GUIDELINES to FAIRify data management and make data reusable

PARTHENOS

ABOUT THIS GUIDE

This guide offers a series of guidelines to align the efforts of data producers, data archivists and data users in humanities and social sciences to make research data as reusable as possible.

The guidelines result from the work of over fifty PARTHENOS project members. They were responsible for investigating commonalities in the implementation of policies and strategies for research data management and used results from desk research, questionnaires and interviews with selected experts to gather around 100 current data management policies (including guides for preferred formats, data review policies and best practices, both formal as well as tacit).

With a focus on (meta)data and repository quality, the PARTHENOS team extracted a set of twenty guidelines which different disciplines have in common.

For easy reference, the team assigned each of the guidelines to making data Findable, Accessible, Interoperable or Reusable. This subdivision is based on the FAIR Data Principles which were first published by FORCE11 (2016) and are intended to guide those wishing to enhance the reusability of research data. Each of the PARTHENOS guidelines is accompanied by specific recommendations for data producers and data users on the one hand and for data archivists on the other hand. The icons below the guidelines visualise which stakeholder is addressed.

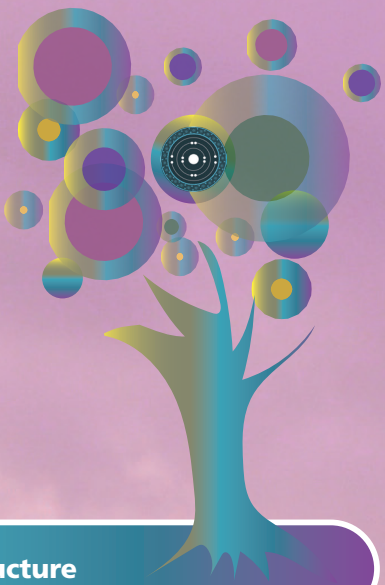
The lamp icon shows recommendations for data producers and data users such as researchers and research communities in history, archaeology, language studies and social science studies.

The wheel icon shows recommendations for research infrastructures and data archives in research institutes and cultural heritage institutions.

PARTHENOS is a consortium of sixteen European research institutions and infrastructures. PARTHENOS members aim to increase the reusability of research data by building bridges between the data life cycles of research communities, data repositories, research infrastructures and cultural heritage institutions in the interrelated fields of the humanities and social science.

20 GUIDELINES

to FAIRify data management and make data reusable



1 Invest in people and infrastructure

An important prerequisite to be able to implement the rest of the nineteen guidelines in this guide, is to invest in data infrastructures and in hiring and educating data experts.

Get acquainted with best practices in research data management. Check out the PARTHENOS training modules on data management or have a look at the CESSDA Data Management Expert Guide.

Invest in hiring and educating data experts and define a budget for making investments in technical infrastructure and staff.

FINDABLE

Research data should be easy to find by both humans and computer systems and based on mandatory descriptions of the metadata that allows the discovery of interesting datasets.

2 Use persistent identifiers

Locating data is a necessary condition for any other step from access to reuse. To be findable, any data object and dataset should be uniquely and persistently identifiable over time with a persistent identifier (PID). A PID continues to work even if the web address of a resource changes. PIDs can take different forms, such as a Handle, DOI, PURL, or URN.

Reference the PID which was assigned to your dataset in your research output.

Select the appropriate form of persistent identification schema and assign a PID to every resource. Use the PID Guide from NCDD to decide on the right PID for your research infrastructure.

3 Cite research data

If research data have a persistent identifier and are cited in accordance with community standards, the corresponding data objects or datasets are more easily found.

Get acquainted with data citation guidelines that are specific to your field or discipline and cite research data accordingly.

Provide information about best practices in data citation to research communities and make it easy for data users to cite data, e.g. by using a standardised button which says 'How to cite this dataset'.

