You can upload metadata about sensitive data, but not the data itself.
- There can be demands or restrictions from others:
  - Grant conditions.
  - Collaborators providing confidential data.
  - Re-use of data.
  - Patents.
  - Data affected by security clearance or critical infrastructure.
  - Other ethical issues such as rare animals or natural resources.



Dette billede efter Ukendt forfatter er licenseret under CC BY

7. Data Sharing

- Example answers

For publication of data, we will be using the following repositories: 1) Zenodo to publish project outputs such as reports and data 2) the domain specific repository XYZ for X 3) The Jupyter notebook will be published in GitHub together with the documentation. A persistent identifier will be assigned by the repositories (using Zenodo and GitHub integration). With these sharing and storing practices in place, we support the FAIR principles.

We do not plan to impose any restrictions on the re-use of data generated and published in the course of this project. In cases where pre-existing licensing restrictions prevent full publications, we will provide at least one publicly available example.

The re-use of dataset #3 is restricted due to the intellectual property rights of industry and health sector partners.

7. Data Sharing

- Questions

- What access conditions apply to the data?
- Can it be used for commercial purposes (which license does it have)?

8.Responsibilities and Resources

- Who will be responsible for data management?

- Outline the roles in the research project.
  - Mostly relevant if you have several collaborators.

Responsibilities

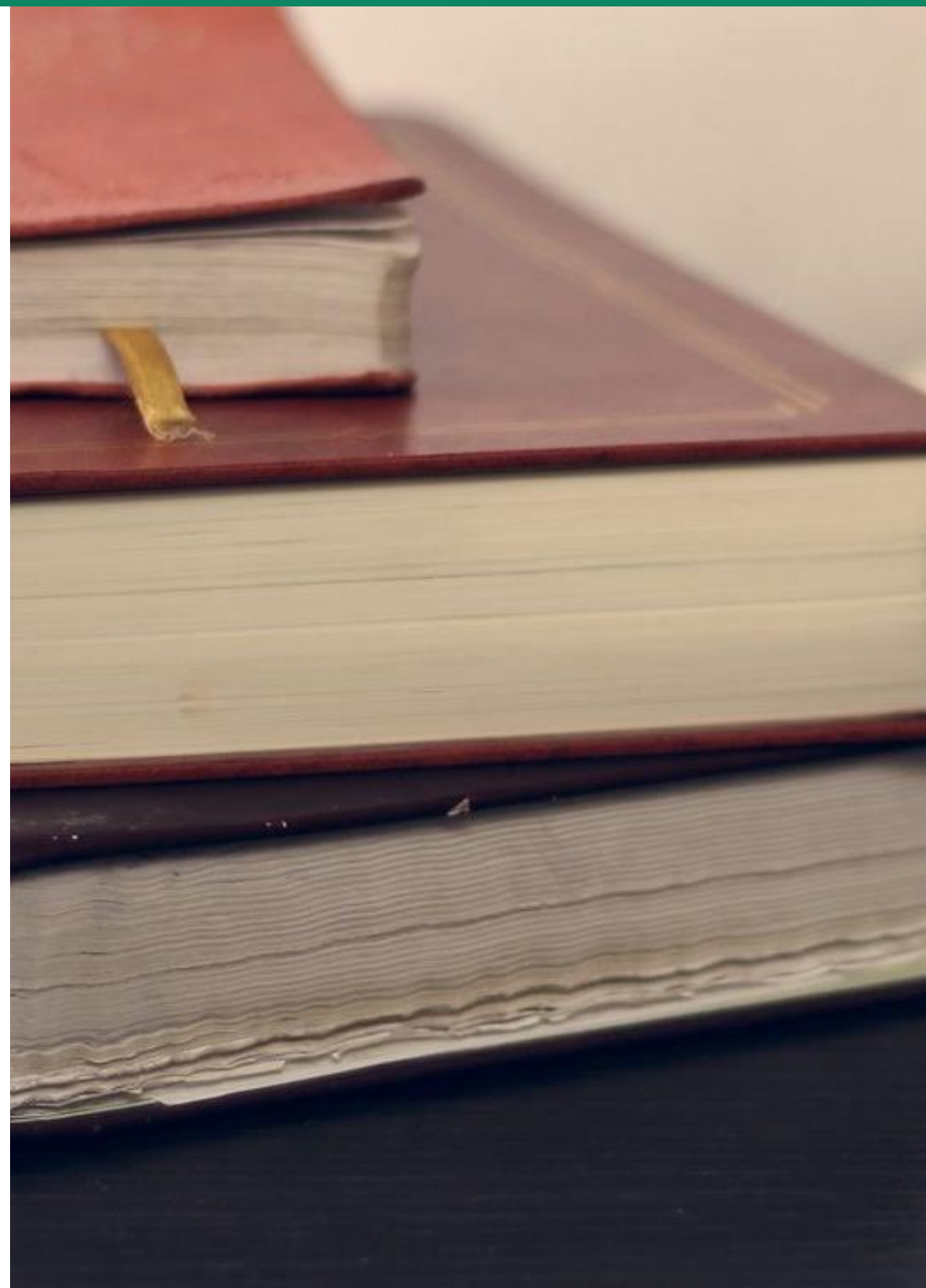
- Who is responsible for creating, reviewing, revising, and implementing the DMP?
- Who will be responsible for metadata?
- Who will be responsible for data quality?
- Who will be responsible for storage and backup?
- Who will be responsible for archiving?
- Who will be responsible for determining the conditions for reuse?
- Will there be a need for you or another person to be involved long-term?

