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?

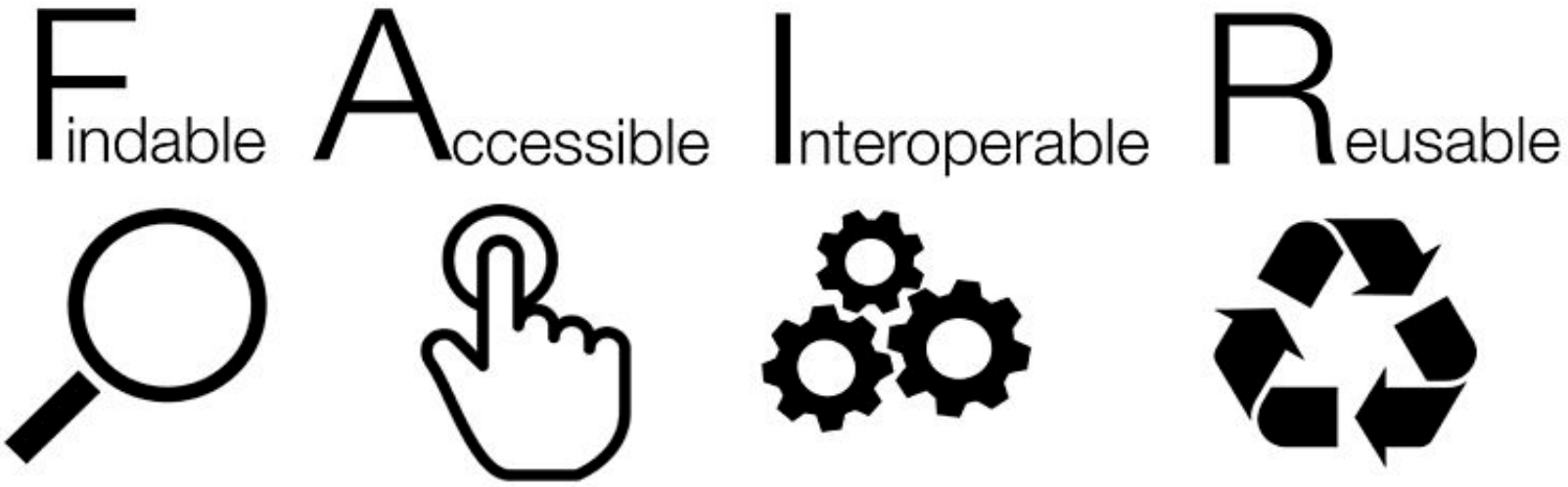


Image CC-BY-SA by SangyaPundir

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;

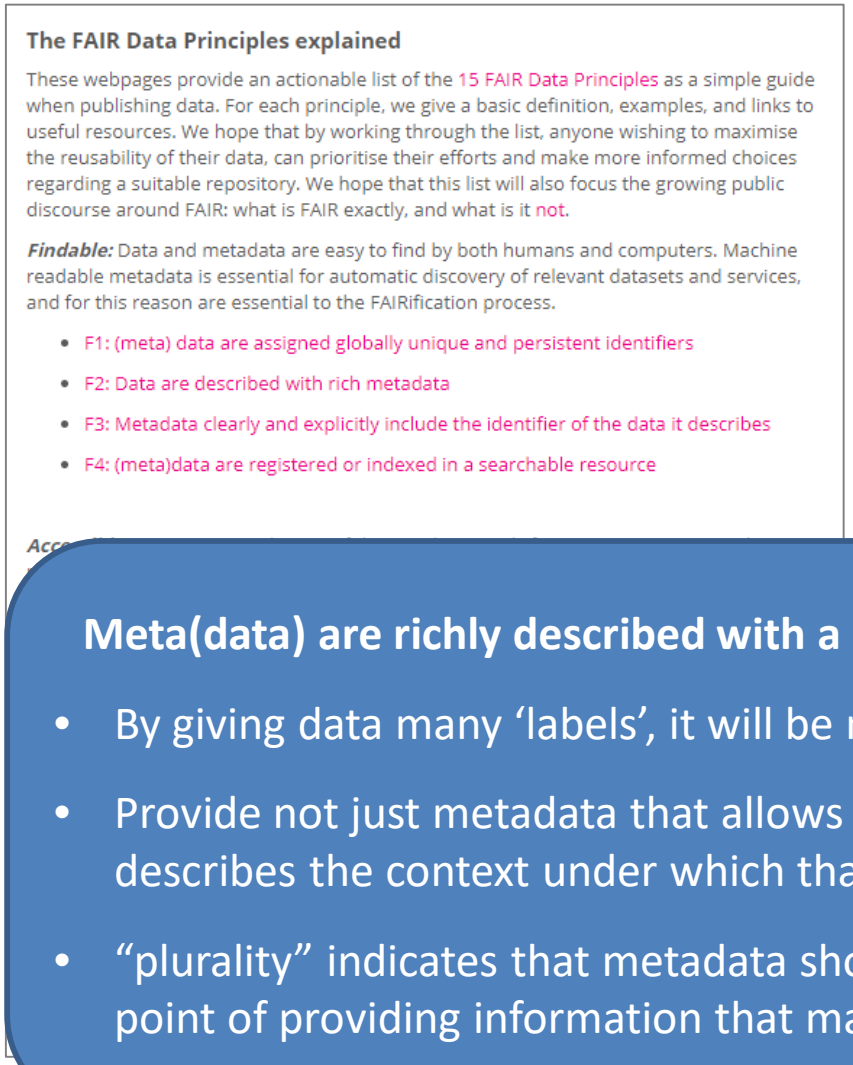
R1.1. (meta)data are released with a clear and accessible data usage license;

