

RESEARCH DATA MANAGEMENT

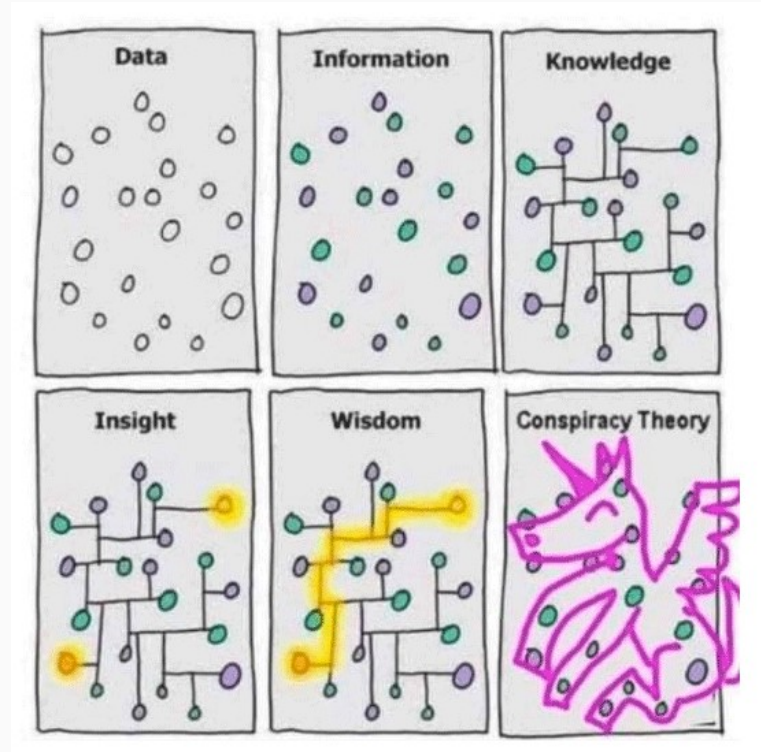
Basic Concepts –Lecture
Autumn 2022

Maija Paavelainen
Information Specialist
Data Support

TODAY

Today's session is in English, but you can get help also in Finnish.

Please feel free to interrupt for questions anytime, out loud or writing in chat.



TODAY

Motivation and context

Recognize your data

Data Documentation

Personal and Sensitive Data

Intellectual Property

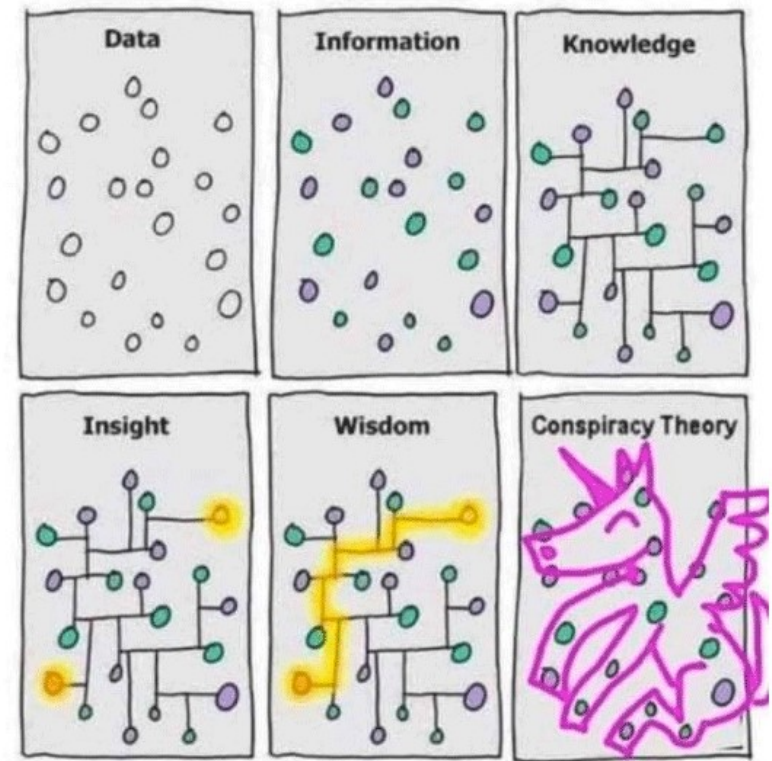
- short break 10min -

Data Storage (IT Services!)

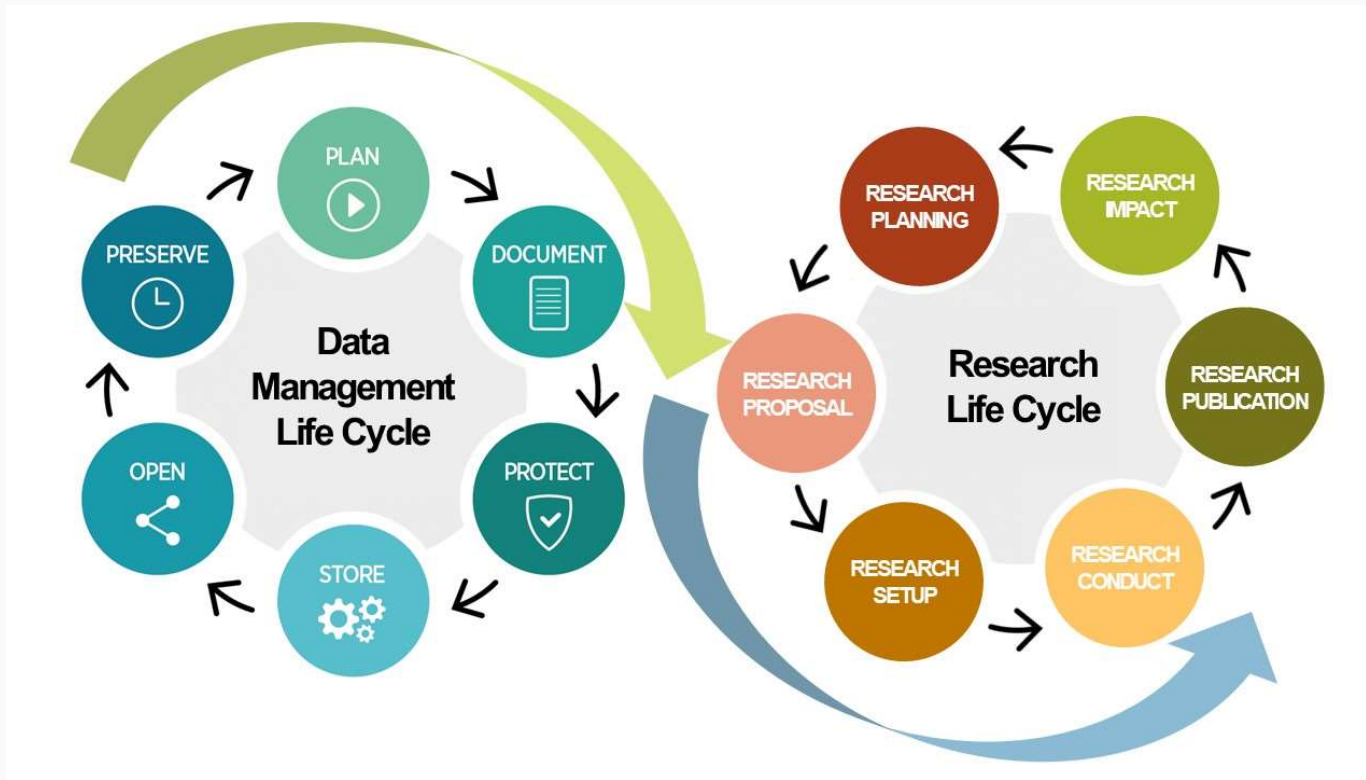
Archiving Data

Resources

Further Support



DATA MANAGEMENT – WHY?





