
Research Data Management
Advanced
Datatuki
datasupport@helsinki.fi
Helsingin yliopisto

SESSION 1. DOCUMENTATION

Why Document Your Data?

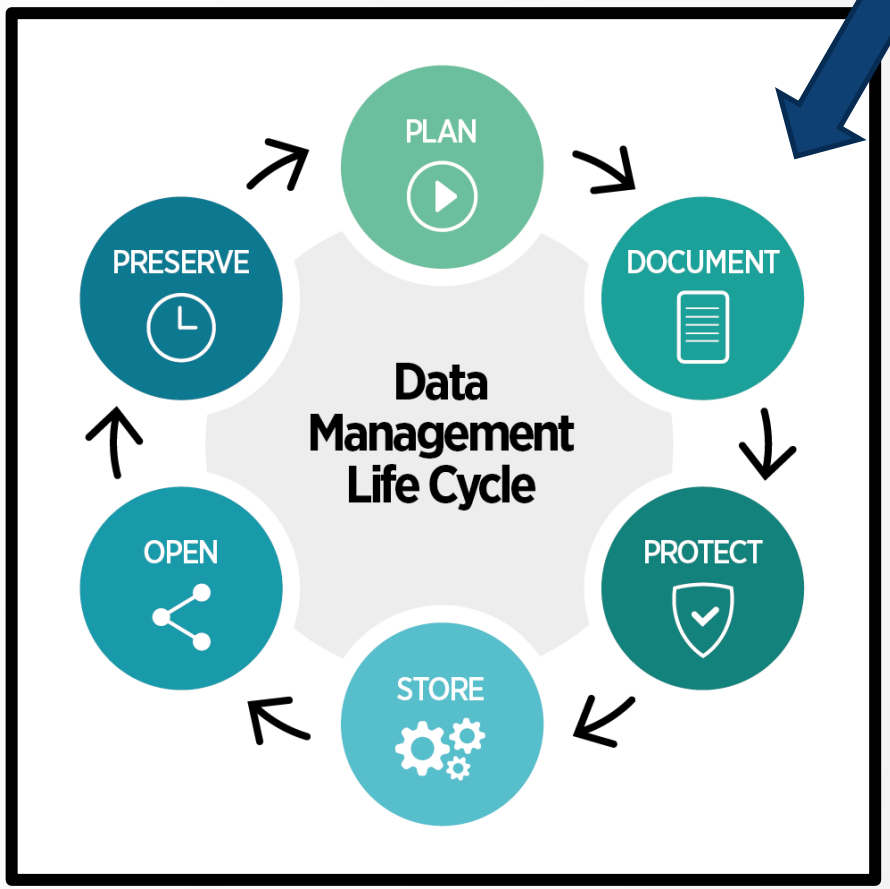
Know Your Data

Metadata

Different Ways to Describe Your Data

Metadata Standards

Tips for Documentation



WHY DOCUMENT YOUR DATA?

DOCUMENTATION

- Documentation means describing the data,
  i.e., these documents explain **what** data the project has and **where** the data originates from.
- Documentation includes data dictionaries (explaining variables and codes) and readme files.
- Other important issues include file naming conventions, version control, and directory structure.
- There are standard methods available for documentation called metadata standards, which should be used if suitable for the data. These will increase the value of the data by making it easier to reuse.

WHY DOCUMENTATION IS IMPORTANT?

- Other people can understand and use your data
- It is easier to share, open and archive data
- It minimizes the risk that your data is misused or misinterpreted
- Metadata makes it possible for others to find your data by using different kinds of search criteria
- Having invested in documentation during the project, will **save time** upon publishing the dataset.

KNOW YOUR DATA

WHAT DATA?

What data types you will have?

Where is the data from?

CATEGORIZATION OF DATASETS

General descriptions of data

- *What kinds of data are collected or reused? In what file formats will the data be?*

Describe/list all datasets and material which are discussed later in the plan, e.g.

A) Data **collected** by yourself

various locations; raw, non-catalogued private collections
various types

B) Data **reused** in your project

ready-made dataset in an archive
remember to cite the original creator or collector in your work!