8.Responsibilities and Resources

- What resources will it require to carry out your research project?

Resources

- Do you require hardware or software?
- Will you require technical expertise?
- Do you or anyone involved need guidance or education?
- Will charges be applied by data repositories?
- Personal costs?
- Do you have sufficient personnel to fulfill all tasks? E.g., future tasks concerning data availability?
- Legal advice?



8.Responsibilities and Resources

- Example answers

The [PI] [data manager XY] [Partner XY] will direct the data management process overall, with the research assistant responsible for ensuring metadata production, day-to-day crosschecks, back-up and other quality control activities are maintained.

The [PI] [data officer] is responsible for the secure storage and preservation of the generated digital research data together with the institute's IT representative.

Each partner in the work packages are responsible for uploading their data and connected metadata in our project community in Zenodo.

Additional resources will be needed to prepare data for deposit and to cover charges from data repositories. According to the cooperation contract between xyz and XYZ, such costs will be covered by …..

No additional resources are needed for data management and storage. All costs are included in the project budget.

Tips for writing a good DMP

- Start as early as possible writing your DMP.
- The DMP will grow with the research project – it is okay to start with ideas, thoughts and keywords and you do not have to start with question 1.
- Update your DMP regularly throughout the project.
- Include all information on how your research data will be handled and argument for your choices.
- Each DMP is unique - the content, composition, and structure can vary greatly, so keep it simple.
- Using a DMP-template (required by funder or AAU) will make it more structured.
- Contact CLAAUDIA Aalborg Universitet - Service Portal, if you need guidance.

ORCID

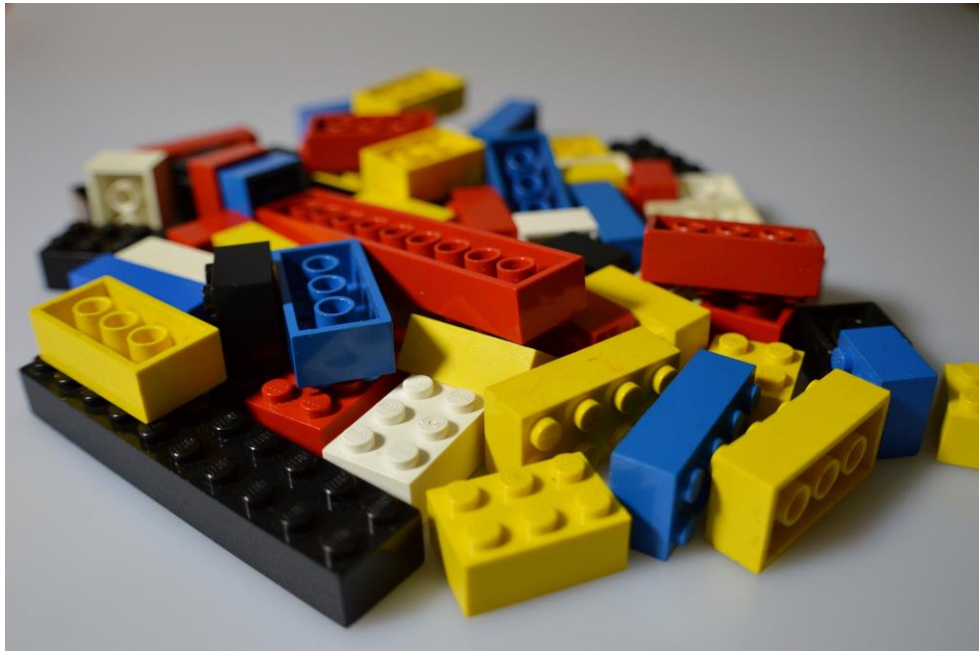
- A unique persistent identifier (PID) for individuals to use when engaged in research activities.

