figshare

Figshare and the FAIR data principles

December 2018

Patrick Splawa-Neyman

Contents

1. Introduction

The purpose of this document is to demonstrate how Figshare applies the FAIR data principles and that Figshare is more than just a publishing platform. The bulk of this paper looks in detail at the 15 FAIR data principles and how Figshare implements them. Figshare is seeking feedback on how we see our product interact with the FAIR data principles.

1.1 Why FAIR is important to Figshare

Figshare was launched in January 2012 and started as a data sharing platform for researchers to add their digital files that would otherwise go unpublished. One of Figshare's early goals was to encourage researchers to publish their negative results and give them a platform to disseminate those results to reduce duplication of work as well as encourage citations of the data itself. Scientific knowledge will accrue at a faster rate if all results are published, not just those results that will be held in high esteem.

Since those early days Figshare has continued to evolve and is now a next generation repository platform. However our ethos has not changed and even before the formalisation of the FAIR data principles we believed that data should be findable, accessible, interoperable and reusable.

As an extension of those principles we also believe in tracking public research outputs via Altmetrics and Dimensions to further encourage making data openly available. We continue to support academic publishers, preprint platforms, researchers, conferences and labs.

Our core beliefs that work alongside the FAIR data principles include:

- Academic research outputs should:
    - be as open as possible, and as closed as necessary
    - never be behind a paywall
    - be human and machine readable / query-able
- Academic infrastructure should be interchangeable
- Academics should not have to provide the same information into multiple systems
- Identifiers are mandatory for all entities
- The impact of research is independent of where it is published and what type of output it is

2. The FAIR data principles in detail

2.1 Findable

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

This principle is potentially the most important of the principles as it underpins all the others. A globally unique and persistent identifier disambiguates a dataset or any other grouping of digital files from all the others much the same way that ORCIDs (another globally unique identifier) disambiguate researchers from each other. Other identifiers include PURLs for web resources, and GRIDs (Global Research Identifier Database) for research organisations.

Differentiating datasets makes them more visible, more available and citable. A DOI (digital object identifier) in Figshare points to the landing page that displays the metadata description and allows access to the data itself. Clicking on the DOI on a Figshare landing page is akin to refreshing the page because the DOI is pointing to the Figshare landing page URL that the user is already on.

DOIs in Figshare can also be reserved. In this instance the DOI is reserved but is inactive meaning that no other person or institution can mint that DOI making it globally unique. An advantage of this is that the reserved yet inactive DOI can be included in a journal publication so that upon publication of the journal article the Figshare item can be published thereby activating the DOI. The end user can then click on the Figshare DOI in a journal article and be taken to the data and its description. Conversely, if the journal article has a persistent identifier, it can be included in the Figshare item as a Reference so the user can link from the data to the journal and back ad infinitum.

Institutions that sign up to Figshare can request that DOI minting be done by Figshare in which case Figshare mints the DOIs through DataCite. Alternatively, an institution can choose to mint DOIs through a DataCite registrant such as ARDS, the Australian Research Data Commons which is also a DataCite member. Either method creates a globally unique persistent identifier by means of algorithms to confirm newly minted DOIs are unique and will not be replicated elsewhere.

The second part of this principle is to ensure the identifier is persistent and will not easily disappear. A globally unique identifier is one that once assigned to one entity cannot be assigned to another. But equally importantly persistence means that resources must be allocated to the entity to ensure that reference rot (comprising link rot and content drift) do not occur.

In order to mint DOIs registry services such as the Australian Research Data Commons (formerly the Australian National Data Service) need to guarantee to DataCite that they themselves will persist in the long-term. All items published through Figshare receive a DOI which is accessed by clicking on the Cite button on a Figshare landing page.

F2. data are described with rich metadata

In order for research outputs to be widely available they need to be accompanied by a generous description that makes clear to potential users what is contained in them. Ideally the description will be contextual, comprehensive and of high quality. Providing a comprehensive description allows computers to sort and prioritise research outputs which must be done by a person if the description is inadequate. When a user manages to find a potentially useful dataset they are then required to ascertain if it is useful depending on how thoroughly it has been described by the originating researcher. Rich metadata is a subjective term but a broad principle to follow is to be generous and include as much detail as time allows irrespective of whether you think others will find it useful. This principle is about providing the correct context for interpretation to help users locate and reuse the data. Each dataset is described comprehensively and includes identifying descriptions including how the data was generated, and under what licence the data can be released.