2. RECOGNIZE YOUR DATA

HELSINGIN YLIOPISTON KIRJASTO
HELSINGFORS UNIVERSITETS BIBLIOTEK
HELSINKI UNIVERSITY LIBRARY

KUVA: LINDA TAMMISTO

RECOGNIZE YOUR DATA

Data collected by various methods

Samples, measurements, surveys, interviews, imaging techniques, curated collections etc.

Data produced during the research project

Analysis results, sequences, **field diaries**, physical or electronic lab journals, copies of physical artifacts, source code, algorithms, software, etc.

Data reused from various origins

Biobank samples, archive materials, data from repositories, codes etc.

DESCRIBE YOUR DATA

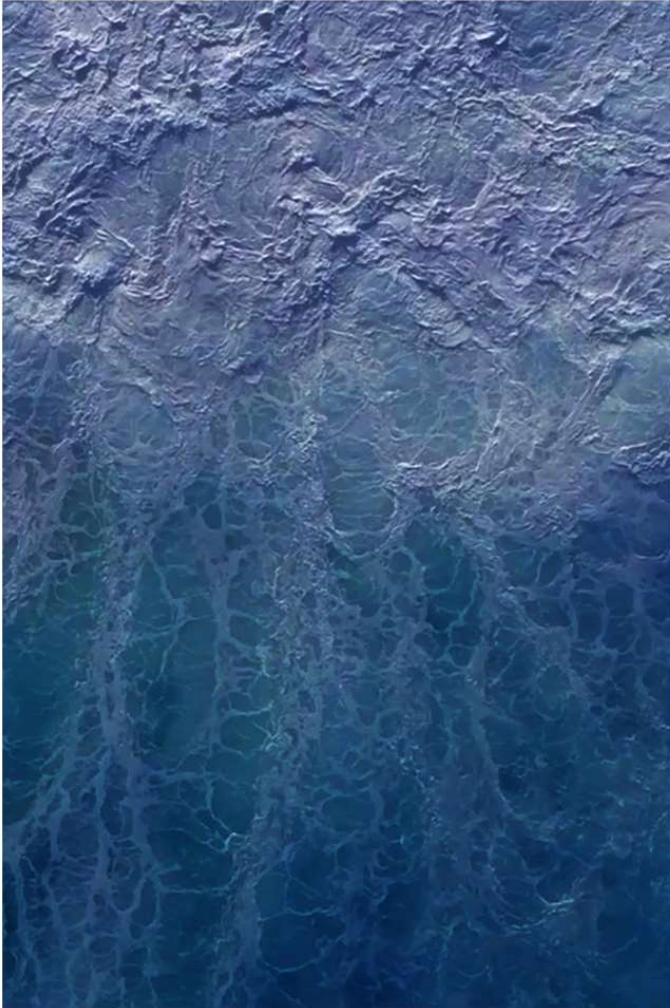
What types?

What kind of equipment
and software you
need?

How much?

Do you need to think
about storage space?

What are your file
formats?



DATA TABLE

- DATA Sheet Model
https://wiki.helsinki.fi/display/RDMforum2014/Data+management+course+materials?preview=/223985293/418336447/Datasheet_model_2022.xlsx
- Data type
- Data source
- File format (software needed?)
- Size estimate
- Personal, sensitive data? Who controls?
- Ownership, other agreements?
- Documentation
- Storage during project

Data type	Source	File format	Personal/ sensitive	Ownership/ agreements	Size estimate
Questionnaire	Collected	electronically captured, .csv	Yes	PI, informing	N = 16, VAR = 35
Analyzed questionnaire	Produced	.csv, .xlsx,	Anonymized	PI and me	
DNA sample	Collected		Yes	UH and me	N = 64
Gene sequences	Produced	FASTA, BAM, .xlsx	Yes	UH and me	2 Gb
Analysed DNA results	Produced	.csv, .xlsx	Anonymized	UH and me	
Statistical data	Reused	Database	Anonymized	Statistics of Finland	
Microscopy images	Collected	.tif	No	Me	5 Gb
Cell countings	Produced	.xlsx, .csv	No	Me	
Analysis codes	Produced	.txt	No	Co- ownership/ research group	
Lab notebooks + Metadata files	Produced	Paper, Scinote-program, .txt, .csv	No	PI and me	100 Mb
Managerial documents (consents, agreements, contracts etc.)	Collected / Produced	Paper, .pdf	No	PI and me	100 Mb



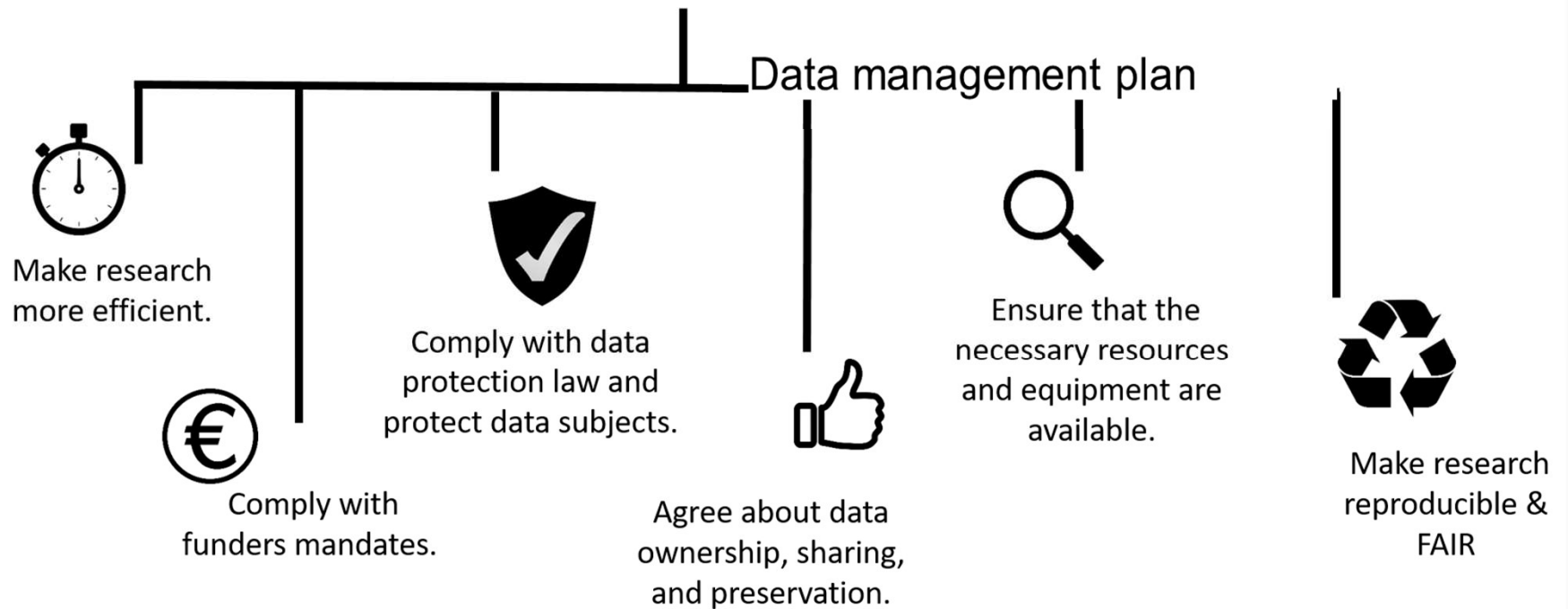
3. HELPFUL TOOLS

KUVA: VEIKKO SOMERPURO

HELSINGIN YLIOPISTON KIRJASTO
HELSINGFORS UNIVERSITETS BIBLIOTEK
HELSINKI UNIVERSITY LIBRARY



DMP: What's in it for me?



