FAIR Data Principles

Idea In Short

FAIR Data Principles (Findable, Accessible, Interoperable, Re-usable) support knowledge discovery and innovation as well as data and knowledge integration, and promote sharing and reuse of data. The principles help data and metadata to be ‘machine readable', supporting new discoveries through the harvest and analysis of multiple datasets.

First published in 2016, the FAIR principles provide guidelines for good data management practices. Each letter refers to four foundational principles and 15 guiding principles that prescribe how FAIRness of data can be achieved through technical implementation. These principles aim to making data FAIR:

1. Findable
2. Accessible
3. Interoperable, and
4. Reusable

The term Data refers all kinds of digital artifacts, such as data, code, software, documents, infrastructure, etc.

FAIR Principles

Although the FAIR principles originate from the life sciences, they can be applied across both, academic and business. Since their publication, the European Union, research institutions and corporations have extended their support for the FAIR principles. This spans from defining policies for data management to technical implementation, tools and infrastructures. Some FAIR implementations strictly adhere to the original definitions, while others are derived from the foundational FAIR principles.

Findable

Metadata are structured information about the collected data material. This information describes the material on various levels, for example where and by whom it was created; on which occasions and with which methods the data were collected; what a variable means and which values it can take. Metadata are not the same as documentation; what signifies metadata is that they are structured in a way that makes them readable by both humans and computers. Findable means that the data is discoverable by both, humans and machines. Data can be exposed for meaningful interactions through programmatic access through APIs and technical interfaces as well as human-friendly interfaces. Data catalogue - a collection of metadata, combined with data management and search tools - can greatly help improve data discovery and findability. When data is enriched with metadata to describe the data and identified using unique and persistent identifiers [PIDs], such as DOIs or Handles using a common ontology, both humans and computers could easily find and consume data. Machine-readable metadata are essential for automatic discovery of datasets and services, so this is an essential component of the FAIR process. The guiding principles are:

- F1. [Meta]data are assigned a globally unique and eternally persistent identifier
- F2. Data are described with rich metadata (defined by R1 below)
- F3. Metadata clearly and explicitly include the identifier of the data they describe
- F4. [Meta]data are registered or indexed in a searchable resource

Accessible

Once the user finds the required data, they need to know how they can be accessed, possibly including authentication and authorization. Accessible means that the data are persisted in appropriate storage, so it can be leveraged using standard technical and operational procedures. This does not mean that the data have to be openly available for everyone. However, information on the proper data access mechanism should be to be available. For example, sensitive data should be properly marked with the pertinent sensitivity (public, confidential, highly confidential, restricted, etc.) class and security level (e.g. 1-5 with 1 being mission critical and 5 being public access). The access procedure should also be properly documented, for example __Access only with explicit approval from the data owner / data steward_ and include the contact details. To share or make a data material accessible is not the same thing as sharing the data freely so that they can be accessed and used by everyone. If the material contains sensitive personal data, or special category data, for example, a confidentiality assessment needs to be made before the material can be released to anyone. Metadata, however, are not sensitive, so even if the

data cannot be made freely accessible, you can use metadata to show that the material exists and under which conditions you may access and reuse it. Ideally, the information about data accessibility should also be read by machines through machine-readable standard licenses. The guiding principles are:

- A1. [Meta]data are retrievable by their identifier using a standardized communications protocol
- A1.1 The protocol is open, free, and universally implementable
- A1.2 The protocol allows for an authentication and authorization procedure, where necessary
- A2. Metadata are accessible, even when the data are no longer available

Interoperable

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. Interoperable means that the data can be exchanged and used across different applications and systems — also in the future, for example, by using open file formats. It also means that the data can be integrated with other data and systems, both internal and external to an organization. This can be achieved through using metadata standards, standard ontologies, nomenclature, and governed vocabularies as well as meaningful links between the data [semantics] and related digital business processes.

The main responsibility for this rests on the organization that makes the data accessible. But, it also means that data producers and consumers should use standardized ways to enter information such as dates, time periods, and geographic coordinates, etc. Data stakeholders should use a widely adopted vocabulary to describe categories and code variables according to an accepted standard. If possible, you should save the data in a widely used file format that is supported by common operating systems and can be opened in several programs, or use software that can export data in such file formats when the project is finished. The guiding principles are:

- I1. [Meta]data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- I2. [Meta]data use vocabularies that follow FAIR principles
- I3. [Meta]data include qualified references to other [meta]data

Reusable

The ultimate goal of FAIR is to optimize 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. Reusable means that the data are well documented and curated and provide rich information about the context of data creation. The data should conform to community standards and include clear terms and conditions on how the data may be accessed and reused, preferably by applying machine-readable standard licenses. This allows others either to assess and validate the fitness of purpose [Utility] and use [Warranty] of the collected data.