4 Use persistent author identifiers

A persistent author identifier (e.g. VIAF, ISNI or ORCID) helps to create linkages between datasets, research activities, publications and researchers and allows recognition and discoverability.

Distinguish yourself from any other researcher or research group. Apply for an author identifier if you do not already have one and reference it in your dataset.

Reference author identifiers in the metadata.

5 Choose an appropriate metadata schema

Metadata is essential in making data findable, especially the metadata which is used for citing and describing data. A metadata schema is a list of standardised elements to capture information about a resource, e.g. a title, an identifier, a creator name, or a date. Using existing metadata schemas will ensure that international standards for data exchange are met.

To enable the discovery of content, describe research data as consistently and completely as possible. Include enough information for the data to be accessed and understood later on. If possible, use an existing metadata schema which fits the type of data object or dataset you are describing.

Clearly state which metadata schema you apply and recommend to the research community. To enrich datasets at data deposit, consider having a data submission form which collects additional metadata, e.g. about the provenance of the data.

ACCESSIBLE

Research data should be easily accessible and retrievable with well-defined access conditions using standardised communication protocols.

6 Choose a trustworthy repository

A certified repository offers a trustworthy home for datasets. Certification is a guarantee that data are stored safely, and will be available, findable and accessible over the long-term. Examples of certification standards are CoreTrustSeal, nestor seal and ISO 16363 certification.

Make your data accessible through a trustworthy repository. In addition, if you follow the repositories' standards (on preferred file formats, metadata schemas etc.) you can make sure that all requirements for making data FAIR are met.

Clearly state the level of certification on your website. If you are not (yet) certified, state how you plan to ensure availability, findability, accessibility and reusability in the long-term.

7 Clearly state accessibility

Access information specifies how a data user may access a dataset. When depositing data in a data repository, it should be clear which access options a data depositor can choose.

When choosing an access option, consider legal requirements, discipline-specific policies and ethics protocols when applicable. Choose Open Access when possible. When you collect personal data, ask yourself whether it contains any information which might lead to participants' identities being disclosed, what participants consented to and which measures you have taken to protect your data. If your data cannot be published in Open Access, the metadata should be, allowing data discovery.

Encourage (meta)data to be published in Open Access. Cleary state restricted access options for sensitive (meta)data that should not be part of the publicly accessible (meta)data. In this case, strive to make the (meta)data available through a controlled and documented access procedure.

8 Use a data embargo when needed

During a data embargo period, only the description of the dataset is published. The data themselves are not accessible. The full (meta)data will become available after a certain period of time.

Clearly state why and for what period a data embargo is needed. Make the (meta)data openly available as soon as possible.

Specify whether a data embargo is allowed and what conditions apply.

9 Use standardised exchange protocols

By using standardised exchange protocols, research infrastructures can make (meta)data publicly accessible and harvestable by e.g. search engines, vastly improving accessibility.

Use standardised protocols such as SWORD, OAI-PMH, ResourceSync and SPARQL. Convert metadata schemas into XML or RDF. Maintain a registry for protocol endpoints, the path at which research data can be accessed, and publish them.

To speed up discovery and uncover new insights, research data should be easily combined with other datasets by humans as well as computer systems.

INTEROPERABLE

10 Establish well documented machine-actionable APIs

Well documented and machine-actionable APIs - a set of subroutine definitions, protocols, and tools for building application software - allow for automatic indexing, retrieval and combining of (meta)data from different data repositories.

Document APIs well and make it possible to deliver the schema of the (meta)data model. Consider showing examples of how to successfully mine data from different endpoints and combine them into new data sets usable for new research.

11 Use open well-defined vocabularies

The description of metadata elements should follow community guidelines that use open, well defined and well known vocabularies. Such vocabularies describe the exact meaning of the concepts and qualities that the data represent.

Use vocabularies relevant to your field, and enrich and structure your research output accordingly from the start of your research project.

Give examples of vocabularies the research community may use, based on research domain specifics.