https://orcid.org/

- Today you get a fictive PID

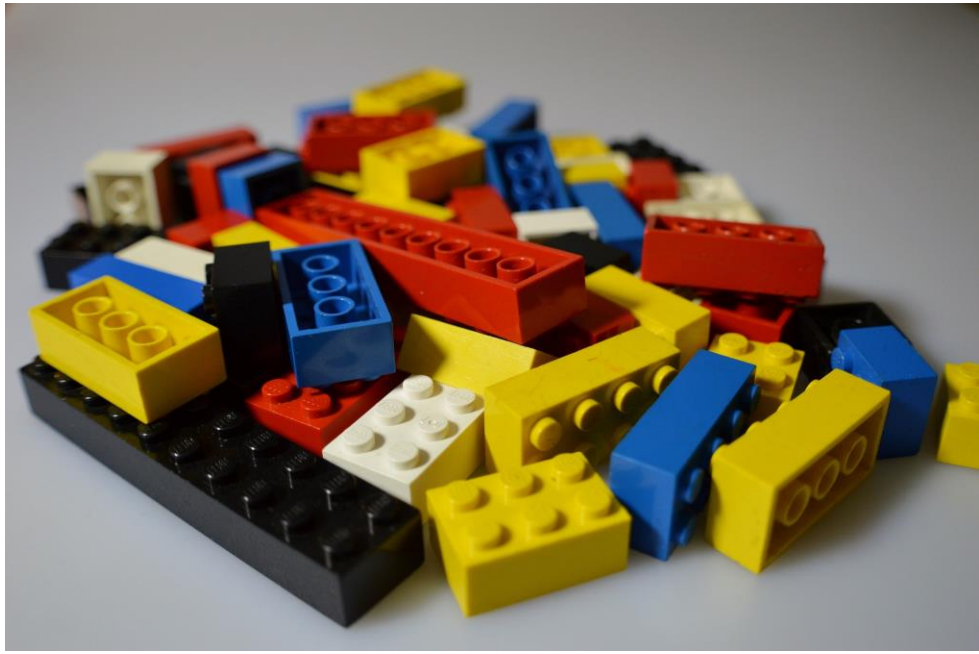
LEGO exercise (Intro)

- In your project you are to build a product.
- You should build it out of existing components from acknowledged research and well documented scientific results.
- Think of the product you are to build and what parts you need to find:
  - A theory
  - A document
  - An interview
  - Software
  - A script
  - One piece of information
  - Collection of empiric results



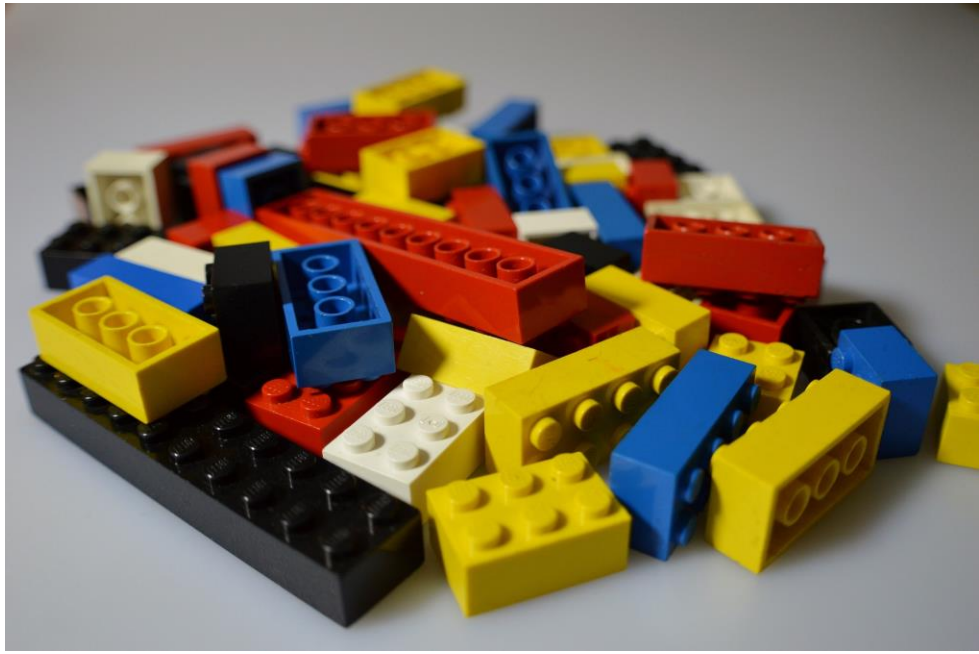
LEGO exercise (part 1)