DATA MANAGEMENT PLANNING SOFTWARE



Home

Public DMPs

Public DMP templates

Help

Welcome

Data management planning tool DMPTuuli helps you to create, review, and share data management plans that meet institutional and funder requirements.

Join the growing number of researchers that have adopted DMPTuuli:



10 725 Users



12 612 Plans



50 Organisations

You can download funder templates without logging in, but from DMPTuuli you will find tailored guidance from many research organisations, including universities and service providers like the Finnish Social Science Data Archive. Why not sign up for an account and try it out?

Sign in

Create account

* Email

* Password

[Forgot password?](#)

Remember email

Sign in

- OR -

Sign in with institutional credentials (HAKA only)

DATA MANAGEMENT PLANNING SOFTWARE

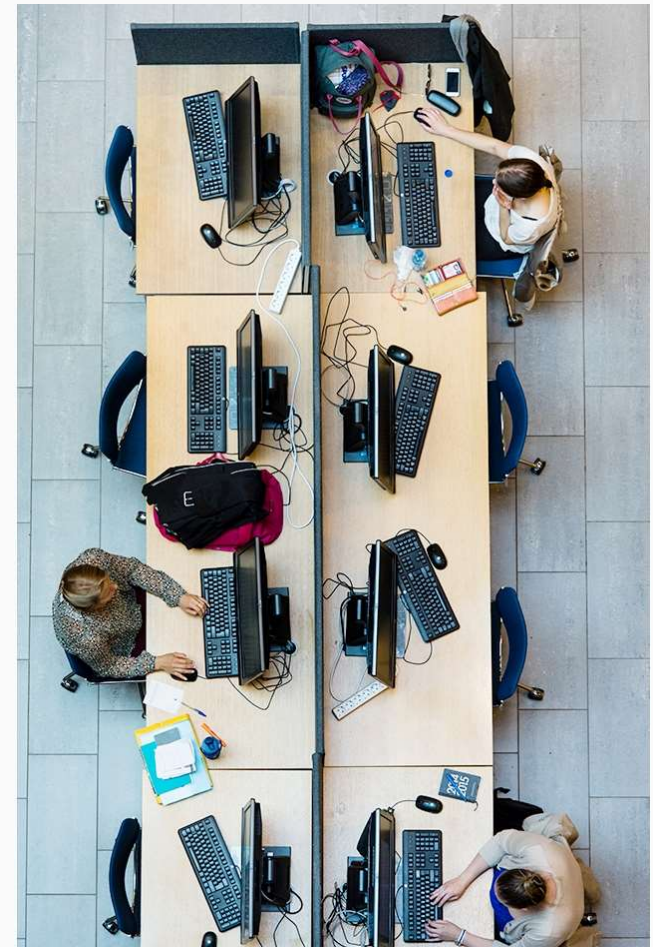
Enables working together

Funder templates, institution templates

Questions direct to describe your process

Example plans

<https://www.dmptuuli.fi/>



Answer questions
one by one

Detailed explanations
by institution

<https://www.dmptuuli.fi/>

What file formats will the data be in?

B *I* [List Icon] [List Icon] [Link Icon] [Table Icon]

Save

Comments & Guidance

Guidance

Comments

FSD

Data format

File format is a primary factor in accessing and reusing your data in the future. List the file formats you use for saving the data, such as CSV, TXT or MP3.

Tips for best practices

- Check the [digital file formats](#) used at the FSD for different types of data.
- Be aware that some statistical formats impose restrictions on the handling of missing information, the number of variables, the length of variable names, and characters allowed in variable and value labels.
- Do not include content or structural information in data files by formatting the text (i.e. using doc-format for bolding or italicization) because formatting disappears in conversions to other formats. If formatting is indispensable to understand the data, use the PDF/A file format.



4. DOCUMENTATION AND METADATA

HELSINGIN YLIOPISTON KIRJASTO
HELSINGFORS UNIVERSITETS BIBLIOTEK
HELSINKI UNIVERSITY LIBRARY

KUVA JUSSI MANNISTO



KUVA: LINDA TAMMISTO

HELSINGIN YLIOPISTON KIRJASTO
HELSINGFORS UNIVERSITETS BIBLIOTEK
HELSINKI UNIVERSITY LIBRARY

DOCUMENTATION AND METADATA

- How do you describe what your data is about?
- Do you have a system for file naming?
- Can you locate the latest version of your data?

DATA DOCUMENTATION

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.

DOCUMENTATION – METHODS

Data dictionaries and Code books	Dictionaries explain variables used in a dataset. Codebooks are collections of codes, algorithms and calculations used
Directory structure	Create a folder structure to suit your project needs
Tagging files	Tags are keywords assigned to files, which enable organizing and searching files easier
File naming conventions	Create a meaningful but brief system with unique names
Version control	Automatic version control system preferred
Readme-files	Readme-files are text documents (e.g. format.txt) providing information about data files to ensure they are interpreted correctly

METADATA STANDARDS



A model to describe data in a controlled way

Field and disciplinary specific

Typically repositories use a standard and assist with the documentation

Find standards: RDA Research Data Alliance, DCC Digital Curation Centre

Qvain <https://qvain.fairdata.fi/>

EXAMPLE: NAMING TIPS FOR FILES

Balance with the **amount of elements** in the name: too general vs. too many. Limit the name to 32 characters or less.

Use meaningful **abbreviations**

Order the elements **from general to specific**

Use the underscore (_) as element delimiter and hyphen (-) or capitalizer to delimit words within an element.

Time should be ordered (YYYYMMDD) (HHMMSS)

For **version control** use the letter V followed by two digits (V06) and extend it if needed for minor changes (V06-02).

Write a **read me file about the naming system** to explain abbreviations



METADATA RESOURCES

Making a research project understandable (Zenodo)

<https://doi.org/10.5281/zenodo.1914401>

Qvain <https://qvain.fairdata.fi/>

DisciplinaryMetadata / Digital CurationCentre DCC

<https://www.dcc.ac.uk/guidance/standards/metadata>

Metadata StandardsbySubject / ResearchData Alliance

RDA <https://rdamsc.bath.ac.uk/>

<https://rd-alliance.org/metadata-principles-and-their-use.html>

General ResearchData/ Digital CurationCentre DCC

<https://www.dcc.ac.uk/resources/subject-areas/general-research-data>



KUVA: VEIKKO SOMERPURO

5. ETHICAL ISSUES



HELSINGIN YLIOPISTON KIRJASTO
HELSINGFORS UNIVERSITETS BIBLIOTEK
HELSINKI UNIVERSITY LIBRARY

PERSONAL INFORMATION

Personal information = all identifiers from which the person is identifiable *directly* or *indirectly*.

