ICARDA

Science for resilient livelihoods in dry areas

F.A.I.R. Data Principles and Practices

Valerio Graziano

Big Data in Agriculture

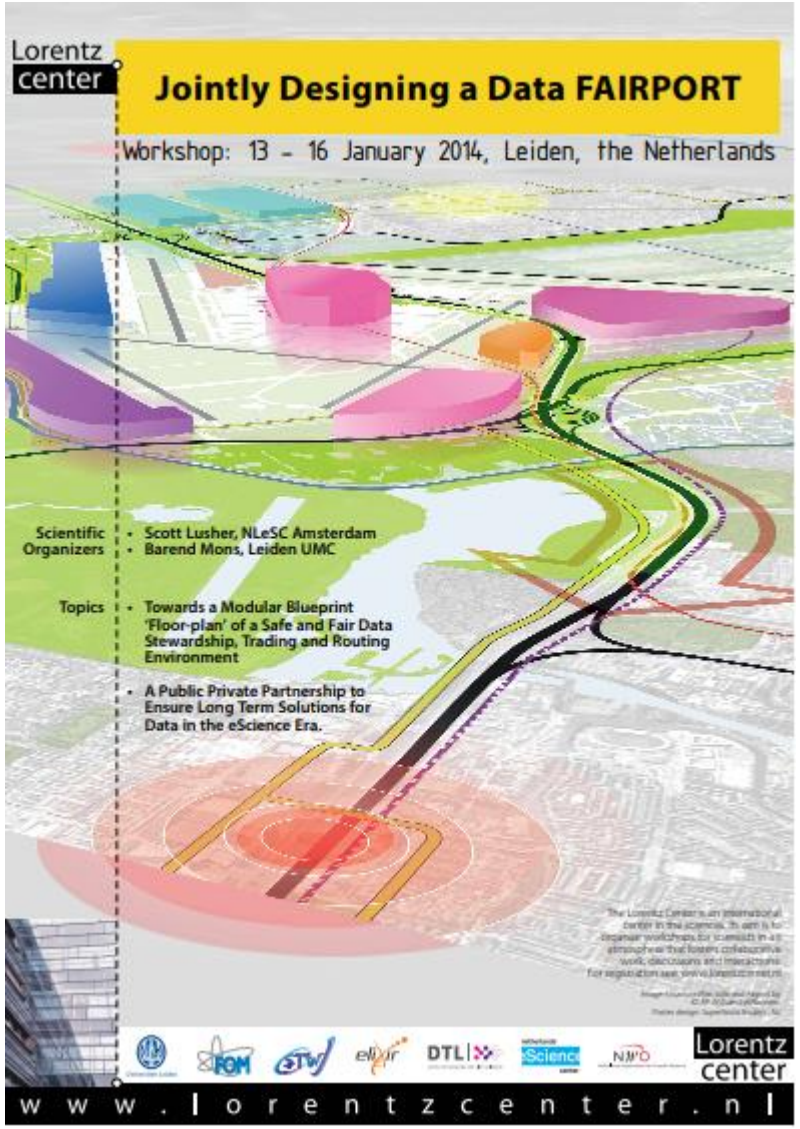
10-14 December, Rabat, Morocco

FAIR Origin

Coined in 2014 at “Jointly Designing a Data Fairport” Lorentz Workshop in Leiden (Netherlands, 2014). Further developed by FORCE11 members and other scholars. **The principles were first published in 2016** and a related design framework for metrics in 2018.

- The acronym spelled out:

Findable, **A**ccessible, **I**nteroperable, **R**eusable.



