FAIR Data Principles

FAIR Principles

- How do we promote reuse of data
- How do we promote reproducibility
- FAIR Principles ensure that data are shared in a way that enables and enhances reuse by humans and machines

The FAIR Data Principles

- to be **F**indable
  - The first step in (re)using data is to find them. Metadata and data should be easy to find for both humans and computers. Machine-readable metadata are essential for automatic discovery of datasets and services,
- to be **A**ccessible
  - Once the user finds the required data, she/he needs to know how can they be accessed, possibly including authentication and authorization.
- to be **I**nteroperable
  - The data usually need to be integrated with other data. In addition, the data need to interoperate with applications or workflows for analysis, storage, and processing.
- to be **R**eusable
  - The ultimate goal of FAIR is to optimise the reuse of data. To achieve this, metadata and data should be well-described so that they can be replicated and/or combined in different settings.

Parties Involved

- Data contributor
    - The entities or organizations who generate the data
- The Data Repository
    - The entity or organization who has responsibility for data “stewardship”

To be Findable

- Data and metadata should be easy to locate, both by humans and by computer systems. Basic machine-readable descriptive metadata enable the discovery of interesting datasets and services.

F1. (Meta)Data are assigned a globally unique and eternally persistent identifier

The context	FAIR DATA – The role of scientists	FAIR Repository – The role of the repository
Each dataset is assigned a globally unique and persistent identifier (PID), e.g. a Digital Object Identifier (DOI). These identifiers make it possible to locate and cite the dataset and its metadata. Principle F1 can be considered as the most important, because without persistent identifiers it is difficult to fulfill the other properties of the FAIR principles.	Scientists should ensure and be aware of the fact that each (published) dataset is assigned a globally unique and persistent identifier. Certain repositories automatically assign a PID (e.g. a DOI) to the dataset published there. If this is not the case, a different repository should be considered, or the respective repository operator should be consulted.	A repository must be able to assign a globally unique PID to a dataset when it is published. Furthermore it can be useful to notify the researcher of the PID before the actual data publication, so that it can be included in a corresponding manuscript as a reference to the research data, for example.

F2. Data are described with rich metadata (see R1.)

The context	FAIR DATA – The role of scientists	FAIR Repository – The role of the repository
Each dataset should be described with detailed (see R1.) metadata: The metadata show how the dataset was generated, who collected / edited / published the data and under which conditions (license) a dataset may be re- used. Metadata thus provide the necessary context information for the correct interpretation of research data. This information must also be machine-readable.	Researchers should describe each dataset carefully and as completely as possible with metadata. The metadata should contain descriptive information about the context, quality and condition or characteristics of the data. Other researchers (including researchers from different disciplines) should be able to understand the context of the data from the metadata.	The repository provides a metadata schema that enables researchers to specify relevant metadata (general metadata and / or subject-specific metadata).

F3. (Meta)Data are registered or indexed in a searchable resource

The context	FAIR DATA – The role of scientists	FAIR Repository – The role of the repository
Metadata are used to create easily searchable indexes of records. These indexes allow researchers to search for existing datasets, which is similar to the way a search for an article or within an information platform is performed.	Scientists should ensure that they provide precise and complete metadata whenever possible (see F2. and R1.).	The repository supports a structured input of metadata, e.g. by providing specific submission forms or XML Schema that facilitates the storage of PID, author names, subject areas, etc. The repository should facilitate the creation of indexes.

F4. Metadata specify the data identifier

The context	FAIR DATA – The role of scientists	FAIR Repository – The role of the repository
The metadata and the dataset they describe are often separate files. The association between the metadata and the dataset is established by identifying the PID in the metadata.	Scientists should ensure that in the course of publication or archiving a PID is assigned to each dataset.	The repository allows researchers to upload metadata for each dataset and assigns a corresponding PID.

To be Accessible

- Data and metadata should be archived long-term and made available in such a way that they can be easily retrieved by machines and humans, or be used locally with the help of standard communication protocols.

A1. (Meta)Data are retrievable by their identifier using a standardized communications protocol

The context	FAIR DATA – The role of scientists	FAIR Repository – The role of the repository
Data can be retrieved from the internet by a click via a high-level interface to a low-level protocol called TCP. As a result, the data is displayed in the user's web browser (via FTP or HTTP(S)). Principle A1. declares that the data retrieval can be carried out without specialized tools. To achieve this, it must be clearly defined who can access the actual data and which prerequisites must be fulfilled.	Most scientists will use HTTP (S) or FTP as the communication protocol. For sensitive/protected data for example, a mechanized protocol may not guarantee secure access to the data. In such cases the FAIR requirements are fulfilled if an e-mail or other contact information of a person/data manager is given, with whom access to the data can be discussed. These contact details have to be explicitly stated in the metadata.	(Meta) Data archived in the repository can be accessed via a standardized protocol. Access barriers should be avoided. Furthermore, if sensitive data (e.g. with an embargo) is available in the repository, clear contact and responsibility information should be displayed in the metadata.

