

# Bring Your Own Data Management Plan

Diana Pilvar  
Data Manager  
Institute of Computer Science  
University of Tartu

Training coordinator  
ELIXIR-Estonia  
[diana.pilvar@ut.ee](mailto:diana.pilvar@ut.ee)

# Schedule

- Start
  - Introduction of resources
  - Initial brainstorming task and QA
  - Planning, storage, organisation, vocabularies
  - Writing session and QA
  - READMEs
- Coffee & Snacks
  - GDPR and ethics
  - Writing session and QA
  - Data sharing and FAIR
  - Writing session and QA
- End

# What is data?

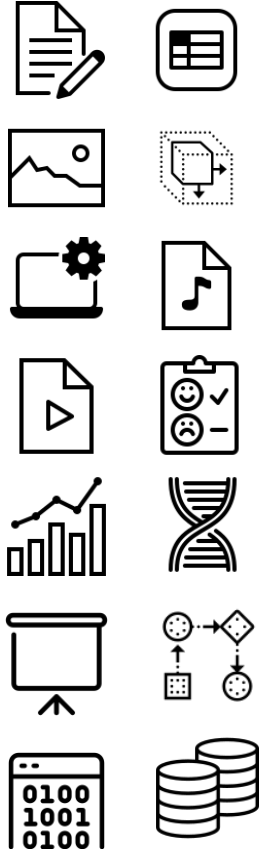
**Data** is collection of discrete values that convey information, describing quantity, quality, fact, statistics, other basic units of meaning, or simply sequences of symbols that may be further interpreted.

Examples: text, spreadsheets, images, 3D models, software, audio files, video files, reports, surveys, patient records, abstract ideas, measurements, statistics, raw biological sequence (DNA, RNA, amino acid), slides, workflows, algorithms, codes, databases

**Data** is all digital resources that have been produced/used during and at the end of the project.

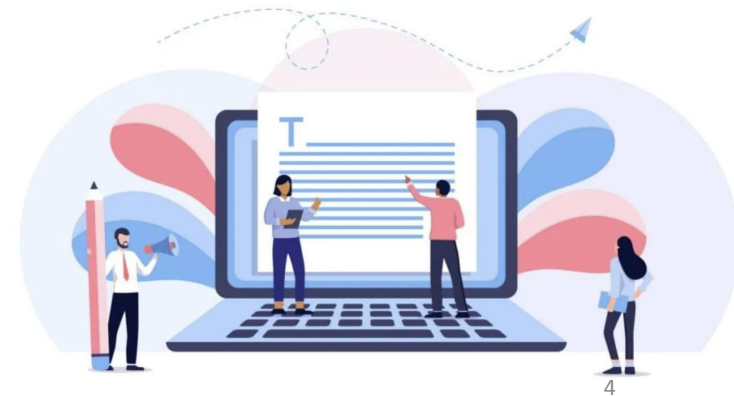
**Metadata** – all of the descriptive information that helps to find/reuse/interpret the digital resource. “Data about data”

Examples: image files contain metadata about the date picture was taken, resolution, size, what equipment was used etc.



# Data Management Plan

- Data management should be planned in the early stages of a research project (before data collection).
- DMP is a **living document** and should be updated as the research project progresses to match e.g. an update of the infrastructures, research softwares or a novel collaboration.
- Depending on the length of the project, DMP should be updated at least once a year



**Funders require data  
management plan**

**Thinking through your  
project helps avoiding risks**

**Plan necessary resources  
and budget**

**Promotes data sharing,  
reuse, and preservation**



**Defined roles and  
responsibilities**

**Everyone is on the same  
page about data handling**

**Onboarding for new  
employees is easier**

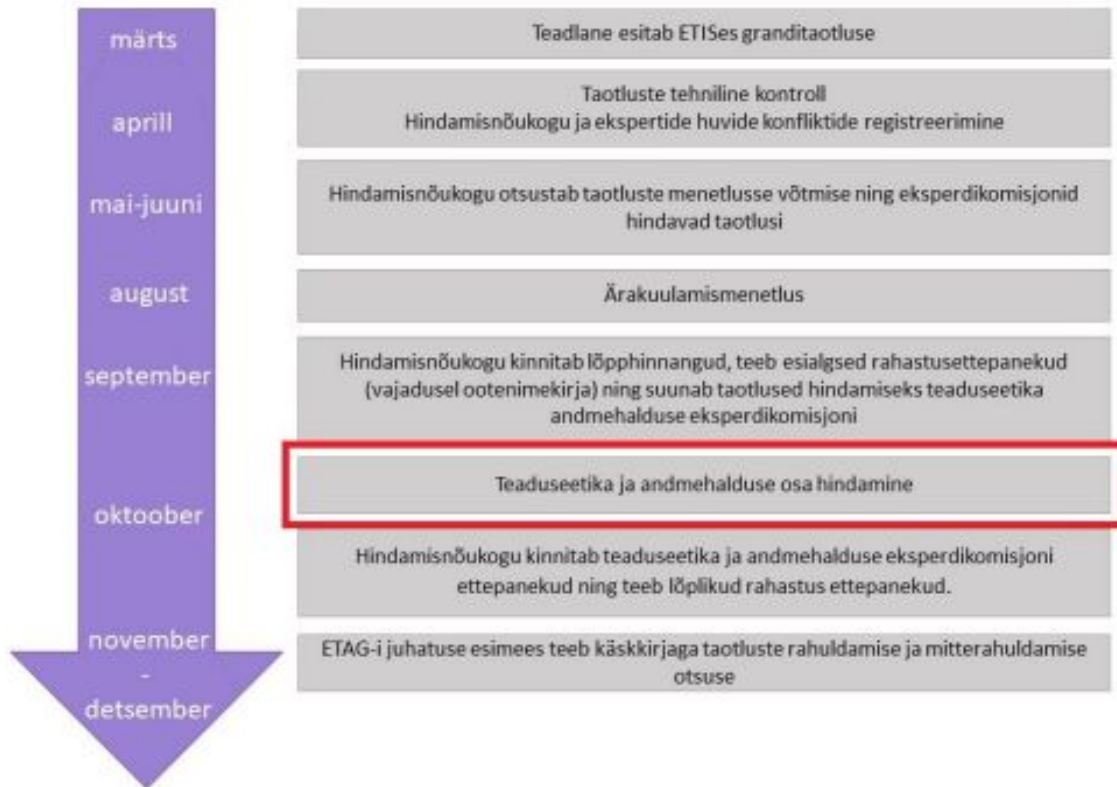
**Students/employees cannot  
leave projects with data**

# Last Year's DMP Evaluation Timing Changes

Previously your funding depended on good DMP evaluation.

Last year initial funding was decided before DMP was assessed.

**This year ERC/ ETAg doesn't even require DMPs.**



# Data life cycle



# Planning





# Data Management Plan writing tools

You can use Microsoft Word or Google Docs, if you wish, but there are better options.

- DMPonline <https://dmponline.dcc.ac.uk/plans>
- Data Stewardship Wizard  
<https://ds-wizard.org/data-management-plans>
- ARGOS <https://argos.openaire.eu/home>

Comparison between the tools <https://ds-wizard.org/comparison>

University of Tartu and TalTech recommend DMPonline. DSW is Elixir tool.

# DMP templates

Depending on your funder you might prefer one of these:

- Digital Curation Centre (DCC template)
  - Favored by ETAg
  - ETAg is not listed in DMPonline - tick the box funder not listed. DCC is default template
- European Commission
  - Horizon 2020
  - Horizon Europe

Always check if your institution has its own template and or policies! At the moment only NICPB has [data management policy](#) (but no template).

# Data management guidances: Estonia

- Taltech DMP guide <https://taltech.ee/en/library/data-management-plan>
- University of Tartu Library DMP guidance  
<https://utlib.ut.ee/en/content/data-management>
- University of Tartu Library course materials (both in Eng and Est):  
<https://sisu.ut.ee/andmekursus>
- Estonian University of Life Sciences:  
<https://library.emu.ee/en/data-management-plans>
- Tallinn University <https://www.tlu.ee/en/research-data>
- ETAg (Ethics and data checklist file)  
<https://etag.ee/en/funding/research-funding/personal-research-funding/call-2025-2/>

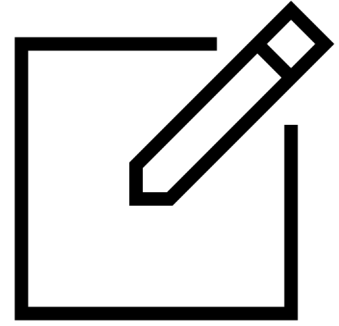
# DMP guidances: International

- European Commission
  - [https://ec.europa.eu/research/participants/data/ref/h2020/grants\\_manual/hi/oa\\_pilot/h2020-hi-oa-data-mgt\\_en.pdf](https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf)
  - <https://enspire.science/wp-content/uploads/2021/09/Horizon-Europe-Data-Management-Plan-Template.pdf>
- Science Europe  
<https://scienceeurope.org/media/411km040/se-rdm-template-3-researcher-guidance-for-data-management-plans.docx>
- The Research Data Management toolkit for Life Sciences <https://rdmkit.elixir-europe.org/>
- Data Management Expert Guide (DMEG) (social science)  
<https://dmeg.cessda.eu/Data-Management-Expert-Guide>
- Digital Curation Centre DMP guidance <https://www.dcc.ac.uk/dmps>
- Making Qualitative Data Reusable - A Short Guidebook For Researchers And Data Stewards Working With Qualitative Data <https://doi.org/10.5281/zenodo.8160880>

