CYVERSE™

Transforming Science Through Data-driven Discovery

Using cyberinfrastructure to make life sciences data FAIR: Lessons learned

Ramona Walls

rwalls@cyverse.org @Ramona_Walls

25 October, 2017

Sixth Annual Workshop of the Clinical and Translational Science Ontology Group

Ann Arbor, MI (presented remotely)

NSF

The University of Arizona

TACC Texas Advanced Computing Center

CSH Cold Spring Harbor Laboratory

UNCW

CYVERSE

Cyber Universe

Vision: Transforming science through data driven discovery

Mission: Design, deploy and expand national cyberinfrastructure for life sciences research, and to train scientists in its use.

Usage: ~50K users, 3PB of data, 100s of publications, workshops, courses, and discoveries

NSF

Next Generation Data Management

- F.A.I.R. data principles
- Leverage semantics
- Let the data do the work
- Decentralize management

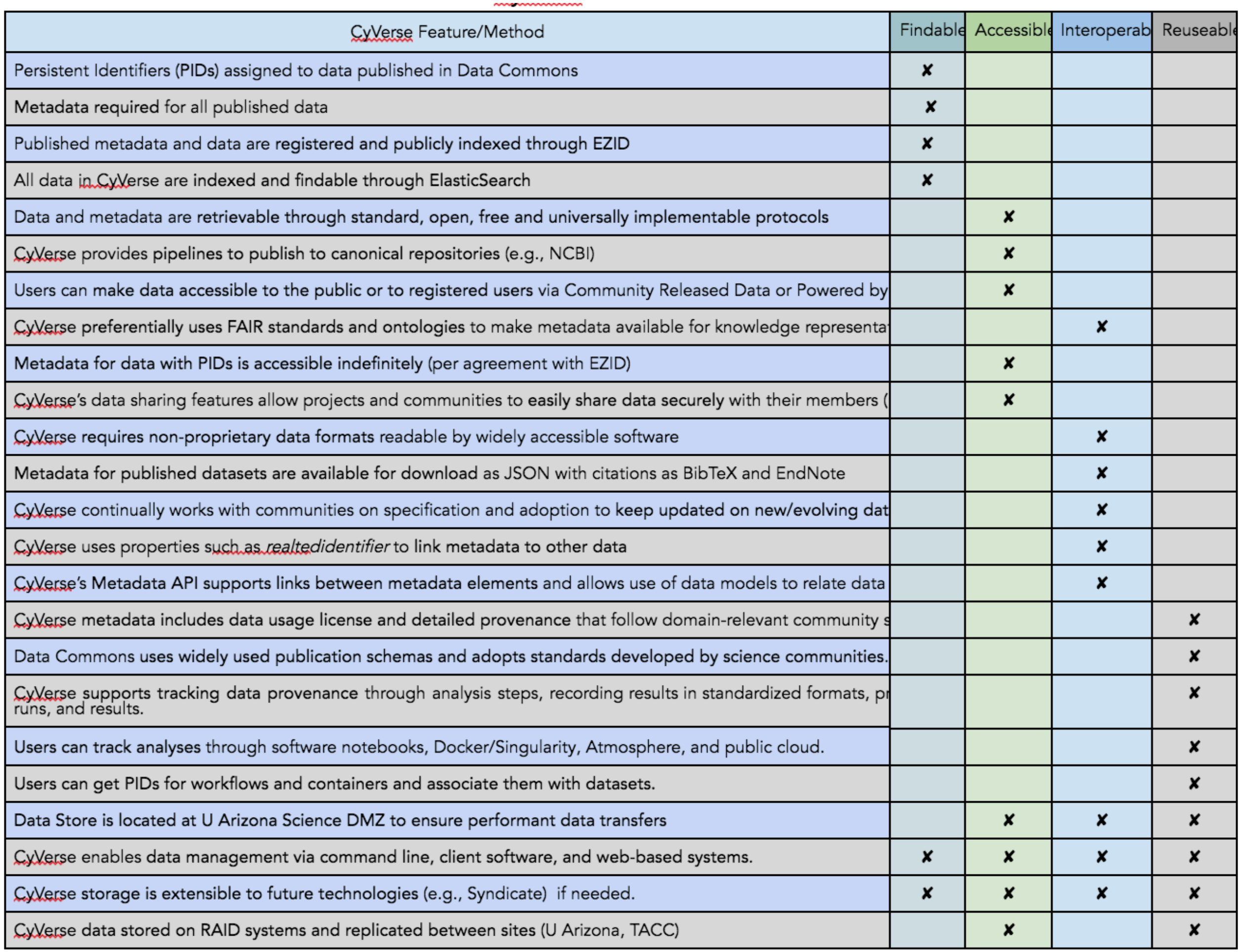
CyVerse Data Commons

A space where data can live as a searchable, discoverable, and reusable resource

- Data management
- Project management
- Data publication
- Permanent Identifiers
- Metadata and Ontologies
- Community Standards
- Data discovery