- *Direct identifiers*: name, phone number, social security number, picture, voice, fingerprint, dental chart
- *Indirect identifiers*: gender, age, education, profession, nationality, work history, system log history, marital status, residence information, car license number



Melody Sisters –yhtye, Helsingin kaupunginmuseo CC BY 4.0

IF YOU HANDLE PERSONAL INFORMATION



..be ready to prepare:

Data management plan where you describe your workflows and assess the risks

Privacy notice & informed consent agreements for research subjects

Impact assessment

Ethical review if the handling of personal info contains high risks to the study subjects

SENSITIVE DATA

Some **personal data is sensitive** and processing it is only allowed within the grounds of [specific exceptional bases set by the GDPR.](#) (EU General Data Protection Regulation)

The researcher is responsible for identifying any data that, if revealed, could harm the data subjects.

Political opinion
Racial or ethnic origin
Religion or beliefs
Trade union membership
Genetic or biometric data processed for identifying
Health information
Sexual behaviour or orientation
Criminal convictions and offences

WORKING WITH PERSONAL DATA

Anonymous data

Data subject is no longer identifiable with reasonable effort, when personal data has been made anonymous.

Pseudonymous data

An individual data unit cannot be **re-identified** based on the **pseudonymised data** without **additional, separate information**.

1. Original dataset with personal information

name	fav food	shoe size
Matrix, John	sandwich	11,5
Deckard, Richard	noodles	9
Standish, Claire	candy	7
Baines, Lorraine	pizza	7,5
Macready, Randall	ice cream	9,5

2. Pseudonymization. Removing personal information and replacing it with non-personal coding

person 1		
person 2		
person 3		
person 4		
person 5		
	fav food	shoe size
	sandwich	11,5
	noodles	9
	candy	7
	pizza	7,5
Matrix, John	cream	9,5
Deckard, Richard		
Standish, Claire		
Baines, Lorraine		
Macready, Randall		

3. Pseudonymized dataset with separate pseudonymization key

name	fav food	shoe size
person 1	sandwich	11,5
person 2	noodles	9
person 3	candy	7
person 4	pizza	7,5
person 5	ice cream	9,5

pseudonymization key	
person 1 =	Matrix, John
person 2 =	Deckard, Richard
person 3 =	Standish, Claire
person 4 =	Baines, Lorraine
person 5 =	Macready, Randall

EASY MISTAKES

Using free
consumer versions
of cloud/survey
services

Anonymisation vs.
pseudonymisation

Wrong definition of
personal data

Informing
participants

Missing
agreements



KUVA: VEIKKO SOMERPURO

HELSINGIN YLIOPISTON KIRJASTO
HELSINGFORS UNIVERSITETS BIBLIOTEK
HELSINKI UNIVERSITY LIBRARY

GUIDANCE ON WORKING WITH PERSONAL DATA

Flamma on [data protection and informing subjects](#)

Flamma on the [need of ethical review](#)

Useful guidance by [Office of the data protection ombudsman](#) and by [Finnish social science data archive](#).



6. INTELLECTUAL PROPERTY

HELSINGIN YLIOPISTON KIRJASTO
HELSINGFORS UNIVERSITETS BIBLIOTEK
HELSINKI UNIVERSITY LIBRARY

KUVA: VEIKKO SOMERPURO

INTELLECTUAL PROPERTY - MAKE WRITTEN AGREEMENTS ABOUT

Who created
different data
types?

Who created
codes and
software?

Who has usage
right to your
data?

Who issues the
right to reuse it?

Have you
agreed on
authorship?

Who decides
about opening
and
preservation?



HELSINGIN YLIOPISTON KIRJASTO
HELSINGFORS UNIVERSITETS BIBLIOTEK
HELSINKI UNIVERSITY LIBRARY

INTELLECTUAL PROPERTY – DATA OWNERSHIP

No simple guidelines – discuss with the PI

Affected by

- Source of funding
- Cooperation with other parties
- Employment relationship

INTELLECTUAL PROPERTY – TRANSFER OF RIGHTS



Undertaking the transfer of rights agreement

Many funding agencies require data ownership to be transferred to the university.

→UH **does not** own your data unless specifically agreed upon.

→ PI's are responsible for contracts

Permanentti Helsingin kaupunginmuseo 1986 CC BY 4.0

INTELLECTUAL PROPERTY - LICENSING DATA

Every type of data in your table probably has different type of ownership and following from that, different user rights

If you own the data, you can license it

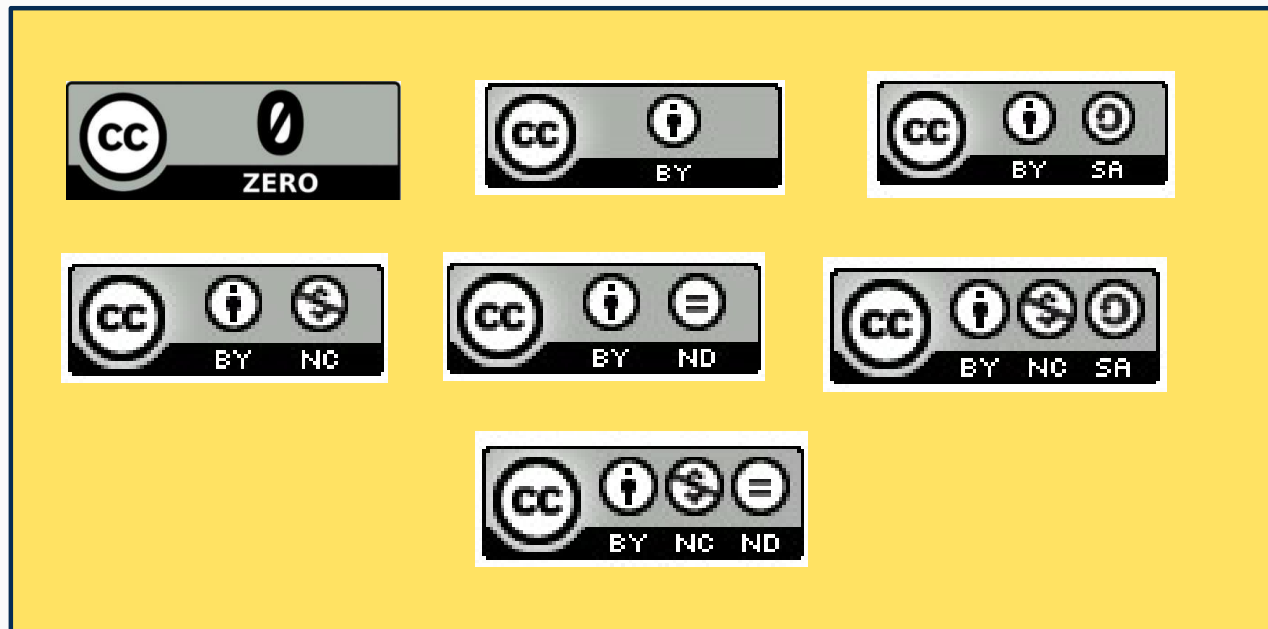
Licensing guide

<https://libraryguides.helsinki.fi/oa/eng/licensing>

Take into account any personal / sensitive data issues



INTELLECTUAL PROPERTY – LICENSING DATA



<https://creativecommons.org/>



PHOTO BY VEIKKO SOMERPURO

HELSINGIN YLIOPISTON KIRJASTO
HELSINGFORS UNIVERSITETS BIBLIOTEK
HELSINKI UNIVERSITY LIBRARY

GUIDANCE ON LEGAL ISSUES

[Data protection guide for researchers by UH \(Flamma\)](#)