C) Data **produced** during your project

notebooks, research diaries, field notes, comments, annotations, coding, and register a PID for your datasets so others can cite you

D) Managerial documents, agreements, contracts etc.

DATA SHEET MODEL

	Data type	Source of the data	File Format	Size estimate	Sensitivity / Personal data + controller	Owner / other agreements?	Documentation	Storage during project	Opening	Long term archiving
1										
2										
3										
4										
5										
6										
7										
8										
9										
10										
11										

https://wiki.helsinki.fi/download/attachments/223985293/Meilahti_RDM_data_spring-20.xlsx?version=1&modificationDate=1584953369397&api=v2

DATASETS TABLE EXAMPLE

Data type	File format	Personal or sensitive data	Storing data and backups during the project	Documentation and metadata	Ownership and Agreements	Opening or publishing data after the project
Measurements	.xls .csv	No	Personal storage at UH (home folder)	readme.txt codebooks	UH and LUKE Agreement	Opening via publication at DRYAD or Zenodo
Gene sequences	.txt fasta	No. Collecting only from plants.	Group storage space	.FASTA	PI	NCBI Genome
Programme codes	.xml ASCII R-code	No	GitLab & Shared network drive hosted by UH	GitLab & readme.txt	Co-ownership of the research group	Via publication and Zenodo
Microscopy images	.tiff	No	Server storage space	OME-TIFF	PI	Electron Microscopy Public Image Archive (EMPIAR)
Lab notes	.doc .txt .pdf	No Patenting or commercializing?	Electronic lab notebook Scinote Cloud service	Programme generates metadata by itself	PI and me	No
Samples (applying from THL Biobank)	.xls	Anonymization will be done by Biobank	Freezer at the Institute of Biotechnology (PI's lab)	Unique identifier code	THL Biobank-licence Research agreement DMP	Samples discarded one year after publishing the results.
Questionnaire forms	Paper forms	Yes Data Controller UH	Locked filing cabinets in PI office.	codebook readme.txt	PI Informing participants	No, only metadata will be open in FSD. Forms discarded 2 years after project ends.
Spatial data about land use and forest stand	.tiff, Coloured	No, open data	Datacloud at UH (service coming soon)	Supplements at Etsin	National land survey of Finland: license CCBY	Processed data at Zenodo

METADATA

METADATA

Metadata is “data that provides information about other data”, but not the content of the data
(Merriam Webster Dictionary 2019, Wikipedia 2023)

”DATA ABOUT DATA”

DESCRIPTIVE METADATA

DATA DOCUMENTATION

Ted Eytan: https://www.flickr.com/photos/taedc/

Makes data…

understandable, findable and usable

METADATA TELLS YOU…

Relevant information about the data:

- What kind of data it is (name, description, discipline, format)
- Who created the data (creator, organisation, distributor)
- How the data was created (methods, equipment, software)
- What has been done to the data (processing, editing)
- Where the data locates (storage place, identifiers)
- How the data can be reused (terms of use, licenses)

DATA DOCUMENTATION

Make your data describe itself!

What the data is?
Where it came from?
How can it be reused?

Understandability
"User manual" of the dataset

Makes the dataset self-explanatory and usable for others

File naming conventions, explain variables, codebooks, use tags, readme-files + administrative documents, licenses, etc.

Discoverability
"Label" of the dataset

Describes what the dataset contains. Should be available even if you cannot open data itself.

Title, description, creator, persistent identifier, etc.

DIFFERENT LEVELS OF DESCRIPTION

On what level do you have to describe your data?

