FAIR data

Jean-Baka Domelevo Entfellner (ILRI)
BHKi hybrid seminar series
19th July 2022

Genesis of the FAIR paradigm for scientific data

Four fundamental principles first described in a **2016 paper** (MD Wilkinson et al, *The FAIR Guiding Principles for scientific data management and stewardship*, Scientific Data 3:160018. DOI: 10.1038/sdata.2016.18)

Goals:

- Enhance paths to discovery through better data management
- Provide steps towards machine-actionability of data
- Enforce openness and enable reproducibility of science

The Four Guiding Principles

- **Findability**: the ability for a dataset to be found through keyword-based searches, browsing repositories, querying metadata, etc
- **Accessibility**: metadata and if possible the data itself, are accessible by humans and machines through an open, free format (API)
- **Interoperability**: controlled vocabularies enable the dataset to be integrated into data workflows, and integrated in part or in full into other pieces of research
- **Reusability**: rich metadata, detailed provenance of the dataset parts, domain-relevant standards are enforced so that third parties can reuse the data

| The FAIR Guiding Principles

To be Findable:
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:
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

To be Interoperable:
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

To be Reusable:
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

Enablers for FAIR: resource locators

Findability and **Accessibility** through **stable resource locators**: handles (http://handle.net/) and Digital Object Identifiers (https://www.doi.org/)

- Provide persistent identifiers for digital objects e.g. datasets
- Just any digital object can receive a handle
- Resolution services freely available
- DOIs are based on the handle system

Enablers for FAIR: data formats and ontologies

Interoperability through standardized, machine-operable **(meta)data formats** (XML, JSON) and controlled vocabularies (aka ontologies)

- JSON format the new de facto standard for data sharing
- JSON both human- and machine-readable
- FAO's AGROVOC (https://www.fao.org/agrovoc/) vocabulary for agricultural research
- Gene Ontology (http://geneontology.org/) for gene function, localization, etc, is a standard for gene annotation sharing

FAIRER better than FAIR

Adding two requirements, FAIRER is FAIR plus:

- **Ethical** data: informed consent to be enforced and forms to be included as metadata
- **Reproducible** research: datasets to be linked with workflows (in a standardized language e.g. CWL), so that the research based on the data can be re-run and the results therefore verifiable