[Data protection yammer group \(UH, Yammer\),](#)

ask questions and find relevant documents (e.g. informing participants)

[Instruction on concluding an agreement tutkimuksenjuristit@helsinki.fi](#)

[FSD: Informing Research Participants](#)



5. STORAGE SOLUTIONS DURING THE PROJECT

KUVA: LINDA TAMMISTO

HELSINGIN YLIOPISTON KIRJASTO
HELSINGFORS UNIVERSITETS BIBLIOTEK
HELSINKI UNIVERSITY LIBRARY

Teksti

6 . PUBLISHING AND ARCHIVING

KUVA: MIKA HUISMAN

HELSINGIN YLIOPISTON KIRJASTO
HELSINGFORS UNIVERSITETS BIBLIOTEK
HELSINKI UNIVERSITY LIBRARY

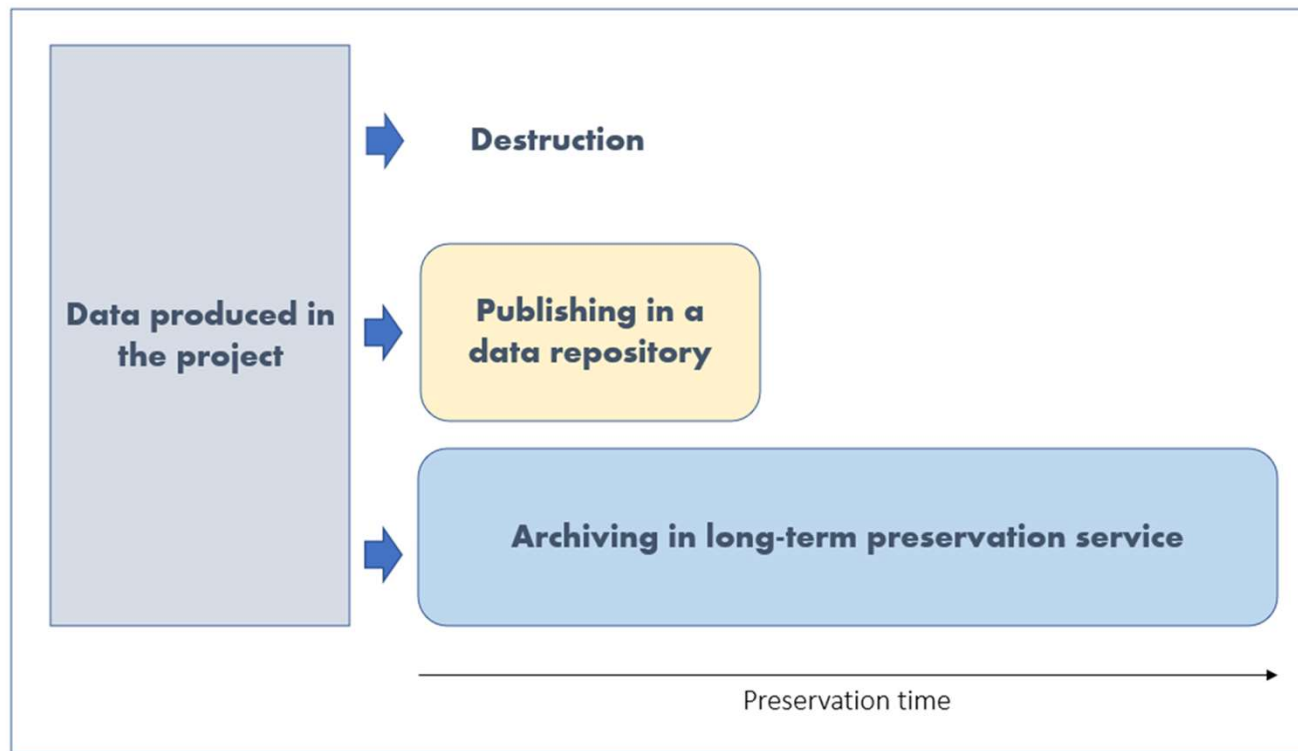
PUBLISHING AND ARCHIVING

- What happens to the data after the active phase of the project?
- How, when, where and to whom will the data be made available?
- How and where will data with long-term value be made available?

Data publishing needs to support data discovery, referencing to data, access to data and reuse of data



DESTRUCT, PUBLISH, ARCHIVE?



WHY SHARE DATA?