- Whole project
- Documents
- Datasets
- Files

README.txt
metadatastandards

- Variables
- Abbreviations

Codebooks
Dictionaries

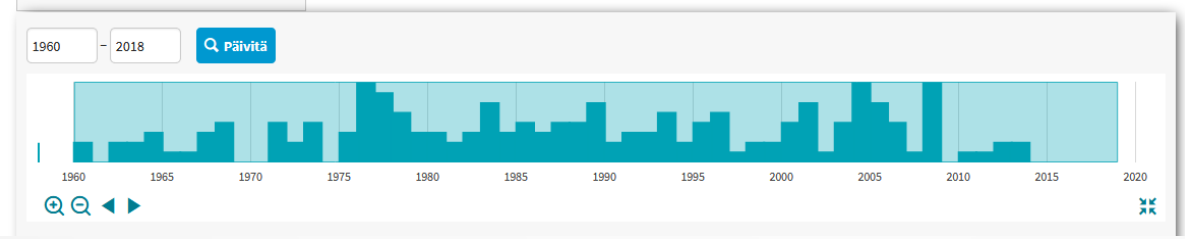
DO YOU STILL REMEMBER…

Do you yourself remember the meaning of all the markers or variables after six months?
AMC, GHD34, GFP …

Think of all the different datatypes without metadata…

Data without description is unusable.

H	I	J	K	L	M
3,25	2,25	19	20	18	15
3,25	2,25	19	20	18	15
3,25	2,25	19	20	18	15
3,25	2,25	19	20	18	15
3,25	2,25	19	20	18	15
3,25	2,25	19	20	1	
3,25	2,25	19	20	1	
3,25	2,25	19	20	1	
3,25	2,25	19	20	1	
3,25	2,25	19	20	1	
3,25	2,25	19	20	1	
3,25	2,25	19	20	1	
3,25	2,25	19	20	1	
3	4,13	24	25	2	
3	4,13	24	25	2	
3	4,13	24	25	2	
3	4,13	24	25	2	

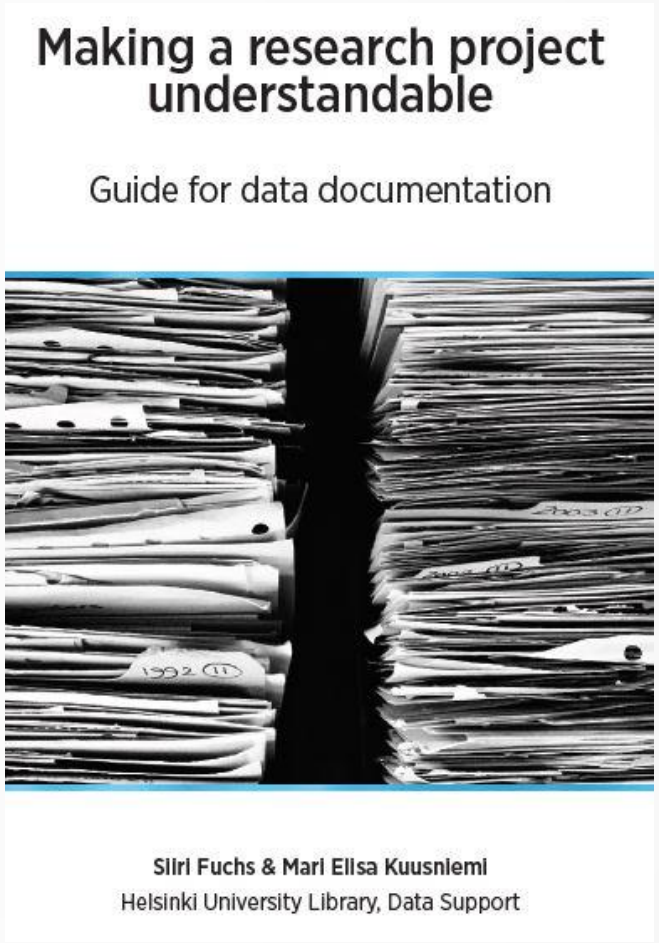


DIFFERENT WAYS TO DESCRIBE YOUR DATA

ELEMENTS OF DATA DOCUMENTATION

Documentation methods
Naming of the files
Directory structure
Version control
Data dictionaries and codebooks
Readme files
Laboratory notebooks
Metadata standards