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



The FAIR Data Principles explained

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.

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

Acc

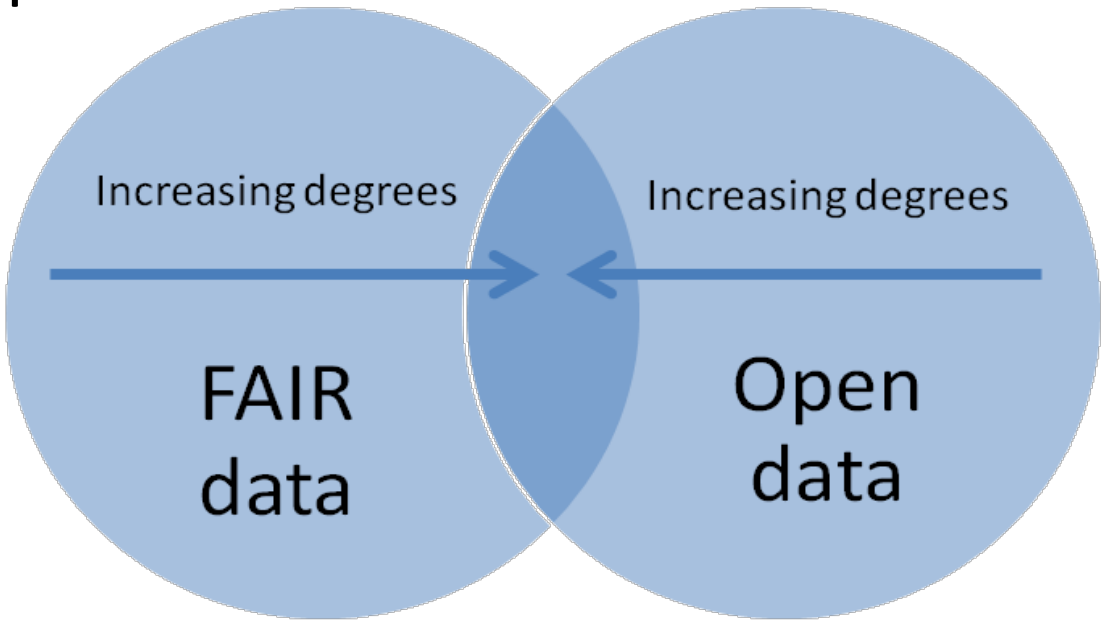
- Clarifications from the Dutch Techcentre for Life Sciences
- Each principle is a link to further clarification, examples and context

