CYVERSE™

Transforming Science Through Data-driven Discovery

Managing large collaborative datasets for easier storage, analysis, and discovery

Webinar for PIRE EHT

March 20, 2019

Ramona Walls – Senior Scientific Analyst

University of Arizona

rwalls@cyverse.org [@Ramona Walls]

THE UNIVERSITY OF ARIZONA

TACC TEXAS ADVANCED COMPUTING CENTER

CSH Cold Spring Harbor Laboratory

NSF

Topics

- Brief introduction to CyVerse
- What are FAIR data and why should you make your data FAIR?
- Some CyVerse tools for creatingFAIR data:
  - Upload, download, and organize data on CyVerse
  - Share data with collaborators
  - Add and edit CyVerse metadata -- three ways
  - Using and creating metadata templates
- By the end of the webinar, you should be able to:
  - Upload a file to CyVerse
  - Share a file with a colleague
  - Add metadata to a file

Overview

Vision:

Transforming science through data-driven discovery

Mission:

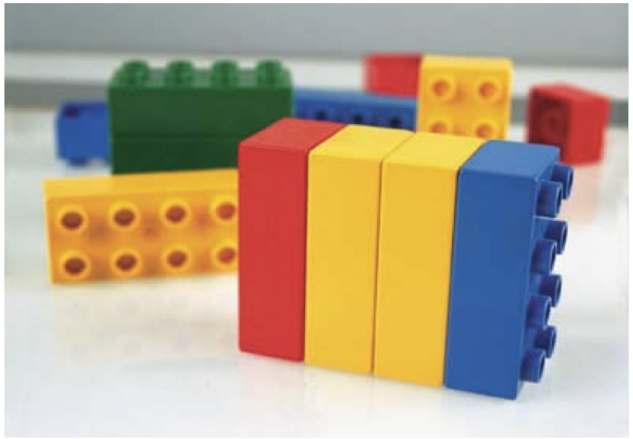
Design, develop, deploy, and expand a national cyberinfrastructure for life science research, and train scientists

Usage:

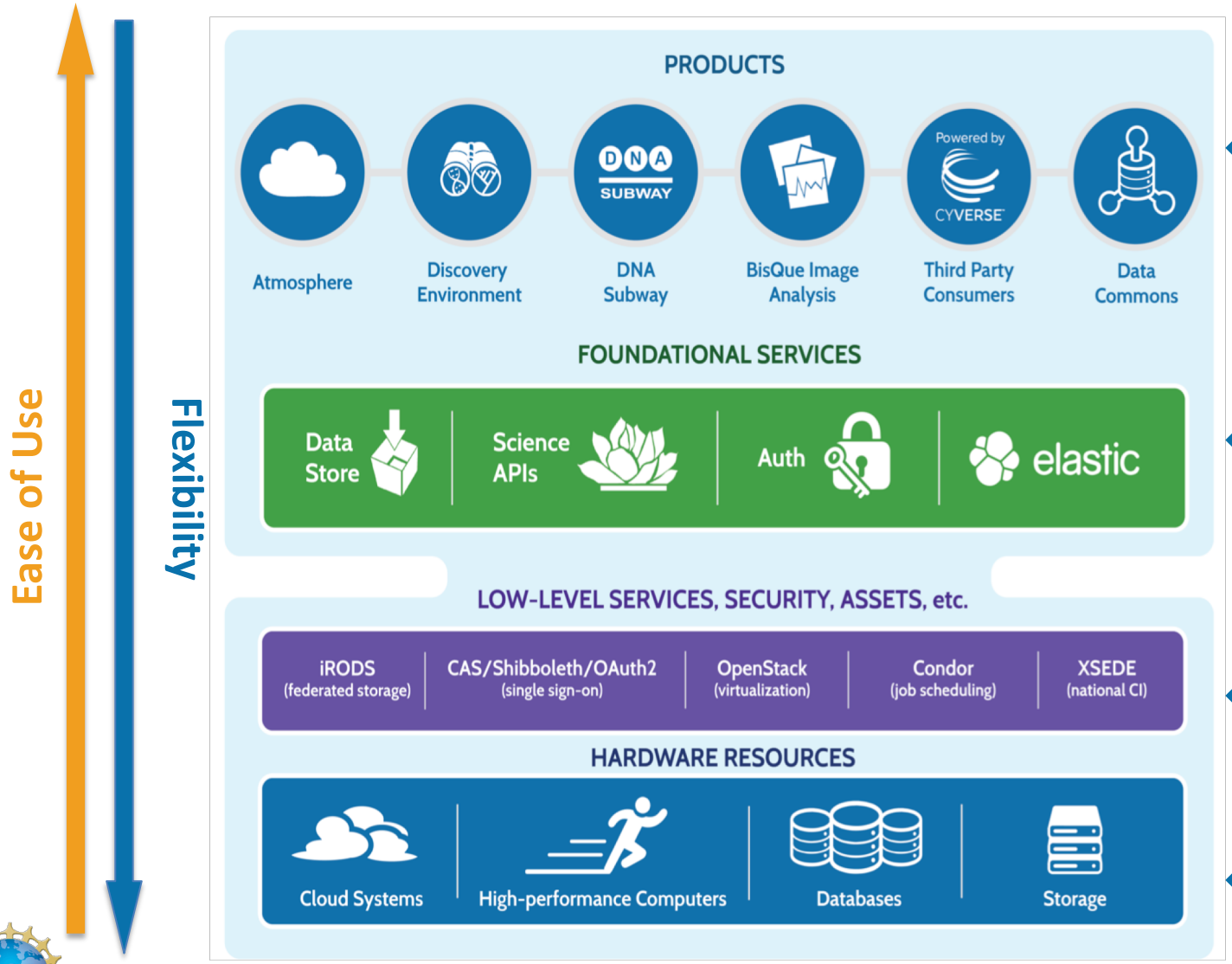
More than 50K users, PB of data, and hundreds of publications, courses, and discoveries

