FAIR Principles Baseline Comments

A FAIRsFAIR WP4 Discussion Document

It’s noticeable in various FAIR-related work that the same comments and questions related to the original Principles are repeatedly referenced. Rather than do the same thing for FAIRsFAIRWP4 we will retain the baseline issues and comments in this document and refer back to them periodically to see if they’ve been addressed either by our own work or by others.

This text seeks to consider the issues around the FAIR data Principles, particularly as they apply to the notion of a Trustworthy Digital Repository (TDR). Issues here must be answered (or at least acknowledged) for us to provide an aligned approach to FAIR-enabled Trustworthy Digital Repositories. We can progress without all of these questions being addressed, but clarifying them will ensure a better overall solution.

NB: The detailed challenges of defining FAIR indicators, developing tests against those indicators, defining the details of TDR-FAIR evaluation processes (including assessment methods such as CMMI) will be addressed elsewhere. The focus here is a narrow review of the principles themselves.

Part of the goal is to identify and record these baseline questions here, and not to include them in the various other pieces of ongoing work.

`Monospaced text` is taken from https://www.nature.com/articles/sdata201618. This matches the sequence used by the FAIR Data Maturity working Group on RDA[1]. But note that the often referenced Force 11 principles page[2] transposes F3 and F4. As far as I'm aware there is no canonical FAIR Principles object with a DOI…

The FAIR Guiding Principles

To be Findable:

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

In many repositories, the digital object (data plus metadata) share a single identifier. Granular identifiers may also be assigned to 'parts' of an object including files or structures within a file (one identifier per variable for instance). We must not assume that local assumptions about 'object models' are universal.

Persistent Identifier services which support the minting and resolution of identifiers and the submission and management (including versioning within a PID and relationships between PIDs) such as DataCite can provide assurances that identifiers are globally unique. But they can only provide a system which supports persistence, they cannot guarantee it. Persistence in terms of maintained metadata and resolution to the correct maintained object remains dependent on the data steward.

Repositories not using a 'recognised' persistent identifier service may be able to offer assurances that their own identifiers are unique and that their procedures ensure persistence.

[1] https://github.com/RDA-FAIR/FAIR-data-maturity-model-WG/blob/master/results%20of%20preliminary%20analysis/v0.02/FAIR_Principles_Findable_v0.02.pdf

[2] https://www.force11.org/group/fairgroup/fairprinciples

But they must provide evidence for this and could be expected to do so in more detail than a repository using a well known PID service.

When repositories are seeking to enable this principle their persistent identifier service providers share some responsibility in providing evidence for uniqueness and for best practices in enabling and ensuring persistence

Clarify: **persistent**?

NB: some texts[3] refer to 'eternal' persistence, which seems desirable but optimistic.

Do we envisage a registry of PID providers? Would they be 'approved' and who by?
Could repositories or other data stewards not using the "Big" PID providers register themselves and provide details of their approach to ensuring uniqueness and persistence? What would the minimum criteria for PID service persistence be?

F2. data are described with rich metadata (defined by R1 below)

Richness is clearly desirable but hard to define, or to quantify at a generic level. There are challenges here in developing clear requirements and defining acceptable evidence. If particular metadata standards are to be deemed 'rich enough' (e.g. for a particular data type or discipline) this implies a need for community consensus and/or some form of authoratitative body and perhaps a metadata standard registry[4] which defines which standards and which levels of 'richness' are sufficient.

See related comments under R1

Clarify: **rich**?

F3. metadata clearly and explicitly include the identifier of the data it describes

One Implication here is that the metadata and the data it describes may be different 'objects' in different locations. There is a clear need to be able to find/resolve/locate the data *from* its metadata.
An identifier might be buried in prose and therefore unclear, or it may not be explicit that an identifier string is to the associated data (e.g. if there are numerous identifiers in the metadata).

[3] https://www.force11.org/group/fairgroup/fairprinciples
[4] https://fairsharing.org/standards/

An identifier in bold at the top of the record might be sufficient for a human, but a machine might require an identifier bound in appropriately defined structural metadata elements.

Clarify: **clearly?**
Clarify: **explicitly?**

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

Can the search (resource discovery) system be available to a limited group i.e. not the general public?

‘Registered’ implies a *push* action by the data steward (e.g. repository) while ‘indexing’ implies a *pull* action by the resource discovery system (which could be the repository again, or could be a search engine crawler over which the repository has no control). In a push scenario there may be more control over timely updates to information than in a pull scenario.

Do we envisage a registry of acceptable resource discovery systems?

To be Accessible:

Accessibility implies that some metadata has been accessed during the finding process, but it cannot imply that the data itself must be accessible to all (cf sensitive personal data). So Accessibility implies a process for assessing the rights of the requester to gain access to the data and either granting access or justifying rejection.

A1. (meta)data are retrievable by their identifier using a standardized communications protocol

Metadata and data may share an identifier as a single conceptual digital object.

Metadata should be retrievable (resolvable?) via an identifier but circumstances (justified or unjustified) may mean that the data is not retrievable (cf: A2).

In the most open interpretation of a communications protocol calling a repository on the telephone, quoting an identifier and receiving a copy of the data and metadata in the post would qualify as a communications protocol if both parties understand the inputs, outputs and parameters of the exchange. Is a more restrictive definition envisaged?

Is there some element of standardisation that needs to be clarified? Is this protocol documented, approved, adopted, managed?

Do we envisage a registry of acceptable communications protocols?

Clarify: **communications protocol?**
Clarify: **standardized communications protocol?**

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

- Open as in open source?
- Free as in Libre or Free as in no cost? Cf: FOSS vs FLOSS
- What does universally cover?
- Implementation by whom?

Clarify: **open?**
Clarify: **free?**
Clarify: **universally?**
Clarify: **implementable?**
Clarify: **universally implementable?**

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

It is helpful to acknowledge that some digital objects must be protected for legal and ethical reasons. This is a good starting point to ensure that digital objects, and the repositories that hold them, are not penalised for a perceived lack of openness.

Is it reasonable to assume that some metadata *must* always be available without an authentication/authorisation procedure?

How do we define the acceptable scope of ‘necessary’ within the boundaries of FAIR?
E.g.

- Sensitive personal data only accessible to approved users (OK)
- Restricting data access to minimise sharing (Not OK)

Clarify: **where necessary?**

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

This meets the ‘principle’ that principles should be empirically demonstrable.
One could meet this principle by ensuring that the metadata remains accessible ‘somewhere’.

Would it be reasonable to require that the data steward ensures that any persistent identifiers they have minited continue to resolve to some relevant information (e.g. Landing page) or, at a minimum that the PID resolves to the metadata record (e.g. DataCite)?

Clarify: **accessible**?

To be Interoperable:

Interoperable between?

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

Clarify: **formal**
Clarify: **accessible**
Clarify: **shared**
Clarify: **broadly applicable**
Clarify: **language for knowledge representation**

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

Is the implication here controlled vocabularies and ontologies?
This principle is recursive in nature.
Though CV’s values are implemented as values within metadata the vocabulary itself is conceptually a digital object. Are the FAIR principles expected to be applied identically to vocabularies as they are to digital object data and metadata?

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

- Metadata to metadata
- Data to data
- And between data and metadata

Clarify: **qualified references**

To be Reusable:

Reusable by?

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

The sentiment here is clear and justified, but there are a range of areas where clarification is needed and these may depend on context. For example whether the (meta)data are rich, plural (sufficiently broad and granular) and relevant seems to depend on whether a “domain-relevant community standard” can be identified, this implies a registry of standards associated with domains, communities (disciplines).

The reference here to “meta(data)” is in contrast to “(meta) data” elsewhere. It’s assumed that this is intended to indicate that this principle is expected to focus less on data and more on the metadata side.

Clarify: **richly**
Clarify: **plurality**
Clarify: **accurate**
Clarify: **relevant**

Are we dependent on the domain-relevant community standards (R1.3) to provide the context which would offer clarification here? Is such a standard also a dependency for R1.1 and R1.2.

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

Both the metadata and data must have usage (rights) information accessible to the user at the point of use.

Clarify: **clear**
Clarify: **accessible**
Clarify: **data usage licence**

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

Both the metadata and data must identify which activity they are generated by/used by and which human actor/machine agent they are attributed to.

Clarify: **detailed**
Clarify: **provenance**

Can these be clarified at the generic level or are they dependent on local context? In one environment an author and publisher might be sufficiently detailed provenance. In others a detailed history of all derived objects and which agents and processes generated them may be required to ensure trust in the resultant data.

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

This principle indicates that we must either clarify terms at a generic level and apply them to the full range of (meta)data scenarios, or we must acknowledge that clarification is only possible with further contextual information. Community standards could include approved file formats, approved metadata schemas or detailed instructions on how objects should be structured. This implies a need to define domain-relevant communities and to define the level of formality needed to ‘approve’ their standards. It certainly implies a need for a (FAIR) registry of such standards.

Clarify: **domain**
Clarify: **domain-relevant**
Clarify: **community**
Clarify: **standards**