# Data management contacts in Estonia

- University of Tartu
  - Research Data Senior Specialist
    - Tiiu Tarkpea [tiiu.tarkpea@ut.ee](mailto:tiiu.tarkpea@ut.ee)
  - Senior Specialist of Data Protection
    - Terje Mäesalu [terje.maesalu@ut.ee](mailto:terje.maesalu@ut.ee)
  - Specialist of Data Protection
    - Tuuli Randver [tuuli.randver@ut.ee](mailto:tuuli.randver@ut.ee)
  - Legal Advisor in Matters of Intellectual Property
    - Reet Adamsoo [reet.adamsoo@ut.ee](mailto:reet.adamsoo@ut.ee)
  - Chief Information Security Officer
    - Risto Rahu [risto.rahu@ut.ee](mailto:risto.rahu@ut.ee)
  - Tartu University Library data administrator, DataDOI
    - Evelin Arust, [evelin.arust@ut.ee](mailto:evelin.arust@ut.ee)
  - HPC Information Security Adviser
    - Tommy Tomson [tommy.tomson@ut.ee](mailto:tommy.tomson@ut.ee)
- Estonian University of Life Sciences
  - Consultations on research data and DMPs
    - Kersti Laupa [kersti.laupa@emu.ee](mailto:kersti.laupa@emu.ee)
- ELIXIR Estonia
  - DMP consultation (Life Sciences)
    - Diana Pilvar [diana.pilvar@ut.ee](mailto:diana.pilvar@ut.ee)
    - Heleri Inno [heleri.inno@ut.ee](mailto:heleri.inno@ut.ee)
- Tallinn University of Technology
  - Head Bibliographer of Data Management
    - Janelle Kirss [janelle.kirss@taltech.ee](mailto:janelle.kirss@taltech.ee)
- Tallinn University
  - Management of research data
    - Kaja Jakobson, [kaja.jakobson@tlu.ee](mailto:kaja.jakobson@tlu.ee)
  - Advice on data management plans
    - Xavier Dubois [xavier.dubois@tlu.ee](mailto:xavier.dubois@tlu.ee)
- Research data repository
  - <https://datadoi.ee/page/mission?locale-attribute=en>

# Brainstorming exercise



# Brainstorming exercise

- What data shall you be using?
- Where are you going to keep the data?
  - Are there any restrictions? E.g., sensitive data
  - What files should be backed-up and how often
  - Version control
  - File formats
  - Do you need extra disc space?
- Sensitive data
  - Who gets access to the data and how is it controlled?
- Where shall you upload the data?
  - Can you even share the data?
  - What repository to use?
  - What license to add?
- How shall you document the process?
  - README file
- Who is responsible for the tasks?
- Do you need to allocate extra money for data management?

[https://docs.google.com/spreadsheets/d/1InmNY0nU67k\\_uwjDN0UZPFz5FIL5E7EsU5y6LAVwECg/edit?usp=sharing](https://docs.google.com/spreadsheets/d/1InmNY0nU67k_uwjDN0UZPFz5FIL5E7EsU5y6LAVwECg/edit?usp=sharing)

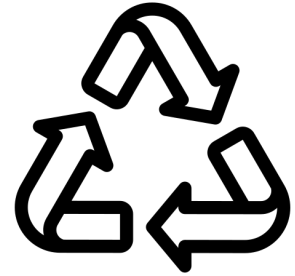
# Collect





# Collect

- Create/collect new data
- or reuse already existing data
  - Which version of existing data?
  - Check licenses and restrictions
  - Check machine readability, interoperability
  - what if the author of the data uploads a new version



# Collect

- Metadata standards

- Will help you describe your data
- By stating that appropriate standards have been used adds credibility to your work.
- Where to find standards:
  - Check the repository you use the most or plan to upload to. See their about page. What standards to they use?
  - Look from:
    - <https://fairsharing.org/>
    - <https://rd-alliance.github.io/metadata-directory/standards/>
    - <https://www.dcc.ac.uk/guidance/standards/metadata/list>
- Domain agnostic: Dublin Core Metadata Standards, Data Documentation Initiative
- Bio: Darwin Core, Bioschema.org
- Humanities: Text Encoding Initiative

# Administrative metadata

- Dublin Core Metadata generator
- [https://nsteffell.github.io/dublin\\_core\\_generator/](https://nsteffell.github.io/dublin_core_generator/)
- Short (15 fields) and long (>20 fields)
- Output: XML, HTML

## Item submission

Describe Access Upload Review CC License License Complete

### Describe Item

#### Author(s): \*

Last name

First name

Add

The name(s) of the author(s) of this dataset.

#### Title: \*

The title of the dataset.

#### Date of Issue: \*

Year

Month

Day

You can leave out the day and/or month if they aren't applicable.

#### Publisher: \*

An entity responsible for making the resource available. For example University of Tartu, Tallinn University, an institute or a department etc.

#### Abstract: \*

Add

Give a short abstract of the dataset.

# Data organisation - The file catalogue

The foundation to good data organisation is well organised files.

- Easy to understand where to find information- Intuitive
- Planned early
- Consistency - Assign a person to this
- Self-explanatory names
- Unique names
- Add README file

# The file catalogue

## A) Organized by File type

### Dataset.A

- Code
  - Step.1
  - Step.2
- Data
  - Processed
  - Raw
- Results/
  - Figure.1
  - Figure.2
- Models
- readme.txt

## B) Organized by Analysis

### Dataset.B

- Figure.1
  - Code
  - Data
  - Results
- Figure.2
  - Code
  - Data
  - Results
- Table.1
  - Code
  - Data
  - Results
- readme.txt

```
graph TD
    ENBIOproject[ENBIOproject] --> Data[Data]
    ENBIOproject --> ConsumerSurvey[ConsumerSurvey]
    ENBIOproject --> StakeholderSurvey[StakeholderSurvey]
    ENBIOproject --> Documentation[Documentation]
    ENBIOproject --> Methodology[Methodology]
    ENBIOproject --> Method_ConsumerSurvey[Method_ConsumerSurvey]
    ENBIOproject --> Method_StakeholderSurvey[Method_StakeholderSurvey]
    ENBIOproject --> Questionnaires[Questionnaires]
    ENBIOproject --> QuestionnaireConsumerSurvey[QuestionnaireConsumerSurvey]
    ENBIOproject --> QuestionnaireStakeholderSurvey[QuestionnaireStakeholderSurvey]
```

What is an alternative  
Approach for this case?

**Dryad FAIR Data** practices to organize files in a logical schema.

[https://datadryad.org/stash/best\\_practices#organize](https://datadryad.org/stash/best_practices#organize)

<https://www.essda.eu/Training/Training-Resources/Library/Data-Management-Expert-Guide/2.-Organise-Document/File-naming-and-folder-structure>

## 5 levels is good practice

### Projects

.....Project1\_SMPHS

.....01\_Admin\_doc

Grant documents, financial documents, DMP

.....02\_Research / 02\_Phase1

.....01\_Raw\_data

At least one copy of raw data, read-only

.....02\_Analysis

All the analysis and results, along with the README file about the analysis process

.....03\_Figures

Can be separate folder or in analysis folder

.....04\_Dissemination

Articles, presentations, reports

.....03\_Additional\_materials

Articles connected with the research

.....README.txt

How the project is organised, **how to name files**, who to contact if there are any problems

.....Project2\_GTYS

- 00 [redacted]
- 00 LEPING
- Ajutine kirjutamine
- Articles and reports
- Deliverables
- [redacted] MINISTRI\_KASKKIRI
- [redacted] MINISTRI\_KASKKIRI- [redacted]
- [redacted] MUUDATUS\_MINISTRI\_KASKKIRI
- [redacted] MUUDATUSE\_EELNOU [redacted]
- Ettevalmistus
- Fotod
- [redacted] Eetikakomitee load
- GRAND\_PLAN
- JAGATUD\_MATERJALID
- Koduleht

- Koosolekud
- Majutus ja haldus eelarvestamine
- Meeskond/Töökuulutused
- Mudelid
- [redacted] detailanalüüs
- [redacted] genereerimine
- [redacted] ettevalmistusnädal
- Presentations
- Projektijuhi materjalid
- [redacted]
- [redacted]
- Uudiskiri
- UX UI topics
- Mittefunktsionaalsed nõuded 👤
- Nõuded rakendustele ja riskid.docx 👤

- Documentation
- Joonised
- Muud süsteemid
- 
- SWOT
- TEHIK
- Testimine
- Tööstusmagistrantide testülesanded
- Arendatavate süsteemide kirjeldused
- Arvutuskeskond
- \_eelarve
- \_eelarve\_partneritega
- \_eelarve\_partneritega
- \_eelarve\_partneritega\_TAIELIK.xlsx
- \_eelarve\_partneritega.xlsx
- tulemuste vahendamise teenus

