Improve the quality and impact of your research through data management

A guide for making your data FAIR

Open Science

Contents

What is research data?

Research data refers to data that is generated and/or utilised in research.

Data lifecycle refers to all the stages of existence of digital data, from creation to digital preservation or destruction.

Managing digital data requires planning and know-how. Unmanaged data can become corrupted by themselves, or they can be messed up or lost.

Responsible research requires openness and making it possible to repeat or review any work carried out. Data and software are an important part of research.

What are the FAIR principles?

Findability, *Accessibility*, *Interoperability* and *Reusability* are objectives that should be taken into account to support good data management, as well as the quality and impact of research

- More on the FAIR principles: https://www.go-fair.org/fair-principles/ and https://www.fairdata.fi/tietoa-fairdatasta/fair-periaatteet/

The FAIR principles are intended for all forms of data, both quantitative and qualitative, and their objective is to ensure that the data is machine-readable and interoperable.

- The principles also apply to metadata with which data is described and made findable.

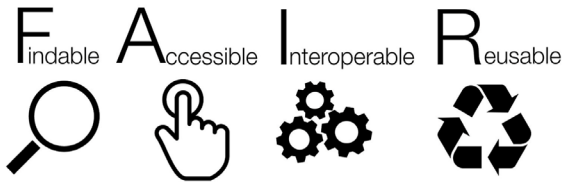
FAIR data is understandable to people and can also be processed programmatically. For example, a table that is understandable to people in a PDF file is not easy to edit, while a CSV file is. Well-structured data can be combined with other data in the same format, for example, or searched.

FAIR principles and the researcher

The researcher can approach the subject by thinking about how well they could repeat the work carried out: can the data and methods be found, are they understandable and usable to others?

Research funders and scientific publishers may require adherence to the FAIR principles and administration of data.

The purpose of the instructions provided in this document is to support researchers in the planning of their own work so that adhering to the FAIR principles is as easy as possible. The end of each section features a researcher's checklist regarding the subject matter of the section.



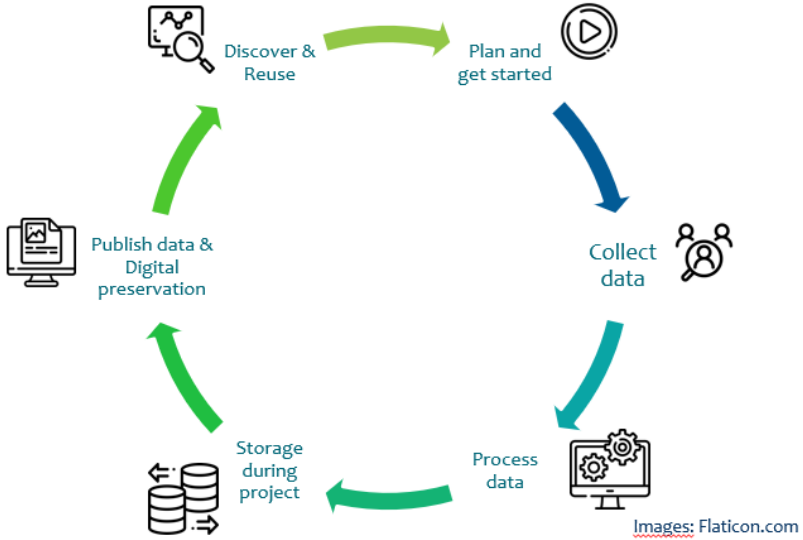
Data management as part of research

Making all of the research FAIR

The FAIR principles must not be separate from the research process. They should be adhered to throughout the entire process, and they can be applied to all research outputs.

Reproducibility can be a criterion used to decide what is stored. In such cases, it is imperative that the researchers also document and store thei workflow.

‘FAIR by design’ thinking involves aiming to plan t entire research process so that, for example, metadata is accumulated automatically and the necessary identifiers are easy to create.



Data management begins in the research planning phase (DMP – Data Management Plan)

The FAIR principles are to be implemented throughout the entire research process. So, be sure to plan

- the entire lifecycle of the data and take it into account in the planning of your research (e.g. versions to be stored, digitally preserved and destroyed)
- the technical specifications related to the data and matters related to rights
- the different forms of documentation (e.g. source code, code book, guidelines) that will contribute to openness.

Responsible research entails making the process open, i.e. explaining what has been done to the data and how the end result was achieved. This must be taken into account in the data management plan, which must be created in the project planning phase and upheld as the research progresses.