FAIR Guidance

*"These high-level FAIR Guiding **Principles precede implementation choices**, and do not suggest any specific technology, standard, or implementation-solution; moreover, the Principles are not, themselves, a standard or a specification. **They act as a guide to data publishers**..."*

*"Good data management … is the key conduit leading to knowledge discovery and innovation, and … **reuse by the community** after the data publication process."*

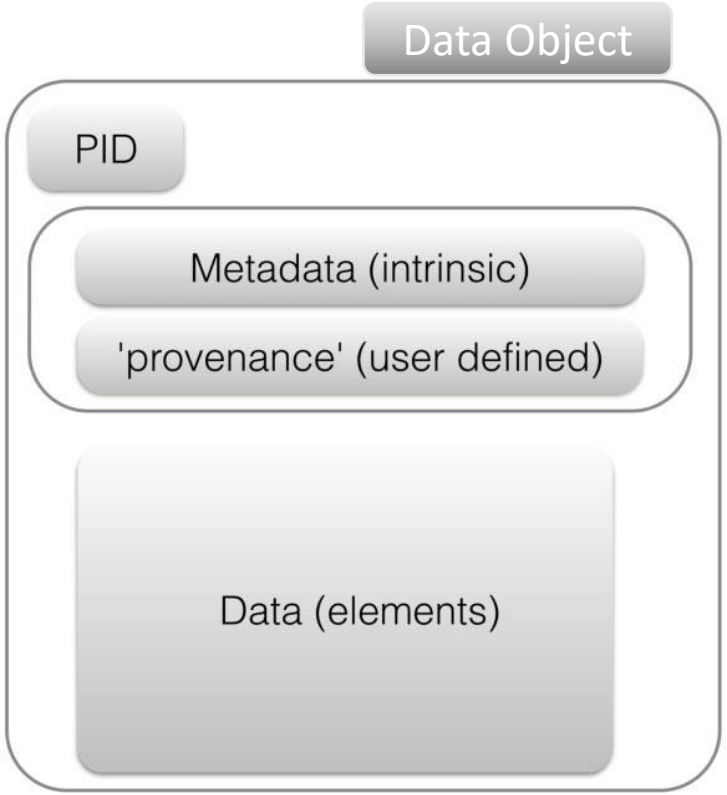
*"… the FAIR Principles put specific emphasis on **enhancing the ability of machines** to automatically find and use the data, in addition to **supporting its reuse by individuals**."*

Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 3:160018 doi: 10.1038/sdata.2016.18 (2016).

FAIR Basics

The principles refer to **Data Objects**, a machine (first) and human (second) intelligible resource of information constituted by:

- **Data**: in the form of **digital object** (i.e. file).
- **Metadata**: **information** about that digital object.
- **Persistent Identifier** (PID).



It is defined as **FAIRport** any "machine-oriented data repository" that:

- Contains FAIR Data Objects.
- Provides accessibility for Data Objects re-use.
- Has a full and open description of all technologies, controlled vocabularies and formats used.

https://www.force11.org/fairprinciples

Questions

F for Findable

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

Examples: dx.doi.org, hdl.handle.net
Reason: PID ensures the Findability of the data object

F2. Data are described with rich metadata

Examples: Dublin Core, CG Core
Reason: your data object can be found through its metadata (i.e. DSpace repositories)

F3. Metadata clearly and explicitly include the identifier of the data they describe

Examples: cg.identifier.doi, cg.identifier.uri
Reason: it links metadata, data and identifier composing the data object

F4. (Meta)data are registered or indexed in a searchable resource

Examples: Dataverse, CGSpace, MELSpace
Reason: the identifier itself does not ensure visibility, while a repository (optimally a FAIRport) might

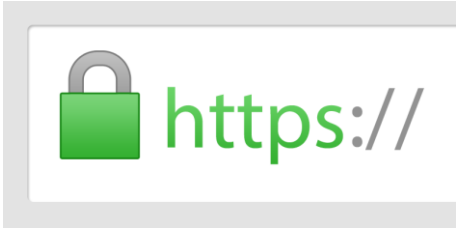
https://www.go-fair.org/fair-principles/

A for Accessible

A1. (Meta)data are retrievable by their identifier using a standardised communications protocol

Examples: http, https
Reason: up-to-date protocols ensure connection safety and are a Search Engine Optimization (SEO) asset

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



Examples: http, https
Reason: free and open protocols represent no obstacle to the user

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

Examples: http, https
Reason: limited access data is still FAIR and its condition of accessibility must be safeguarded by the protocol

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

Examples: when a repository is abandoned due to unsustainable costs, metadata should be left available online
Reason: metadata are valuable and allow the user to contact the source of the data object to request it

https://www.go-fair.org/fair-principles/

I for Interoperable

« L'interopérabilité est la capacité que possède un produit ou un système, dont les interfaces sont intégralement connues, à fonctionner avec d'autres produits ou systèmes existants ou futurs et ce sans restriction d'accès ou de mise en œuvre. » - AFUL, French speaking Libre Software Users' Association

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

Examples: Dublin Core, CG Core, JSON
Reason: metadata are meant to be intelligible and adopting a clear language will ensure so

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

Examples: ISO, AGROVOC
Reason: just as for metadata, quality vocabularies supported by communities are the basis of mutual readability

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

Examples: CG Core specifies that is built upon Dublin Core
Reason: by providing extensive information on the nature of your ontology, you allow for better integrations

https://www.go-fair.org/fair-principles/

R for Reusable

R1. Meta(data) are richly described with a plurality of accurate and relevant attributes

Examples: datasets metadata should include details about versioning, typology, format but also title, description…
Reason: metadata is content and rich metadata weight on the Findability, while also allowing advanced evaluations

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

Examples: Creative Commons
Reason: it is essential to 0 liabilities by defining the data object license

R1.2. (Meta)data are associated with detailed provenance

Examples: metadata on acquisition date, authors, original URI, versioning
Reason: it will increase the credibility of the resource and foster sharing

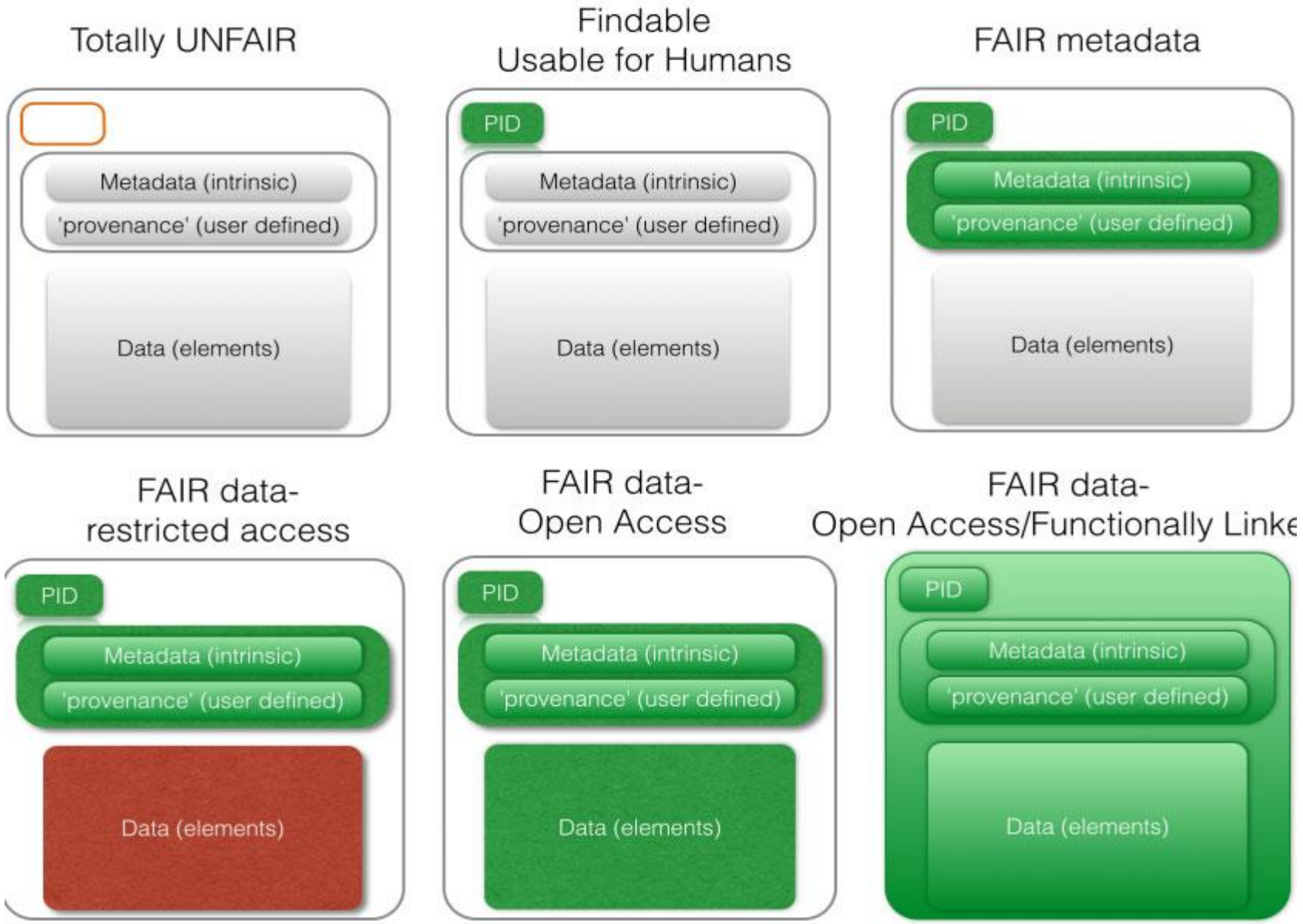
R1.3. (Meta)data meet domain-relevant community standards

Examples: “CGIAR Open Access and Data Management Policy” in relation to CG Core
Reason: meeting the minimum metadata requirements within your community ensures reusability at least within it

https://www.go-fair.org/fair-principles/

Questions

Data as increasingly FAIR Digital Objects



Provided, Open

Provided, Limited

Not Provided

1) Data Object not even Findable

2) Data Object only Findable.

3) Data Object Findable, FAIR Metadata

4) Data Object is FAIR although restricted

5) Data Object is fully FAIR

6) Data Object is FAIR and optimized

https://www.force11.org/fairprinciples

FAIR Example: The Dataverse® Project

- ✓ **Public Digital Object Identifier** (DOI) and other persistent identifiers (Handles). **[F]**
- ✓ Landing page providing access to **indexed and searchable metadata**, data files, dataset terms, waivers and licenses, version information. **[F, A, R]**
- ✓ Deposits include any **complementary files** (such as documentation or code) needed to understand the data and analysis. **[R]**
- ✓ **Public metadata**. **[F, A]**
- ✓ This **metadata is offered at three levels** 1) data citation metadata, which maps to DataCite schema or Dublin Core Terms, 2) domain-specific metadata, 3) file-level metadata. **[I, R]**
- ✓ **Public machine-accessible interfaces** to search the data, access the metadata and download the data files, using a token to grant access when data files are restricted. **[A]**

Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 3:160018 doi: 10.1038/sdata.2016.18 (2016).

FAIR Practices

Dataset Curation (PID, File Format, Vocabulary):

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

Search Engine Optimization (SEO) (sitemap, link building, protocol):

F1. (Meta)data are assigned a globally unique and persistent identifier
F4. (Meta)data are registered or indexed in a searchable resource
A1. (Meta)data are retrievable by their identifier using a standardised communications protocol
A1.1 The protocol is open, free, and universally implementable
A1.2 The protocol allows for an authentication and authorisation procedure, where necessary

Thank you