- 
- kodeerimiskeskus
- riskiskooride tulemuste vahendamise teenus
- \_eelarve\_ \_veidi\_vale.xlsx
- tegevuskava
- Küsimused lahendamiseks
- Mudelite repositoorium
- Nõuded riskimudeli tarkvarale
- 
- Online koodikeskus
- tegevuskava\_kinnitatud juhtrühmas
- tegevuskava\_kinnitatud juhtrühmas\_
- 
- Sündmuste ja tegevuste logi
- Tulemuste IS
- Uldskeem
- Üldine



# Data organisation - Naming the files

## File name

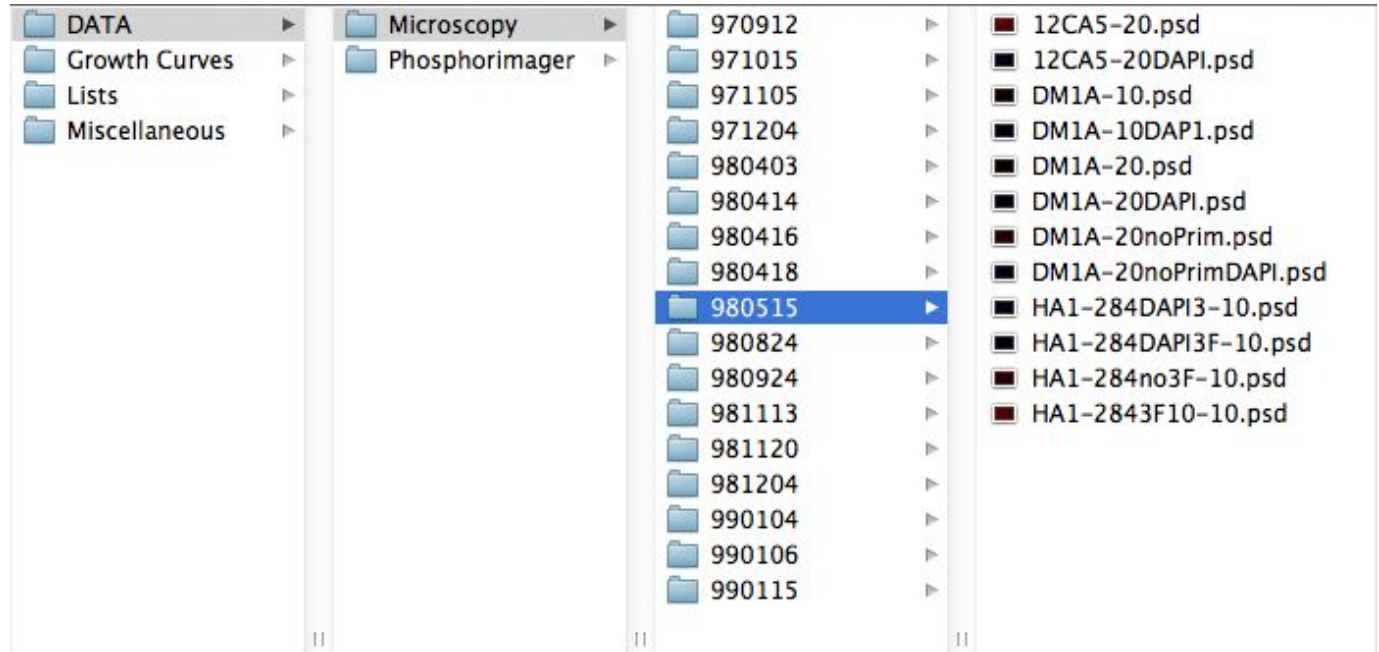
- Machine readable, human readable and plays well with default order of the files
- About 25 characters long
- Use abbreviations if possible (clarifications in the README file)
- Use standard abbreviations for countries, languages, units, methods
- Use “\_” or “ - ” instead of spaces
- Don't use extra symbols (&%€!)
- Add version number, date of creation, name of creator

TIME\_DocumentTitle\_Version

PREFIX\_DocumentTitle\_Version\_Time\_PersonUpdating

# Example

Science data librarian  
Amy Hodges example



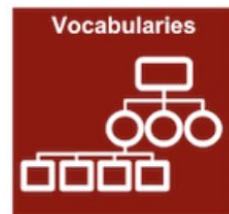
<https://guides.library.stanford.edu/data-best-practices/name-files>

# Vocabularies

Used vocabularies should be noted and linked. This information should be easily found and should be near the dataset.

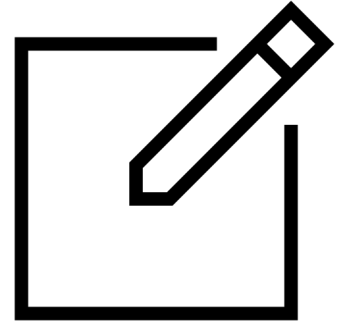
Resources for vocabularies:

- <https://fairsharing.org/>
- <https://www.ebi.ac.uk/ols/index>



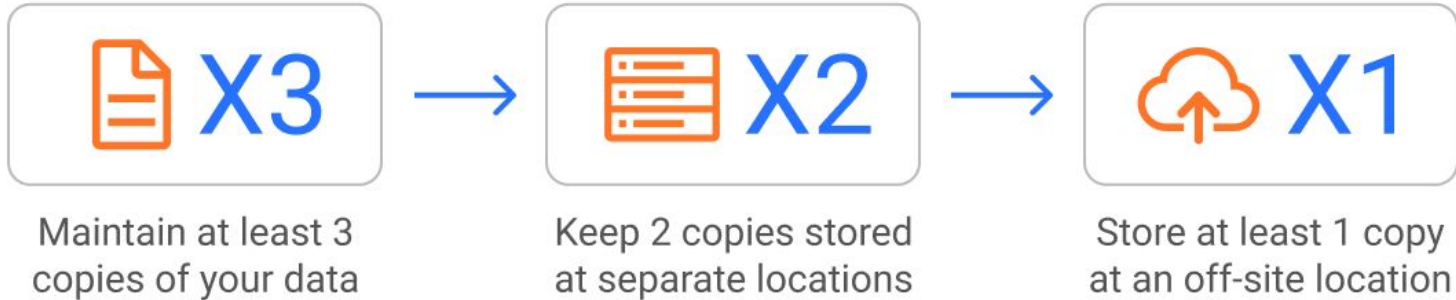
Tip! Find a database/repository in your field and check what vocabularies and standards they use.

# Writing session



# Backing up data

## 3-2-1 Backup Rule



**It is a good practice to have a copy of the original raw data in a separate location, to keep it untouched and unchanged (not editable)**

# Data Storage

- Know the volume and format of your data. How long will you store raw data, pipelines, analysis workflows?
- Access rights during project
- Data transfer.

If the data comes from an external facility or needs to be transferred to a different server, you should think about an appropriate and safe data transfer method

[Tartu Ülikooli teadusarvutuste keskuse infoturbe nõunik Tommy Tomson: "Andmete varundamisest"](#)

# Process



# Version control

Version control – giving **provenance** to your work

- Know who and when made the edits (think of co-authored paper)
- Can always go back to an earlier version
- Enables multiple persons to work on the same file simultaneously
- Manual version control ( $v1.1 \rightarrow v1.2$  /  $v1 \rightarrow v2$ ; *add name of the last editor*)
- Automatic version control (Zenodo, Git, Datalad, iRODS)





## Git (GitHub and GitLab)

Git is a **free and open source** distributed version control system designed to handle everything from small to very **large projects** with speed and efficiency.

Git is easy to learn and has a tiny footprint with lightning fast performance. Usually used for software development, but you can use it for whatever you want.

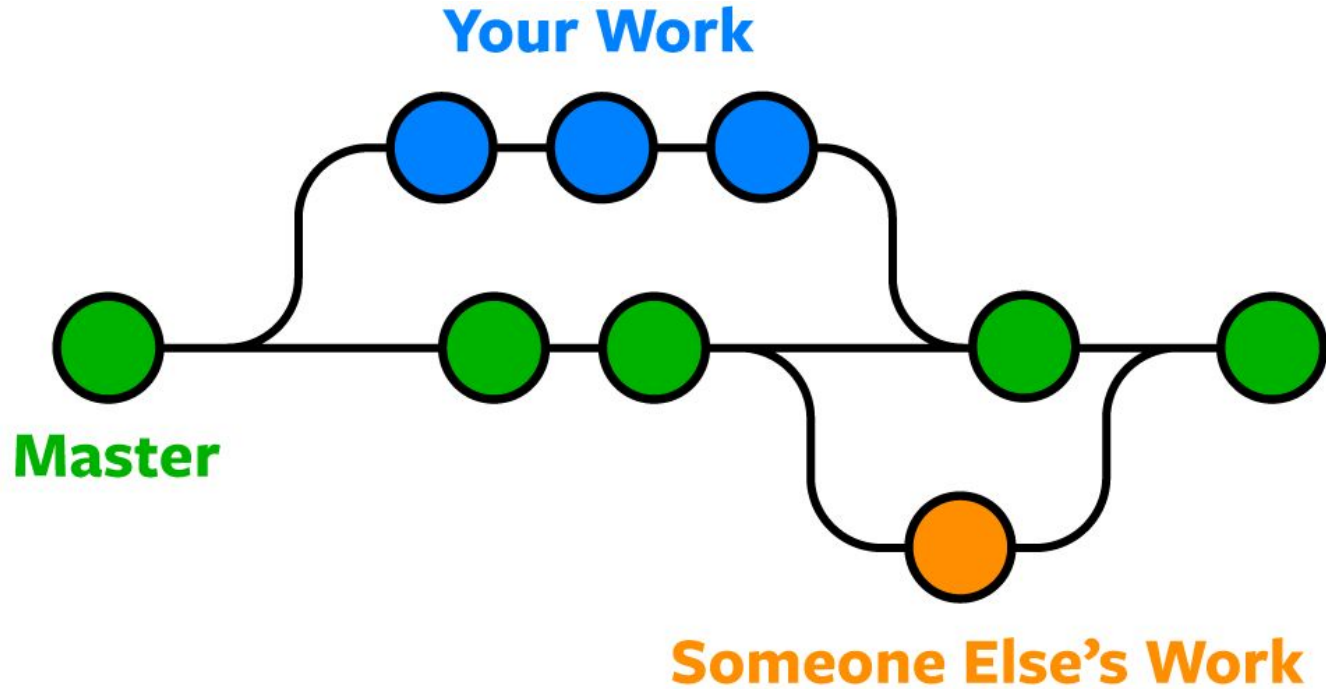
Many people can edit the files simultaneously, version control, backing up the data. However, does not work for all kinds of data.