Lego blocks of cyberinfrastructure

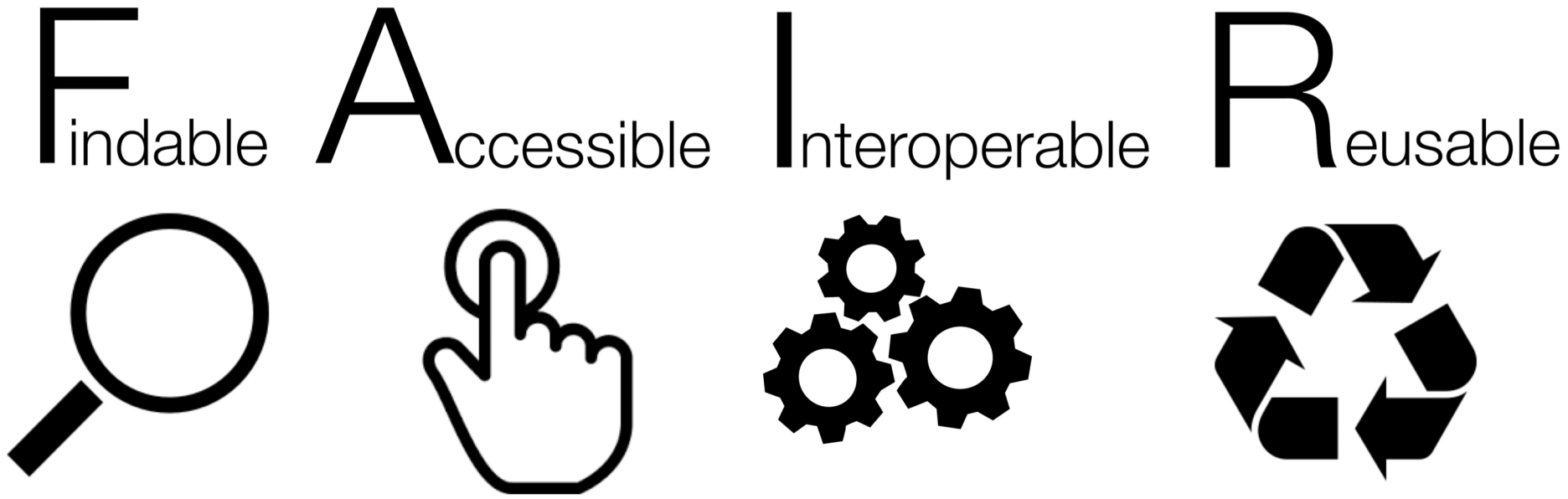
- Danish 'leg godt' - '**play well**'
- Also translates as '**I put together**' in Latin
- If a solution is not available you can craft your own using CyVerse components