Siiri Fuchs, & Mari Elisa Kuusniemi. (2018, December 4)
Making a research project understandable
- Guide for data documentation (Version 1.2).
Zenodo. http://doi.org/10.5281/zenodo.1914401

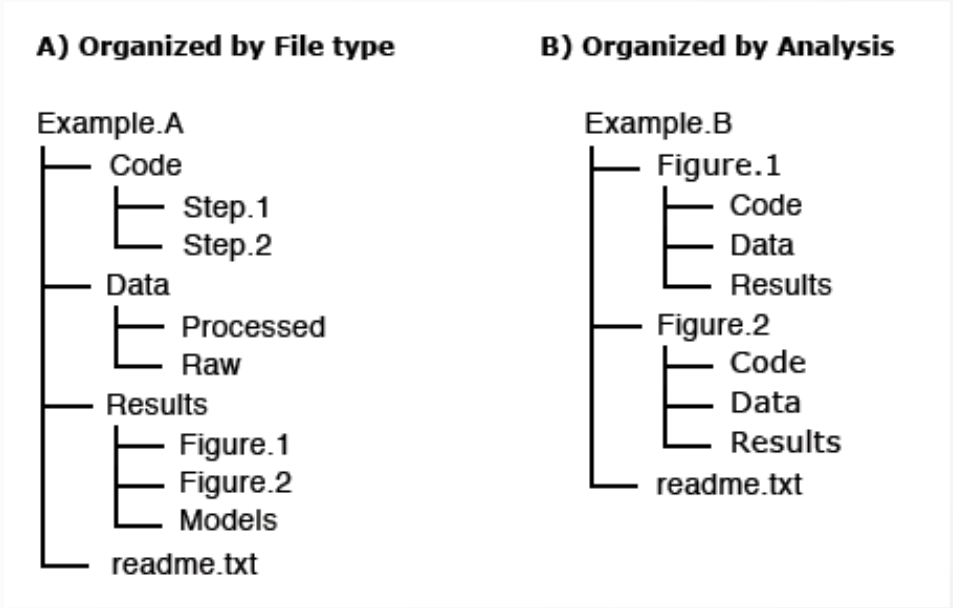


NAMING OF THE FILES

- Plan naming **at the beginning of the project.**
- Guiding principles: **consistency and clarity**
- Good file names are **constructed logically** (e.g. by date) and they **inform on the content** of the files (Purdue University)
- The date is always in the same form: yyyy-mm-dd -> files are organized chronologically
- https://www.fsd.uta.fi/aineistonhallinta/en/processing-qualitative-data-files.html#naming
- File naming convention helps you stay organized, contain quickly information from the title, and assists others in navigating in your directories.
- Use unique file names so that files can be recognized without folder name

DIRECTORY STRUCTURE

- Folder structure should be based on the needs of the project
- Clear folder structure helps with access control (e.g. when you have sensitive data)
- Right balance with shallow and deep folder hierarchy helps with finding the correct file
- Key words and tags help with finding the correct files



Source: DRYAD

VERSION CONTROL

Favour automatic version control:

- Wiki
- GitHub, GitLab
- OneDrive for UH
- Google Docs (on US server!)

Name different versions clearly:

- Good practice: *_V02-03.doc*
- Avoid: *_draft, _final, _final3, _finalfinal*

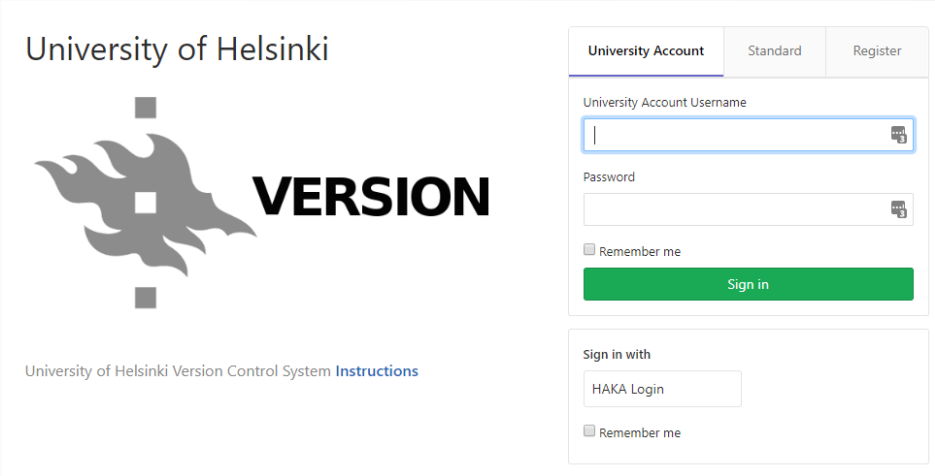
Generate an archive folder for old versions.

