Explanation of the [FAIR data principles](#)

Wilkinson et al. (2016), The FAIR Guiding Principles for scientific data management and stewardship, *Scientific Data* 3, [doi:10.1038/sdata.2016.18](#)

Principle		In other words	Researcher's responsibility	Requirements to be fulfilled by the repository
To be findable: Data and metadata should be easy to find by both, humans and computer systems. Basic machine readable descriptive metadata allows the discovery of interesting data sets and services.	F1. (meta)data are assigned a globally unique and persistent identifier	Each data set is assigned a globally unique and persistent identifier (PID), for example a DOI, ARK, RRID… These identifiers allow to find, cite and track (meta)data.	Ensure that each data set is assigned a globally unique and persistent identifier. Certain repositories automatically assign identifiers to data sets as a service. If not, researchers must obtain a PID via a PID registration service.	A repository needs to have a predictable way to assign a PID to each component of a dataset (e.g. each file or nanopublication), in order to be able to include these identifiers into the corresponding metadata before the submission.
	F2. data are described with rich metadata (defined by R1 below)	Each data set is thoroughly (see below, in R1) described: these metadata document how the data was generated, under what term (license) and how it can be (re)used, and provide the necessary context for proper interpretation. This information needs to be machine-readable.	Fully document each data set in the metadata, which may include descriptive information about the context, quality and condition, or characteristics of the data. Another researcher in any field, or their computer, should be able to properly understand the nature of your dataset. Be as generous as possible with your metadata (see R1).	Allow researchers to upload metadata for each data set.
	F3. metadata clearly and explicitly include the identifier of the data it describes	The metadata and the data set they describe are separate files. The association between a metadata file and the data set is obvious thanks to the mention of the data set's PID in the metadata.	Make sure that the metadata contains the data set's PID.	Allow researchers to upload metadata for each data set.
	F4. (meta)data are registered or indexed in a searchable resource	Metadata are used to build easily searchable indexes of data sets. These resources will allow to search for existing data sets similarly to searching for a book in a library.	Provide detailed and complete metadata for each data set (see F2).	Request and store part of the metadata in a structured way, for example by providing a form with specific fields to be completed or by providing an XML schema to be used by the researchers. For example the storing of PID's, author names, disciplines, etc. will facilitate the creation of indexes. However, it must remain possible to provide arbitrary metadata in addition.

Principle		In other words	Researcher's responsibility	Requirements to be fulfilled by the repository
To be accessible: Data and metadata should be stored for the long term such that they can be easily accessed and downloaded or locally used by machines and humans using standard communication protocols.	A1. (meta)data are retrievable by their identifier using a standardized communications protocol.	If one knows a data set's identifier and the location where it is archived, one can access at least the metadata. Furthermore, the user knows how to proceed to get access to the data.	Clearly define who can access the actual data, and specify how. It is possible that data will actually not be downloaded, but rather reused *in situ*. If so, the metadata must specify the conditions under which this is allowed (sometimes versus the conditions needed to fulfill for external usage/"download").	(Meta)data archived on the repository is accessible using a standardized protocol.
	A1.1 the protocol is open, free, and universally implementable	Anyone with a computer and an internet connection can access at least the metadata.	--	The repository does not rely on a proprietary or commercial communication protocol.
	A1.2 the protocol allows for an authentication and authorization procedure, where necessary	It often makes sense to request users to create a user account on a repository. This allows to authenticate the owner (or contributor) of each data set, and to potentially set user specific rights.	--	Provide a way for authentication and authorization of users, including machine-users.
	A2. metadata are accessible, even when the data are no longer available	Maintaining all data sets in a readily usable state eternally would require an enormous amount of curation work (adapting to new standards for formats, converting to different format if specifically needed software is discontinued, etc.). Keeping the metadata describing each data set accessible, however, can be done with much less resources. This allows to build comprehensive data indexes including all current, past and potentially arising data sets.	Provide detailed and complete metadata for each data set (see below in R1).	Archive metadata "for ever" and ensure it always fulfills criterion A1. To ensure the long-term preservation of metadata beyond the lifetime of a repository, consider possibilities to easily extract and move metadata to another repository. In particular, ensure that metadata and data are physically separate files. Furthermore, repositories should have a 12 month contingency plan.

