Using CyVerse to make your data FAIR

What is FAIR Data?

FAIR data principles (Wilkinson et al. 2016) provide a set of basic requirements for making research data **Findable, Accessible, Interoperable, and Reusable**.

The FAIR Guiding Principles (Box 2 from Wilkinson et al. 2016)

To be FAIR...

To be Findable:

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

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

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

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

To be 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

To be 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

To be 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

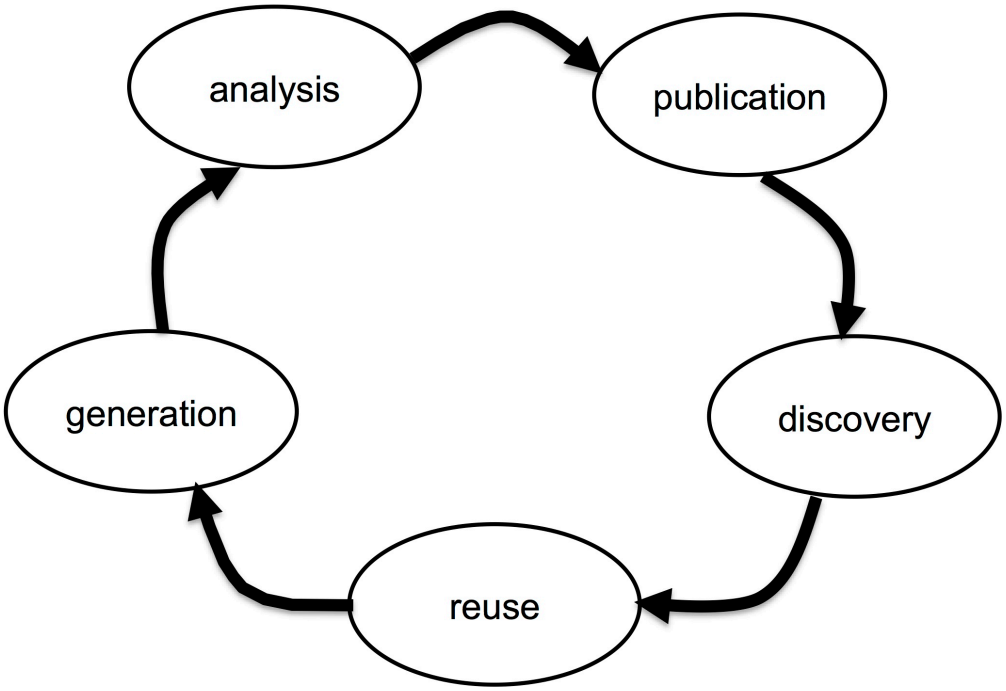
The CyVerse Data Commons is working to ensure that each of these principles is met for datasets published through our CyVerse Curated Data program (datasets with DOIs), and is working with community members to make Community Released Data as FAIR as possible.

FAIR data throughout the data life cycle

Most scientists are used to generating and analyzing data, but more and more, scientists are publishing their data or discovering and reusing published data. Although the FAIR principles provide guidance on what features published data should have to be findable, accessible, interoperable, and reusable, the best way to make data FAIR is to plan for it at every stage of the data life cycle.

CyVerse has features that support data management throughout the life cycle. This section provides links to some of those features.

The Data Life Cycle:



Data Generation

Most CyVerse users generate their data externally and bring it into CyVerse for analysis. For information on how to bring data into CyVerse, see the wiki page on Downloading and Uploading Data.

Through CyVerse analysis tools, you may generate new data on the CyVerse Data Store.

For analyses run in the Discovery Environment, the data is stored in the Analyses folder in your home directory (unless you specify a different output directory). You can use the Analyses Window to view the parameters associated with the any output data created in the Discovery Environment (see Using the Analyses Window).

Data Analysis

CyVerse offers several platforms for analyzing data, each with features that support FAIR data.

For beginners, have a look at the Discovery Environment Manual or the Atmosphere Manual.

Discovery Environment (DE) features for reproducible science include:

- The DE stores metadata on every analysis run, including input and output files, time run, who ran it, and all parameters. Any analysis can be relaunched using the same parameters.

Data Publication

See the page on Publishing Data through the Data Commons.

Data Discovery

Search:

All data on the CyVerse Data Store are indexed using ElasticSearch. There is an advanced search interface in the Discovery Environment, which will return results for all data you have permission to see (your own data, data shared with you, and public data). There is a simple search interface in the Data Commons, which will return results from all public data in the /iplant/home/shared directory, that is, all data in the Data

Commons.

Metadata:

The best way to make your data more discoverable is to use metadata. If you publish data with a DOI through the Data Commons, you are required to add the DataCite metadata template, but you can also add custom metadata. If you are the owner of a Community Released Data folder, you are required to add the Dublin Core metadata template to the parent folder, but you should also add it to any relevant sub-folders.

Data Reuse

External Resources:

If you must use Excel spread sheets, read this first: https://www.tandfonline.com/doi/full/10.1080/00031305.2017.1375989

https://dataoneorg.github.io/Education/

https://www.dataone.org/best-practices

https://www.force11.org/group/joint-declaration-data-citation-principles-final

https://www.nsf.gov/pubs/2018/nsf18041/nsf18041.jsp

http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005510

https://www.biorxiv.org/content/early/2018/09/16/418376

http://terraref.org/articles/existing-data-standards-and-tools/