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

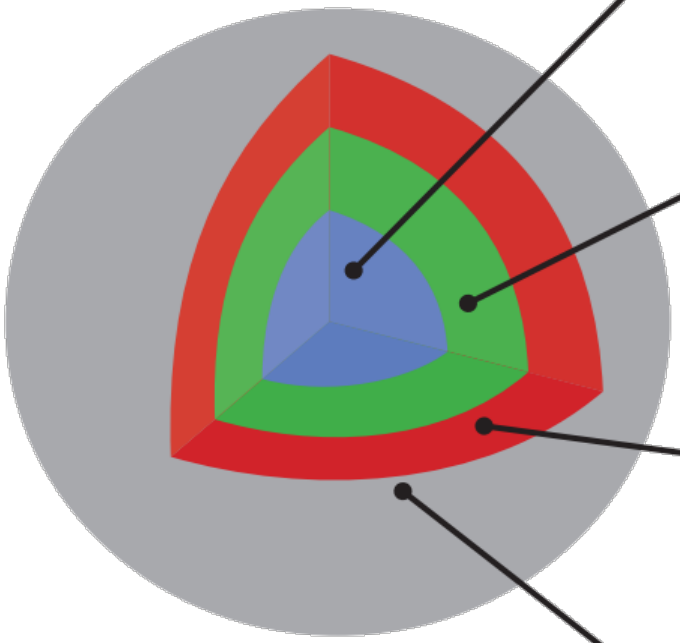
- By giving data many ‘labels’, it will be much easier to find and reuse the data.
- Provide not just metadata that allows discovery, but also metadata that richly describes the context under which that data was generated
- “plurality” indicates that metadata should be as generous as possible, even to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



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 use 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.

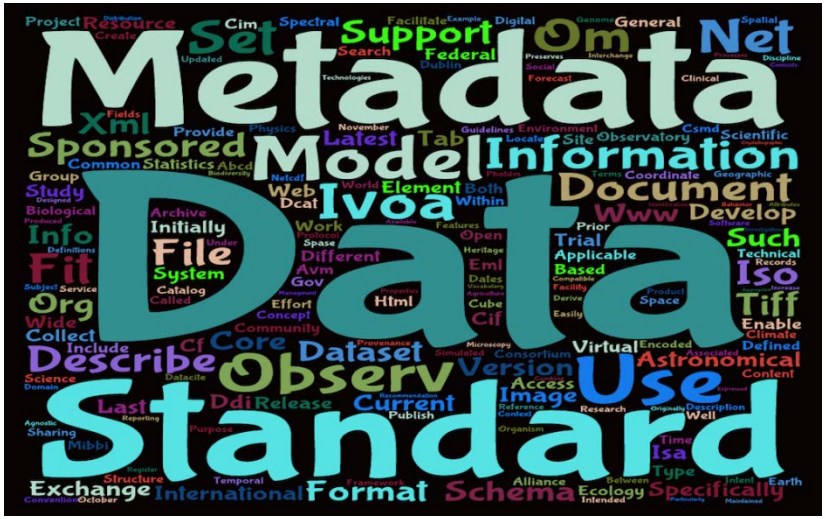
Type	Recommended	Avoid for data sharing
Tabular data	CSV, TSV, SPSS portable	Excel
Text	Plain text, HTML, RTF PDF/A only if layout matters	Word
Media	Container: MP4, Ogg Codec: Theora, Dirac, FLAC	Quicktime H264
Images	TIFF, JPEG2000, PNG	GIF, JPG
Structured data	XML, RDF	RDBMS

