



Research data management

Taha Ahmed, research data advisor (KTH Library)

<http://solarchemist.se/kth>

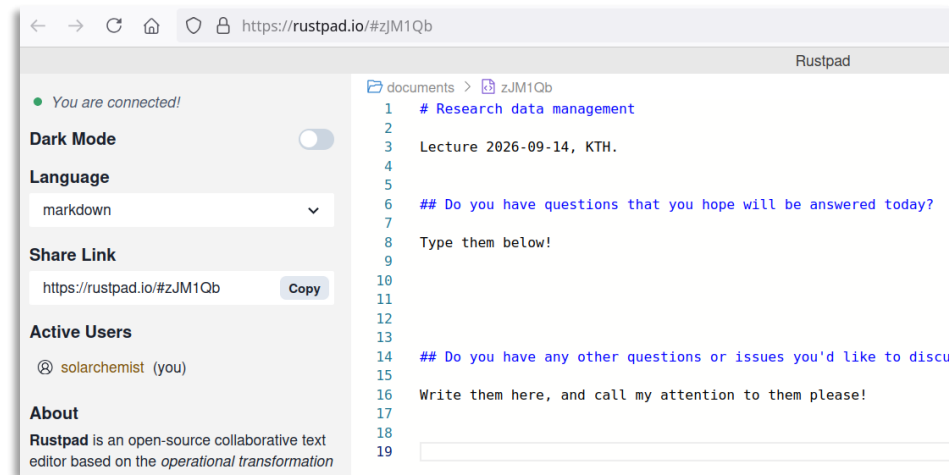
Outline

1. Presentation round
2. What is research data?
3. Open data and FAIR data
4. Open science and reproducible research
5. The data management plan (DMP)
6. Research data management
7. Resources for using and reusing data
8. Want to know more? Links to additional resources

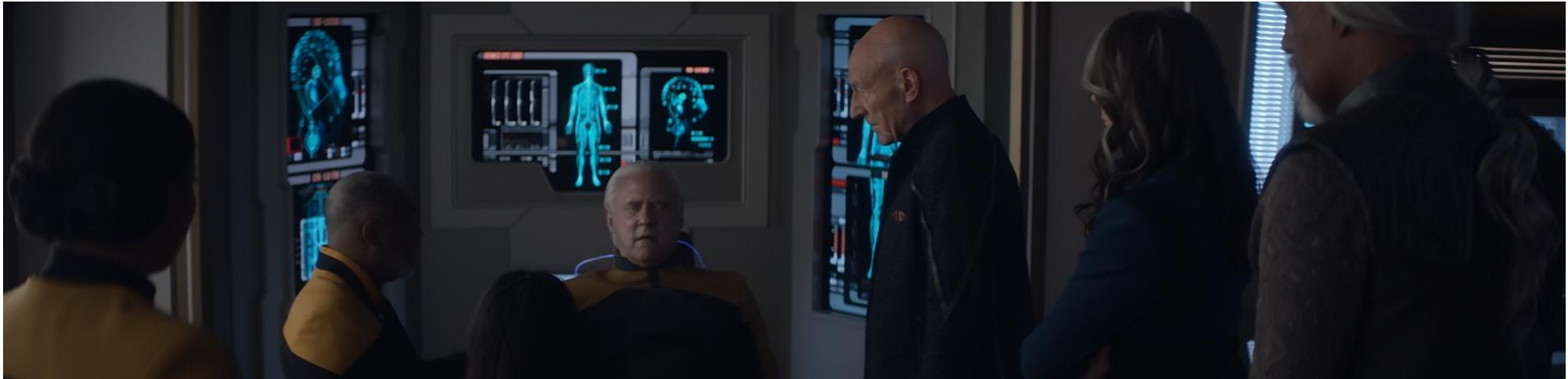


What questions do you have?

- Do you have questions that you hope will be answered today?
- Please write what you want to know more about in this pad:
<https://rustpad.io/#zJM1Qb>



Do you work with data?



- What is your research topic?
- Does your research involve the collection, analysis or generation of data?
- What tools do you use to work with data?



Suggested reading

- Research data management, KTH Library
<https://www.kth.se/en/biblioteket/publicera-analysera/hantera-forskningsdata>
- Manage data, Swedish National Data Service
<https://researchdata.se/en/manage-data>
- Data with personal information, Swedish National Data Service
<https://snd.se/en/doris-researchers/data-personal-information-doris>
- Ethics in research, VR, with hyperlink to European Code of Conduct for Research Integrity <https://www.vr.se/english/mandates/ethics/ethics-in-research.html>
- KTH Policy for managing research data, 2021 (PDF)
https://intra.kth.se/polopoly_fs/1.1037530.1714460492!/Riktlinje-om-hantering-av-forskningsdata.pdf

What is data?

Definition Wikipedia (computer science):

Data requires [interpretation](#) to become [information](#). [Digital data](#) is data that is represented using the [binary number](#) system of ones (1) and zeros (0), instead of [analog](#) representation. In modern (post-1960) computer systems, all data is digital.

Definition according to EOSC:

reinterpretable digital representation of information in a formalized manner suitable for communication, interpretation, or processing.