CyVerse product stack



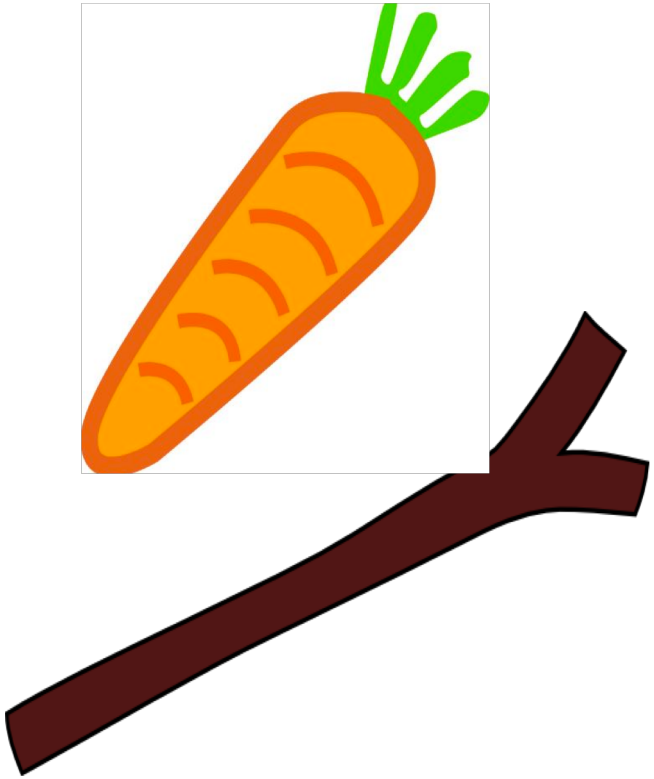
What are FAIR data?



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

Why make your data FAIR?

- Make your life easier
- Improve your reputation
- Meet funder requirements
- Practice reproducible science
- Support the common good



http://worldartsme.com/

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:

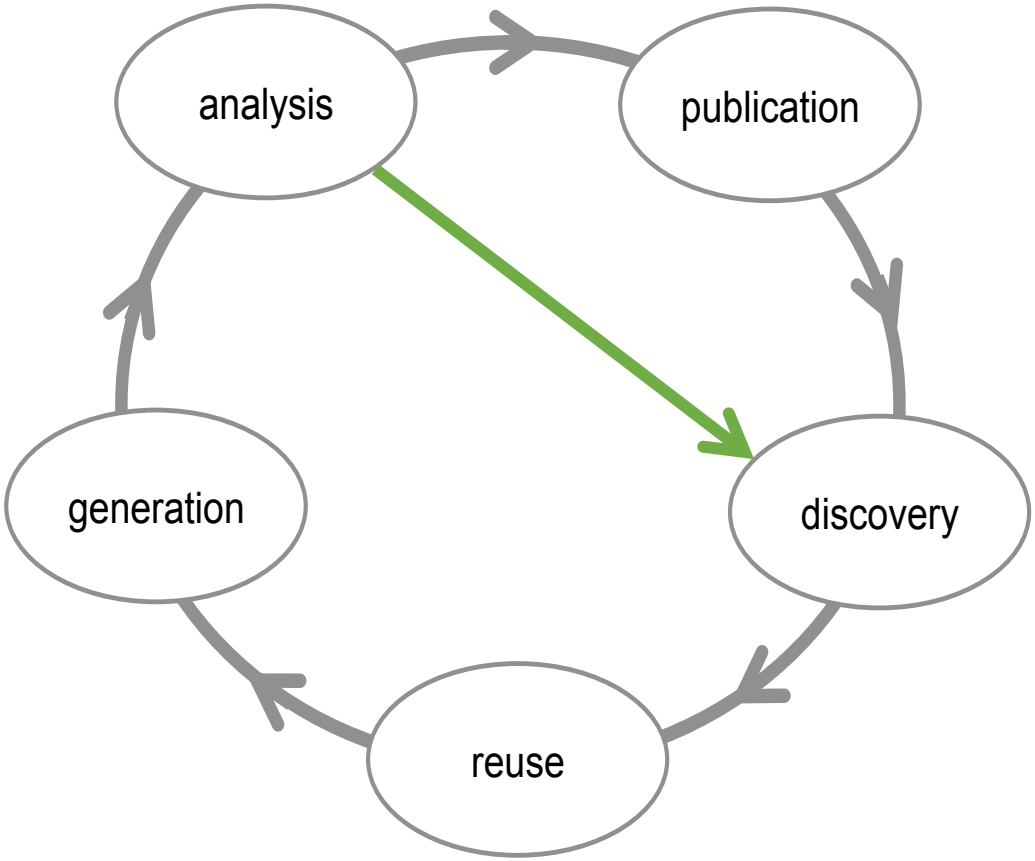
- I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- I2. (meta)data use vocabularies that follow FAIR principles
- I3. (meta)data include qualified references to other (meta)data