- Pick 5-7 LEGO bricks.
  - The bricks are your components to build your product.
- Describe your bricks in writing so that others would be able to identify them.
  - This is the metadata for the components for your product.
- Build a figure with your chosen bricks.
  - The figure is your product.
- Describe the figure in writing so that others would be able to build it.
  - This is the metadata for your product.
- Write if others are allowed to play with the figure, and if they can rebuild it and/or use it for collaboration.
  - This is metadata on your product and licenses on if and how it can be used.
- Place your figure into an envelope and write your PID and Figure ID on metadata.
  - This is you preparing your data to archive it.
- When you upload your product it gets a PID (Persistent Identifier).
  - Just like a book have an ISBN number, digital objects can get a DOI (digital Object ID).
- Place the envelope in the marked are and place the descriptions beside.
  - This is the publishing of your product in the repository.



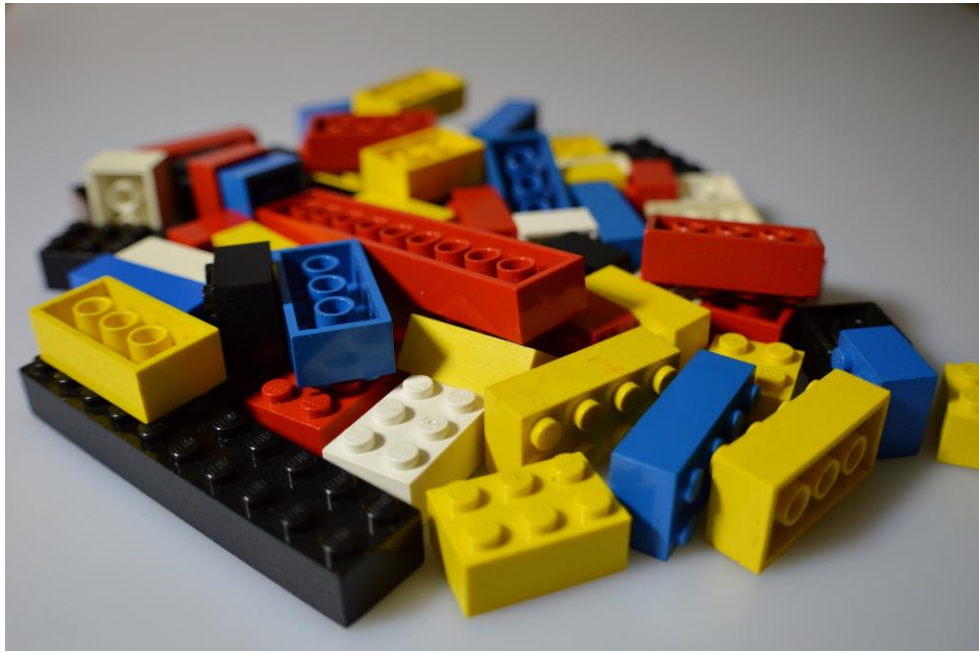
LEGO exercise (part 2)

- Get a figure ID and find it and its metadata.
- Take the metadata (description) of the product – let the envelope with the product stay in the repository.
- Read the description of bricks and try to find them.
  - If you cannot find the exact same, try your best to find substitutes.
- Read the description of the figure and try to build it.
- Present the figure to its original creator.
  - Use the “ORCID” ID to source



