CYVERSE®

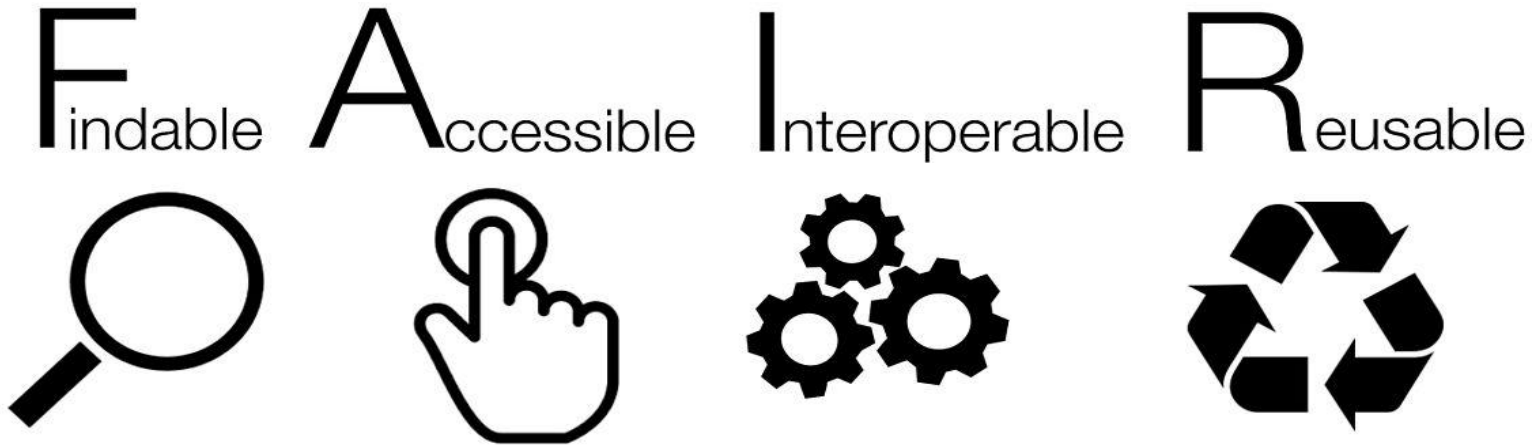
Transforming Science Through Data-Driven Discovery

CyVerse Features for FAIR Data

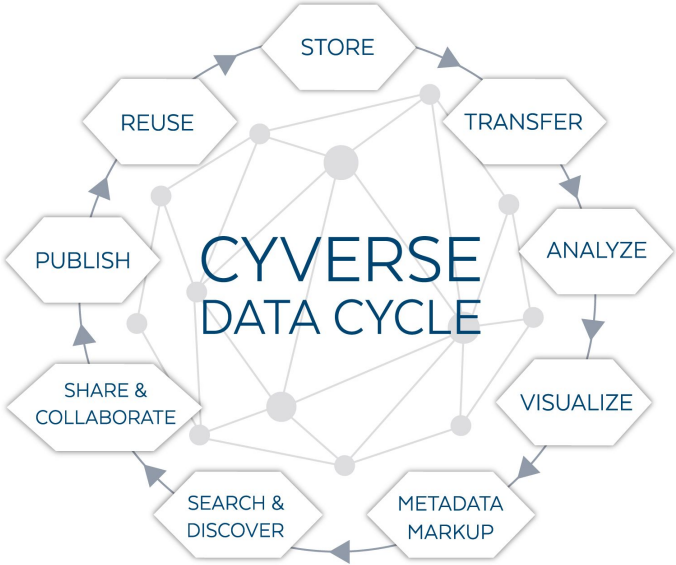
Amanda Cooksey

November 5, 2021

CyVerse is supported by the National Science Foundation under Grant Nos. DBI-0735191, DBI-1265383, and DBI-1743442



FAIR Basics



- FAIR is a set of guidelines
- many ways to be FAIR
  - good, better, best
  - worst thing to do is nothing at all
- CyVerse provides **opportunities** to make your data more FAIR
- plan for ‘FAIRness’ at every stage of the data life cycle
- metadata, metadata, metadata!!!

‘F’ is for 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

https://www.nature.com/articles/sdata201618

‘A’ is for 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

https://www.nature.com/articles/sdata201618

‘I’ is for 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

https://www.nature.com/articles/sdata201618

‘R’ is for 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 standard

https://www.nature.com/articles/sdata201618

Licensing

- Open data licenses
- Commonly used open licenses CC0 and ODC-PDDL
    - don’t require attribution or share-alike
    - all rights waived
- Some licenses require attribution or share-alike
- Institutions and funding agencies may have licensing requirements

- **Elasticsearch** in Discovery Environment **(F)**

- **CyVerse metadata**
  - AVU format for metadata **(I)**
  - templates use community standard schema and standardized vocabularies **(I,R)**
  - metadata can include refs to other identifiers **(I,R)**
    - BioSample ID
    - DOI of paper about the dataset
    - DOI of a related dataset
    - provenance information
  - analyses have associated metadata **(R)**

- **CyVerse curated data (DOI)**
  - globally unique identifier, included in metadata **(F)**
  - uses DataCite metadata schema **(F,I,R)**
  - use CC0 or ODC-PDDL licenses **(R)**
  - data and metadata accessible **(A)**
    - CyVerse Data Commons via HTTPS
    - iCommands or CyberDuck for large datasets
    - metadata also available at DataCite Commons

- **SRA, WGS submission pipelines**
  - NCBI standard metadata schema **(F,A,I,R)**
  - globally unique ID, included in metadata **(F)**

Resources

- FAIR guidelines
- Assess the FAIRness of your data
- About licenses
- SRA submission pipeline
- WGS submission pipeline
- DOI request
- CyVerse Portal

CYVERSE®

Transforming Science Through Data-Driven Discovery

Executive Team

The University of Arizona	TACC	Cold Spring Harbor Laboratory
Parker Antin	Matt Vaughn	Doreen Ware
Nirav Merchant		Dave Micklos
Eric Lyons		

CyVerse is supported by the National Science Foundation under Grant Nos. DBI-0735191, DBI-1265383, and DBI-1743442