Keep the original raw data separate and safe.

- Makes it possible to return to an older version of data.
- Can save you from losing the data.
- Version control can be done automatically or manually.

UH VERSION CONTROL SYSTEM

- GitLab based version control system for projects
- Internal and shared projects
- Ideal for research collaboration
- Check out: https://version.helsinki.fi/
- Instructions: https://wiki.helsinki.fi/x/tASBDQ
- A blog post (only in Finnish): https://blogs.helsinki.fi/thinkopen/versionhallinta-on-valttamaton-tyokalu-tutkimukselle/



DATA DICTIONARIES AND CODEBOOKS

- Dictionaries explain variables used in a dataset
- A data dictionary explains all the variable names and values in your spreadsheet.

Variable

Variable name (e.g. DOB, AGE)

Measurement units

Allowed values

Definition of the variable

HOW TO MAKE A DATA DICTIONARY?

Sheet_1

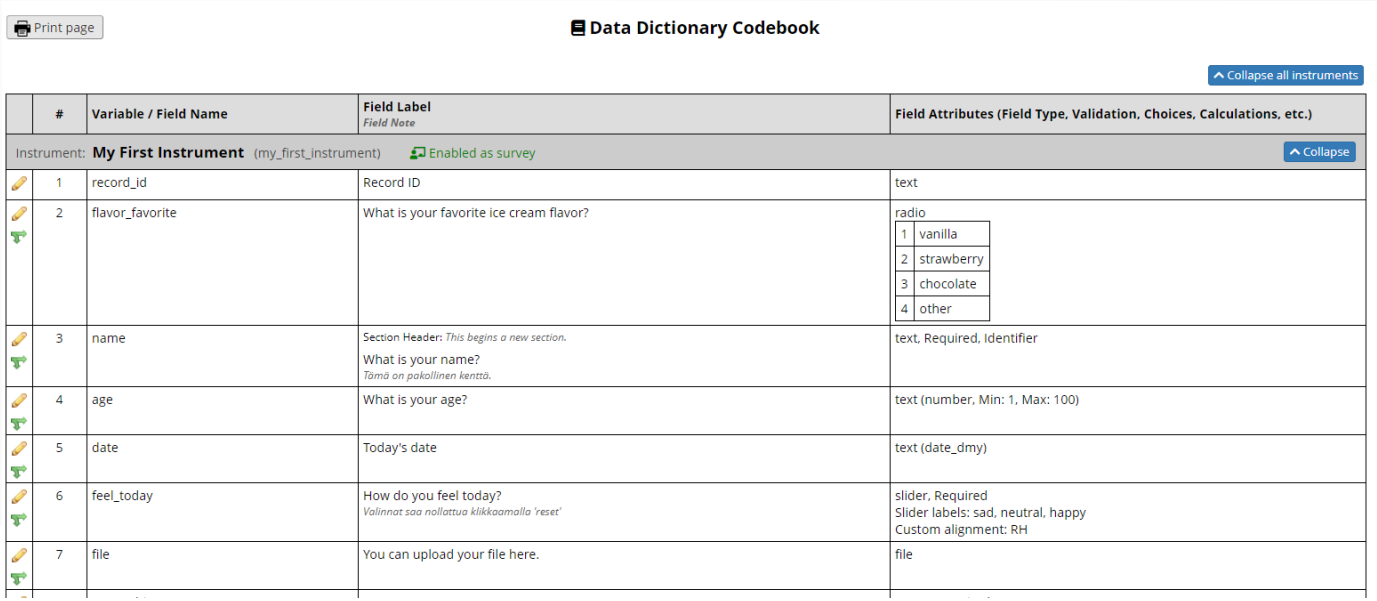
Show rows with cells including:

Variable	Variable name	Measurement unit	Allowed values	Description
Participant ID number	ID	Numeric	001-999	ID number assigned to participant in sequential order
Group number	GROUP	Numeric	1-30	Group assigned to participant based on ID number
Age in years	AGE	Numeric	18.0-65.0	Age of participant in years
Date of birth	DOB	mm/dd/yyyy	1-12/1-31/1951-1998	Participant's date of birth
Gender	SEX	Numeric	1 = male 2 = female	Participant's gender
Date of survey	SURVEY	mm/dd/yyyy	01/01/2015 – 01/01/2016	When the participant completed the survey
Self-reported consumer spending	SPEND	Numeric	0-100,000,000	Self-reported average yearly expenditure
Market sentiment	SENTIMENT	Numeric	1 = negative 2 = neutral 3 = positive	Sentiment towards US domestic economy
Actual GDP growth	GDP	Numeric	-5.0-5.0	Average US yearly GDP growth

Source: How to Make a Data Dictionary, https://help.osf.io/m/bestpractices/l/618767-how-to-make-a-data-dictionary, OSF Guides

SOFTWARES THAT CREATES METADATA

- REDCap is an example of software that creates metadata
  - https://redcap.helsinki.fi/redcap/
- Note: many manufacturers have software that produces its metadata or formats, making the data incompatible with other programs.
- However, usually manufacturers do have the standard metadata version/format available.



Print page

Data Dictionary Codebook

Collapse all instruments

	#	Variable / Field Name	Field Label *Field Note*	Field Attributes (Field Type, Validation, Choices, Calculations, etc.)
Instrument: **My First Instrument** (my_first_instrument) Enabled as survey				Collapse
	1	record_id	Record ID	text
	2	flavor_favorite	What is your favorite ice cream flavor?	radio 1, vanilla 2, strawberry 3, chocolate 4, other
	3	name	Section Header: *This begins a new section.* What is your name? *Tämä on pakollinen kenttä.*	text, Required, Identifier
	4	age	What is your age?	text (number, Min: 1, Max: 100)
	5	date	Today's date	text (date_dmy)
	6	feel_today	How do you feel today? *Valinnat saa nollattua klikkaamalla 'reset'*	slider, Required Slider labels: sad, neutral, happy Custom alignment: RH
	7	file	You can upload your file here.	file

BRIEFLY ABOUT REDCAP

Application for building and managing online *surveys and database*: https://projectredcap.org/

- For collecting sensitive and personal data
- Compliant with GDPR requirements when used properly

University of Helsinki REDCap:

- Anyone with UH credentials can access and use it
- Installed on the UH's servers → data stored on the UH's servers

README FILES

- Readme-files are text documents (e.g. README.txt)
- Provide information about data files to ensure they are interpreted correctly.
- These become especially important when sharing and publishing data
- They are also helpful to your future self.

README FILES

Write down everything related to your project:

- Names of the files and file formats
- How the data is organized (directory structure)
- How the data is produced (containing equipment used and software)
- How the data has changed or how it’s been processed/edited
- Explain the codes, abbreviations or variables used in the naming of the files
- At a minimum, save this information in a README.txt file and save it with the actual data.