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

What can you do to make your data FAIR?

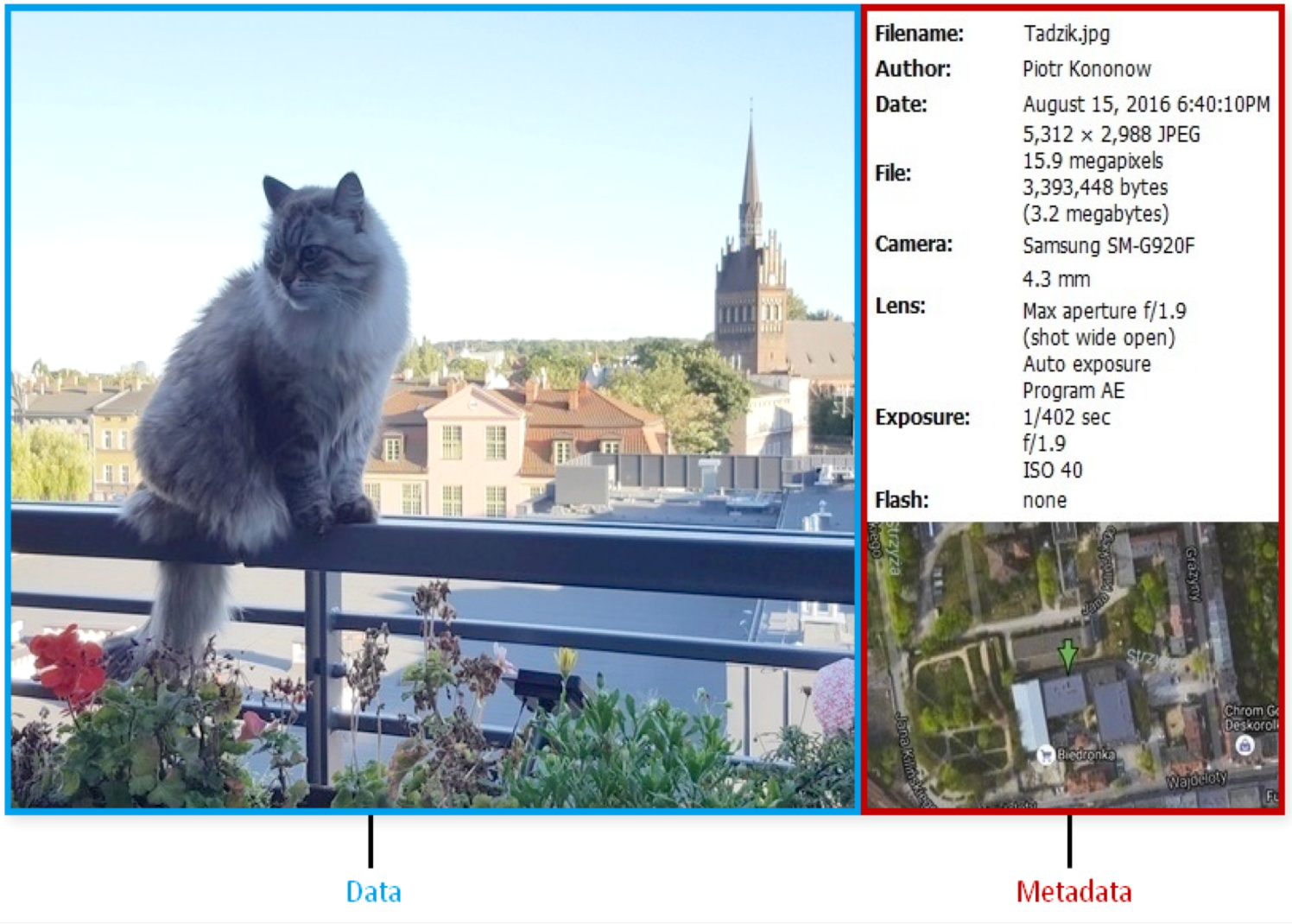
Understand the data life cycle



Learn about data management

- Ten simple rules for creating a good data management plan:
  - https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004525
  - Use https://dmptool.org/
- Use spreadsheets appropriately:
  - https://www.tandfonline.com/doi/full/10.1080/00031305.2017.1375989
- Check out Data Carpentry:
  - https://datacarpentry.org/
- and DataONE:
  - https://www.dataone.org/best-practices
- Best practices for scientific computing:
  - https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1001745

Use metadata



Follow best practices for data organization

- Folder and file naming conventions are useful, but restrictive
  - Metadata can be used to subset data as needed
- Avoid unnecessarily deep hierarchies
- Write down rules for organizing and naming data
- Have a strategy for version control
- Avoid special characters and spaces
- See: https://guides.library.upenn.edu/datamgmt/fileorg

Practice reproducible science

- Record metadata at the same time you collect data (not when you are ready to publish)
- Keep track of workflows and data provenance
- Create a data dictionary if needed
- Use non-proprietary data formats
- Publish your data in an open repository

PIRE-EHT data example