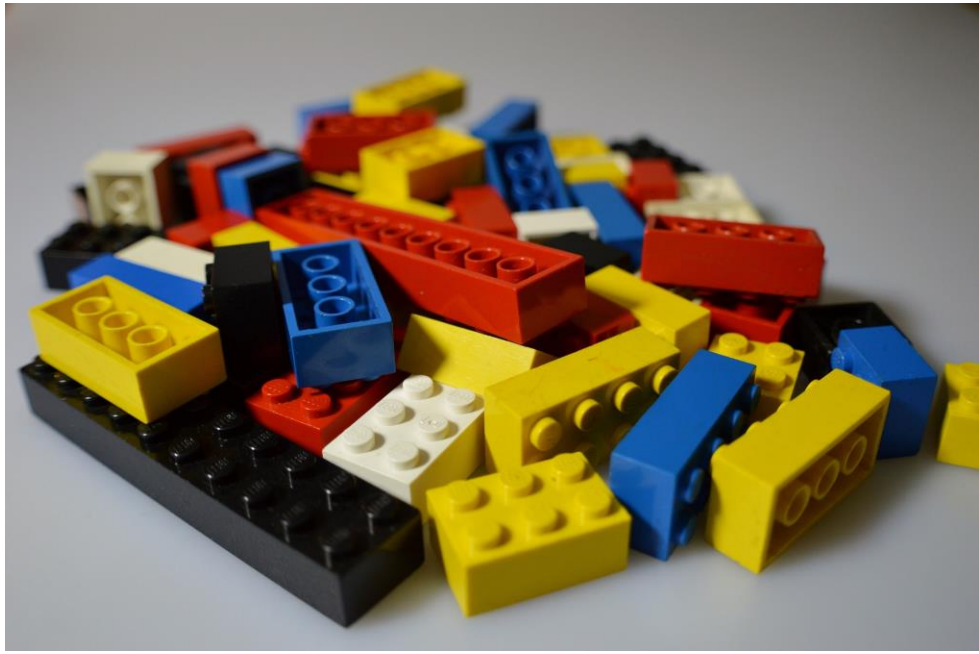
LEGO exercise (part 3)

- Does the figure look like the original?
- Did you find the description understandable?
- Did you experience any challenges?



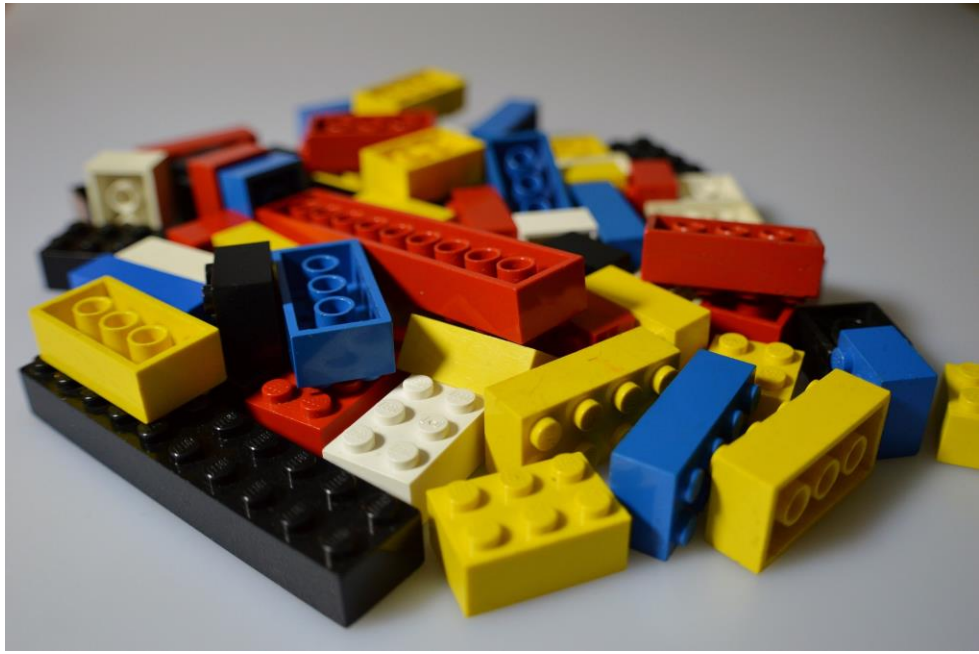
LEGO exercise (part 4)

- Take the envelope with the original product.
- Open the envelope with the original figure.
- Read the description of how to use the product.
- Are you able to use the product as it is?
- Are you allowed to make changes or build it into new projects?
- Would it be possible to connect the different figures and make collaborations?
  - Be aware of licenses and what do you need to connect them.