A1.1 The protocol is open, free, and universally implementable

The context	FAIR DATA – The role of scientists	FAIR Repository – The role of the repository
In order to maximize data re-use and facilitate data retrieval, the protocol should be free (of charge) and open source and therefore globally implementable. Every user with a computer and internet connection should be able to access the metadata.	As a researcher one should inquire whether the protocol used in a research data repository corresponds to the FAIR principles (free, open and implementable).	The repository uses an open (no proprietary or commercial) communication protocol. Examples: HTTP(S) FTP SMTP

A1.2 The protocol allows for an authentication and authorization procedure, where necessary

The context	FAIR DATA – The role of scientists	FAIR Repository – The role of the repository
The “A“ in FAIR specifies the exact conditions under which the data is accessible, which does not necessarily means that data has to be “open“ and / or “free“. It should be noted that even highly protected data can be FAIR data. The access conditions to data should be transparent. These conditions should be (automatically) understood and executed by computer systems, so that users can be made aware of them.	The creation of a user account enables the researcher to authenticate himself / herself as the author / creator of a dataset and to grant further user-specific rights. In consequence, this criterion may also influence the choice of repository to which researchers submit their data.	The repository should offer a role and rights management that supports the authentication and authorization of users, including machine-operated accesses. Examples: OAuth HTTPS FTPS

A2. Metadata are accessible, even when the data are no longer available

The context	FAIR DATA – The role of scientists	FAIR Repository – The role of the repository
The provision of datasets for an indefinite period of time requires a large amount of curation effort by the repository operators. In addition, the readability of the datasets may be limited or even non-existent after a longer period of time, as file formats and software programs are continuously changing and evolving. Furthermore it is possible that data publications are withdrawn. Facing these facts, it is necessary to provide metadata which are available long-term, and which describe the dataset in a human- and machine-readable way.	Metadata provide valuable information in the planning of research, especially for replication studies. Researchers should be aware that metadata can be used to trace the data and to recognize authors, institutions or publications in connection with the original research, even if the original data is no longer available.	With the fulfilment of the criteria listed under A1., a sustainable availability of all metadata is often secured. It is important that repository operators indicate this, e.g. within a documentation. In addition, the repository should have an exit strategy that ensures that (meta)data are preserved and accessible even when the repository ends its services.

To be Interoperable

- Data should be available in such a format that it can be exchanged, interpreted and combined in a (semi-)automated manner with other data, to be carried out by man and machine operations.

I1. (Meta)Data use a formal, accessible, shared, and broadly applicable language for knowledge representation

The context	FAIR DATA – The role of scientists	FAIR Repository – The role of the repository
People exchange information through the use of common languages. This also applies to computers. Therefore, data should also be available in a representation that is comprehensible for machines. When (meta)data is searched, computer systems must be able to decide whether the contents of the searched records are relevant. Controlled vocabularies / ontologies / thesauri and a clearly defined framework, e.g. in the sense of the Semantic Web, are required for the creation and application of such metadata.	Researchers should provide as precise and complete metadata as possible for a dataset and the files contained within it (see R1.).	The repositories provide machine-readable data and metadata with a well- established formalism where possible. In particular, data and metadata should be structured, using vocabularies / ontologies / thesauri which are commonly used in the disciplines addressed. Examples: RDF, OWL DAML+OIL, JSON LD

I2. (Meta)Data use vocabularies that follow FAIR principles

The context	FAIR DATA – The role of scientists	FAIR Repository – The role of the repository
Controlled vocabularies / ontologies /thesauri, which are used to describe datasets, must be properly identified and documented. They should also be accessible, interoperable and carefully described (FAIR). In addition, domain- based mappings may help to develop a common understanding of the data and improve data interoperability and findability.	In a FAIR data world, researchers are able to refer to specific metrics which help them to evaluate the FAIRness of a controlled vocabulary / ontology / thesaurus in their field of expertise. However, these metrics often do not (yet) exist or are currently developed, which makes it difficult for researchers to address and meet this requirement.	I2 can be difficult for repository operators to meet, because the actuality of the recommended vocabularies / ontologies / thesauri has to be guaranteed long-term. In addition, the quality of these can vary greatly depending on the discipline or the evaluation criteria applied.