EXAMPLE OF README CONTENT

What you should include in the documentation of your data

TITLE: Name of the dataset or research project that produced it

CREATOR: Names and addresses of the organization or people who created the data

IDENTIFIER**:** Number used to identify the data, even if it is just an internal project reference number (e.g. URN:123abc)

DATES: Key dates associated with the data, including project start and end date, data modification, data release date, and period covered by the data

SUBJECT: Keywords or phrases describing the subject or content of the data

FUNDERS: Organizations or agencies that funded the research

RIGHTS: Any known intellectual property rights held for the data

LANGUAGE: Language(s) of the intellectual content of the resource, when applicable

LOCATION: Where the data relates to a physical location, record information about its spatial coverage

METHODOLOGY: How the data was generated, including equipment or software used, experimental protocol, other things you might include in a lab notebook

LABORATORY NOTEBOOKS

- For specific disciplines, the most important method for documenting research (sometimes data itself)
- Usually, physical notebooks kept in labs
- Clear rules on how to keep a notebook
- Handwritten notebooks are often chaotic and always unsearchable. Source: Nature

"Experiments and sample metadata will be documented in detail in laboratory notebooks (in paper and in electronic format)."

ELECTRONIC LABORATORY NOTEBOOKS (ELN)

- A documentation method
- The program creates metadata automatically
- Notes stay up to date
- Safe data storage and access control
- Makes data sharing and cooperation possible
- Reporting is easy

labguru

SCINOTE

RSpace COMMUNITY

WHICH ELN PROGRAM?

- At the moment, there are over 100 different programs available
- Splice-bio has listed all the best ELN programs

- UH hasn’t acquired (yet) any specific program

METADATA STANDARDS

METADATA STANDARDS

- Metadata standard is a format for describing a dataset in a controlled way.
- There are general or disciplinary-specific formats and standards.
- Favor metadata standards instead of an uncontrolled description if possible

WHY?

- Many open data archives require a specific data description format.
- If you want to deposit your data, you should find out the required format already at the beginning of your project.
- Choosing a repository makes it easier to select a metadata schema.
- Are you collecting all the necessary information?
- Using standards and coherency in documentation will make data more understandable, facilitate its reuse and make combining datasets possible.
- By using metadata standards, you improve the **interoperability**, **findability** and also **machine readability** of your research data.

FINDING METADATA STANDARDS

OGS

miappe

JHS

EML

- Disciplinary Metadata / Digital Curation Centre DCC
- Metadata Standards by Subject / Research Data Alliance RDA
- General Research Data / Digital Curation Centre DCC.

FAIRsharing.org standards, databases, policies

- FAIRsharing.org - search for metadata standards by repositories

- Metadata Tools / Standford University Library

ISO TC/211

ONTOLOGIES AND VOCABULARIES

Controlled vocabularies to describe your data and to make data machine readable and searchable

- EMBL-EBI Ontology
  - •Lookup service for biomedical ontologies that aims to provide a single point of access to the latest ontology versions.
- Data vocabularies (Tietomallit),
  - •A service for managing and publishing data vocabularies.
  - •It contains data component libraries, i.e. data specifications for harmonizing information used jointly by different actors.

PERSISTENT IDENTIFIERS

- **A long-lasting reference to a digital resource**. An identifier is a label which gives a unique name to an entity: a person, place, organization. (ORCID Support, 2019)
- Unlike URLs, which may break, a persistent identifier reliably points to a digital entity as long as possible.
- For an individual researcher: https://orcid.org/
- For organizations: https://ror.org/
  - University of Helsinki: https://ror.org/040af2s02
  - ror-identifier is linked to other identifiers also

DESCRIBE YOUR DATA DOCUMENTATION

In your DMP:

"Metadata will be collected according to minimum Information About a Microarray Experiment (MIAME)."

"We will deposit sequencing reads from putative commercial cell lines in standard formats to the Gene Expression Omnibus (GEO, NIH, USA)."

GEO is a public functional genomics data repository supporting MIAME-compliant data submissions.

TIPS FOR DOCUMENTATION

DATA DOCUMENTATION

If possible, use metadata standards and controlled vocabularies.

If available, use data management software, to make documenting easier.

At minimum, store this documentation in a readme.txt file or the equivalent, together with the data.

START EARLY

- Make a plan about documentation as early as possible when the project starts
- **Start in time**: the earlier you start to describe the data, the easier it is.
- The quality of metadata diminishes if it is produced afterwards.
- For data that cannot be freely shared, at least metadata will be publicly provided.