Further examples:
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

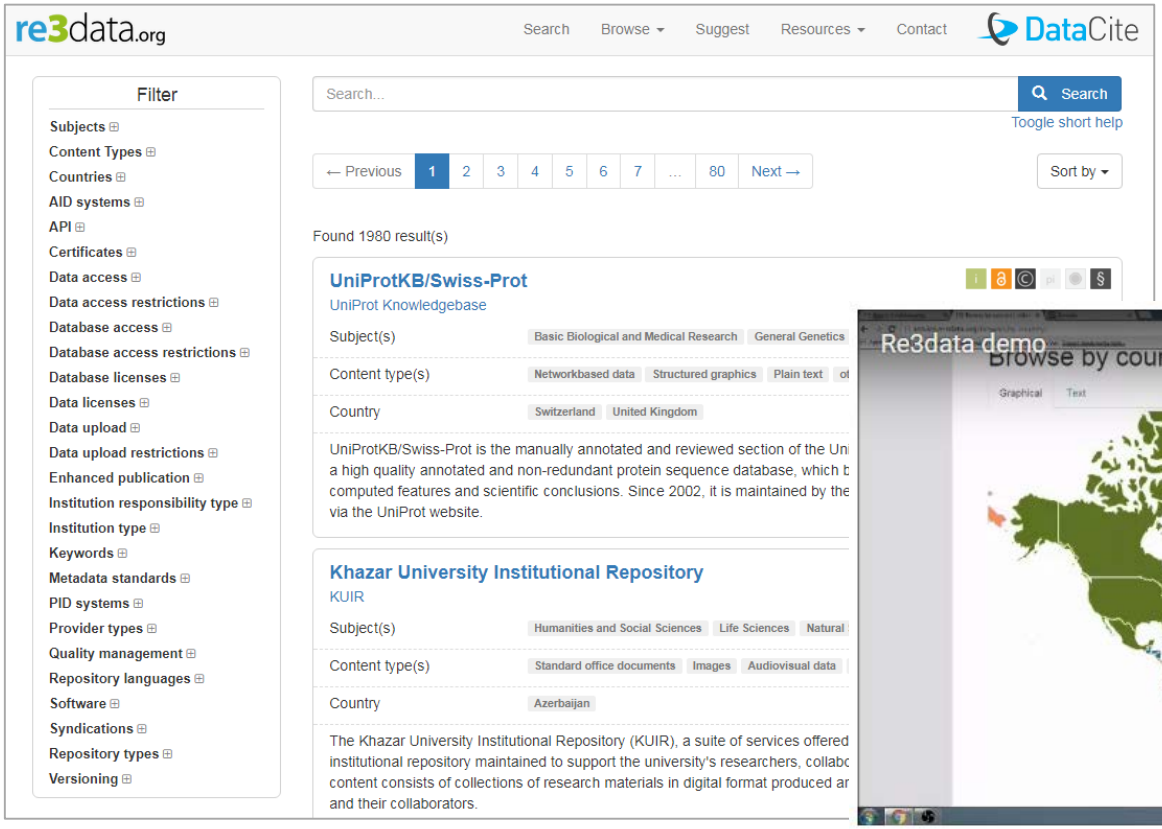
What do you want to allow others to do with your data?