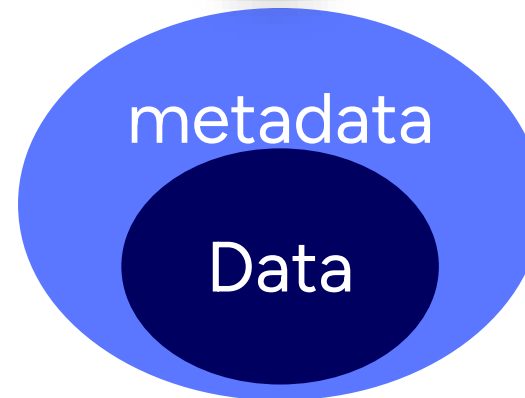


101
011

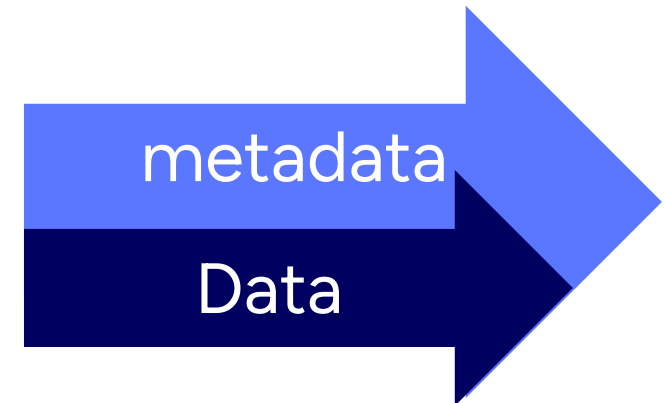
Bit sequence = 1 byte data

What is metadata?

Metadata = data about data, often the information needed to interpret data in a meaningful context, (technical, administrative, descriptive metadata)



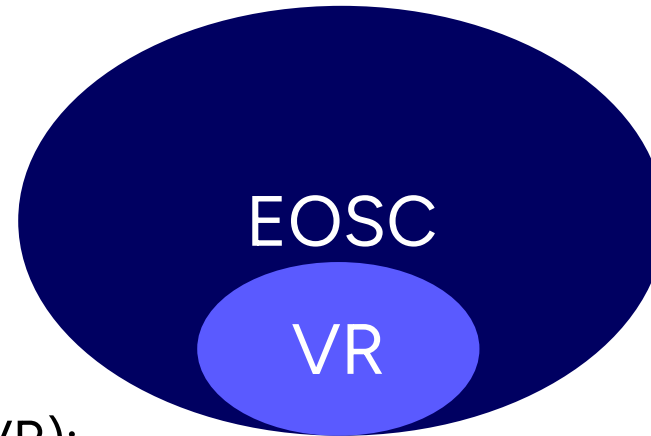
Data-object



Data-flow



What is research data?



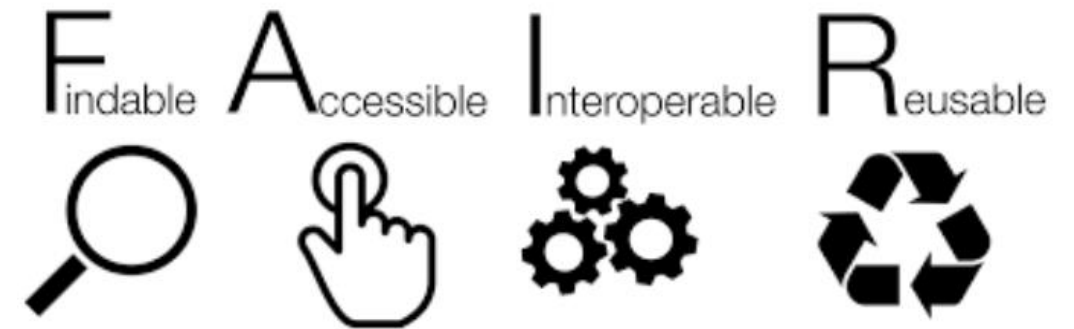
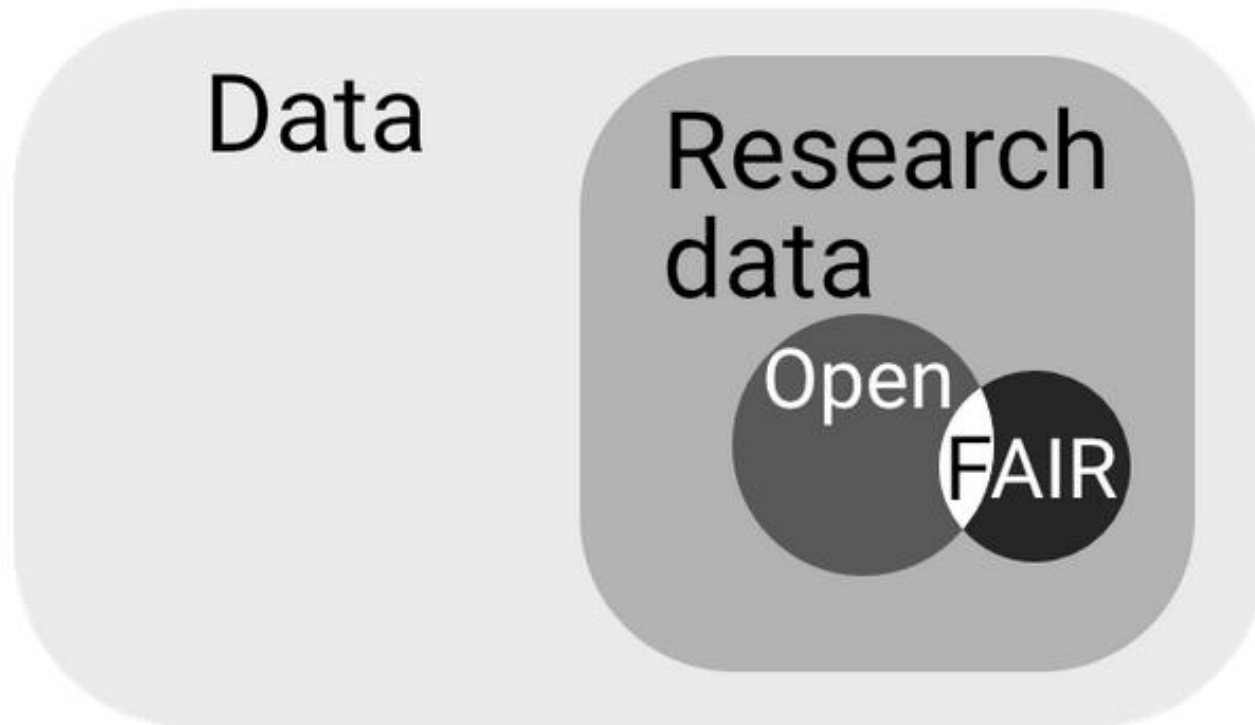
Swedish research council (VR):