LEGO exercise (part 5)

- The big box of LEGO is the internet.
- The marked area is a repository.
- How was it different to recreate the product to “download” it from the repository
- This is an example of the importance of:
  - FAIR
  - Data management
  - Use of existing research
  - Licenses
- Reflections on the exercise?



Findable

Humans and machines can find out the data exits.

Because you:

- Publish searchable metadata.
- Assign a unique persistent identifier.

The FAIR Guiding Principles for scientific data management and stewardship says:

F1: Data, and/or metadata, assigned a globally unique and persistent identifier

F2: Data are described with rich metadata

F3: Metadata clearly and explicitly include the identifier of the data they describe

F4: Data, and/or metadata, are registered or indexed in a searchable resource

Dette billede efter Ukendt forfatter er licenseret under CC BY

Accessible

Humans and machines can find out how to get access to the data.

Because you:

- Upload to public data repository.
- Define access conditions for data and metadata.

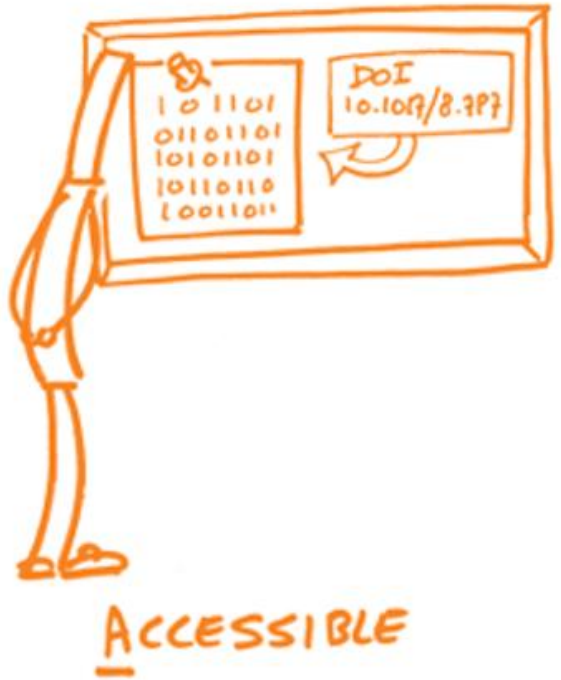
The FAIR Guiding Principles for scientific data management and stewardship says:

A1. Data, and/or metadata, are retrievable by their identifier using a standardised communications protocol

A1.1 The protocol is open, free, and universally implementable

A1.2 The protocol allows for an authentication and authorization procedure, where necessary

A2. Metadata are accessible, even when the data are no longer available



Dette billede efter Ukendt forfatter er licenseret under CC BY

Interoperable

Humans (and machines) can open the data and work with it.

Because you:

- Use open file formats.
- Use community standards, keywords and ontologies.

The FAIR Guiding Principles for scientific data management and stewardship says:

I1. Data, and/or metadata, use a formal, accessible, shared, and broadly applicable language for knowledge representation.

I2. Data, and/or metadata, use vocabularies that follow FAIR principles

I3. Data, and/or metadata, include qualified references to other Data, and/or metadata

Dette billede efter Ukendt forfatter er licenseret under CC BY

Reusable

Humans and machines understand how the data was created and how to reuse it.

Because you:

- Attach sufficient documentation.
- Add a usage license.

The FAIR Guiding Principles for scientific data management and stewardship says:

R1. Data, and/or metadata, are richly described with a plurality of accurate and relevant attributes

R1.1. Data, and/or metadata, are released with a clear and accessible data usage license