On the data lifecycle (1): The maturity level of data

Raw data refers to data in its original form, as it was obtained directly from the source and before it was processed in any way.

Primary data refers to data collected or created and maintained for a specific purpose, e.g. a ‘master copy’ of the data used by the researcher, for which the author is responsible.

Published research data refers to a referenceable data entity that can grow through an managed quality, documentation and versioning process. At a minimum, its metadata is publicly available and any restrictions to its availability are determined in a machine-readable manner. It can consist of data from several primary sources.

On the data lifecycle (2): Creation and preparation

The creation of data can involve different instruments or collection methods that are important to document.

From raw data to analysable data

- In the first stage, the data collected for research is so-called raw data. Raw data is the starting point of data processing, and it should always be possible to return to it.
- For analyses, the raw data is **validated** and **described**, and it is sometimes enriched and classified before it can be analysed.

Various commands or other types of code are often used in the preparation of data

- **Program code** means the tool with which data is produced. For example, the models and the analysis and data processing code used (e.g. R scripts) should be preserved, documented and published if there are no grounds for non-publication.

On the data lifecycle (3): Refinement and management

Code book refers to expression of variables, entities or categories in the data. Such dataset-specific code books must be documented and described together with the data. Aim to use existing standards and/or code books in order to improve understandability and reproducibility.

Workflow refers to a series of functions that can be used to standardise the work process and ensure consistent results. Workflows structure processes and improve efficiency. A workflow facilitates the recording and publication of the data lineage and provenance.

When a data object is created or made into a modified version, it is assigned a version number. **Version management** involves keeping a record of stored versions and facilitating reversion to them, which is important in terms of the traceability of data, modification tracking and error correction.

Instructions for the lifecycle of data and its documentation

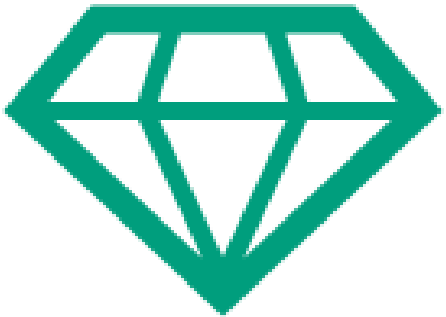
- Store the raw data if its preservation is possible and appropriate.
- Document the data sources. Use persistent identifiers, if possible (e.g. for source datasets, infrastructures, sensors, protocols).
- Document the processing of data and its principles in detail, such as the validation process, whether spelling errors have been corrected, conversions, etc.
- Plan the versioning, digital preservation and destruction of raw data as well.
- Make sure that you can revert to earlier versions, if necessary, by ensuring that a sufficient number of versions is created over the course of the work. Ensure that the versions can be identified unambiguously.
- Ensure file integrity by calculating checksums for the files and remember to make backup copies. Record these actions as well.
- Think about what is necessary in terms of the repeatability of the research.
- Detailed documentation is important, particularly with regard to data to be preserved digitally, e.g. in terms of the devices and software used.

CHECKLIST: Can your research be reproduced?

1. Is your work steered by a data management plan throughout the entire data lifecycle so that all the data processing procedures are open and sufficiently documented?
2. How have you taken into account the openness of data and usage restrictions throughout the process?
3. Have you utilised shared practices, such as standards and glossaries, in the metadata and the actual data?
4. Have you systematically documented the lifecycle of the research data and is the description accurate? Are as many data processing stages as possible automated and is the code stored? Are the software and settings used documented (technical documentation)?
5. Have you versioned the data and other outputs?
6. Have you stored the data and its documentation in a referenceable form (persistent identifiers and metadata)?

When reproducibility is ensured, the FAIR principles are implemented naturally as part of the research process and there is no need to create data or documents separately in the publication phase of the article or data.

Assessing the value and quality of data



FAIR principles and the value of data

The reusability of data increases its value to the scientific community and society. Researchers are responsible for determining the value of data on their part and take care preserving its value.

Managing data in accordance with the FAIR principles facilitates responsible reuse of the data and increases the impact of the research.

The more detailed the documentation of the origin of the data is, the easier the data is to reuse. This is why the workflow through which raw data is made into published research data must be recorded.

The researcher may collect the data in order to establish an answer to a question that they have posed within their own discipline, but the data may also have value as part of a larger whole, e.g. in the organisation’s database, as comparison data, as part of a meta analysis or within another discipline.