“Data som har **skapats under forskningsprocessen** och som ligger till grund för vetenskapliga publikationer. Om **redan befintliga data** från andra aktörer har använts, såsom exempelvis olika typer av myndighetsdata, **utgör dessa forskningsdata endast** om de har analyserats, bearbetats eller på annat sätt **förädlats under forskningsprocessen**. Redan befintliga data som endast använts i sin ursprungliga form och som redan förvaltas och tillgängliggörs av en annan aktör omfattas inte”

European Science Cloud (EOSC):

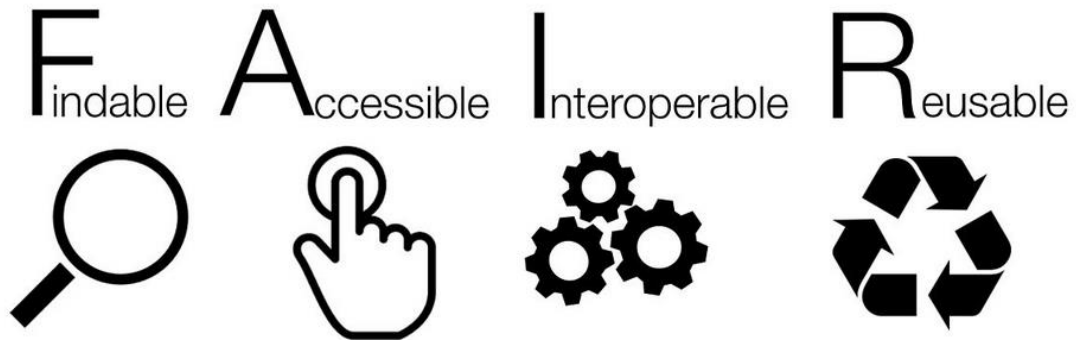
“**data collected or produced** in the course of scientific research activities and **used as evidence in the research process, or** commonly accepted in the research community as **necessary to validate research findings and results.**”

Open data, FAIR data, Research data



Metadata and data should be findable, accessible* in an interoperable way so that the data can be re-used.

What is FAIR data?



FAIR and Open don't necessarily overlap 1:1!

Also see FAIR for Research Software (FAIR4RS).

"As open as possible, as restricted as necessary."

- Findable:
 - Data are described with rich metadata
 - Data are indexed in a searchable resource
 - Data are assigned a persistent identifier (e.g. doi)
- Accessible:
 - Data or metadata is retrieveable by PID.
 - Data is retrieveable by an open protocol
 - Metadata is readily accessible (open) even when the data in itself cannot be.
- Interoperable
 - Data uses a formal, broadly applicable language for knowledge representation
 - Data uses a FAIR vocabulary
 - Data includes references to other data
- Reuseable
 - Metadata are richly described with relevant attributes
 - Data are released with a clear use licence
 - Data follows community standards

[Wilkinson et al. Sci Data 3, 160018 \(2016\).](#)



What data do you collect and use?

- What data do you collect and use?
- How do you protect sensitive data?
- Can data be aggregated and will this affect anonymity?
- How do you make sure that it's possible to verify your results?
- What data should be shared for others to use?
- Large datasets makes new things possible, but!
 - aggregation easily leads to identification
 - using historical data to create models will emphasize historical biases

Reasons to make data Open

- Transparency
- Reproducibility & replicability
- Quality of research
- Reuse
- Efficiency and collaboration
- Impact
- Government requirements and policy by funding bodies and institutions

Reasons to keep data Restricted

Legal and ethical reasons

- Personal Integrity (GDPR)
- Violation of the rights of indigenous people (CARE principles)
- Breach of Intellectual property rights (IP-law)
- Export Control (e.g., military applications)
- National security reasons

Pragmatic reasons

- Embargo or fear of getting scooped
- Lack of appropriate tools
- Lack of time
- Unclear responsibility – who is in charge of documenting and sharing the data?