Such information also ensuring data reproducibility or to design new projects based on the original results. In other words, data reuse encourages collaboration and avoid data duplication (data silos). Additional conditions for reusability are that the data are described with sufficient and relevant metadata, that both humans and computers can read the metadata, and that there is clear information about, for instance, the purpose of the collected data, the context for the data collection as well as which equipment and software were used for the data collection and analysis. You should also clearly specify the conditions for how the data may be used. The guiding principles are:

- R1. [Meta]data are richly described with a plurality of accurate and relevant attributes
- R1.1. [Meta]data are released with a clear and accessible data usage license
- R1.2. [Meta]data are associated with detailed provenance
- R1.3. [Meta]data meet domain-relevant community standards

The FAIR principles apply to both data and metadata. Implementing FAIR principles will lead to the creation of a data ecosystem - the Internet of FAIR Data and Services - that is responsive to changing market conditions and automatically adapts to new emergent scenarios around data standards, exchange interfaces, protocols, identification mechanisms, availability, etc.

The FAIR principles are not rules or standards. The FAIR principles must not be mistaken for rules or standards that you can use to evaluate tools, data, policies, etc. This would soon make the principles out-of-date and inapplicable across multiple disciplines. Adopting the FAIR principles will often be a gradual adaptation of work routines – but it could also be a huge leap, where you replace one type of infrastructure with another. It will be up to the

different research areas and research communities to make the FAIR principles work in their respective contexts.

FAIR Practices

Applying the FAIR principles depends on specific disciplines and industry structures. However, there are different activities that organizations should consider when developing business processes and workflows to make enterprise data FAIR-compliant. For instance, documenting data using an enterprise-wide data glossary in a data catalog, choosing appropriate file formats, adding metadata, provisioning and governing data access to data consumers [users], licensing data (data monetization) or adding a persistent identifier (data classification) will help enhance FAIR compliance.

Documentation adds rich context to your data and makes the data easier to understand and reuse in the future. File formats determine how data can be used. It is important to decide what file formats to use for data collection, data processing, data archiving, and long-term archival. File formats are important to consider, when you want to combine datasets or make data readable by machines. Metadata are data about data. Research data need metadata to become findable, accessible, interoperable and reusable - by humans and machines. Access to data means that you determine who you make your data available for, how you provide access, and under which conditions. To make your data accessible and easy to find, you must provide your data and metadata with a persistent identifier [PID]. A PID is a long-lasting reference to a digital resource and provides the information required to reliably identify, verify and locate your research data. A data license is a legal arrangement between the data creator and the data user that specifies what users can do with the data, such as Acceptable Usage Policy [AUP]. It is one of the most effective ways to communicate permissions to potential data users.

Community

Although started by a community operating in the life science, the FAIR principles have rapidly been adopted by publishers, funders, and pan-disciplinary infrastructure programmes and societies. Many groups and organisation are working to define guidance and tools to help researchers and other stakeholders (like librarians, funders, publishers, and trainers) make data more FAIR. If you are interested in participating in these communities there are two global initiatives that act as umbrella organizations and reference

points for many discipline-specific efforts: GOFAIR and the Research Data Alliance [RDA].

- Under GOFAIR, there are many Implementation Networks [INs] committed to implementing the FAIR principles
- Under the RDA, there are several groups tackling different aspects relevant to the RDM life cycle. Among these, FAIR Data Maturity Model Working Group is reviewing existing efforts, building on them to define a standard set of common assessment criteria for the evaluation of FAIRness

This working group brings together stakeholders from different scientific and research disciplines, the industry and public sector, who are active and/or interested in the FAIR data principles and in particular in assessment criteria and methodologies for evaluating their real-life uptake and implementation level.

Summary

The FAIR principles are a set of community-developed guidelines to ensure that data or any digital object are Findable, Accessible, Interoperable and Reusable. The FAIR principles specifically emphasize enhancing the ability of machines to automatically find and use data or any digital object, and support its reuse by individuals. Standards for the description, interoperability, citation etc. are at the core of these principles. Following FAIR principles enables machines and humans to find, access, interoperate, and reuse data. That additional value stems from machine-actionability, which lacking in data management initiatives. The FAIR Principles are intended as guidelines to enhance the enterprise value of data assets.