GitLab for University of Tartu

NB! Git does not give out DOI! You need to connect it with a repository (for example Zenodo) to get a DOI.



# Versioning: Git



# Data quality and consistency

- Assign tasks
  - who will generate data, analyse data and manage data;
- for data collection: define data dictionary
- automated quality monitoring through tools, pipelines, dashboards;
- training of study participants and researchers, surveyors or other staff involved;
- adopting standards;
- instrument calibrations;
- post collection data curation;
- Data entry validation

# Analyze and Process

Write everything down in README file



# README file

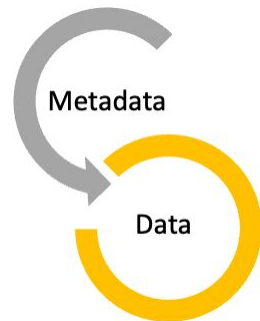
Document describing the dataset (or any other file).

You can use standards and make it machine-readable (Git, Markdown), but it is not requirement. A simple \*.txt file is perfect.

Name should be: README.txt (or README.md)

Best practices on writing a README:

- Metadata and the dataset are usually in a separate file.
- It is important to connect them to each other (link).
- Each dataset get its own README file
- Be consistent
- Standardized date formats, ISO8601 (YYYY-MM-DD)
- Follow scientific conventions for your discipline (taxonomy, geospatial names, keywords)



# README file

## Administrative metadata / general information

- Project name, nr, grant, funder
- Authors, Creators (institutions, address)
- List of included/associated files
- License
- Citation
- Keywords

## Descriptive metadata

- Dataset authors and headlines
- Abstract
- Related publications
- Used protocols
- Used equipment



# README file

## Structural metadata

- Methods of data collection (references)
- Methods for data analysis (references)
- Any instrument/software specific information (also versions)
- Standards and calibrations used
- Quality assurance process
- *Symbols/codes used (low quality/outliers)*
- Persons involved in each of the processes
- File structure and file naming conventions

## Data-specific information (data vocabulary)

- Variable list (full names, definitions, column headers)
- Units of measurements
- Codes/symbols for missing data
- Abbreviations used








# Example 1

<http://dx.doi.org/10.23673/re-300>

### HPC Cloud traces for better cloud service reliability

Dehury, Chinmaya Kumar; Chhetri, Tek Raj; Lind, Artjom; Srirama, Satish  
Narayana; Fensel, Anna

📄 Export ▾

Name	Size	Description
 README.txt	1.446Kb	Information about this data, project, and contributors.
 Anonymised_data_V1.zip	108.3Mb	Contains the actual data in CSV and the code to anonymise the sensitive data.
 README_DATA.txt	2.665Kb	Readme for the actual data present in "Anonymize_data_V2" zip file.
 Source_code_V1.zip	10.28Kb	Contains the source code used in the paper "A Combined Metrics Approach to Cloud Service Reliability using Artificial Intelligence".
 README_SOURCE_CODE.txt	2.200Kb	Readme for the source code present in "Source_code_V1" zip file.

No Thumbnail

#### Date

2021

#### URI

<https://datadoi.ee/handle/33/425>

<http://dx.doi.org/10.23673/re-300>

#### Metadata

[Show full item record](#)

#### Abstract

This data is in support of the research on "A combined system metrics approach to cloud service reliability using artificial intelligence" (doi: 10.20944/preprints202111.0548.v1)

#### Keyword

Failure Prediction; Fault-tolerance; Cloud computing; Artificial Intelligence; Reliability

#### Item type

info:eu-repo/semantics/dataset

#### Collections

[Arvutiteaduse andmed](#)



# Example 1

This repository contains necessary data and code to reproduce the result provided in "A Combined Metrics Approach to Cloud Service Reliability using Artificial Intelligence" paper.

This paper is under consideration to publish.

## Files & Folders

=====  
Anonymised\_data\_V1.zip : Contains the actual data in CSV and the code to anonymise the sensitive data.  
README\_DATA.txt : Information about the actual data present in "Anonymize\_data\_V2" zip file.  
Source\_code\_V1.zip : Contains the source code used in the paper "A Combined Metrics Approach to Cloud Service Reliability using Artificial Intelligence".  
README\_SOURCE\_CODE.txt : Information about the source code present in "Source\_code\_V1" zip file.

## Paper Under consideration

=====  
<https://www.preprints.org/manuscript/202111.0548/v1>

## Citing

=====  
If you find this code and data useful, please consider citing:

```
---latex
---
@article{chhetri2021combined,
  title={A Combined Metrics Approach to Cloud Service Reliability using Artificial Intelligence},
  author={Chhetri, Tek and Dehury, Chinmaya Kumar and Lind, Artjom and Srirama, Satish Narayana and Fensel, Anna},
  year={2021},
  publisher={Preprints}
  doi={\url{https://www.preprints.org/manuscript/202111.0548/v1}}
}
---
```

## Contact

=====  
The corresponding author is  
Chinmaya Kumar Dehury,  
Assistant Professor of Distributed System  
Mobile & Cloud Lab  
Institute of Computer Science  
University of Tartu  
<https://kodu.ut.ee/~dehury/>

# Example 3

<https://doi.org/10.5281/zenodo.4542915>

February 16, 2021

Dataset Open Access

## Analytic Electron Density Representation for Electron Density Learning

Bruno Cuevas-Zuñiría

Analytic Electron Density Representation for Electron Density Learning

Electron density learning is a new path for machine-learning in chemistry. In this dataset, we provide a benchmark of analytic wave-functions of small organic molecules that can be used to teach machine learning-algorithms. All densities have been computed at the DFT level (wB97X/6-31+G(d)).

The readme.md includes information about how to read it.

Note: The all20set is not uploaded because of its size, exceeding the 50GB. If you need it, you can contact us at [bruno.czuviria@upm.es](mailto:bruno.czuviria@upm.es)

Preview

Files (10.3 GB)

Name	Size	
<a href="#">dimers.index</a>	42.2 kB	<a href="#">Download</a>
md5:29c4617a9414184489ba2e9500aef4c1		
<a href="#">dimers.wfn.h5</a>	1.3 GB	<a href="#">Download</a>
md5:3277fd0bc2de9270c17c6e4f01c2c87a		
<a href="#">gdb7.index</a>	83.0 kB	<a href="#">Download</a>
md5:4ef39e62678cbb4e46f9e214ca58ae5c		
<a href="#">gdb7.json</a>	98 Bytes	<a href="#">Preview</a> <a href="#">Download</a>
md5:991aeeab7c125171a0992e46336bab8		
<a href="#">gdb7.wfn.h5</a>	3.6 GB	<a href="#">Download</a>
md5:e55d22db91a4a526d4869c99681d7bce		
<a href="#">gdb8.index</a>	123.4 kB	<a href="#">Download</a>
md5:5a7328caab462b24bccb51578f11b67		
<a href="#">gdb8.wfn.h5</a>	5.2 GB	<a href="#">Download</a>
md5:f55cd0811a359f1dbe0f60dc3002aeae		
<a href="#">README.md</a>	1.9 kB	<a href="#">Preview</a> <a href="#">Download</a>
md5:83d05bca7849e410db7fa1ac90264f9		

50

views

64

downloads

[See more details...](#)

Indexed in

OpenAIRE

Publication date:

February 16, 2021

DOI:

DOI: [10.5281/zenodo.4542915](https://doi.org/10.5281/zenodo.4542915)

Keyword(s):

[Quantum-Chemistry](#) [Machine-learning](#) [DFT](#)

License (for files):

[Creative Commons Attribution 4.0 International](#)

Versions

Version 1

Feb 16, 2021

[10.5281/zenodo.4542915](https://doi.org/10.5281/zenodo.4542915)

Cite all versions? You can cite all versions by using the DOI [10.5281/zenodo.4542914](https://doi.org/10.5281/zenodo.4542914). This DOI represents all versions, and will always resolve to the latest one. [Read more.](#)

Share



Cite as

Bruno Cuevas-Zuñiría. (2021). Analytic Electron Density Representation for Electron Density Learning [Data set]. Zenodo. <https://doi.org/10.5281/zenodo.4542915>

Start typing a citation style...