What is open science?

Commonly depicted as the “open science umbrella”, encompassing at least:

- Open data
- Open source
- Open standards

“Open science is about making research open, inclusive, and useful for everyone. The work involves different ways of making the processes that generate, evaluate, explore, and preserve scientific knowledge, practices, and perspectives more accessible.”

<https://openscience.se/en/what-is-open-science-0> (SUHF and SND)

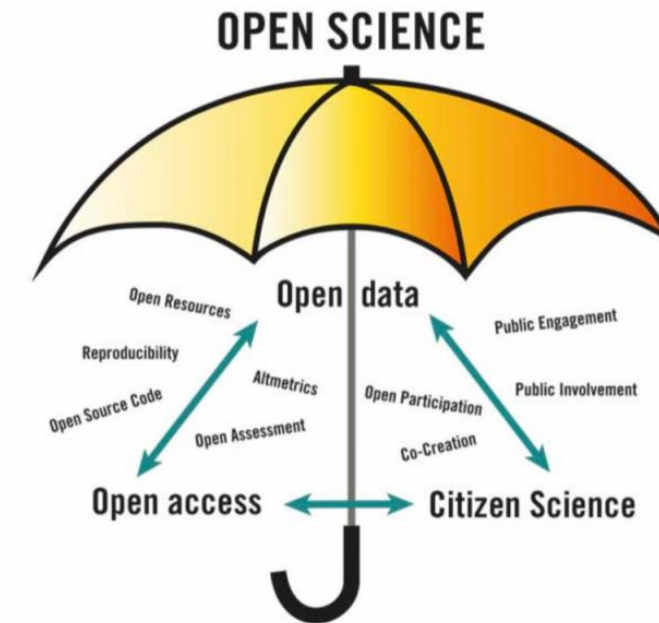


Illustration: Lotta Tomasson/VA CC BY-NC 2.0

Research should be reproducible

- Knowledge should not be only for the privileged
- Data, code, and other digital artefacts are an important research output and should be made open whenever possible

Reproducible research also has many other positive benefits, both to science and society.
Such as...?

To do reproducible research, you will need to use a **reproducible workflow**, ideally by employing **open science practices**.

"A reproducible workflow", 2 min video

<https://www.youtube.com/watch?v=s3JldKoA0zw>

Open science policies and requirements

- Described in international, national and EU guidelines and policy documents
 - [EU's Open Science policy \(EC\)](#)
 - [Swedish government guideline](#): Open access from 2021, all data published as FAIR by 2026.
- Requirements typically come from funding agencies
 - [Horizon 2020 Annotated Model Grant Agreement](#)
 - [VR Requirements, terms and conditions - Swedish Research Council \(vr.se\)](#)
 - [Science Foundation Ireland Open Research policy](#)
- Universities can have their own policies, and are often tasked with enforcing requirements from funding agencies
 - [KTH policy](#) : Every project must have a person responsible for RDM. Data needs to be FAIR.
 - [KTH Digitalization policy and strategy](#)
- Most large journals also require that data is made available and FAIR in a data repository when publishing or submitting a manuscript.

Make a Data Management plan – reflect on data management early on

- Data Management Plans describe how data will be handled in a project.
- They usually follow a template (typically funder provided) where common headings are:
 - **Data description** (What data is collected?)
 - **Documentation and Data Quality** (How will data be QC'd and what meta-data will be collected?)
 - **Storage and Backup** (How will data be protected?)
 - **Legal and ethical requirements** (Is the data sensitive and if so in what way?)
 - **Data sharing and preservation** (How will data be shared and archived?)
 - **RDM responsibilities** (Who is responsible, and who has which role in the project?)





Make a Data Management plan – reflect on data management early on

- Data Management Plans describe how data will be handled in a project.
- They usually follow a template (typically funder provided) where common headings are:
 - **Data description** (What data is collected?)
 - **Documentation and Data Quality** (How will data be QC'd and what meta-data will be collected?)
 - **Storage and Backup** (How will data be protected?)
 - **Legal and ethical requirements** (Is the data sensitive and if so in what way?)
 - **Data sharing and preservation** (How will data be shared and archived?)
 - **RDM responsibilities** (Who is responsible, and who has which role in the project?)
- Ideally a living document that is updated during the research process
- Exact requirements can vary (check your funding agreement!):
 - VR: DMP needs to be written before project begins. Only needs to be sent in if requested.
 - EU (e.g. Horizon Europe):
 - A **concise** DMP is sent in together with the project proposal.
 - A **comprehensive** DMP is sent in as a deliverable typically 6 months after project start.
 - For longer projects (>12 months) an **updated** plan needs to be sent in.
 - At the end of the project a **final** DMP is to be sent in. This often gets published.