The value of data is usually increased by its uniqueness and the amount of resources used on its production. The data also often has value from the perspective of the openness of the research, evaluation and peer reviews.

Changes in the value of data

The value of data can decrease or increase over time: data collected in order to solve a current problem may only be used once, but it may gain new significance, e.g. as a part of a time series, as a reference for change or as a part of the cultural heritage. Research data may gain value as cultural heritage over time

- More on the cultural heritage value of data (in Finnish)

It is difficult to predict what data will be valuable in the future, so careful selection and ensuring the understandability of the data to be preserved are an essential part of the research process.

- Example: monitoring data regarding the thawing of the Aura River in Turku, collected from archival data and old newspaper information alike (full research article). Now, 100–200 years later, the data can be used to analyse the impact of climate change.
- Preservation can also increase the value of raw data. For example, seemingly inconsistent observations can be the firsts signs of a change starting to take place. Because of this, raw data can be worth storing even if it appears to contain inexplicable errors at the time of collection.

Published instructions for assessing the value and quality of data

There are many methods and criteria for determining the value of data. For example, its value can be assessed based on its integrity, comprehensiveness, reliability and consistency. If not all of the data can be preserved, determining its value and quality supports the selection of data for digital preservation in particular.

Next, we will present the main points of two published guidelines that researchers can utilise and apply when assessing the value and quality of their data:

- The University of Helsinki's guidelines for determining the value of data for digital preservation: Krister Talvinen. (2019). Digital Preservation (Fairdata-PAS): Guidelines for UH Evaluators. Zenodo.
- Statistics Finland's information material quality assurance framework

The guidelines of the National Archives of Finland also contain a chapter that focuses on research materials and value assessment from an archival perspective in particular,

Perspectives on assessing the expected reuse of data

- Is your data, as a whole, comprehensive enough, so that it can be used in different ways in the future? Comprehensiveness can refer to a certain phenomenon, period of time, etc.
- As for the technical requirements of your data, is the data sound enough to enable varying usage in the future? Or if the data has to be corrected programmatically, does the expected value of data outweigh the costs arising from improving the data?
- Is your data usable as complementing other data?
- Is your data usable as a point of reference in comparison with other data?
- Is your data only partially analyzed?
- Is it reasonable to expect that with future research methods your data can be utilized further?

Krister Talvinen. (2019). Digital Preservation (Fairdata-PAS): Guidelines for UH Evaluators. Zenodo.

Perspectives on assessing the future value of data

- Is your data crucial for the future progression of the discipline or some areas of it?
- Can further utilization of your data lead to significant scientific discoveries or publications?
- Is your data scientifically or culturally unique?
- Can further utilization of your data lead to commercial applications, business collaboration or patents?
- Does your data have significant educational value, so it could be utilized e.g. in researcher training?

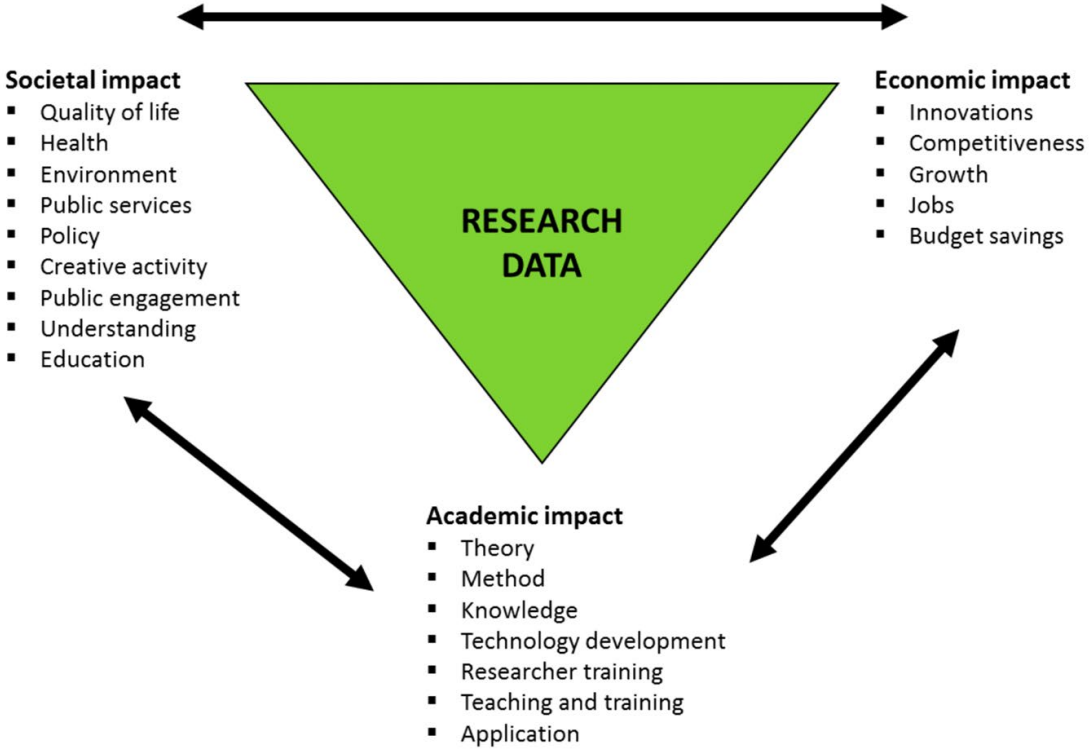
Krister Talvinen. (2019). Digital Preservation (Fairdata-PAS): Guidelines for UH Evaluators. Zenodo.