# Example 3

```
1 # Analytic Electron Density Representation for Electron Density Learning
2
3 ## Summary
4
5 **The peptide dataset won't be uploaded to the repository because of its size. We'll provide it upon request**
6
7 Electron density learning is a new path for machine-learning in chemistry. In
8 this dataset, we provide a benchmark of analytic wave-functions of small
9 organic molecules that can be used to teach machine-learning algorithms.
10 All the densities have been computed at the DFT level (wB97X/6-31+G(d)).
11
12 The dataset contains four parts:
13
14 - Small organic molecules: "GDB7/GDB9".
15 - Capped peptides: "all20set"
16 - Dimers: "s66x8"
17 - Single-molecule drugs: "fdaset".
18
19 Our learning experiments use a combination of all these data excepting the fdaset, which serves as validation set.
20
21 ## How to read it?
22
23 The wave-functions are available in HDF5 format, which is handy for loading
24 large arrays of numbers. However, it requires the user to understand the
25 structure of data. All HDF5 files are structured as:
26
27 name of the molecule
28   attributes
29     Number of atoms -> (integer)
30     Number of primitives -> (integer)
31     Total charge -> (integer)
32     coordinates (in Bohrs) -> (real matrix)
33     atomic symbols (H, C, ...) -> (char vector)
34     symmetry index -> (integer vector)
35     primitive centre index -> (integer vector)
36     exponents -> (real vector)
37     occupancies -> (real vector)
38     density matrix -> (real matrix)
39
40 We store density matrix instead of the orbital coefficients because it saves
41 computing time at electron density computing. Therefore, all orbital
42 information is lost.
43
44 A2MD library allows to interact easily with these files and to compute
45 electron density in training time using PyTorch and CUDA acceleration.
46
47 ## Contact information
48
49 Don't hesitate to contact us! And please, cite us:
50
51 bruno.czuviria@upm.es
52
53
54
```

# Example 4

*This DATSETNAMereadme.txt file was generated on YYYY-MM-DD by NAME  
<help text is included in angle brackets, and can be deleted before saving>*

## *GENERAL INFORMATION*

### *1. Title of Dataset:*

### *2. Author Information*

#### *A. Principal Investigator Contact Information*

*Name:*

*Institution:*

*Address:*

*Email:*

#### *B. Associate or Co-investigator Contact Information*

*Name:*

*Institution:*

*Address:*

*Email:*

#### *C. Alternate Contact Information*

*Name:*

*Institution:*

*Address:*

*Email:*

### *3. Date of data collection (single date, range, approximate date) <suggested format YYYY-MM-DD>:*

### *4. Geographic location of data collection <latitude, longitude, or city/region, State, Country, as appropriate>:*

### *5. Information about funding sources that supported the collection of the data:*

# Example 4

## *SHARING/ACCESS INFORMATION*

- 1. Licenses/restrictions placed on the data:*
- 2. Links to publications that cite or use the data:*
- 3. Links to other publicly accessible locations of the data:*
- 4. Links/relationships to ancillary data sets:*
- 5. Was data derived from another source? yes/no*  
*A. If yes, list source(s):*
- 6. Recommended citation for this dataset:*

## *DATA & FILE OVERVIEW*

- 1. File List:*  
*<list all files (or folders, as appropriate for dataset organization) contained in the dataset, with a brief description>*
- 2. Relationship between files, if important:*
- 3. Additional related data collected that was not included in the current data package:*
- 4. Are there multiple versions of the dataset? yes/no*  
*A. If yes, name of file(s) that was updated:*
  - i. Why was the file updated?*
  - ii. When was the file updated?*

# Example 4

## METHODOLOGICAL INFORMATION

1. Description of methods used for collection/generation of data:

<Include links or references to publications or other documentation containing experimental design or protocols used in data collection>

2. Methods for processing the data:

<describe how the submitted data were generated from the raw or collected data>

3. Instrument- or software-specific information needed to interpret the data:

<include full name and version of software, and any necessary packages or libraries needed to run scripts>

4. Standards and calibration information, if appropriate:

5. Environmental/experimental conditions:

6. Describe any quality-assurance procedures performed on the data:

7. People involved with sample collection, processing, analysis and/or submission:

## DATA-SPECIFIC INFORMATION FOR: [FILENAME]

<repeat this section for each dataset, folder or file, as appropriate>

1. Number of variables:

2. Number of cases/rows:

3. Variable List:

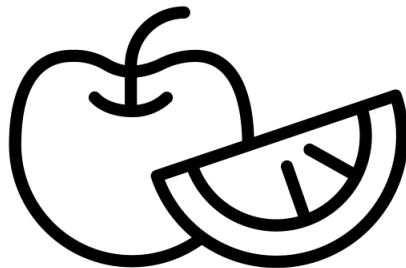
<list variable name(s), description(s), unit(s) and value labels as appropriate for each>

4. Missing data codes:

<list code/symbol and definition>n

5. Specialized formats or other abbreviations used:

Coffee  
Tea  
Snacks



Personal and  
sensitive data





# Personal Data under the GDPR



[youtube link](#)

# Important documents: Laws EU and Estonia

- EU: General Data Protection Regulation (Euroopa Liidu isikuandmete kaitse üldmäärus)
  - <https://gdprinfo.eu/> (English and Estonian)
- EST: Personal Data Protection Act (Isikuandmete kaitse seadus)
  - <https://www.riigiteataja.ee/akt/104012019011?leiaKehtiv> (Estonian)
  - <https://www.riigiteataja.ee/en/eli/507112023002/consolide> (English)
- EST: Data Protection Inspectorate (Andmekaitse Inspeksioon)
  - [Isikuandmete töötleja üldjuhend](#)
  - [Andmekaitse ABC](#)
  - [Vastutav ja volitatud töötleja](#)

# (Sensitive) personal data in GDPR

Personal data: Any information that relates to an identified or identifiable **living** individual (not a legal entity). **The processing of special category personal data is prohibited by default** unless one of the conditions set out in Article 9(2) GDPR is met.

## Regular Personal data (Isikuandmed)

- Name
- Address
- Phone nr
- Bank / Credit card nr
- Email address
- IP address
- Cookies
- Online identifiers
- Rare occupation

## Special category personal data (Erilisi isikuandmed)

- Biometric data
- Genetic data
- Health data, disability
- Race and ethnicity
- Political opinions
- Philosophical views
- Religious beliefs
- Trade union membership
- Sexual orientation
- Sex life

## Not listed in GDPR: Sensitive data (Tundlikud andmed)

- Confidential data (industry)
- Ecological data about rare and endangered species

# Personal data and processing

viewing

collection

consultation

retrieval

recording

storage

granting  
access

erasure or  
destruction

alteration  
(incl. adding  
a grade)

disclosure

communication

closure

The processing of personal data is any operation with personal data from the moment it is received to the moment it is destroyed, whether digitally or on paper.

12

# GDPR

Article 13. When the personal data is being collected, at the time of the collection, the data controller needs to provide the data subject with the following information (contact of the controller, purpose, data, transfer information, etc.).

Article 17. Right to erasure ('right to be forgotten').

Article 25. Data protection by design and by default.

Article 30. Records of processing activities.

Article 32. Security of processing.

Article 83. General conditions for imposing administrative fines. The total amount of the administrative fine shall not exceed the amount specified for the gravest infringement. Depends on what the infringement was, the fine can go up to 10 000 000 EUR / 20 000 000 EUR or up to 2% / 4% of the total worldwide annual turnover of the preceding financial year, whichever is higher.

# Principles of personal data processing

## Lawfulness

- Processing is lawful only when there is a valid legal basis

## Purpose limitation

- Data can be processed collected for specified purposes only

## Minimisation

- Collecting as little data as possible for the set purpose. It is not allowed to collect additional data just in case

## Storage limitation

- Obligation to store data until the purpose has been fulfilled, and arising from the law

## Integrity and confidentiality

- Unauthorised and unlawful use (sending or sharing information about someone by someone not authorised to do so)

# UniTartu: Researchers are invited to enter data on research projects that involve personal data processing

<https://siseveeb.ut.ee/en/announcement/researchers-are-invited-enter-data-research-projects-involve-personal-data-processing>

**From 1 October, researchers are invited to enter data on planned or ongoing research projects which involve the processing of personal data on the webpage [ak.ut.ee](https://ak.ut.ee).**

The principal investigator or an employee appointed by the principal investigator, who must have the university's user account, must enter data about the research project involving the processing of personal data on this webpage **before they start to collect the data**. They must complete a 14-question form, which takes about 15–30 minutes, depending on the research project. Data must be entered about all research carried out while employed by the university, regardless of where the funding comes from or whether the ethics committee approval is sought. The research team can modify or supplement the entered data later, as well as add links to agreements or documents related to the research in the document management system (DHIS).

For any questions about entering data on research studies, please contact Tuuli Randver ([tuuli.randver@ut.ee](mailto:tuuli.randver@ut.ee), 737 5139). You may also invite a data protection specialist to your unit to help enter the data.

# (Sensitive) personal data

Data security must be implemented for all types of data, whether in electronic form, on paper, or in any other format. The objectives of data security are:

- Prevent unauthorized access to data processing equipment.
- Prevent unauthorized reading, copying, deletion, etc., of data.
- It must be possible to retroactively identify who accessed, stored, modified, or deleted what, and to whom, when, and why personal data was transmitted.
- Ensure that each person has access only to the data and processing methods allowed for their role.
- Organize work to meet data protection requirements.
- Provide appropriate training for individuals processing personal data under one's supervision.