- 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

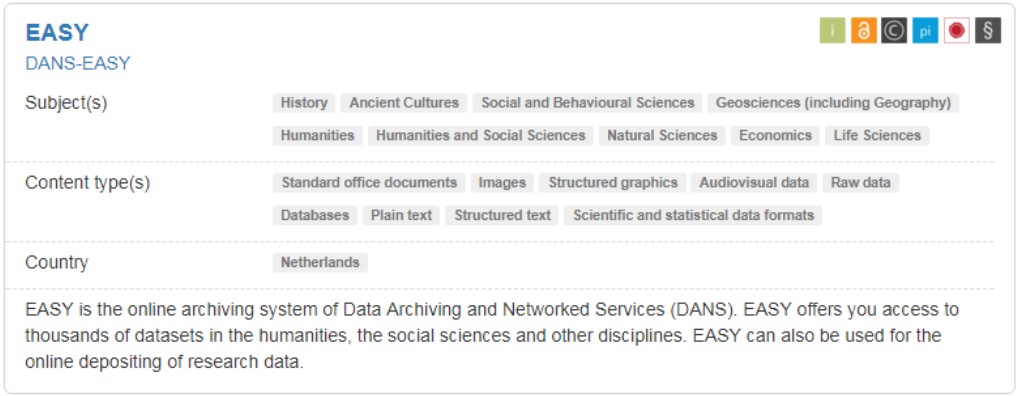


www.re3data.org

www.fosteropenscience.eu/content/re3data-demo

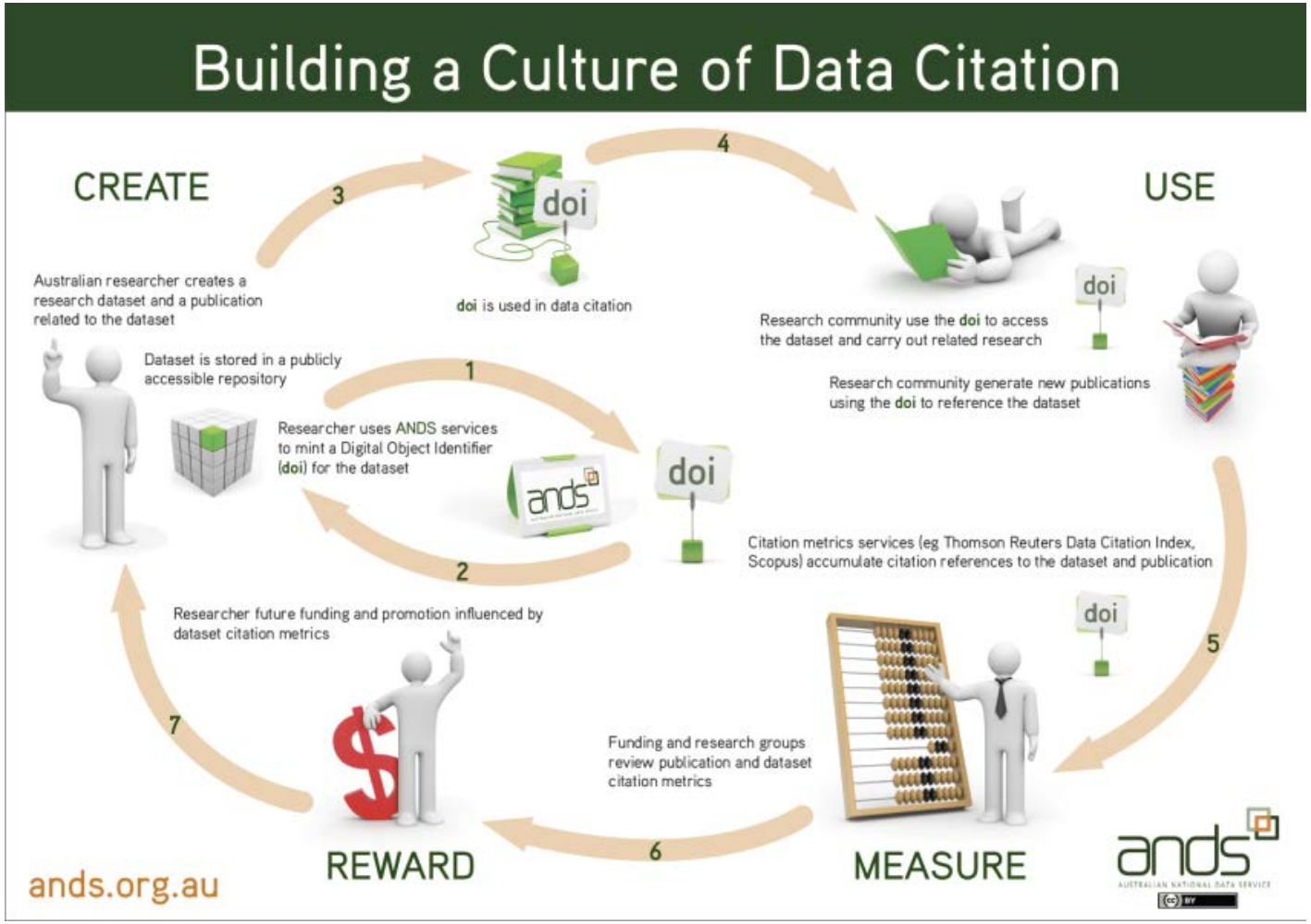
Considerations selecting repositories

- 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?

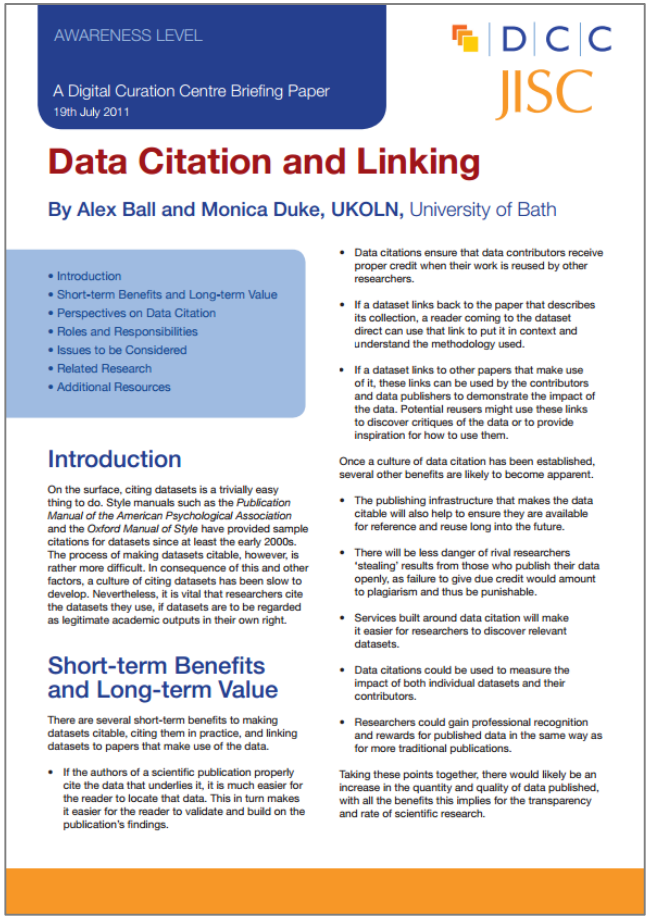


http://ands.org.au/cite-data

How to cite data

Key citation elements

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



AWARENESS LEVEL

A Digital Curation Centre Briefing Paper
19th July 2011

DCC JISC

Data Citation and Linking

By Alex Ball and Monica Duke, UKOLN, University of Bath

- Introduction
- Short-term Benefits and Long-term Value
- Perspectives on Data Citation
- Roles and Responsibilities
- Issues to be Considered
- Related Research
- Additional Resources

Introduction

On the surface, citing datasets is a trivially easy thing to do. Style manuals such as the *Publication Manual of the American Psychological Association* and the *Oxford Manual of Style* have provided sample citations for datasets since at least the early 2000s. The process of making datasets citable, however, is rather more difficult. In consequence of this and other factors, a culture of citing datasets has been slow to develop. Nevertheless, it is vital that researchers cite the datasets they use, if datasets are to be regarded as legitimate academic outputs in their own right.

Short-term Benefits and Long-term Value

There are several short-term benefits to making datasets citable, citing them in practice, and linking datasets to papers that make use of the data.

- If the authors of a scientific publication properly cite the data that underlies it, it is much easier for the reader to locate that data. This in turn makes it easier for the reader to validate and build on the publication's findings.
- Data citations ensure that data contributors receive proper credit when their work is reused by other researchers.
- If a dataset links back to the paper that describes its collection, a reader coming to the dataset direct can use that link to put it in context and understand the methodology used.
- If a dataset links to other papers that make use of it, these links can be used by the contributors and data publishers to demonstrate the impact of the data. Potential reusers might use these links to discover critiques of the data or to provide inspiration for how to use them.

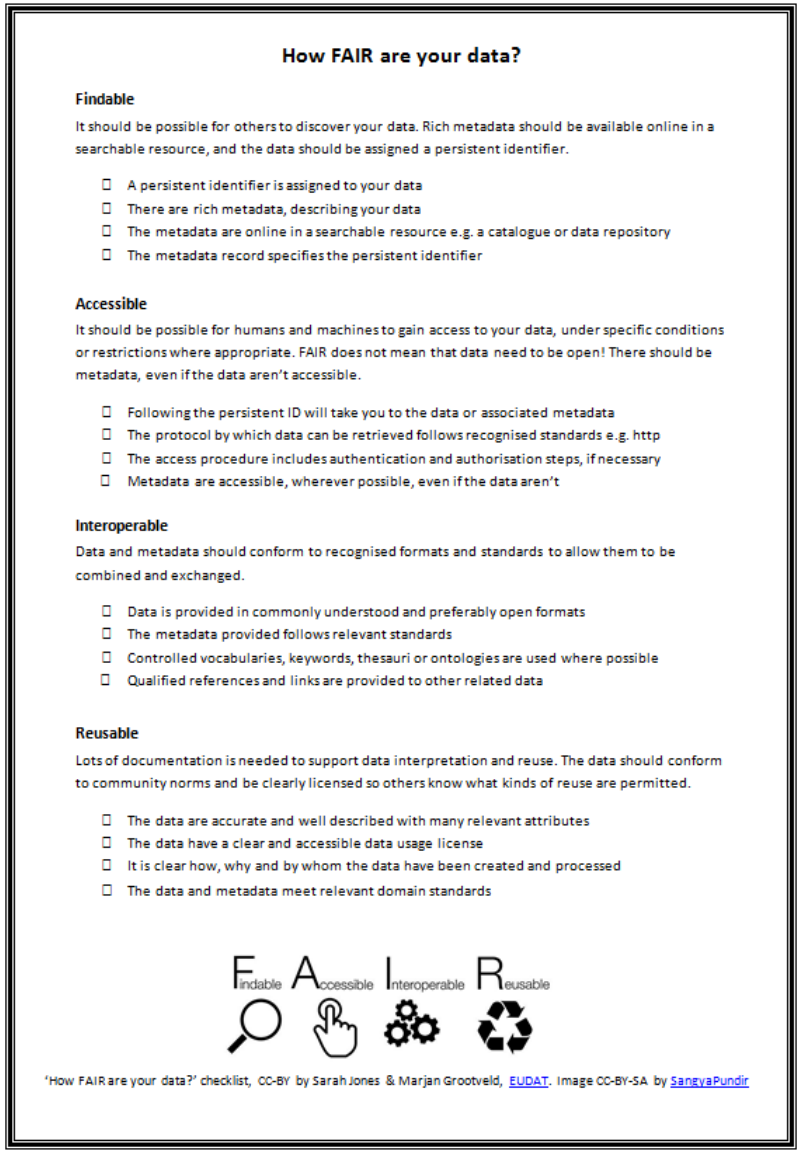
Once a culture of data citation has been established, several other benefits are likely to become apparent.

- The publishing infrastructure that makes the data citable will also help to ensure they are available for reference and reuse long into the future.
- There will be less danger of rival researchers 'stealing' results from those who publish their data openly, as failure to give due credit would amount to plagiarism and thus be punishable.
- Services built around data citation will make it easier for researchers to discover relevant datasets.
- Data citations could be used to measure the impact of both individual datasets and their contributors.
- Researchers could gain professional recognition and rewards for published data in the same way as for more traditional publications.

Taking these points together, there would likely be an increase in the quantity and quality of data published, with all the benefits this implies for the transparency and rate of scientific research.

www.dcc.ac.uk/resources/briefing-papers/introduction-curation/data-citation-and-linking

How FAIR are your data?



How FAIR are your data?

Findable

It should be possible for others to discover your data. Rich metadata should be available online in a searchable resource, and the data should be assigned a persistent identifier.

- ☐ A persistent identifier is assigned to your data
- ☐ There are rich metadata, describing your data
- ☐ The metadata are online in a searchable resource e.g. a catalogue or data repository
- ☐ The metadata record specifies the persistent identifier

Accessible

It should be possible for humans and machines to gain access to your data, under specific conditions or restrictions where appropriate. FAIR does not mean that data need to be open! There should be metadata, even if the data aren't accessible.

- ☐ Following the persistent ID will take you to the data or associated metadata
- ☐ The protocol by which data can be retrieved follows recognised standards e.g. http
- ☐ The access procedure includes authentication and authorisation steps, if necessary
- ☐ Metadata are accessible, wherever possible, even if the data aren't

Interoperable

Data and metadata should conform to recognised formats and standards to allow them to be combined and exchanged.

- ☐ Data is provided in commonly understood and preferably open formats
- ☐ The metadata provided follows relevant standards
- ☐ Controlled vocabularies, keywords, thesauri or ontologies are used where possible
- ☐ Qualified references and links are provided to other related data

Reusable

Lots of documentation is needed to support data interpretation and reuse. The data should conform to community norms and be clearly licensed so others know what kinds of reuse are permitted.

- ☐ The data are accurate and well described with many relevant attributes
- ☐ The data have a clear and accessible data usage license
- ☐ It is clear how, why and by whom the data have been created and processed
- ☐ The data and metadata meet relevant domain standards

'How FAIR are your data?' checklist, CC-BY by Sarah Jones & Marjan Grootveld, EUDAT. Image CC-BY-SA by SangyaPundir

- 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?