Perspectives on assessing proven significance

- Has your data been used in some particularly significant publication or scientific discovery?
- Is your data crucial for national or global research infrastructures?
- Has your data been used in significant research collaboration between various organizations?
- Has the production of data, in itself, required significant investments and resources?
- Has your data already previously been assessed by e.g. an Ethics Committee? Ethics Committee reports may be useful for decision-making, and the preservation of data can sometimes be justified for ethical reasons.
    - For example, guidelines for ethical review in human sciences (in Finnish)

Krister Talvinen. (2019). Digital Preservation (Fairdata-PAS): Guidelines for UH Evaluators. Zenodo.

The dimensions of research impact



Krister Talvinen. (2019). Digital Preservation (Fairdata-PAS): Guidelines for UH Evaluators. Zenodo.

High-quality data corresponds with reality

Correctness
means how well the data corresponds with reality. Examining the correctness of data may also lead to discoveries of systematic distortions in the information material.

Accuracy
means how well the information of the data corresponds with the objective. Accuracy indicates how truthful the data is.

(Up-to-dateness)
if significant

Consistency
means that the data is uniform and self-consistent. Consistency can also indicate consistency between different datasets.

Comprehensiveness
means the intended temporal and areal comprehensiveness of the data, as well as the intended target units and characteristics. Comprehensiveness also indicates the extent to which the data contains the information sought.

https://www.stat.fi/org/vuosiohjelma/tietoaineistojen-laatukriteerit.html

High-quality data is comprehensively described

Originality
means that the data and any changes to the information therein can be traced. The origin of the data is known.

The understandability of metadata
means how comprehensively the metadata describes the data and helps to understand the content.

Compliance with recommendations
means that the data and its characteristics comply with established standards, practices and regulations, which are also indicated in connection with the data.

https://www.stat.fi/org/vuosiohjelma/tietoaineistojen-laatukriteerit.html

High-quality data can be reused

Machine readability means whether the data has been structured so that it can be processed mechanically and in different information systems.

Usage rights mean what usage rights have been assigned to the data and what can be done with the data, i.e. what purposes the information material can be used for.

Timeliness means that the data can be used at the time stated and with sufficiently frequent complementation.

As open as possible, as restricted as necessary

https://www.stat.fi/org/vuosiohjelma/tietoaineistojen-laatukriteerit.html

How valuable is your data?

Every research data entity is unique in its own context. Describing your data and any research conducted based on it, as well as opening the data, increases its value to research. However, not all data can be preserved forever, and someone must bear long-term responsibility for preserved data.

When determining the value of data, you as the researcher and the organisation responsible for the data can consider:

- The prospects of reusing the data
  - the significance to research now and in the near future, or in the long term
  - usage within the researcher's discipline vs. broader usage as complementary data or data that facilitates comparison
  - The amount of unanalysed information contained in the data
- The quality of the data and documentation
  - how it supports the repeatability of the research or quality assurance
- The monetary value of the data
  - investments and work resources, commercial value
  - the interests of the research funder
- The significance of the data to society through research impacts
  - e.g. its significance in resolving humanity's abject problems.
- The historical and cultural significance of the data, its scientific and cultural uniqueness
  - as part of long time series for researchers and society alike.