Use the Data Management plan – to document and plan your RDM

- The DMP should reflect how you handle data in your research project.
- Opportunity to create common rules and agree on conventions within the project
- Especially useful for collaborations and projects with great variability in datatypes.
- Conventions that are unproductive can be amended or changed
- Helps you plan for requirements that come later in the project (e.g. when publishing)



Make a Data Management plan – DMPonline

Welcome to DMPonline at KTH Royal Institute of Technology

This instance of DMPonline is provided by KTH Royal Institute of Technology to help you write data management plans. Create an account to access guidance, advice and templates from KTH Royal Institute of Technology. Once you have an account you can link this to your university account and log in with your KTH credentials. KTH DMPonline will store and use the information you enter according to the [terms of use](#).

Do not enter information about high security projects into KTH DMPonline. If you need a DMP for such a project, feel free to log in, create a test plan and download the questions from the appropriate template. Use the preferred template and reply to the questions in order to write a DMP using your preferred writing tool and storing it in a secure location.

The information you enter in KTH DMPonline will not become available to the public unless you chose to make it so. However, KTH Research Data Services can access the information (general information such as project description, participating researchers, as well as more specific information about needs for support and storage). This information can be used for support purposes, but also for reuse in administrative systems to simplify and provide better services.

If you don't agree with this, please log in and create an empty plan to download the questions from your preferred template. You can then use the template and write down your answers to the questions outside the system.

If you have questions about research data management at KTH, please contact researchdata@kth.se

<https://dmp.kth.se>

Sign in

Create account

* Email

* Password

[Forgot password?](#)

☐ Remember email

Sign in

- or -

Sign in with your institutional credentials

Research data lifecycle



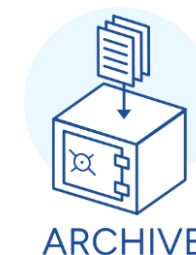


Share your code and/or data



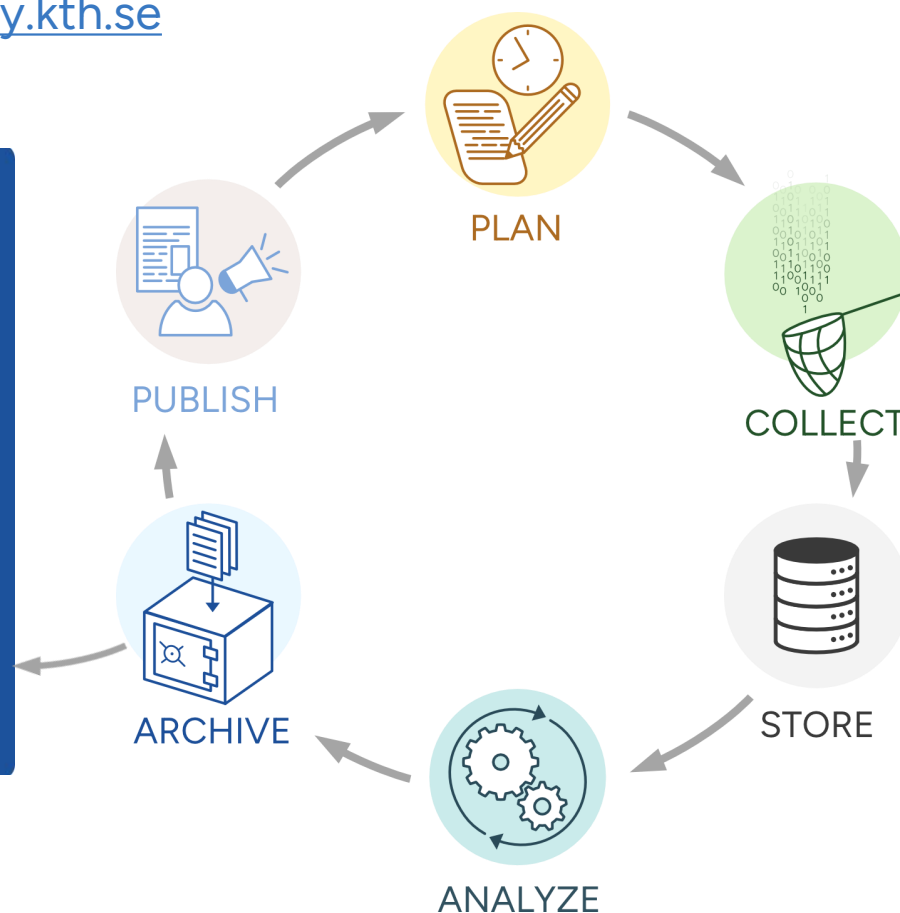
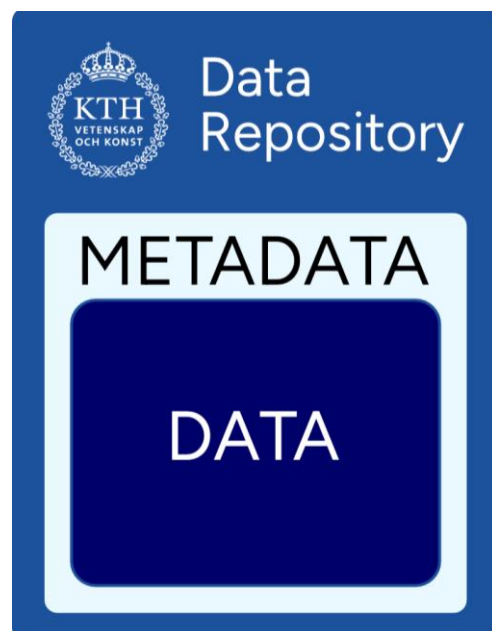
PUBLISH