# University of Tartu intranet

## Isikuandmete töötlemine

## Personal data processing

### Consent

#### ▼ Isikuandmete töötlemine

- Meelespea
- KKK
- Andmekaitse koolitusmoodul
- Andmekaitsest ülikoolis
- Andmekaitse teadustöös
- Andmekaitse ABC
- Isikuandmete töötlemine teadusuuringutes
- Isiku andmete väljastamine
- Isikuandmete töötlemine üliõpilaste lõputöös
- Juhendid ja õigusaktid
- Juhised isikuandmete kaardistajale
- Juurdepääsupiirangu seadmine kirjavahetuse registreerimisel
- Koostöövõrgustiku liikmed
- Loengus ja praktikumis filmimine
- Lõputööde kaitsmised ja isikuandmed
- Mõisted
- Mõjuhinang
- Nõusolek ja näidistekstid
- Pseudonüümide kasutamine
- Rikkumisest teavitamine
- Täiendusõppe isikuandmed
- Ürituste korraldamine ja listikirjad
- Õiguslik alus andmete väljastamiseks

#### ▼ Personal data processing

- Cameras in classroom
- **Consent**
- Data protection at the university
- Data protection in research
- Definitions
- FAQ
- Impact assessment
- Memorandum
- Organising events and sending emails to mailing lists
- Personal data breach
- Personal data in continuing education
- Personal data processing in research
- Pseudonymised data
- Releasing personal data
- Thesis defences and personal data

# Data protection impact assessment (DPIA)

Where **a type of processing** in particular using new technologies, and taking into account the nature, scope, context and purposes of the processing, **is likely to result in a high risk** to the rights and freedoms of natural persons, the controller shall, prior to the processing, carry out an assessment of the impact of the envisaged processing operations on the protection of personal data. A single assessment may address a set of similar processing operations that present similar high risks.

- <https://gdprinfo.eu/en-article-35>
- [University of Tartu intranet page on Impact assessment with an example](#)
- [How to conduct a Data Protection Impact Assessment \(template included\)](#)
- [Template](#)
- [GDPR DPIA Example: Perfect Examples of DPIAs](#)
- [Andmekaitse inspeksioon: Mõjuhindang: tüütu kohustus või elupäästev tööriist?](#)

# DPIA

concrete examples of the types of conditions that would require a DPIA:

- If you're using new technologies
- If you're tracking people's location or behavior
- If you're systematically monitoring a publicly accessible place on a large scale
- If you're processing personal data related to “racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation”
- If your data processing is used to make automated decisions about people that could have legal (or similarly significant) effects
- If you're processing children's data
- If the data you're processing could result in physical harm to the data subjects if it is leaked

# 10 steps towards privacy compliance in research

<https://doi.org/10.5281/zenodo.10417514>

## 10 steps towards privacy compliance in research



### Keep the GDPR in mind when designing your research

Do you need to collect personal data, why, and how much?

### Make sure you have a legal basis to use personal data

E.g., public interest or consent

### Arrange ethics review

Ethics review makes sure that you have also taken ethical implications into account

### Document privacy risks and privacy-related decisions

E.g., in a Data Management Plan, privacy scan, or Data Protection Impact Assessment

### Inform participants properly

E.g., in an information letter, oral script, privacy statement

### Protect your data with organisational measures

E.g., access control, agreements with external parties, data protection policies, researcher training

### Enable participants to exercise their rights

E.g., right to data access, correction, objection, erasure

### Protect your data with technical measures

E.g., anonymise, pseudonymise, encrypt your data, use safe storage

### FAIR data: balance risks and Open Science principles

E.g., share under restricted access, or only share metadata and materials

### Ask for help when you need it!

Contact your privacy officer or data steward for support

# University of Tartu: Intellectual property legislation, documents

## Regulations and documents:

### Regulation for Processing Development Projects

- 9.7. deed of assignment of intellectual property rights that are created in the course of the project to the university.
  - 10. 10. By virtue of the deed of assignment, referred to in clause 9.7, the principal investigator and other implementers of the project assign to the university the economic rights to the intellectual property they develop as a result of the project, including the right to apply for a patent or to register a utility model and to become the owner of the patent or utility model. The requirement to sign the deed of assignment does not apply to employees of the university whose employment contract includes an agreement about the assignment of economic rights.
  - NB! Students don't have employment contract but still need to sign the deed of assignment
- Procedure for Managing Intellectual Property Created at the University of Tartu
  - Contact Technology Transfer Specialist Alina Paas [alina.paas@ut.ee](mailto:alina.paas@ut.ee)

# Instructions for Using Intellectual Property Rights

Excerpts from the intellectual property rights instructions conducted by UT lawyer Reet Adamsoo. These excerpts are recommended to use in data management plan:

- The data belong to the University of Tartu. Persons employed for filling the grant will assign the proprietary rights to the results of the research (including the data) performed under the grant agreement to the University with the Employment Contract (academic employees) or with another written document (Act of Assignment of the Intellectual Property Rights)
- Data will be disclosed under the Creative Commons license [CC-BY 4.0](#)
- A third party, whose data have been used for creating the results of the grant, may set restrictions to the usage of the data. In this case those restrictions must be considered while the data are being licensed, i.e. the university can give the license for the data usage only in the scale of rights allowed by the third person (i. e. the scale of rights that university has received from the third persons)
- If the University or a third person, whose data have been used for creating the results of the grant, wants to submit a patent or a utility model application, the publishing of the data has to be postponed until the submission of the application

# Examples to be inspired by

We are not the owners of the data. The owners have the permission to share this data with us for the purpose of this work. Since we are not the owners of the data, we are not allowed to preserve the data for future purposes. All the received data is pseudonymised and we do not have to key to de-pseudonymised the data (the data owners will have that information). The transfer of the data will follow all the required laws and will be handled with care and security.

For all the data involved, we have obtained Estonian Ethic Committee allowance to analyse the data. All the sensitive data will be held in secure virtual servers, managed by HPC. All the raw data concerning this project will be deleted by the HPC at University of Tartu at the end of the permission to use the data. The access to the data is controlled. Only the HPC system administrator, the PI and the researchers who need, are given access to the data.

We will abide by the GDPR and IKS rules when dealing with all the data.

We will follow the Estonian Code of Conduct for Research Integrity.

# Important documents: University level

- University of Tartu
  - [Instructions for applying the Code of Conduct for Research Integrity](#)
  - [Guide for data protection in research](#)
- University of Tallinn
  - [Andmekaitse juhend: „Isikuandmete töötlemine teadustöös“](#)
- [https://elixir.ut.ee/data\\_management/sensitive-data/](https://elixir.ut.ee/data_management/sensitive-data/) (English and Estonian)



# Important documents: Ethics and Research Integrity

- EU: The European Code of Conduct for Research Integrity (Euroopa teaduse eetikakoodeks)
  - <https://www.akadeemia.ee/wp-content/uploads/2024/01/euroopa-teaduse-eetikakoodeks.-kolmas-taiendatud-versioon.-16.01.24.pdf> (Estonian)
  - <https://allea.org/wp-content/uploads/2023/08/Feedback-to-Stakeholders-on-2023-ECoc-Revision-1.pdf> (English)
- EU: Ethics in Social Science and Humanities
  - [https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/guidance/ethics-in-social-science-and-humanities\\_he\\_en.pdf](https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/guidance/ethics-in-social-science-and-humanities_he_en.pdf) (English)

# Important documents: Ethics and Research Integrity

- Estonian Code of Conduct for Research Integrity (Hea teadustava)
  - [https://ut.ee/sites/default/files/inline-files/code\\_of\\_conduct\\_for\\_research\\_integrity\\_eng\\_1.pdf](https://ut.ee/sites/default/files/inline-files/code_of_conduct_for_research_integrity_eng_1.pdf) (English)
  - <https://www.etag.ee/wp-content/uploads/2017/02/HEA-TEADUSTAVA.pdf> (Estonian)
- EST: Code of Ethics of Estonian Scientists (Eesti Teadlaste Eetikakoodeks)
  - <https://etag.ee/teadusagentuur/dokumendid/eetikakoodeks2002/> (Estonian)
  - [https://www.akadeemia.ee/wp-content/uploads/2020/06/code\\_ethics2002-3.pdf](https://www.akadeemia.ee/wp-content/uploads/2020/06/code_ethics2002-3.pdf) (English)
- EST: Estonian Research Council Guidelines for Completing Your Ethics Self-Assessment for Application of Personal Research Funding
  - [https://www.etag.ee/wp-content/uploads/2019/03/Eetika\\_Tabel\\_ENG.pdf](https://www.etag.ee/wp-content/uploads/2019/03/Eetika_Tabel_ENG.pdf) (English)

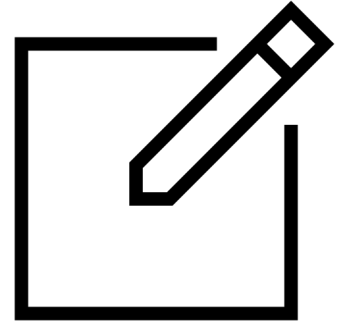
# Ethics

Information should be provided about:

- Humans
- Personal data
- Human embryos and/or fetuses
- Human cells and/or tissues
- Animals
- Genetic resources and/or associated traditional knowledge
- Low income countries
- Environment, Health, and Safety
- Potential Misuse of Research Results
- Other Ethics Issues

**If the project has no contact with the ethical issues mentioned in the guide, it should also be described in the application.**

# Writing session



# Preserve



# Data lifecycle

PRESERVE - for long time period (10 years in Estonia)

- Ensure data safety and integrity.
- Change the file format and update software to make sure that they do not become outdated or obsolete.
  - Prefer open source formats
- Change hardware and other storage media (such as paper, magnetic tape, etc) to avoid degradation.
- Ensure that data is organised and described with appropriate metadata and documentation to be always understandable and reusable.
- Make sure you have the right to preserve it (you are owner of the data)

**It is much easier to upload data to a repository in which case the preservation duty falls on them.**

# Data preservation

- Data, that should be preserved:
  - Funder, publisher and institution policies (usually, data should be preserved for at least 5 or **10 years** after the end of the project).
  - Legal or ethical requirements (e.g. clinical trial data).
  - Unique data or that cannot be easily re-generated (e.g. raw data, analysis workflow).
  - Data that will probably being reused in the future.
  - Data of great value for society (scientifically, historically or culturally).
- Choose trustworthy research data repositories or deposition databases, based on your data type. Repositories could be publicly accessible and allow you to also publish your data.