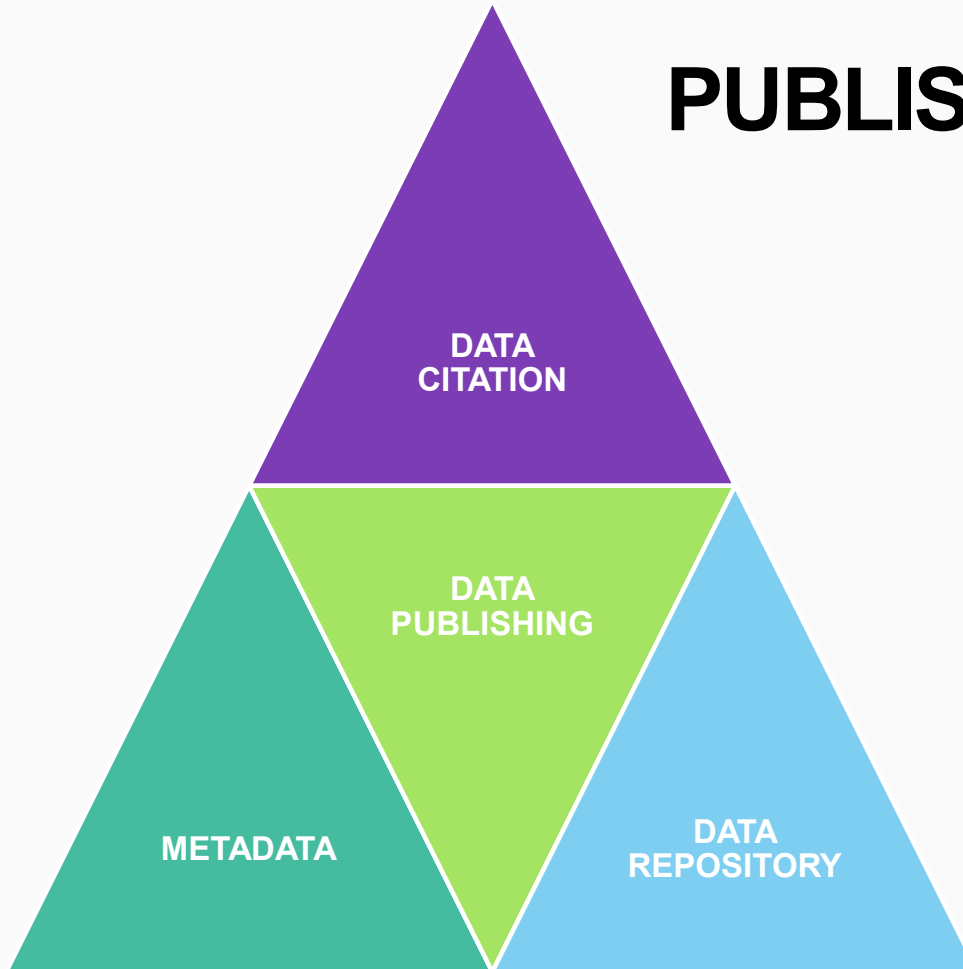
Individually

- Visibility
- Contacts and joint publications
- Merit

Collectively

- Institutional merit
- Transparency of data and methods
- Efficiency

PUBLISHING AND ARCHIVING



Metadata

- Information about the data for discovery and use

Persistent identifier

- Access, formal data citation

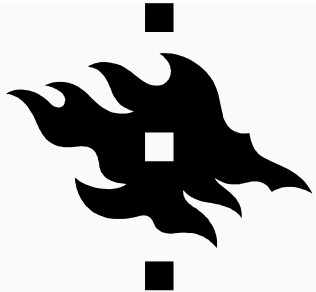
Data repository

- Access, long term archival

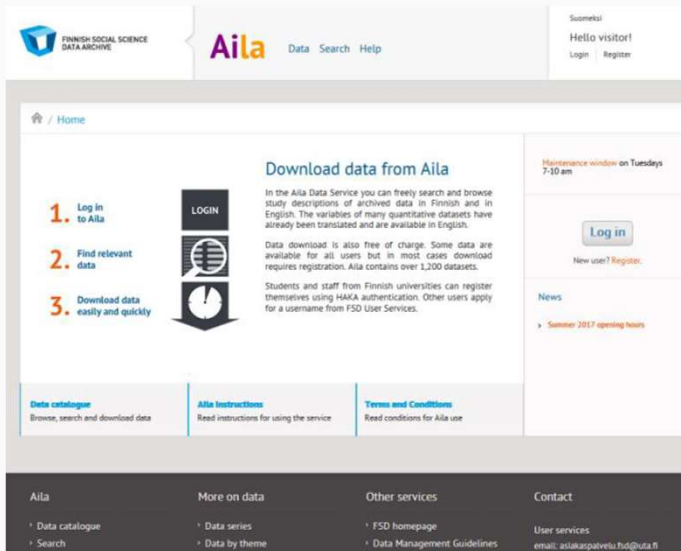
PUBLISHING AND ARCHIVING

Online Data
Repository

Data
Journal



CHOOSE A REPOSITORY



Many others, such as IDA, Zenodo, Figshare etc for publishing and archiving data

CHOOSE A REPOSITORY

- Global registry of research data repositories
- Permanent storage and access of data sets to researchers, funding bodies, publishers and scholarly institutions.
- Easy identification of appropriate repositories - curated archives in your field of research.



<https://www.re3data.org/>

PUBLISH AT LEAST YOUR METADATA

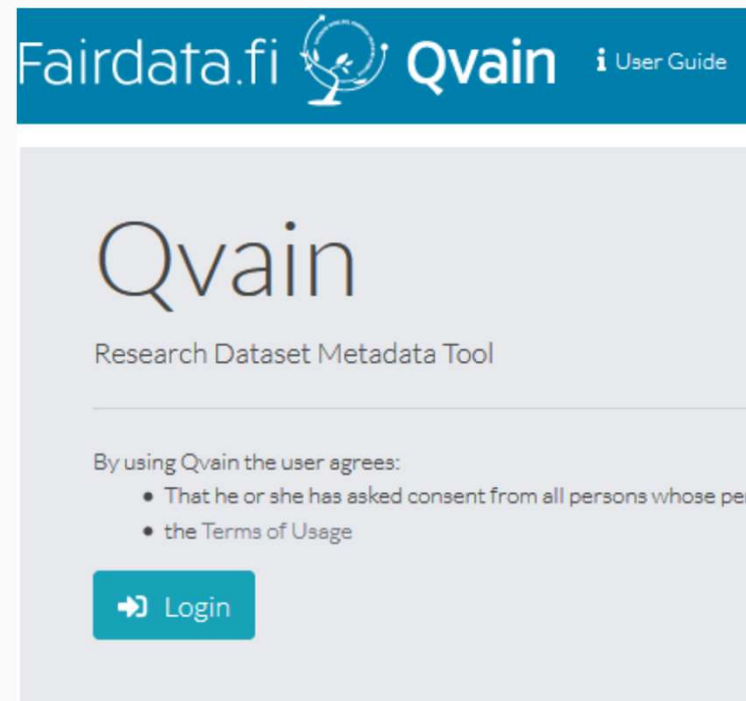
Qvain is a metadata tool, provided by CSC.

Create metadata and publish it to **Etsin** -a database for documentation of datasets.

User guide available:

<https://www.fairdata.fi/en/qvain/qvain-user-guide/>

Qvain Can use files from IDA and other repositories and includes an automated metadata checking tool.



DATA PUBLICATION

The screenshot displays the homepage of the Journal of Open Humanities Data. The navigation bar includes links for Home, About, Contact, Content, and Research Integrity, along with a search bar and options for Log In and Register. The main header features the journal's logo and buttons for 'Start Submission' and 'Become a Reviewer'. Below the header, there's a section for 'Data Papers' with a featured article titled 'A Social Network of the Prosopography of the Neo-Assyrian Empire' by Heidi Jauhiainen and Tero Alstola. The article's abstract describes a social network of over 17,000 individuals from the Neo-Assyrian period. Keywords include social network analysis, prosopography, social structure, Assyria, Assyriology, and cuneiform. The 'How to Cite' section provides the full citation and DOI. At the bottom of the article, statistics show 428 Views, 110 Downloads, and 10 Twitter shares. A sidebar on the right offers navigation options like 'JUMP TO', 'DISCUSSIONS', and 'RELATED CONTENT' with a list of links including Abstract, Overview, Method, Dataset Description, Reuse Potential, Notes, Acknowledgements, Funding Statement, Competing Interests, Author Contributions, and References.

Home About Contact Content Research Integrity Search... Log In Register

Journal of Open Humanities Data Start Submission Become a Reviewer

Reading: A Social Network of the Prosopography of the Neo-Assyrian Empire Share: f t S in Download A- A+ [Icons] All Display

Data Papers

A Social Network of the *Prosopography of the Neo-Assyrian Empire*

Authors: Heidi Jauhiainen ✉, Tero Alstola

Abstract

The dataset is a social network of over 17,000 individuals who lived during the so-called Neo-Assyrian period of Mesopotamian history, primarily in the eighth and seventh centuries BCE. The undirected network of individuals connected by co-occurrences in cuneiform documents was semi-automatically extracted from the *Prosopography of the Neo-Assyrian Empire*. In addition to two weighted versions of the one-mode co-occurrence network, the dataset also contains a two-mode person-text network and rich metadata for each individual. For the first time, the dataset allows large-scale computational analysis of social structures in the Assyrian Empire. The data is primarily stored as plain text and CSV files, inviting scholars to further expand and enrich it. The scripts and files used for creating and standardizing the data are also available in the Zenodo repository.