- Funders and journals are increasingly requiring that research data is shared in conjunction with publication of results.
- Code is usually shared in code forges (eg. GitHub, Codeberg, GitLab, Sourcehut).
- Research data is usually shared in *data repositories*.
- Data repositories can be generalist or domain specific.
- Examples of generalist repositories are [Zenodo](#) (CERN) and [DORIS](#) (SND)
- To find a suitable domain-specific repository search [re3data.org](#)
- Data is typically shared under a CC0 (public domain) or CC-BY license.
- Repositories can also be used to find relevant **datasets to re-use!**
- Some journals have requirements that certain datatypes are shared in certain, domain-specific repositories.
 - [Nature publishing group requirements](#)
 - [PLoS recommendations](#)
 - [Cell press guide](#)



Share your data and/or code on the KTH Data Repository

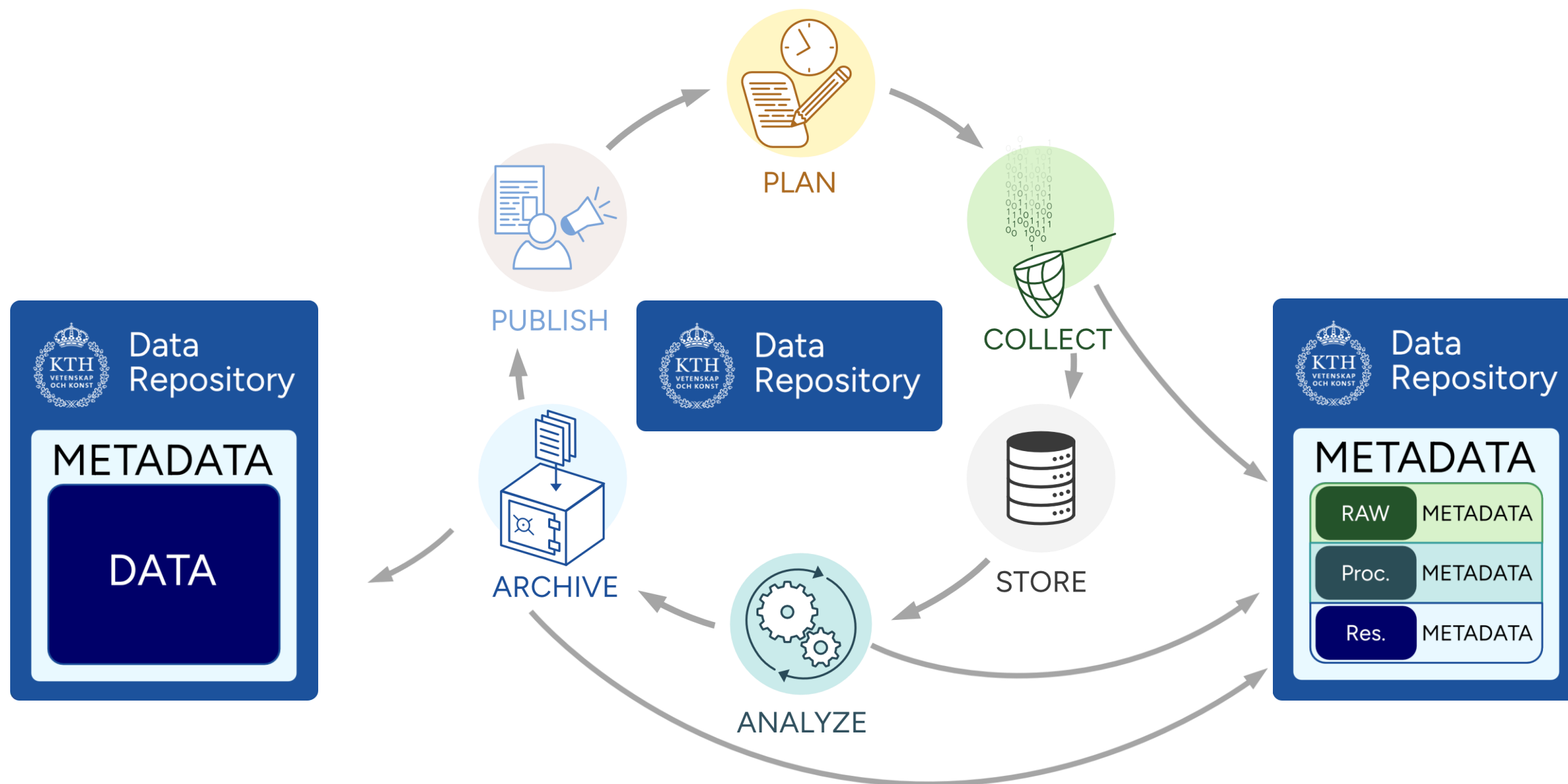
<https://datarepository.kth.se>



- A service to researchers that need to meet increased requirements on research data being accessible according to **FAIR principles**.
- You can get support and feedback on how to meet requirements on FAIR data
- For archiving research data for at least ten years after the end of a research project
- For open data & for data that need to be stored and made accessible according to the FAIR principles but with restricted access.