Principle		In other words	Researcher's responsibility	Requirements to be fulfilled by the repository
To be interoperable: Data should be ready to be exchanged, interpreted and combined in a (semi)automated way with other data sets by humans as well as computer systems.	I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.	Interoperability typically means that each computer system has at least knowledge of the other system's formats in which data is exchanged. If (meta)data are to be searchable and if compatible data sources should be combinable in a (semi)automatic way, computer systems need to be able to decide if the content of data sets are comparable. Obvious issues arise when different languages are used to describe the data or when spelling errors make the comparison of descriptions and variable names more difficult. It is critical to use controlled vocabularies and a well-defined framework to describe and structure (meta)data in order to ensure findability and interoperability of datasets.	Provide machine readable data and metadata in an accessible language, using a well-established formalism. In particular, data and metadata are annotated with resolvable vocabularies/ontologies/thesauri that are commonly used in the field. The RDF extensible knowledge representation model is a way to describe and structure datasets. You can refer to the Dublin Core Schema as an example.	Support the upload of machine readable data and metadata provided in an accessible language, using a well-established formalism. In particular, ensure that computer systems will be able to distinguish the metadata from the data file.
	I2. (meta)data use vocabularies that follow FAIR principles	The controlled vocabulary used to describe data sets needs to be documented. This documentation needs to be easily findable and accessible by anyone who uses the data set.	The vocabularies/ontologies/thesauri are themselves findable, accessible, interoperable and thoroughly documented, hence FAIR. Researchers can refer to metrics assessing the FAIRness of a digital resource (if available).	Ideally, provide a FAIRness score for each digital resource.
	I3. (meta)data include qualified references to other (meta)data	If the data set builds on another data set, if additional data sets are needed to complete the data, or if complementary information is stored in a different data set, this needs to be specified. In particular, the scientific link between the data sets needs to be described. Furthermore, all data sets need to be properly cited (i.e. including their persistent identifiers).	Properly cite relevant/associated data sets, in particular by providing their persistent identifiers, in the metadata, and describe the scientific link/relation to your data set.	Ideally provide a structured way, for example by providing a form with specific fields to be completed, to declare references to other (meta)data. Requesting specific formats for some entries (e.g. URL, scientific link) will enhance interoperability.

Principle		In other words	Researcher's responsibility	Requirements to be fulfilled by the repository
To be reusable: Data and metadata are sufficiently well-described to allow data to be reused in future research, allowing for integration with other compatible data sources. Proper citation must be facilitated, and the conditions under which the data can be used should be clear to machines and humans.	R1. meta(data) are richly described with a plurality of accurate and relevant attributes	Description of a data set is required at two different levels: (1) metadata describing the data set (intrinsic): what does the data set contain, how was the data generated, how has it been processed, how can it be reused … (2) metadata describing the data (submitter-defined): any needed information to properly use the data, such as definitions of the variable names	Provide complete metadata for each data file. Some points to take into consideration (non-exhaustive list): - Scope of your data: for what purpose was it generated/collected? - Particularities or limitations about the data that other users should be aware of. - Date of the data set generation, lab conditions, who prepared the data, parameter settings, name and version of the software used. - Is it raw or processed data? - Variable names are explained or self-explanatory (i.e. defined in the research field's controlled vocabulary). - Version of the archived and/or reused data is clearly specified and documented.	Allow researchers to upload metadata for each data set.
	R1.1. (meta)data are released with a clear and accessible data usage license	The conditions under which the data can be used should be clear to machines and humans. This has to be specified in the metadata describing a data set.	Include information about the license in the metadata. If a particular license is needed, you have to provide it along with the data set. Where possible it is suggested to use common licenses, such as CC 0, CC BY, etc., which can be referred to by URL.	Allow license files to be uploaded or referred to. Ideally foresee a structured way, for example by providing a form with specific fields to be completed, to declare the license. Ensure that computer systems will be able to distinguish the metadata from the data file.
	R1.2. (meta)data are associated with detailed provenance	Detailed information about the provenance of data is necessary for reuse: this will, for example, allow researchers to understand how the data was generated, in which context it can be reused, and how reliable it is. Provenance is a central issue in scientific databases to validate data.	The metadata to thoroughly describe the workflow that led to your data: Who generated or collected it? How has it been processed? Has it been published before? Does it contain data from someone else, potentially transformed or completed? Ideally the workflow is described in a machine-readable format. Criterion I3 is closely linked to this issue when reusing published data sets.	Allow the separation between intrinsic, submitter- and user-defined metadata. In particular, allow annotation of data by others than the original submitter (e.g. to comment specific entries of a data set).
	R1.3. (meta)data meet domain-relevant community standards	It is easier to reuse data sets if they are similar: same type of data, data organized in a standardized way, well-established and sustainable file formats, documentation (metadata) following a common template and using common vocabulary. If community standards or best practices for data archiving and sharing exist, they should be followed. Note that quality issues are not addressed by the FAIR principles. How reliable data is lies in the eye of the beholder and depends on the foreseen application.	Prepare your (meta)data according to community standards and best practices for data archiving and sharing in your research field. There might be situations where good practice exist for the type of data to be submitted but the submitter has valid and specified reasons to divert from the standard practice. This needs to be addressed in the metadata.	Repositories, in particular when they are specialized on a specific research field, may implement minimal standards regarding the uploaded metadata or data. Different certifications exist for repositories, see for example the Data Seal of Approval standards.