http://datacommons.cyverse.org/

How CyVerse facilitates FAIR data



CyVerse Feature/Method	Findable	Accessibl	Interoperab	Reuseable
Persistent Identifiers (PIDs) assigned to data published in Data Commons	✘			
Metadata required for all published data	✘			
Published metadata and data are registered and publicly indexed through EZID	✘			
All data in CyVerse are indexed and findable through ElasticSearch	✘			
Data and metadata are retrievable through standard, open, free and universally implementable protocols		✘		
CyVerse provides pipelines to publish to canonical repositories (e.g., NCBI)		✘		
Users can make data accessible to the public or to registered users via Community Released Data or Powered by		✘		
CyVerse preferentially uses FAIR standards and ontologies to make metadata available for knowledge representa			✘	
Metadata for data with PIDs is accessible indefinitely (per agreement with EZID)		✘		
CyVerse's data sharing features allow projects and communities to easily share data securely with their members (		✘		
CyVerse requires non-proprietary data formats readable by widely accessible software			✘	
Metadata for published datasets are available for download as JSON with citations as BibTeX and EndNote			✘	
CyVerse continually works with communities on specification and adoption to keep updated on new/evolving dat			✘	
CyVerse uses properties such as *realtedidentifier* to link metadata to other data			✘	
CyVerse's Metadata API supports links between metadata elements and allows use of data models to relate data			✘	
CyVerse metadata includes data usage license and detailed provenance that follow domain-relevant community s				✘
Data Commons uses widely used publication schemas and adopts standards developed by science communities.				✘
CyVerse supports tracking data provenance through analysis steps, recording results in standardized formats, pr runs, and results.				✘
Users can track analyses through software notebooks, Docker/Singularity, Atmosphere, and public cloud.				✘
Users can get PIDs for workflows and containers and associate them with datasets.				✘
Data Store is located at U Arizona Science DMZ to ensure performant data transfers		✘	✘	✘
CyVerse enables data management via command line, client software, and web-based systems.	✘	✘	✘	✘
CyVerse storage is extensible to future technologies (e.g., Syndicate) if needed.	✘	✘	✘	✘
CyVerse data stored on RAID systems and replicated between sites (U Arizona, TACC)		✘		✘

Finable:

- Persistent Identifiers (PIDs) assigned to data published in Data Commons.
- Metadata required for all published data, encouraged for other data.
- Published metadata and data are registered and publicly indexed through EZID and DataCite.
- All data in CyVerse are indexed and findable through ElasticSearch.

Accessible:

- CyVerse data and metadata are retrievable through standard, open, free and universally implementable protocols.
- Pipelines to publish to canonical repositories (e.g., NCBI).
- Metadata for data with PIDs is accessible indefinitely.
- CyVerse's data sharing features allow individuals, teams, and communities to easily share data securely with their members pre-publication or share publically.

Interoperable

- Non-proprietary data formats readable by widely accessible software.
- Metadata for published datasets are available for download as JSON,
- Citations available as BibTeX and EndNote.
- Continually work with communities on specification and adoption to keep updated on new/evolving data types.
- Preferentially use FAIR standards and open source ontologies.
- Use properties such as *realtedidentifier* to link metadata to other data.
- CyVerse's Metadata API supports links between metadata elements and data models for relating data elements to one another.

Reusable

- Data usage license and detailed provenance that follow domain-relevant community standards.
- Widely used publication schemas and standards developed by science communities.
- **Supports tracking data provenance through analysis steps, recording results in standardized formats, providing access to scripts, runs, and results.**
- Users can get PIDs for workflows and containers and associate them with datasets.

Technology for FAIR data:

- Data Store is located at U Arizona Science DMZ to ensure performant data transfers.
- CyVerse enables data management via command line, client software, and web-based systems.
- CyVerse storage is extensible to future technologies (e.g., Syndicate) if needed.
- CyVerse data stored on RAID systems and replicated between sites (UA, TACC)

Open Science: Reproducible, scalable analyses support FAIR data

- Containers
- Virtual Machines
- Interactive analyses/notebooks
- Bring your own storage
- Bring your own compute

jupyter

RStudio®

XSEDE
Extreme Science and Engineering Discovery Environment

amazon webservices™

BIOCONDA®

HTCondor
High Throughput Computing

Open Science Grid

Jetstream

NSF

LESSONS
LEARNED

What we would do (are doing) differently

- Focus on asynchronous training, tier 3 support.
- Develop user-friendly UIs for (meta)data management.
- Automate metadata and publication workflows as much as possible

What we did well

- Connect existing, open source technologies to make them accessible to more scientists.
- Focus on urgent user needs as drivers of development.
- Work with community standards organizations.