I3. (Meta)Data include qualified references to other (meta)data

The context	FAIR DATA – The role of scientists	FAIR Repository – The role of the repository
If a dataset is based on another dataset or contains complementary information of another dataset, this must be specified in the (meta)data. In particular, the scientific relation between the datasets should be described.	Scientists should clearly identify relationships between datasets in the metadata, e.g. by naming their persistent identifiers and describing their scientific link to each other (e.g.' is new version of' is supplement to,' relates to', etc.).	The metadata schema provided by repositories should support the referencing between datasets with corresponding metadata fields (e. g. relatedIdentifer, relationType).

To be Reusable

- A good description of data and metadata ensures that the data can be re-used for future research and are comparable to other compatible sources. It must be possible to cite the data properly, and the conditions under which the data can be re-used should be presented in a way that is easy for man and machines to understand.

R1. Meta(Data) have a plurality of accurate and relevant attributes

The context	FAIR DATA – The role of scientists	FAIR Repository – The role of the repository
The provision of data with extensive metadata makes it easier to find and re-use them. However, the goal of R1. differs from F2. R1. focuses on the ability of a user (machine or man) to decide whether the located data (F2.) does actually fit into the requested context. In order to make this decision, the data creator has to provide metadata that covers an extensive description of the data generation process. This may include e.g. experimental protocols, the manufacturer of the machine or sensor used for data creation, the software used for analyses, etc.	The (meta)data creators should be as detailed as possible when adding (meta)data. This can lead to the provision of (context) information that may first appear to be irrelevant. Examples (non-exhaustive list): • Scope: For what purpose was data created / collected? • Date of dataset generation, (laboratory) conditions, parameter settings, name and version of the software used. • Does the dataset contain raw data or processed data or both? • Variable names / parameters are explained or self-explanatory (i. e. defined in the vocabulary of the research field). • The version of the archived and / or re-used data is clearly specified and documented.	The repository provides a metadata schema that enables researchers to specify relevant metadata (general and / or subject-specific). The metadata are provided in a human- and machine-readable format.

R1.1. (Meta)Data are released with a clear and accessible data usage license

The context	FAIR DATA – The role of scientists	FAIR Repository – The role of the repository
The legal conditions under which data may be used should be clearly defined for machines and man. This must be represented in the metadata.	Scientists should enter information on the legal conditions for the subsequent use of a dataset (stating a user license) in the metadata. It is suggested to use open licensing models such as Creative Commons (e.g. CC BY), which can be referenced by specifying a corresponding URL.	Repositories should enable scientists to upload license files or refer to them. Ideally, the license information is stored in an equally machine-readable format.

R1.2. (Meta)Data are associated with their provenance

The context	FAIR DATA – The role of scientists	FAIR Repository – The role of the repository
Provenance information specify how a dataset was generated and by whom, in which context it can be re-used and how reliable it is. Provenance information establish an important criterion for the validation of data in scientific databases.	For the collection of provenance information, researchers should clearly describe their role in the data generation workflow, how they wish to be cited, and who was involved. They should also state whether the dataset contains foreign data. Again, the description should be in a human- and machine- readable format. It should be noted that principle I3. is closely related to this topic, in particular with regard to the re- use of previously published datasets.	The repository provides a metadata schema that enables researchers to specify relevant metadata (which are general and / or subject-specific). Metadata should be provided in a human- and machine-readable format.

R1.3. (Meta)Data meet domain-relevant community standards

The context	FAIR DATA – The role of scientists	FAIR Repository – The role of the repository
As far as standards or best practices for data archiving and publication within a research community are available, they should be used. Datasets which are structured in a similar or comparable manner have several benefits: They usually contain the same type of data, are structured in a standardized way, are available in established file formats, and possess well documented, structured metadata which are based on a common vocabulary. However, it should be noted that several standards may compete within a community or discipline.	Scientists should prepare their (meta)data according to their community standards and best practices for data archiving and publication. Depending on the area of expertise, several standards / best practices may have been established in the research discipline. If no of the existing standards / best practices are used, the reasons should be clearly stated in the metadata.	Repositories are free to implement standards with respect to uploaded metadata or data, especially if they provided discipline-specific services. These standards have to be regularly checked for being up-to-date by the repository operators.