Figshare allows datasets to be described with the following default fields: title, authors, description, keywords, categories, licence, item type, external references and funding. There is also an ability to add an embargo, make public files confidential, share data via a private link and reserve a digital object identifier.

F3. (meta)data are registered or indexed in a searchable resource

It is not enough to have rich metadata descriptions and expect that this will make the data findable. A resource that cannot be found or can only be found with great difficulty will be little used. Indexing is one of the ways that data can be searchable. Googlebot is Google's web crawling spider that crawls the internet in order to find updated and new pages. Figshare and the institutional instances of Figshare are very well indexed by Google meaning that much of the traffic to Figshare comes from Google. Google has also recently launched Dataset Search which uses Schema.org structured data markup. Any repositories including Figshare that use Schema.org will be able to find their content in Dataset Search.

F4. metadata specify the data identifier

Metadata is usually separate from the data it describes. The link between the two needs to be made explicit by clearly stating the globally unique and persistent identifier in the metadata that relates to the data. It's not enough to have an identifier, it must also be clear what that identifier is. This is exactly what Figshare does. A DOI (a globally unique persistent identifier) is minted and associated with the Figshare landing page's URL and all the data uploaded to the item. When the metadata from a Figshare item is harvested to another system such as Research Data Australia (RDA) the separation of data from metadata using the DOI remains. Research Data Australia clearly labels the persistent identifier as a DOI however RDA does not house the data. It acts as a data discovery service and using the Figshare DOI sends the user back to the Figshare landing page to access the data.

2.2 Accessible

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

The only way data can be widely accessible is via a standardised protocol which in the case of Figshare is HTTPS, the hypertext transfer protocol secure version. This requires no specialised tools other than a web browser. HTTPS is ubiquitous, is exceptionally well documented and does not require manual intervention. Data can still be considered FAIR if another protocol is used as long as the data transfer method uses a standardised protocol. Users of Figshare need no other specialised tool than a web browser to use HTTPS to navigate to a public landing page and access the data and metadata, or use the same protocol to privately share data. The top of each Figshare landing page displays the URL with a green padlock icon to show that it is a secure connection.

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

A free, open and universally implementable protocol smooths the way for and encourages data retrieval and transfer. To easily access data all that is required is a computer with an internet connection. Using these users can access the data using the protocol. HTTPS fulfils these criteria and is not proprietary meaning it is not locked down with any particular vendor or company that requires users to use its proprietary communications protocol. Web browsing is pervasive and it is common knowledge that the hypertexts embedded in text move the user from one page to another. Figshare uses the HTTPS protocol to make data open and free through the use of private links. A private link in Figshare can be sent electronically to anyone in the world and as long as they have an internet connection they will be able to access the data whether they have a figshare.com account or not. HTTPS as it relates to FAIR principle A1 is about a standardised protocol. Principle A1.1 refers to that protocol being open and free.

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

The FAIR principles are related to open data but they are not synonymous with it. FAIR does not always mean open, and restricted data can follow the principles of FAIR. This is of paramount importance to researchers. Accessibility in FAIR does not designate openness, instead it provides the circumstances under which the data can be accessed. Sensitive data can be FAIR, as can embargoed data.

Figshare for Institutions can connect with Single Sign-On services in order to facilitate authentication, authorization and account provisioning. Figshare supports integration with services that employ the Security Assertion Markup Language 2.0 (SAML 2.0). It implements a Shibboleth service provider, with current integrations including Shibboleth Identity Provider, Active Directory

Federation Services (AD FS), Okta and Google Suite SAML Applications. Figshare also supports integrations via federations.

A Figshare for Institutions user can invite others to collaborate on a Project. Those users can authenticate via their institution's Single Sign-On but even users without a Figshare for Institutions account can participate in a Project. In this instance the user will need to create a figshare.com account (which takes under two minutes) and using this account can accept an invitation to collaborate on a Project. The Project Manager can determine which of two levels of authorization a user has and always retains the ability to remove users from a Project.

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

Data can deteriorate over time unless it is monitored and maintained on an ongoing basis. Figshare can arrange storage for institutions and uses Amazon S3 to stores files and generate previews. S3 provides durable infrastructure with data being redundantly stored across multiple facilities and multiple devices in each facility. Institutions can also bring their own storage which might be located on site or in the cloud. When bringing their own storage it is important for each institution to ensure that it has a clear understanding with its provider on the durability of its data. Ideally the data and metadata to an institution's files will remain accessible long-term which is why publishing through Figshare mints a DOI thereby providing a globally unique persistent identifier to the data.