Keywords: social network analysis, prosopography, social structure, Assyria, Assyriology, cuneiform

How to Cite: Jauhiainen, H., & Alstola, T. (2022). A Social Network of the *Prosopography of the Neo-Assyrian Empire*. *Journal of Open Humanities Data*, 8, 8. DOI: <http://doi.org/10.5334/johd.74>

428 Views 110 Downloads 10 Twitter

JUMP TO DISCUSSIONS RELATED CONTENT

- Abstract
- (1) Overview
- (2) Method
- (3) Dataset Description
- (4) Reuse Potential
- Notes
- Acknowledgements
- Funding Statement
- Competing Interests
- Author Contributions
- References

<http://doi.org/10.5334/johd.74>

<https://zenodo.org/communities/ane/?page=1&size=20>

A photograph of a modern glass building facade, likely a library or university building. The glass reflects the sky and surrounding structures. A large, white, stylized question mark is overlaid on the center of the image. The text 'OPEN SCIENCE?' is written in white, bold, sans-serif capital letters across the middle of the question mark.

OPEN SCIENCE?

HELSINGIN YLIOPISTON KIRJASTO
HELSINGFORS UNIVERSITETS BIBLIOTEK
HELSINKI UNIVERSITY LIBRARY

OPEN SCIENCE

- Choice based on your values
- Attitude directing your actions; even on the grassroots level
- Not free of charge – allocate time and money!
- Poor title for its details – **Responsible science?**



OPEN SCIENCE



- Opening is a systematic and controlled process comprising of many phases
- Under no circumstances should opening be an accident, but a decision
- Open science is worth striving for
- Closing or securing the data needs to be justified
 - "I do not want to open" is not enough
- Opening is scary!

OPEN SCIENCE – MEET THE REQUIREMENTS

Persistent identifiers

ORCID for researchers <https://tutkijatunniste.fi/tutkijat/>

DOIs or URNs for publications and data <https://www.doi.org/>
<https://www.kansalliskirjasto.fi/fi/palvelut/tiedonkuvailun-asiantuntijapalvelut/urn-tunnukset>

Licences

How to license your own publications?

How licenses direct the (re)use of already published data?

<https://libraryguides.helsinki.fi/oa/lisenssit>



OPEN SCIENCE



Aila

Data Search Help

“As open as possible, as closed as necessary”

Home / Data catalogue

Study descriptions in English

In total 1273 datasets.

Display the data catalogue [in Finnish](#).

Availability

- A ⓘ
- B ⓘ
- C ⓘ
- D ⓘ

Type

- Qualitative
- Quantitative

You can filter the studies in the catalogue by availability status, data file language or type. Study descriptions can be sorted by ID number, title, availability, data type or publication date. Only one sorting criterion is allowed at a time. [More information](#) on availability and conditions.

Dataset availability:

- (A) openly available for all users without registration,
- (B) available for research, teaching and study,
- (C) available for research only (including e.g. Master's, licentiate and doctoral theses),
- (D) available only by permission from the data depositor/creator.

You can browse archived data, for example, by [keyword](#), [topic](#) and [series](#).

https://ec.europa.eu/research/openscience/pdf/openaccess/ord_extension_faqs.pdf



FAIR DATA?

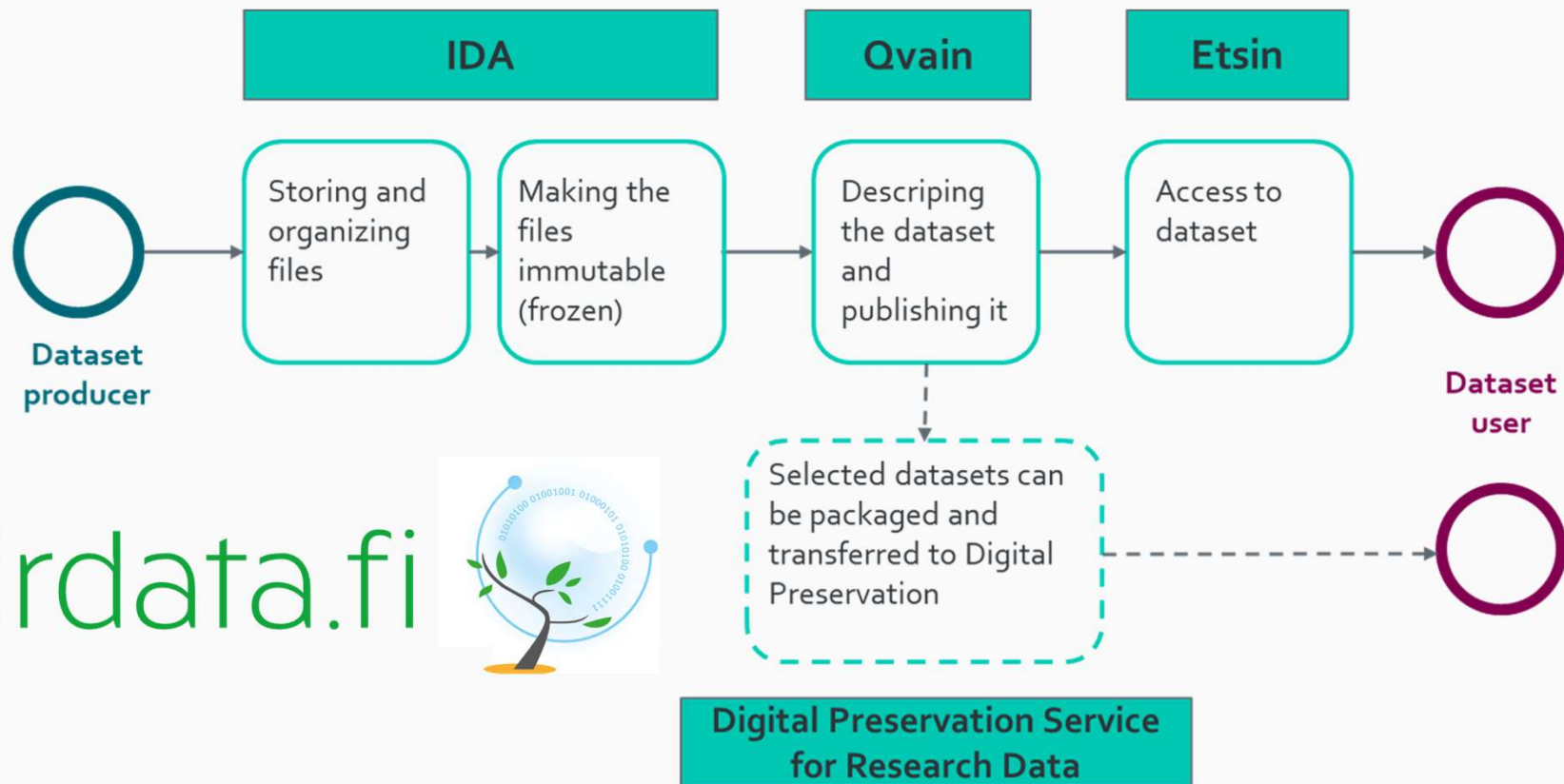
<https://www.fairdata.fi/en/about-fairdata/fairdata-services/>

FAIRDATA Services

Finnish Ministry of Education and Culture

<https://www.ands.org.au/working-with-data/fairdata/training>

FAIR DATA?