# Data preservation

- When preparing data for preservation:
  - Do not include data that are temporary.
  - Ensure well described and self-explanatory documentation.
  - Include information about provenance.
  - Include sufficient licensing information.
  - Ensure that data is well organised.
  - Ensure that a consistent naming scheme is used.
  - Use standard, open source, file formats instead of proprietary ones.
- If you need to preserve non-digital data (e.g. paper), consider whether digitalising the data is feasible or consult with data management support services in your institution.
- If you need to preserve materials, such as micro-organisms, biomaterials or biomolecules, consult with data management support services in your institution to find appropriate centers or biobanks.



# Repositories

- Field specific repository
- Institutional/national repository
  - DataDOI - University of Tartu, University of Tallinn
  - Dspace - Estonian University of Life Sciences
  - TalTech Data - TalTech
- General repositories
  - EOSC - EU
  - Zenodo - EU
  - FigShare - USA

# University of Tartu legal data owning rights example

The data belong to the University of Tartu. Persons employed for filling the grant will assign the proprietary rights to the results of the research (including the data) performed under the grant agreement to the University with the Employment Contract (academic employees) or with another written document (Act of Assignment of the Intellectual Property Rights).

Data will be disclosed under the Creative Commons license CC-BY 4.0

A third party, whose data have been used for creating the results of the grant, may set restrictions to the usage of the data. In this case those restrictions must be considered while the data are being licensed, i.e. the university can give the license for the data usage only in the scale of rights allowed by the third person (i. e. the scale of rights that university has received from the third persons)

If the University or a third person, whose data have been used for creating the results of the grant, wants to submit a patent or a utility model application, the publishing of the data has to be postponed until the submission of the application.

# Data sharing and reuse

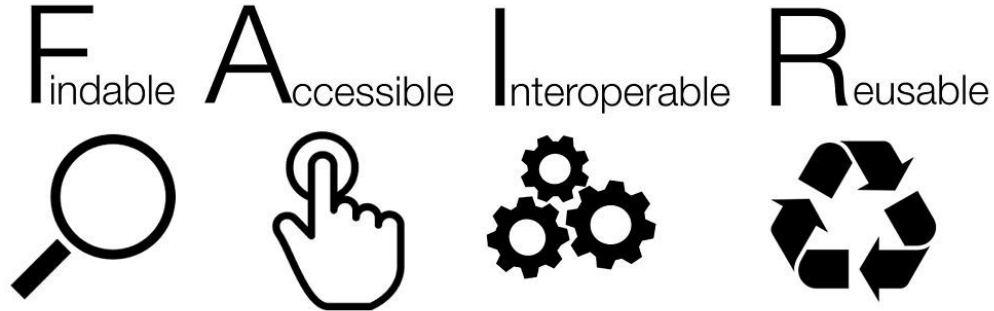


# Share

making your data known to other people

***“As open as possible, as closed as necessary”***

Be FAIR



## Findable

Metadata and data should be findable for both humans and computers

F

A

## Interoperable

Data needs to work with applications or workflows for analysis, storage and processing

I

R

## Accessible

Once found, users need to know how the data can be accessed

## Reusable

The goal of **FAIR** is to optimise data reuse via comprehensive well-described metadata

## Findable

- F1. (meta)data are assigned a globally unique and eternally persistent identifier.
- F2. data are described with rich metadata.
- F3. (meta)data are registered or indexed in a searchable resource.
- F4. metadata specify the data identifier.

## Interoperable

- I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- I2. (meta)data use vocabularies that follow FAIR principles.
- I3. (meta)data include qualified references to other (meta)data.

**doi: 10.1038/sdata.2016.18**

## Accessible

- A1 (meta)data are retrievable by their identifier using a standardized 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.

## Reusable

- R1. meta(data) have a plurality of accurate and relevant attributes.
  - R1.1. (meta)data are released with a clear and accessible data usage license.
  - R1.2. (meta)data are associated with their provenance.
  - R1.3. (meta)data meet domain-relevant community standards.

# The FAIR Principles in practice

## FINDABLE

- Persistent Identifier (DOI, Handle, ORCID, ROR)
- Keywords, metadata

## ACCESSIBLE

- Who has access under which conditions to the data (during and at the end of the project)

OPEN ≠ FAIR

## INTEROPERABLE

- File formats (free and open format)
- Domain specific vocabulary

## REUSABLE

- README file (how data was created)
- License



# F – Findable



**Persistent identifier** (PID) - a permanent reference to a digital resource used to search, identify and link publications and datasets (as well as other digital objects).

This works even if the original location of the object/resource has changed. The PID links to the new location of the object without changing its address.

Examples: DOI, PURL, Handle, URN

ORCID – for researchers <https://orcid.org/>

ROR – for organisations (and funders) <https://ror.org/>



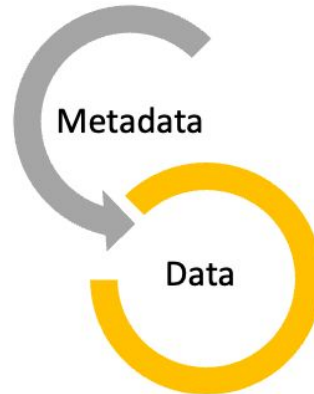


# F – Findable

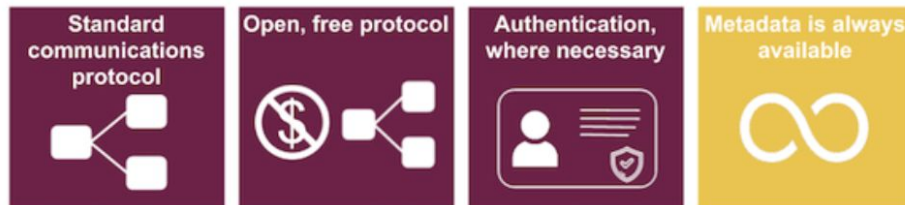
Clarify in the **README** file the dataset's PID, all the other **used resource's PIDs** that are connected with this dataset (workflows, protocols, published articles, etc.).

Just uploading the data to the Internet is usually not enough for it to be findable.

For this,  
**indexing** works,  
or just adding specific  
**keywords**.



# A – Accessible



The important aspects of the accessibility of the data:

- Who can access the data and under which conditions?
- How are the data backed up?
- How is the information stored?
- Who is the owner of the data;
  - and can this put restrictions on the accessibility of the data while in collection phase or at the end of the project?

It is understandable that datasets will expire/ get lost/ get deleted over time. Also, storing the data may be quite costly (both for the storing space and the costs of it).

**However, the corresponding metadata should still be kept, even if the data itself does not exist anymore.**

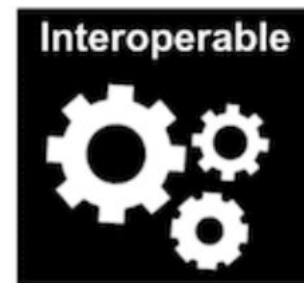
# Accessibility

- Determine the right type of access for you data. Even if the access to the data is restricted, it is good practice to openly and publicly share the metadata of your data.
  - **Open access:** data is shared publicly. Anyone can access the data freely.
  - **Registered access or authentication procedure:** potential users must register before they are able to access the data. Datasets that are shared via registered-access would typically have no restrictions besides the condition that data is to be used for research. Registered access allows the data archive to monitor who can access data, enabling reminders about conditions of use to be issued.
  - **Controlled access or Data Access Committees (DACs):** data can only be shared with researchers, whose research is reviewed and approved by a Data Access Committee (DAC).
  - **Access upon request (not recommended):** in order to manage this type of access a named contact is required for the dataset who would be responsible for making decisions about whether access is granted. The owner of the data must provide his/her contact in the documentation associated with the datasets (metadata). *Metadata about the datasets must be open.*

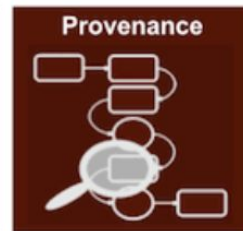
# I – Interoperable

It is recommended to keep one copy of the file in software independent file format.