KTH Data Repository





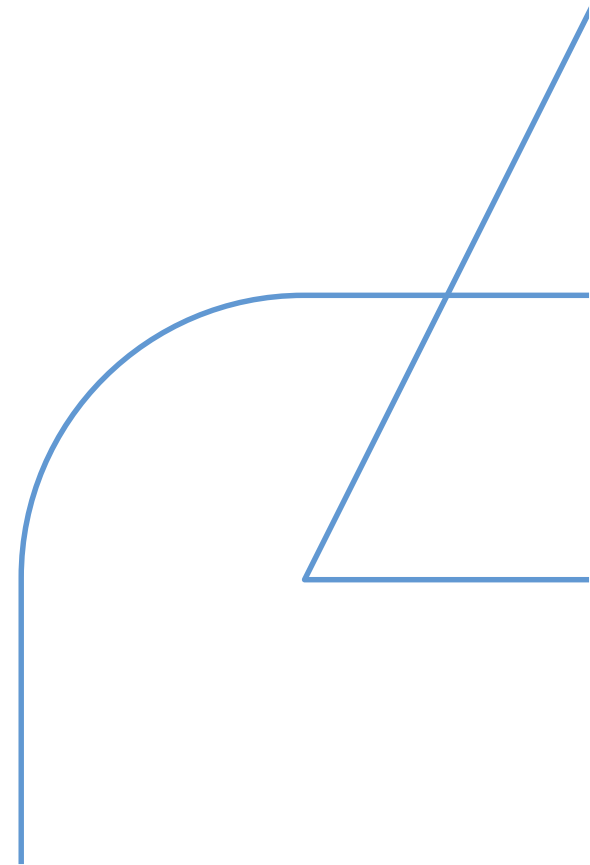
KTH Data Repository documentation & support



- <https://docs.datarepository.kth.se>
- Contact us at researchdata@kth.se with your questions

Discussion (15 min)

- Follow-up on any questions in the pad!
- Discuss in class:
 - Have you shared or, are you planning to share any of your data? Why or why not?
 - If you are sharing data, how or where will you share it?
 - Are there any obstacles to sharing data in your project? If so what?
 - Have you looked for data to re-use from a repository or elsewhere?
- Pick one of the datasets in your group. Discuss who owns the data and whom it belongs to. And who has the right to use the data.



Good open science practices



PLAN

- Write a DMP
- Read the DMP
- Use the DMP



COLLECT

- Document your procedure
- QC and look at your data



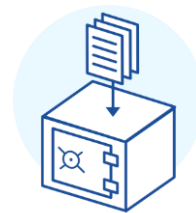
STORE

- Label your files and data
- Structure your storage
- Manage access and encryption



ANALYZE

- Use collaborative tools
- Document your analysis



ARCHIVE

- Use repositories (and PIDs!)
- Add sufficient metadata



PUBLISH

- Plan your publication early
- Read up on requirements
- Publish your (meta)data



Links to useful resources - RDM

- [DMPOnline](#)
Tool to create a data management plan. Used by KTH (dmp.kth.se)
- [Argos](#)
Tool to create a data management plan. By Openaire (EOSC).
- [RDMkit](#)
Collection of useful tools and links for RDM. Aimed at Life science.
- [GoFAIR](#)
More on Fair principles.
- [CARE Principles](#)
CARE principles for indigenous Data Governance
- [Cordis](#)
European Union R&D project database (includes H2020 and HE)
- [SpringerNature guidelines](#)
Publisher policy on data and data repositories
- [KTH web-course on RDM and open science](#)
Open Canvas course on Research Data Management at KTH.
- [KTH Data Repository](#)
Available as a service for KTH researchers and collaborators
- **SND – Swedish National Dataservice:**
 - [Manage data](#)
 - [Research with personal data](#)
 - [DORIS](#) – SND Catalogue / repository
- [Zenodo](#)
Large Generalist data repository . By CERN.
- [Re3data.org](#)
Database of repositories (“repo-of-repos”).
- [Open Science Framework](#)
OSF also contains a repository service.

Links to useful resources - vizualisation

- [DataVizcatalogue.com](https://datavizcatalogue.com)
Contains links to many additional resources and galleries
- [Data-to-Viz](https://data-to-viz.com)
Useful decision trees for selecting graphics
- [Ktheme for R](https://kththeme.com) and [kth-quarto](https://kth-quarto.com)
Unofficial KTH colorschemes and layouts for R and quarto
- [Visual storytelling psychology](https://visualstorytellingpsychology.com) by Nadieh Bremer
Educational slides on the psychology behind visualization
- [InfraVis](https://infravis.org) National Infrastructure for DataViz
Provides support to researchers (a lot of it free)
- **Affinity Designer**
Available through the KTH Software Center, KI sharepoint, SU Serviceportalen
- **Adobe Illustrator**
Academic Licenses available through your university
- [Adobe Effects](https://adobe.com/products/photoshop/elements.html)
Free version available. Great for working with layout.
- [Python graph gallery](https://python-graph-gallery.com)
Contains lots of examples on how to make plots in python.
- [R Graph gallery](https://r-graph-gallery.com)
Contains lots of examples on how to make plots in R.
- [Plotly](https://plotly.com)
Making graphs interactive. For R, python, Matlab and more.
- [Datawrapper color guide](https://datwrappr.com)
What to consider when choosing colors for data visualization
- **Color-tools:**
 - [KTH Visual Identity](https://kth.se/visualidentity): KTH graphical profile pages
 - [SU Visual Identity](https://su.se/visualidentity): SU graphical profile pages
 - [KI Visual Identity](https://ki.se/visualidentity): KI graphical profile pages
 - colors.co: Fast palette generator.
 - palett.es: Generate a palette from one color
 - [Color-blindness tool](https://color-blindness-tool.com): Check your palettes.
 - [Contrast checker](https://contrast-checker.com): Check your contrast
- [Tikz for LaTeX](https://tikz.org)
Create vector-graphics in LaTeX.