Data sharing and storage

On the data lifecycle phases

Data sharing can be started during the research project. In such a case, the dataset can be active. The management and curation of data are carried out as agreed on based on the data management plan so that all parties involved know the rights and responsibilities related to the data, e.g. who is allowed to edit the data, how versioning is carried out and how the data can be used in general. *For example, you can share data via a network drive, a data pool or email, or in a service intended for the sharing of confidential data.*

Data storing involves documenting the data carefully and attaching metadata relevant to its usage and management. *You can store data in a suitable service (data repository or data service).*

Publishing data makes the data referenceable material that must have a persistent identifier and referencing instructions. Published data is carefully documented, and at least metadata that facilitates finding the data has been published. The metadata indicates how the material can be accessed for use and of what quality it is, so that it can be used for repeating a research process or conducting new research. *You can publish data via a suitable service (data repository or data service).*

Digital preservation of data means preserving the understandability and integrity of the data in the long term, across decades or even centuries. It involves taking into account aspects such as changes in different technologies, mediums and formats. Confidential materials can also be preserved digitally in certified services. *You can digitally preserve data at suitable services, some options are presented in this guide.*

Note! Access to the data can be restricted in all phases. Sufficient data protection and security must be ensured!

Assembling a dataset with further use in mind

A dataset is an entity assembled from data. When data is published, a statement must be made on its level of refinement and documentation, as well as aspects such as

- its granularity and resolution, as well as the reference accuracy required (identifiers)
- the file format, how the data is divided into files and how these are documented (structure, identifiers)
- how the published material is potentially versioned or added to in the future
- expressing time series and different variables in the metadata and making them searchable across datasets.

Take the FAIR principles into account when making decisions – think about the referenceability, findability, interoperability and further use of the materials.

- Data is often easier to divide into parts than to reassemble from separate files.
- On the other hand, large files can be cumbersome to transfer and process.
- Try to think of future users and new methods of use.
- Remember that your data can also be used in secondary publications (which are not new versions of the primary publication).
- Adhere to good practices established within your discipline.

Selecting storage and publication services

- The FAIR principles require that the outputs of the research have a persistent identifier and their metadata is findable.
  - In practice, this requires utilising shared services intended for research data publishing.
- Establish your minimum requirements for the publication service.
  - Data security, availability, accessibility, tool support, data models, persistent identifiers
- If possible, utilise discipline-specific services.
  - Consider who could later utilise your data – where can it be found the easiest?
    - machine readability of metadata, search options
  - What are the resources available?
  - Check the different formats, the interoperability of the content, the structure of the content.
  - Remember that copies can be made of the same material for different purposes.
- Determine usage rights for the data.
  - Contractual grounds
  - Policies of the funder and the organisation

Identification of reliable services

Certificates, such as the CoreTrustSeal, are helpful when assessing the reliability of services.

Suitable options can be sought via services such as re3data.org . Select a discipline-specific service, if possible.

Consult the instructions and support services of your own organisation.

In Finland, reliable services suitable for researchers are provided by research organisations, as well as cultural heritage organisations and CSC.

Transferring rights or sharing them with an organisation or service does not remove copyrights, but it facilitates long-term material management.

Publishing static data

Static data is an unchanging entity, e.g. the background materials or data of a research project on which the article published is based. It can also be a sample from other data sources. A reference and a Data & Code Availability Statement are attached to the article. **The dataset and its different versions must be unambiguously identifiable.**

The research evidence can often consist merely of a well-documented file or collection of files or materials if new materials were generated in the research project. In such cases, take software and codes and their sharing into account as well.

Select an openly documented, common and open file format, as this facilitates further use of the data.

Usually, good documentation requires both human-created descriptive metadata on the content and automatically produced metadata on the technical characteristics of the data.

Other documentation is also often required, as well as references to different sources, such as standards, glossaries, methods, code sets, etc.

Publishing changing data

There are different kinds of changing data. **Accumulating data** means that materials are cumulated and no changes are made to already collected data. **Dynamic data** means that retroactive changes can also be made to the data.

In accumulating data, new materials can be added as new files or updates to old files. Accumulating data can be published in some services.

As scientific materials, databases require special documentation, planning and maintenance. However, the same recommendations apply to both the description and the publication of static data.

In order to remain operational, databases require maintenance, which can be costly in the long term. Because of this, their lifecycle and maintenance arrangements after the end of the research project funding must be planned in advance.

Interoperability and utilising standards applied in digital preservation make it easier to transfer data to new environments, if needed.