12 Document metadata models

Clearly documenting metadata models helps developers to compare and make mappings between metadata.

Publish the metadata models in use in your research infrastructure. Document technical specifications and define classes (groups of things that have common properties) and properties (elements that express the attributes of a metadata section as well as the relationships between different parts of the metadata). For metadata mapping purposes, list the mandatory and recommended properties.

13 Prescribe and use interoperable data standards

Using a data standard backed up by a strong community, increases the possibility to share, reuse and combine data collections.

Check with the repository where you want to deposit your data what data standards they use. Structure your data collection in this format from the start of your research project.

Clearly specify which data standard your institution uses, pool a community around them and maintain them especially with a perspective on interoperability. Good examples are CMDI (language studies) and the SIKB0102 Standard (archaeology).

14 Establish processes to enhance data quality

To boost (meta)data quality and, therefore, interoperability, establish (automatic) processes that clean up, derive and enrich (meta)data.

Establish procedures to minimise the risk of mistakes in collecting data. E.g. choose a date from a calendar instead of filling it in by hand.

Invest in tools to help clean up (meta)data and to convert data into standardised and interoperable data formats. Combine efforts to develop workflows and software solutions for such automatic processes, e.g. by using machine learning tools.

15 Prescribe and use future-proof file formats

All data files held in a data repository should be in an open, international, standardised file format to ensure long-term interoperability in terms of usability, accessibility and sustainability.

From the start of your research project think about future-proof file formats. Use preferred formats which are recommended by the data repository and are independent of specific software, developers or vendors.

Encourage the use of formats that are considered suitable for long-term preservation such as PDF-A, CSV and MID/MIF files. Provide an easy-to-find and detailed overview of accepted file formats.

Research data should be ready for future research and future processing, making it self-evident that findings can be replicated and new research effectively builds on already acquired, previous results.

REUSABLE

16 Document data systematically

To make clear what can and what cannot be expected in a dataset or repository, data should be systematically documented. Being transparent about what's in the data and what isn't facilitates trust and, consequently, data reuse.

Provide codebooks, including a description of methodology, a list of abbreviations, a description of gaps in the data, the setup of the database, etc.

17 Follow naming conventions

Following a precise and consistent naming convention - a generally agreed scheme to name data files - makes it significantly easier for future generations of researchers to retrieve, access and understand data objects and datasets.

Consult the policies and best practices for your research discipline or domain to find the most suitable naming convention.

Clearly state best practices to create and apply specific file naming conventions.

18 Use common file formats

By using standardised file formats that are widely used in your community, reusability is increased.

Use current popular file formats next to archival formats to share your data, e.g. Excel (xlsx) and CSV or ESRI Shapefiles next to MID/MIF files.

Publish the data in popular formats next to the archival format if they are not the same.

19 Maintain data integrity

Research data which were collected should be identical to the research data which are accessed later on. To ensure data authenticity, checks for data integrity should be performed.

Implement a method for version control. The guarantee that every change in a revised version of a dataset is correctly documented, is of integral importance for the authenticity of each dataset.

To identify if a file has been modified, it is essential to record provenance - the origin of the data plus any changes made over time - and to compare any copy with the original. A data integrity check can be performed by means of a fingerprint such as a checksum, or by a direct comparison of two files. Provide a mechanism to address different versions, for example by adding the version to the identifier as a search parameter.

20 License for reuse

To permit the widest reuse possible of (meta)data, it should be clear who the (meta)data rights holder is and what license applies.

Make sure you know who the (meta)data rights holder is before publishing your research data.

Communicate the (meta)data license and reuse options transparently and in a machine-readable format. To improve interoperability, try to map your licenses to frameworks which are already widely adopted such as Creative Commons.

PARTHENOS is a Horizon 2020 project funded by the European Commission. The views and opinions expressed in this publication are the sole responsibility of the author and do not necessarily reflect the views of the European Commission.