- Text: TXT, ODT, PDF/A, XML
- Tables: CSV, TSV
- Pictures: TIFF, PNG, JPG 2000, SVG, WebP
- Music files: WAV, FLAC, OPUS
- Video: MPEG2, Theora, VP8, VP9, AV1, Motion JPG 2000 (MJ2)



# R – Reusable



## Recommendations:

- Describe, for what purpose the data was collected
- Indicate the specifics or limitations of the data that other users should be aware of
- Note down the time, place and conditions, parameters, program versions, etc of data collection/creation
- Note if the data is raw, processed or analysed
- Add enough explanation for the variables (using the appropriate vocabularies)
- Indicate clearly which versions of the data has been archived or used

If the resource has no license, then in the Europe, all the rights belong to the owner and the resource should not be used in any capacity without a permission. It is recommended to use CC0/ CC-BY license, which give the users all the right to use the resource.

Always check, what version of the license has been used.

# Simple but typical examples of reuseless\* data

S1Sh.cu			
	A	B	C
1		Group1	Group2
2		Day 0	
3	Sodium	139	142
4	Potassium	3.3	4.8
5	Chloride	100	108
6	BUN	18	18
7	Creatine	1.2	1.2
8	Uric acid	5.5*	6.2*
9		Day 7	
10	Sodium	140	146
11	Potassium	3.4	5.1
12	Chloride	97	108

Unhelpful document name

Meaningless column titles

Undefined abbreviation

No units

Formatting for information that should be in metadata

Special characters can cause text mining errors

\* reuseless

It is associated with bad [programming](#) style in which a piece of code is useless over and over again. It might be some short version of a big code but it is [absurd](#) and reduces [readability](#).



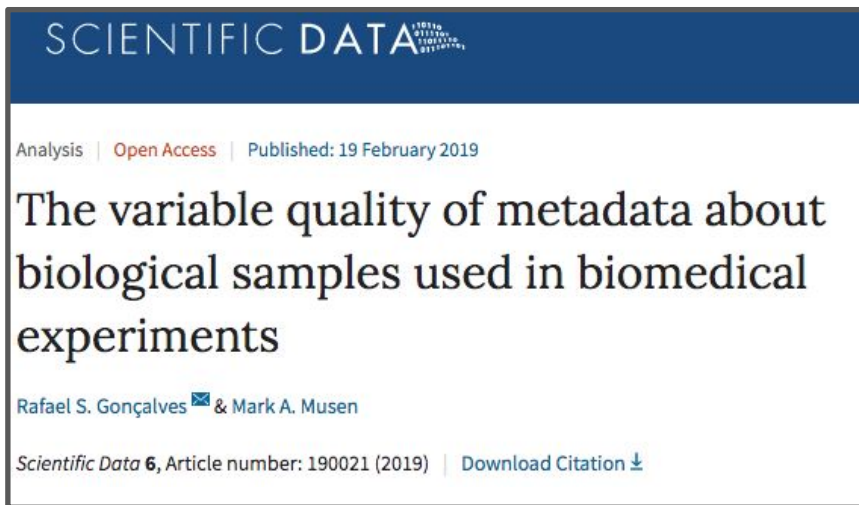
# This is much clearer!

Table_S1_Shanghai_blood.xls						
	A	B	C	D	E	F
1	Parameter	Day	Control	Treated	Units	P
2	Sodium	0	139	142	mEq/l	0.82
3	Sodium	7	140	146	mEq/l	0.70
4	Sodium	14	140	158	mEq/l	0.03
5	Sodium	21	143	160	mEq/l	0.02
6	Potassium	0	3.3	4.8	mEq/l	0.06
7	Potassium	7	3.4	5.1	mEq/l	0.07
8	Potassium	14	3.7	4.7	mEq/l	0.10
9	Potassium	21	3.1	3.6	mEq/l	0.52
10	Chloride	0	100	108	mEq/l	0.56
11	Chloride	7	97	108	mEq/l	0.68
12	Chloride	14	101	106	mEq/l	0.79

# Standardized descriptions matter

*“Most metadata field names and their values are **not** standardized or controlled”*

*“Even simple binary or numeric fields are often populated with inadequate values of different data types”*



age	age [y]
Age	age [year]
AGE	age [years]
`Age	age in years
age (after birth)	age of patient
age (in years)	Age of patient
age (y)	age of subject
age (year)	age(years)
age (years)	Age(years)
Age (years)	Age(yrs.)
Age (Years)	Age, year
age (yr)	age, years
age (yr-old)	age, yrs
age (yrs)	age.year
Age (yrs)	age_years



# R – Reusable

**BY** – attribution – *viita*

**SA** – share alike – *jaga sama litsentsiga*

**ND** – no derivative work – *ei tohi muuta*

**NC** – no commercial use – *ei tohi rahalistel eesmärkidel kasutada*

\* Scientific research? What kind of money is classified under commercial use?

CREATIVE COMMONS

LICENSES

PUBLIC DOMAIN

CC BY

CC BY-SA

CC BY-ND

CC BY-NC

CC BY-NC-SA

CC BY-NC-ND

COPY & PUBLISH

ATTRIBUTION REQUIRED

COMMERCIAL USE

MODIFY & ADAPT

CHANGE LICENSE

✓

✓

✓

✓

✓

✓

✓

✗

✓

✓

✓

✓

✓

✓

✓

✓

✓

✓

✗

✗

✗

✓

✓

✓

✓

✗

✓

✗

✓

✓

✗

✓

✓

✗

✓

You can redistribute (copy, publish, display, communicate, etc.)

You have to attribute the original work

You can use the work commercially

You can modify and adapt the original work

You can choose license type for your adaptations of the work.

commons-photos/

<https://foter.com/blog/how-to-attribute-creative-commons-photos/>

CC-BY-SA



The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**, 160018 (2016). <https://doi.org/10.1038/sdata.2016.18>  
 RDA FAIR Data Maturity Model. Specification and Guidelines <https://zenodo.org/record/3909563#.YORYkUzTX19>  
<https://www.go-fair.org/fair-principles/>

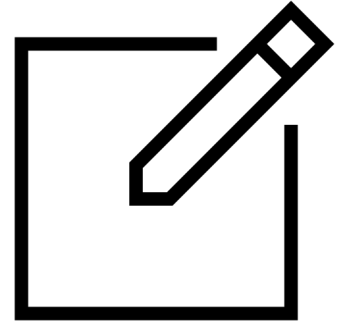
Image credit: [ARDC](#) licensed under a [Creative Commons Attribution 4.0 International License](#)

# Estimation of data management costs

Excel table that might be useful:

Dowling, L., Coffey, A. M., O'Neill, J., Simpson, A., & Straube, A. (2025). Costing Research Data Management: A Framework. Zenodo.  
<https://doi.org/10.5281/zenodo.15465413>

# Writing session



Feedback:

<https://forms.gle/JSstre4L9CQPvx7w5>

# Important links

## The FAIR Principles

- <https://www.howtofair.dk/what-is-fair/>
- <https://www.go-fair.org/fair-principles/>
- <https://www.openaire.eu/how-to-make-your-data-fair>
- <https://fairplus.github.io/the-fair-cookbook/content/home.html>

## Databases and standards

- <https://fairsharing.org/>
- <https://rd-alliance.github.io/metadata-directory/standards/>
- <https://www.dcc.ac.uk/guidance/standards/metadata/list>
- <https://www.ebi.ac.uk/ols/index>

# Important links

## Preferred file formats

- <https://dans.knaw.nl/en/about/services/easy/information-about-depositing-data/before-depositing/file-formats>
- <https://documentation.library.ethz.ch/display/DD/File+formats+for+archiving>

## Licenses

- <https://creativecommons.org/licenses/>
- <https://www.dcc.ac.uk/guidance/how-guides/license-research-data>
- <https://choosealicense.com/>
- <http://ufal.github.io/public-license-selector/>

# Important links

## Metadata

- [https://nsteffel.github.io/dublin\\_core\\_generator/](https://nsteffel.github.io/dublin_core_generator/)
- <https://www.w3.org/TR/2017/REC-dwbp-20170131/#metadata>
- [https://www.andis.org.au/\\_data/assets/pdf\\_file/0004/728041/Metadata-Workinglevel.pdf](https://www.andis.org.au/_data/assets/pdf_file/0004/728041/Metadata-Workinglevel.pdf)
- <https://www.fsd.tuni.fi/en/services/data-management-guidelines/data-description-and-metadata/>
- <https://data.library.arizona.edu/data-management/best-practices/data-documentation-readme-metadata>
- <https://www.fsd.tuni.fi/en/services/data-management-guidelines/data-description-and-metadata/#metadata-standards>
- <https://www.cessda.eu/Training/Training-Resources/Library/Data-Management-Expert-Guide/2.-Organise-Document/Documentation-and-metadata>



# Important links

## README files

- <https://researchdata.wisc.edu/dmp-3-data-documentation/>
- <https://data.research.cornell.edu/content/readme>
- <https://www.makeareadme.com/>
- <https://cornell.app.box.com/v/ReadmeTemplate>
- <https://data.research.cornell.edu/content/readme>

## Tools

- <https://readme.so/>
- <https://stackedit.io/>



Thank you for listening!