Where to get help

RDM @KTH

- researchdata@kth.se
- researchethics@kth.se
- dataskyddsbud@kth.se

KTH Library

- Chat [\(online\)](#), or [e-mail](#) us (or come by the desk)
- [Follow upcoming events in the Calendar](#) 
- [Book a consultation in reference management or information searching](#)
- [Book a consultation in academic writing](#)





Upcoming KTH Library webinars

- [Take control of your research data – get started with a DMP](#), Thu Nov 19 11:00–12:00
- [Manage your research – KTH Data Repository for research data](#), Thu Nov 26 11:00–12:00
- [Zotero – getting started with reference management software](#), Tue Nov 10 12:15–13:00
- [Reference management in Overleaf and LaTeX](#), Thu Nov 12 12:15–13:00
- [Literature search with smart tools](#), Fri Nov 6 11:00–12:00

These are just a selection – follow the [KTH Library Calendar](#) 
and/or the [KTH Calendar](#)  to see all events.



PhD course on “Data management, ...” 7.5 credits

- <https://www.kth.se/student/kurser/kurs/FCA3009>
- FCA3009 Data management, open science and data infrastructures
- Also, follow the new KTH-initiative “Digital Research Collaborative Hub”!
 - <https://kth.se/datahub>
 - <https://kth-library.github.io/kth-datahub-workspace>



Working with data – file system and naming

- Having a well structured file-system with clear file naming can be a huge time-saver especially during submission and publication.
- An example structure:
 - **PROJECT/data/** – contains all the (relevant) data. Also backed up elsewhere.
 - **PROJECT/code/** – contains all the code, backed up and in a version control system.
 - **PROJECT/code/processing/**
 - **PROJECT/code/paper1/**
 - **PROJECT/scratch/** – bulky (temp) data. Can be regenerated if needed from **code+data**
 - **PROJECT/output/** – final outputs.
 - **PROJECT/output/paper1/**
 - **PROJECT/output/paper1/Figures/** – Figures generated
 - **PROJECT/output/paper1/SourceData/** – Source Data for figures
 - **PROJECT/output/paper1/OpenData/** – Curated and described Dataset to be published



Working with data – file system and naming

Another file system example

project/	
code/	code needed to go from input files to final results
data/	raw and primary data (never edit!)
raw_external/	
raw_internal/	
meta/	
doc/	documentation of the study
intermediate/	output files from intermediate analysis steps
logs/	logs from the different analysis steps
notebooks/	notebooks that document your day-to-day work
results/	output from workflows and analyses
figures/	
reports/	
tables/	
scratch/	temporary files that can safely be deleted or lost
README.txt	file and folder description

From: https://rdmkit.elixir-europe.org/data_organisation



Working with data – file system and naming

- The file name is the main identifier of the file
- Brevity is good, but not at the cost of clarity.
- Tips for naming files
 - Balance with the amount of elements: too many makes it difficult to understand vs too few makes it general.
 - Order the elements from general to specific.
 - Use meaningful abbreviations.
 - Use underscore (_), hyphen (-) or capitalized letters to separate elements in the name. Don't use spaces or special characters: ? ! & , * % # ; * () @ \$ ^ ~ ' { } [] < > .
 - Use date format ISO8601: YYYYMMDD, and time if needed HHMMSS.
 - Include a unique identifier
 - Include a version number if appropriate: minimum two digits (V02) and extend it, if needed for minor corrections (V02-03). The leading zeros, will ensure the files are sorted correctly.
 - Write your file naming convention down and explain abbreviations in your **data documentation**.
 - If you need to rename a lot of files in order to organize your project data and manage your files better, it is possible to use applications like [Bulk Rename Utility](#) (Windows, free) and [Renamer4Mac](#) (Mac).



Working with data – file system and naming

- **Examples of good file names**

- Honeybee project, experiment 2 done in Helsinki, data file created on the 2nd Dec 2020
 - File name: `20201202_HB_EXP2_HEL_DATA_V03.xls`
 - Explanation: `Time_ProjectAbbreviation_ExperimentNumber_Location_TypeOfData_VersionNumber`
- Cropped image of an ant head taken on the third of December 2020 by Meg Megson
 - File name: `20201203_MM_HEAD_CROPPED_V1.psd`
 - Explanation: `Time_CreatorData_TypeModification_Version`
- Mass-spectrometry based proteomics file using TMT10 and HiRIEF for a core facility project in Dec 2019
 - File name: `XC-CLL_20190820_TMT10_setH_325ug_IPG3-10_fr72.raw`
 - Explanation: `ProjectID_Time_Quanttype_SetNo_PeptideAmount_FractionantionType_Fraction`