For the sake of the repeatability of the research, pay attention to the need for exact and unambiguous referencing. **A persistent identifier should always direct the user unambiguously to the right data.**

Examples of referring to changing data

A. Refer to a certain data subset by providing the source, an exact demarcation and a persistent identifier

- Hypothetical example: Data Request T. Jansen; SAHFOS; Work published 2014 via SAHFOS ; Area Def: 54-65°N, 0-45°W. Temporal Def: 1980-2012 (April-August) Taxonomic Def: All zooplankton; (dataset). https://doi.org/10.7487/2014.15.1.1

B. Refer to a copy of the data at the given time.

- Hypothetical example: König-Langlo, G., & Sieger, R. (2010). BSRN snapshot 2010-01 as ISO image file (3.75 GB) [Data set]. PANGAEA - Data Publisher for Earth & Environmental Science. (dataset). https://doi.org/10.1594/pangaea.833424

C. Refer to constantly updated data by stating the precise time of reference

NOTE: This method does not necessarily facilitate repeatability, but it is sometimes the only way.

- Hypothetical example: Doe, J. and R. Roe. 2001. The FOO Data Set. Version 2.3. The FOO Data Center. (dataset). https://doi.org/10.xxxx/notfoo.547983 . Accessed 1 May 2011.

D. Refer to a time-stamped database query in a versioned database.

- Hypothetical example: R. Roe. 2017. "The Moo Data Query" created at 2017-07-21 10:25:30 PID https://doi.org/10.xxxx/notmoo.857988 Subset of Moo Database (dataset) PID https://doi.org/10.xxxx/bigmoo.360873

Based on the DataCite guidelines

Usage-restricted data

The use and storage of data can be subject to ethical, legislative or contractual restrictions.

The contractual and process and data security related actions required for usage depend on material-specific restrictions.

Special care must be taken when dealing with data that contain personal data or are owned by other parties and are subject to contracts and conditions.

Contact the data support unit of your organisation if you are unsure about matters related to usage rights. Your organisation's data policies can also provide you with instructions that you can follow in good conscience.

In order for usage-restricted data to be referenceable, it must have enough descriptive metadata, an identifier and a landing page. The metadata must indicate what the restrictions are based on and how the data can be accessed.

- E.g. with a CLARIN ACA licence, in which the material is provided for research use only (contains personal data), e.g. End user license +NC +PRIV +DEP +OTHER v2.1 (in Finnish)

Digital preservation of data in Finland

Digital preservation of research data means that the understandability of the data is preserved essentially permanently. Changes in hardware, software and researcher generations and research paradigms are taken into account by editing the data when necessary.

Data classified as valuable must be preserved for as long as possible. Assessing the value of data involves many dimensions, which were discussed in the previous chapter.

- The value of the data and any questions related to legal matters and ownership must be taken into account when planning the research in order to make the digital preservation transfer process as painless as possible.

In Finland, digital preservation is provided to organisations by operators such as the PAS service of CSC. The Fairdata PAS service was developed for the digital preservation of research data. Transferring data to the Fairdata PAS service always requires an agreement between the organisation and the Ministry of Education and Culture – i.e., from the researcher's perspective, the path towards utilising the service starts with negotiations with the researcher's organisation.

CHECKLIST: Are you up to date on data matters?

1. Plan carefully: What are the valuable outputs suitable for storage that are important in terms of repeatability and responsibility?
2. Are all contractual, maintenance and usage right matters in order for possible future use?
3. How long should different outputs be preserved and when are materials potentially destroyed, e.g. unnecessary versions? The requirements of the funder and background organisations are a factor.
4. Are the publication and storage services of your choosing reliable, suitable for your materials, adhering to the FAIR principles and, preferably, discipline-specific?
5. Have you considered future users when compiling and documenting your materials?
6. Have you considered that the data may require digital preservation in order to facilitate reuse, in accordance with its cultural heritage value or the funder’s requirements?
7. Have you reported having data worth digital preservation to your background organisation?
8. Have you informed the research subjects of how the materials will be shared, stored and published?
9. Have you remembered to refer to the data in the publications in which you have used it?

Authors

Anneli Lehtisalo
Ari Asmi
Hanna Koivula
Heidi Troberg
Jessica Parland-von Essen
Juha Hakala
Katja Laine
Maria Söderholm
Marjut Vuorinen
Mika Virtanen
Nina-Mari Salminen
Pekka Nygren
Saila Huuskonen
Sonja Sipponen
Tanja Lindholm
Tarja Mäkinen
Timo Taskinen
Tomi Rosti
Tuomas Alaterä
Tuula Pääkkönen

Open Science