Overview of the FAIR Data Principles

Amelia Harrison

Presented: April 18, 2022

Overview

- What does FAIR stand for?
- The goal of FAIR data
- Breakdown of the FAIR acronym
- The responsibilities of individual researchers
- Strategies for making your data FAIR

Findable

Accessible

Interoperable

Reusable

What is the goal of FAIR data?

- FAIR was introduced in a 2016 paper to improve the Findability, Accessibility, Interoperability, and Reusability of digital data (Wilkinson et al. 2016)
- Broadly, the goal is to make data more freely available and usable
- More narrowly, there is great benefit for the individual researcher
    - If your data is easy to use, more people will use it
    - Which means more citations

What is metadata?

- Data that describes your data
  - Any data you collect along the way that is not the “main” data
- Examples:
  - Sampling locations
  - Assay type and conditions
  - Bacterial strain
  - Sequencing platform and statistics

Breaking down the FAIR acronym

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 they describe
- 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, when 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 presentation
- 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

What are the responsibilities of individuals in making data FAIR?

Findable

Individuals are responsible for…

- Submitting data to the appropriate database/repository
  - E.g., crystal structures to PDB, sequencing data to the SRA
- Reporting accession numbers
- Ensuring that data and metadata are linked
- Ensuring that all metadata are included and sufficiently detailed
  - Includes metadata collected but not used or discussed in the study

Accessible

Individuals are responsible for...

- Making data available in the free and open resource
- Or, for storing sensitive/private data with the proper protocols
- Keeping metadata, even if primary data can no longer be stored
    - Metadata is often valuable on its own
    - Usually smaller and cheaper to store
- Making custom scripts available in a public repository like GitHub
- Provide submitted manuscripts as pre-prints

Interoperable

Individuals are responsible for…

- Storing (meta)data in standard formats
  - If using proprietary software, export your data in a standard format that can be opened by other programs
- Using ontologies/controlled vocabularies to tag your data
  - Ontologies have structure and standardized descriptions, so all researchers and computers know what is in your (meta)data
  - This also makes your data more **F**indable
- Using standardized language to describe your data when writing/presenting
  - Jargon is a large barrier to entry for people from even slightly different fields

Reusable

Individuals are responsible for...

- Ensuring that other researchers know how your data can be used
  - Release your data with a usage license
- It should be clear how each piece of (meta)data was produced (detailed provenance)
  - Another researcher should be able to tell where each piece of data originated
- Making (meta)data available in a common, expected format
  - The more similar datasets are, the better

Potential obstacles to reproducibility in FAIR data

- Some computational work is *theoretically* reproducible, but has such high barriers that it is not *practically* reproducible
  - “Petabase-scale sequence alignment catalyzes viral discovery”
    - Spent ~$24,000 USD aligning ~5.7 million sequencing runs
- Proprietary software
  - Not every lab has access to the same software
  - Not every person uses the same OS

Questions