In addition to minting a DOI for public datasets Figshare also allows metadata only records to be published with a DOI. A metadata only record that describes a physical collection of rocks in geology is an example of a metadata only record.

2.3 Interoperable

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

The great advantage of modern computer systems is their ability to communicate with each other regardless of which operating system is being used. To be interoperable computer systems need to understand the multitude of formats across systems that enable data exchange. The formal, accessible, shared and applicable languages mentioned in this principle that Figshare uses are Dublin Core, Qualified Dublin Core, DataCite, RDF (the Resource Description Framework) and CERIF XML (the Common European Research Information Format). Using these representations users can pull content out of Figshare in an agreed format at any time for textual or data analysis.

I2 (meta)data use vocabularies that follow FAIR principles

One of the controlled and documented vocabularies that Figshare uses is the Australian and New Zealand Standard Research Classification Fields of Research (FoR) codes. These are listed in full on the Australian Bureau of Statistics website and at the 6-digit level Figshare uses all 1,238 FoR codes as Categories. There are numerous academic classification systems available but the ANZ FoR codes are the most comprehensive across disciplines which is why Figshare has chosen to include them as our categories. Once made publicly available a user can click on any of these Categories and bring up a list of other public items that also include that Category.

Figshare also has a small controlled vocabulary for our item types which currently are figure, media, dataset, fileset, poster, paper, preprint, presentation, thesis and code. The item type links out to a Figshare landing page where the user can obtain more information. These pages show the typical file format in each type, a brief description of each, and an example of how that file will be displayed in the browser when uploaded to Figshare.

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

A qualified reference is one that states the relationship between datasets. More linkages create more context about the research and the relationships between datasets. And of course globally unique, persistent identifiers need to be present for each dataset and collection of datasets as well. In Figshare a Collection is a grouping of individual items. Each Collection has its own title, description, authors, categories, and keywords just like an individual item does. A Collection also has its own DOI but the primary difference is that the Collection pulls together related items. The metadata in the Collection describes the metadata and data in the individual items and how they relate to each other as well as version control. This is what qualified references to other data refers to in this principle.

2.4 Re-usable

R1 meta(data) have a plurality of accurate and relevant attributes

Principle F2 states that data should be described with rich metadata. Principle R1 extends that notion by creating more metadata fields to make the description more valuable to people and machines. This makes the description of the data more complete and it also makes the data more findable. Making your metadata fields accurate is very much up to each institution, for example, you might create a new field on the primary work location for your staff and ask them to select the one campus or location they spend the most time at. But researchers might struggle to see how that is relevant to their research. These attributes can also include research protocols, drug dosage, the numbers of participants or anything an institution deems relevant. What is undeclared in this principle is the idea of not trying to anticipate what data a future user might find worthwhile. The idea is to have as rich and as extensive a description as feasible as it is impossible to tell now what data will be valuable in future.

This creates a conundrum for institutions that have the ability to create custom metadata fields. Do institutions leave the default fields as the standard for users and have not as rich a description, or, do they add custom metadata fields and compromise the number of researchers using the service because they will be potentially turned off by the added administrative task? We are still in an environment of journal impact factors and citation rates which drive academic success, and are not yet at the point where open data is the norm. Open data does not necessarily mean well described data or quality data which is why the FAIR principles are so important.

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

Data reuse is managed by the copyright owner licencing his or her data for reuse with conditions upon publication. A clear, easy to understand licence embedded in the metadata facilitates data reuse. How and under what conditions researchers will allow their data to be shared must be made explicit to machines and people.

Figshare allows institutions to add licences to the standard suite and decide which licences they would like to remove. Institutions can also decide which licence they would like to include as the default. The Creative Commons licence logo, for example, is prominently displayed on all public items alongside the text of the licence. Users who click on either of these are taken to the creativecommons.org website where that specific licence is explained in plain English. From here users can navigate to the legal code of each licence. Each licence has three layers: a legal code, a human readable version and a machine readable version, each of which is accessible to the user.

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

In order to reuse and cite data it is imperative that users know who is releasing the data so the originating author can be acknowledged as well as having an understanding of the data usage restrictions. On Figshare public items the citable authors are the originators of the data. From there it is at the discretion of the authors to delineate the provenance of the data itself. This could be done by adding the provenance metadata to the description field or by adding custom metadata fields so that all users have access to the same metadata fields. Figshare also automatically generates a citation based on input from the author. The default citation style is Datacite but Figshare allows thousands of different citations styles to be created at the click of a button.

All users who publish through Figshare have a profile page which they can populate with their job titles, areas of expertise, location, a biography, previous publications, and social media links. These all help to build a profile about the researcher. In addition to those fields users can also synchronise their Figshare accounts with their ORCID accounts. This allows the metadata from a newly published Figshare item to be automatically pushed to and published on their ORCID page, allowing one more layer to provenance.

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

Using standardised and well established conventions makes data reuse easier. DOIs can be minted through Figshare or through another organisation such as the Australian Research Data Commons, both of which are DataCite members. DOIs are minted by Datacite via one of its members. As well as minting DOIs DataCite also uses the DataCite Metadata Schema which is a list of core metadata properties that provides an accurate and consistent identification of a resource so that resources can be cited and retrieved. As a minimum FAIR data needs to meet that standard. While typically a dataset is being identified, any type of resource can be included. Using community standards makes data more easily reused.

The DataCite Metadata properties comprise three levels: mandatory, recommended and optional. By default when a user on a Figshare landing page exports the metadata using the DataCite export function they are exporting the mandatory fields which are identifier (DOI), Creator (authors), Title (title), Publisher (institutional client of Figshare), PublicationYear (date) and ResourceType (item type).

3. Next steps

Now that Figshare has outlined the FAIR data principles and how we interact with them we are releasing this paper to the Figshare community. We ask our community of users to read and analyse our views and provide feedback by emailing us at info@figshare.com. We have done our best to ensure that this document accurately reflects the FAIR data principles and Figshare functionality.

4. References

Allen, R & Hartland, D 2018, ‘FAIR in practice: Jisc report on the Findable Accessible Interoperable and Reuseable Data Principles’, *Joint Information Systems Committee,* https://doi.org/10.5281/zenodo.1245568.

Australian National Data Service 2018, *The FAIR Data Principles*, viewed 20 September 2018, https://www.ands.org.au/working-with-data/fairdata.

Boeckhout, M, Zielhuis, GA & Bredenoord, AL 2018, ‘The FAIR guiding principles for data stewardship: fair enough?’, *European Journal of Human Genetics*, vol. 26, pp. 931-936.

DataCite Metadata Schema 2018, *Metadata Schema 4.1, viewed 24 September 2018, https://schema.datacite.org/meta/kernel-4.1/doc/DataCite-MetadataKernel_v4.1.pdf.*

Data FAIRport 2018, *The FAIR Data Principles,* viewed 7 October 2018, http://www.datafairport.org/index.html.

F.A.I.R. 2018, *Policy Statement on F.A.I.R. Access to Australia’s Research Outputs,* viewed 5 October 2018, *https://www.fair-access.net.au/__data/assets/pdf_file/0029/74198/F.A.I.R.-Statement_Jan_2017.pdf*

FORCE11 2017, *The FAIR Data Principles*, viewed 3 October 2018, https://www.force11.org/group/fairgroup/fairprinciples.

GO FAIR 2018, *FAIR Principles*, viewed 11 October 2018, https://www.go-fair.org/fair-principles/.

Hodson, S, Jones, S *et al* 2018, 'FAIR Data Action Plan. Interim recommendations and actions from the European Commission Expert Group on FAIR data', *European Commission,* https://doi.org/10.5281/zenodo.1285290.

Kraft, A 2017, 'The FAIR Data Principles for Research Data', *TIB BLOG*, viewed 8 October 2018, https://blogs.tib.eu/wp/tib/2017/09/12/the-fair-data-principles-for-research-data/

Mons, B, Neylon, C, Velterop, J, Dumontier, M, da Silva Santos, LOB & Wilkinson, MD 2017, 'Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud', *Information Services & Use,* vol. 37. pp. 49-56.

Swiss National Science Foundation 2018, *Explanation of the FAIR data principles,* viewed 28 September 2018, http://www.snf.ch/SiteCollectionDocuments/FAIR_principles_translation_SNSF_logo.pdf.

Wilkinson, MD, *et al* 2016, 'The FAIR Guiding Principles for scientific data management and stewardship' *Scientific Data*, 3:160018, https://doi.org/10.1038/sdata.2016.18.