R1.2. Data, and/or metadata, are associated with detailed provenance

R1.3. Data, and/or metadata, meet domain-relevant community standards

Dette billede efter Ukendt forfatter er licenseret under CC BY

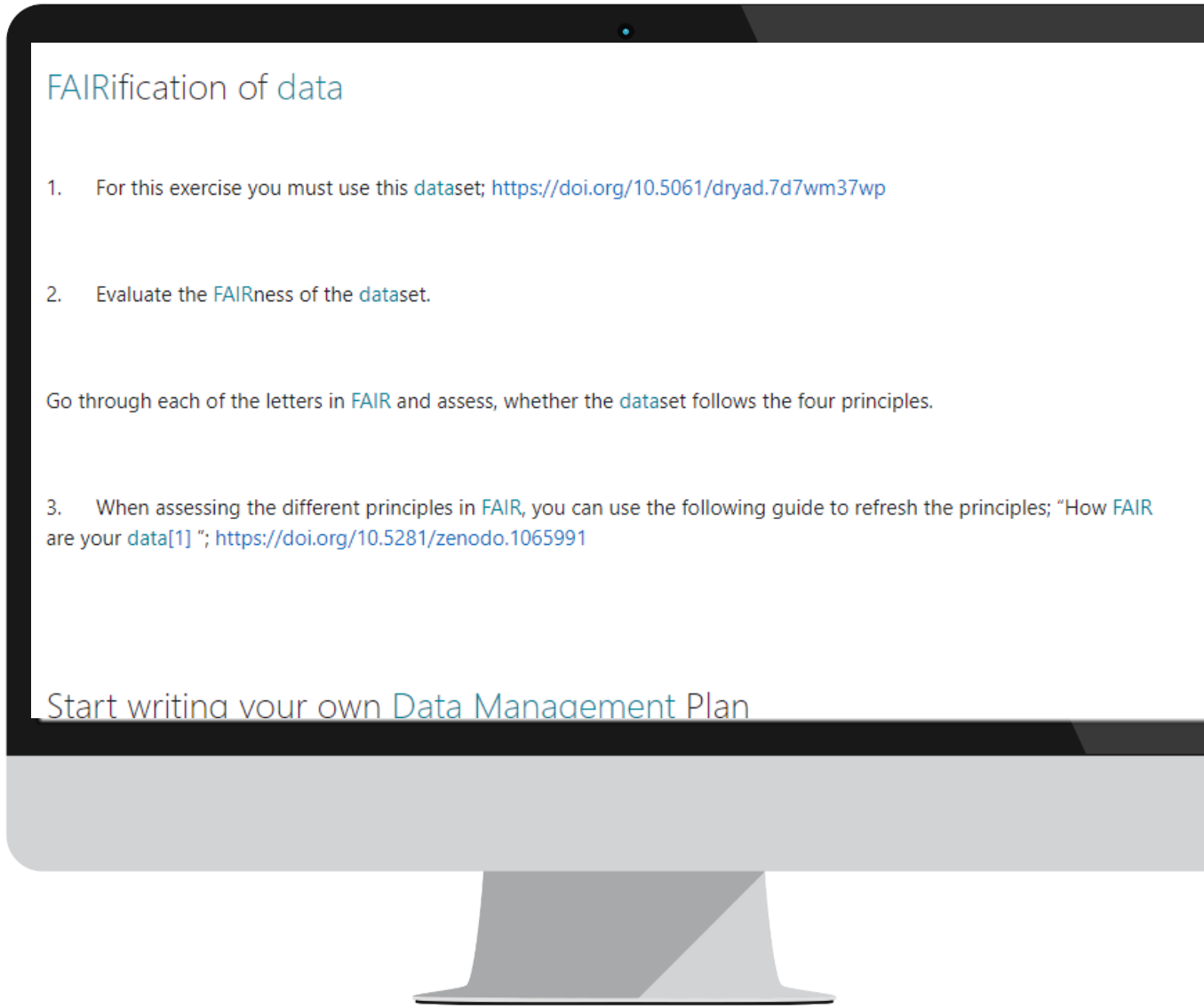
Todays most important points

- Be FAIR - "As open as possible and as closed as necessary".
- A DMP is a great way to structure your data management and make informed choices and changes throughout the project.
- CLAAUDIA is here to help you.

Introduction to homework

- FAIRification – exercise
  - For this exercise you must use the dataset on Moodle.
  - Evaluate the FAIRness of the dataset by using the guide “How FAIR are your data”, also on Moodle.

- Start writing your own DMP
  - Start writing a DMP on your PhD. Project. Use the AAU generic template to write your DMP, which you can find at dmp.deic.dk.
  - Remember, your DMP is a living document, this first version does not need to be perfect or done.
  - Next time, you will discuss your data management plans and considerations in groups.
  - You do not have to hand in your DMP or anything.



Links for inspiration

- AAU FAIR data management
- DMP guidance

- **DMP examples**
- DMP; Opening access to economic data to prevent tobacco related diseases in Africa
- DMP; Empowering Indigenous Peoples and Knowledge Systems Related to Climate Change and Intellectual Property Rights
- DMP; REPAIR

Contact info

serviceportal.aau.dk

researcher.aau.dk
HPC AAU

linkedin.com/company/claaudia/

AALBORG UNIVERSITET