Fairdata.fi



Teksti

RESOURCES AND RESPONSIBILITIES

HELSINGIN YLIOPISTON KIRJASTO
HELSINGFORS UNIVERSITETS BIBLIOTEK
HELSINKI UNIVERSITY LIBRARY

RESOURCES AND RESPONSIBILITIES

Who is responsible for data management tasks?

Who is responsible & controls *data protection* and *information security* issues?

Who is data controller, who is data processor?

Are responsibilities allocated to one person or is the whole research group involved?

Is an expert / employee needed?

What resources (time & workload) is needed for data management?

DATA MANAGEMENT - RESOURCES

Tasks	Resources
Data management planning	1 week
Agreements (consortium, transfer of rights)	2-4 weeks
Data privacy (GDPR) <u>admistration</u>	2-4 weeks
Data documentation and cleaning	1-2 hour/week/person (~5% of the project FTE)
Data publishing (include checking the anonymization)	1-2 week(s)/data set (8 main data sets)
Storage space for sensitive data	10 TB = 2 000€/year
Archiving and deleting data	1-2 week(s)/data set (5 unpublished data sets)

Expert help for data management, preservation, and sharing tasks is provided by UH Data Support

DMP, 2019

DATA MANAGEMENT TAKES TIME



Learning new standards, new regulations and new software

Documenting files, doing version control

Recruiting help, buy storage space or new equipment

Preparing privacy notices, agreements, finding out on legal/ethical reviews and consulting lawyers



FURTHER TRAINING

KUVA: JOEL GRANDELL

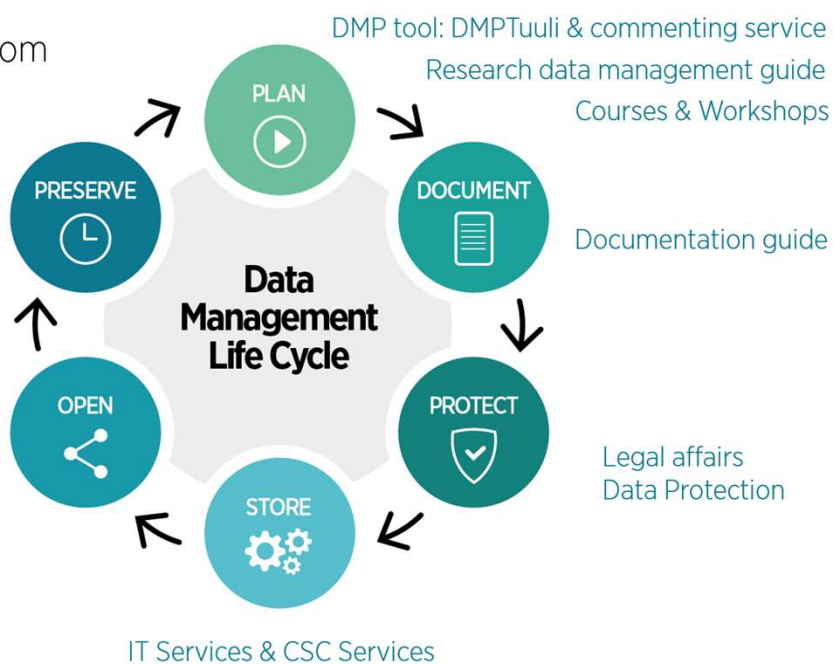
HELSINGIN YLIOPISTON KIRJASTO
HELSINGFORS UNIVERSITETS BIBLIOTEK
HELSINKI UNIVERSITY LIBRARY

DATA SUPPORT SERVICES

Data management services available from Data Support

datasupport@helsinki.fi

Long term preservation (PAS)




TURN TO UH DATA SUPPORT!

IMPROVE YOUR

DMP

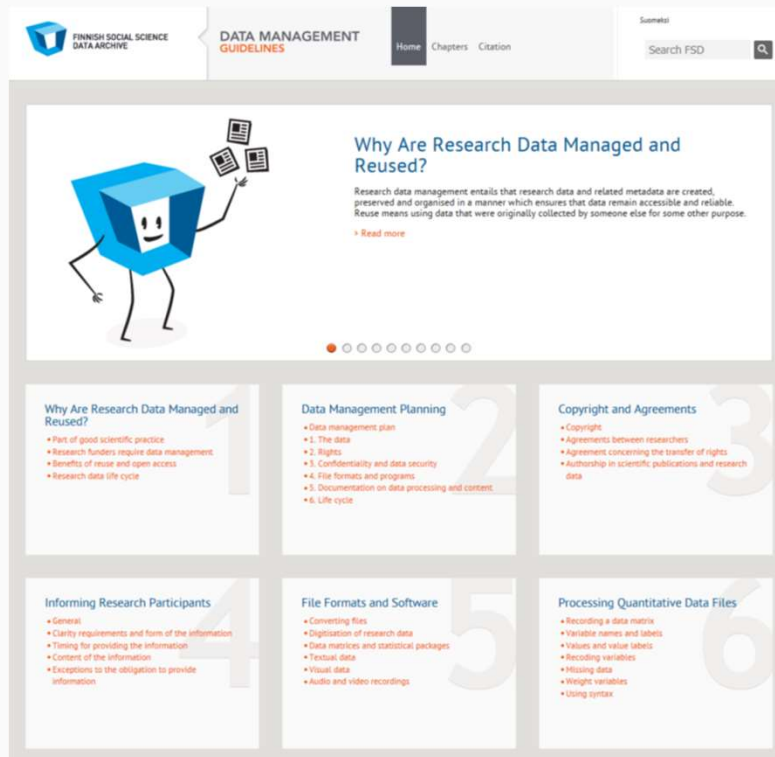
BY SENDING IT TO
DATASUPPORT@HELSINKI.FI

Send your Data Management Plan 10 days
before call deadline.*



*It is not guaranteed
that all DMPs get
comments, plans are
handled in arrival
order.

FSD DATA MANAGEMENT GUIDE



- FSD's Data Management Guidelines
<http://www.fsd.uta.fi/aineistonhallinta/en/>
- Excellent guide on all RDM issues!
- Might be handy to keep the page open in another window while outlining a DMP with Tuuli
- Especially useful on dealing with sensitive data!

DATA SUPPORT ON DIFFERENT LEVELS



<https://www.helsinki.fi/en/research/services-researchers/data-support>

datasupport@helsinki.fi

DATA MANAGEMENT SIMPLIFIED

Know your
data

Describe
your data

Privacy and
intellectual
property

Securing,
storing,
sharing

Archiving,
opening,
publishing

THANK YOU

Maija Paavolainen
Information Specialist
Helsinki University Library
Maija.paavolainen@helsinki.fi
0504150489



@hulib



@HULib



@Helunilib



hulibvideot

#HULIB

www.helsinki.fi/kirjasto

KUVA: TUOMAS UUSHEIMO

HELSINGIN YLIOPISTON KIRJASTO
HELSINGFORS UNIVERSITETS BIBLIOTEK
HELSINKI UNIVERSITY LIBRARY





HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI