FAIR data: an introduction

Sarah Jones
Digital Curation Centre, Glasgow
sarah.jones@glasgow.ac.uk
Twitter: @sjDCC

HKUST Open Science training, 18 March 2019, Hong Kong
Who has heard of FAIR?

Findable  Accessible  Interoperable  Reusable

Image CC-BY-SA by SangyaPundir
What FAIR means: 15 principles

Findable:

F1. (meta)data are assigned a globally unique and persistent identifier;
F2. data are described with rich metadata;
F3. metadata clearly and explicitly include the identifier of the data it describes;
F4. (meta)data are registered or indexed in a searchable resource;

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;

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;

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;

doi: 10.1038/sdata.2016.18
**The FAIR data principles explained**

<table>
<thead>
<tr>
<th>The FAIR Data Principles explained</th>
</tr>
</thead>
<tbody>
<tr>
<td>These webpages provide an actionable list of the 15 FAIR Data Principles as a simple guide when publishing data. For each principle, we give a basic definition, examples, and links to useful resources. We hope that by working through the list, anyone wishing to maximise the reusability of their data, can prioritise their efforts and make more informed choices regarding a suitable repository. We hope that this list will also focus the growing public discourse around FAIR: what is FAIR exactly, and what is it not.</td>
</tr>
</tbody>
</table>

**Findable:** Data and metadata are easy to find by both humans and computers. Machine readable metadata is essential for automatic discovery of relevant datasets and services, and for this reason are essential to the FAIRification process.

- F1: (meta)data are assigned globally unique and persistent identifiers
- F2: Data are described with rich metadata
- F3: Metadata clearly and explicitly include the identifier of the data it describes
- F4: (meta)data are registered or indexed in a searchable resource

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

- By giving data many ‘labels’, it will be much easier to find and reuse the data.
- Provide not just metadata that allows discovery, but also metadata that richly describes the context under which that data was generated
- “plurality” indicates that metadata should be as generous as possible, even to the point of providing information that may seem irrelevant.

FAIR and Open

- Concepts of FAIR and Open should not be conflated.
- Data can be FAIR or Open, both or neither.
- The greatest potential reuse comes when data are both FAIR and Open.

Diagram: Increasing degrees of FAIR data and Open data.
FAIR Digital Objects

DIGITAL OBJECT
Data, code and other research outputs
At its most basic level, data or code is a bitstream or binary sequence. For this to have meaning and to be FAIR, it needs to be represented in standard formats and be accompanied by Persistent Identifiers (PIDs), metadata and documentation. These layers of meaning enrich the object and enable reuse.

IDENTIFIERS
Persistent and unique (PIDs)
Digital Objects should be assigned a unique and persistent identifier such as a DOI or URN. This enables stable links to the object and supports citation and reuse to be tracked. Identifiers should also be applied to other related concepts such as the data authors (ORCIDs), projects (RAIDs), funders and associated research resources (RRIDs).

STANDARDS & CODE
Open, documented formats
Digital Objects should be represented in common and ideally open file formats. This enables others to reuse them as the format is in widespread use and software is available to read the files. Open and well-documented formats are easier to preserve. Data also need to be accompanied by the code used to process and analyse the data.

METADATA
Contextual documentation
In order for Digital Objects to be assessable and reusable, they should be accompanied by sufficient metadata and documentation. Basic metadata will enable data discovery, but much richer information and provenance is required to understand how, why, when and by whom the objects were created. To enable the broadest reuse, they should be accompanied by a ‘plurality of relevant attributes’ and a clear and accessible usage license.
How to be FAIR & encourage reuse

• Choose file formats that are common
• Document your data!
• Use metadata standards
• Share your data via a repository
• Get a persistent identifier (via repository)
• Licence your data, ideally openly
• Cite other people’s data
Choose appropriate file formats

If you want your data to be re-used and sustainable in the long-term, you typically want to opt for open, non-proprietary formats.

<table>
<thead>
<tr>
<th>Type</th>
<th>Recommended</th>
<th>Avoid for data sharing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tabular data</td>
<td>CSV, TSV, SPSS portable</td>
<td>Excel</td>
</tr>
<tr>
<td>Text</td>
<td>Plain text, HTML, RTF</td>
<td>Word</td>
</tr>
<tr>
<td></td>
<td>PDF/A only if layout matters</td>
<td></td>
</tr>
<tr>
<td>Media</td>
<td>Container: MP4, Ogg</td>
<td>Quicktime</td>
</tr>
<tr>
<td></td>
<td>Codec: Theora, Dirac, FLAC</td>
<td>H264</td>
</tr>
<tr>
<td>Images</td>
<td>TIFF, JPEG2000, PNG</td>
<td>GIF, JPG</td>
</tr>
<tr>
<td>Structured data</td>
<td>XML, RDF</td>
<td>RDBMS</td>
</tr>
</tbody>
</table>

Further examples:
[www.data-archive.ac.uk/create-manage/format/formats-table](http://www.data-archive.ac.uk/create-manage/format/formats-table)
Where to find relevant standards?

Metadata Standards Directory
Broad, disciplinary listing of standards and tools. Maintained by RDA group
https://rdamsc.dcc.ac.uk

FAIRsharing
A portal of data standards, databases, and policies
Focused on life, environmental and biomedical sciences, but expanding to other disciplines
https://fairsharing.org
Dataset licensing

What do you want to allow others to do with your data?
• Copy
• Modify
• Remix / reuse

Put as few restrictions as possible.
CC-BY (attribution only)
CC-0 (public domain)
Data repositories

The EC guidelines point to Re3data as one of the registries that can be searched to find a home for data.

www.re3data.org

www.fosteropenscience.eu/content/re3data-demo
Considerations selecting repositories

- Often preferable to use a subject specific repository if available
- Useful if repositories assign a persistent identifier
- Look for certification as a ‘Trustworthy Digital Repository’ with an explicit ambition to keep the data available in long term.
- Generic repositories are also available e.g. Zenodo or institutional repositories

Icons to note open access, licenses, PIDs, certificates...
Citing research data: why?

[Diagram showing the process of building a culture of data citation: CREATE, USE, REWARD, MEASURE.]

http://ands.org.au/cite-data
How to cite data

Key citation elements

- Author
- Publication date
- Title
- Location (= identifier)
- Funder (if applicable)

www.dcc.ac.uk/resources/briefing-papers/introduction-curation/data-citation-and-linking
How FAIR are your data?

- Complete the FAIR data checklist
- Base decisions on how you currently manage and share your data
- Which are the most challenging aspects of FAIR to meet?