NOTICE: The output files contained in this Library directory are the fruit of countless hours of labor, mostly by young scientists. They are provided or the use of the EHT collaboration for EHT-related investigations. Use for any other purpose is not appropriate and not approved by the contributors. If you have any questions about use of the Library please contact one of the Theory WG coordinators.

Cyverse library structure

eht /

M87SimulationLibrary /

GRMHD or GRRMHD /

SANE or MAD / (or other descriptor of dimensionless flux, e.g. semiMAD)

README concerning spin notation - two sig figs, please.

a+0.5 a+0.75 a+0.88 a+0.94 a+0.97 a0

a-0.5 a-0.75 a-0.88 a-0.94 a-0.97 /

At the final level there are many possible naming conventions.

Illinois plan is to write n1xn2xn3_OTHERINFO

where OTHERINFO might include: coordinate system, e.g. mks, or interpolation scheme, e.g. weno

Each "leaf" directory for a simulation should contain, at least,

the dumps, and if available 230 GHz images at an inclination of

approximately 17 deg, and spectra at the same inclination.

So the leaf should contain the directories:

dumps/

images/

spectra/

http://datacommons.cyverse.org/browse/iplant/home/shared/eht/BkupSimulationLibrary/catalog.txt

http://datacommons.cyverse.org/browse/iplant/home/shared/eht/BkupSimulationLibrary/LibraryStructure.txt

/BkupSimulationLibrary
//MAD
///a+0.25
////192x96x96_IHARM
/////dumps
//////dumps_000000000.h5
…
//SANE
///a+0.5
////192x192x192_IHARM
/////dumps
//////dumps_000000000.h5
…

CyVerse Basics

- Create an account at user.cyverse.org
- Upload, download, and organize data on CyVerse
- Share data with collaborators
- Add and edit CyVerse metadata -- three ways
- Using and creating metadata templates

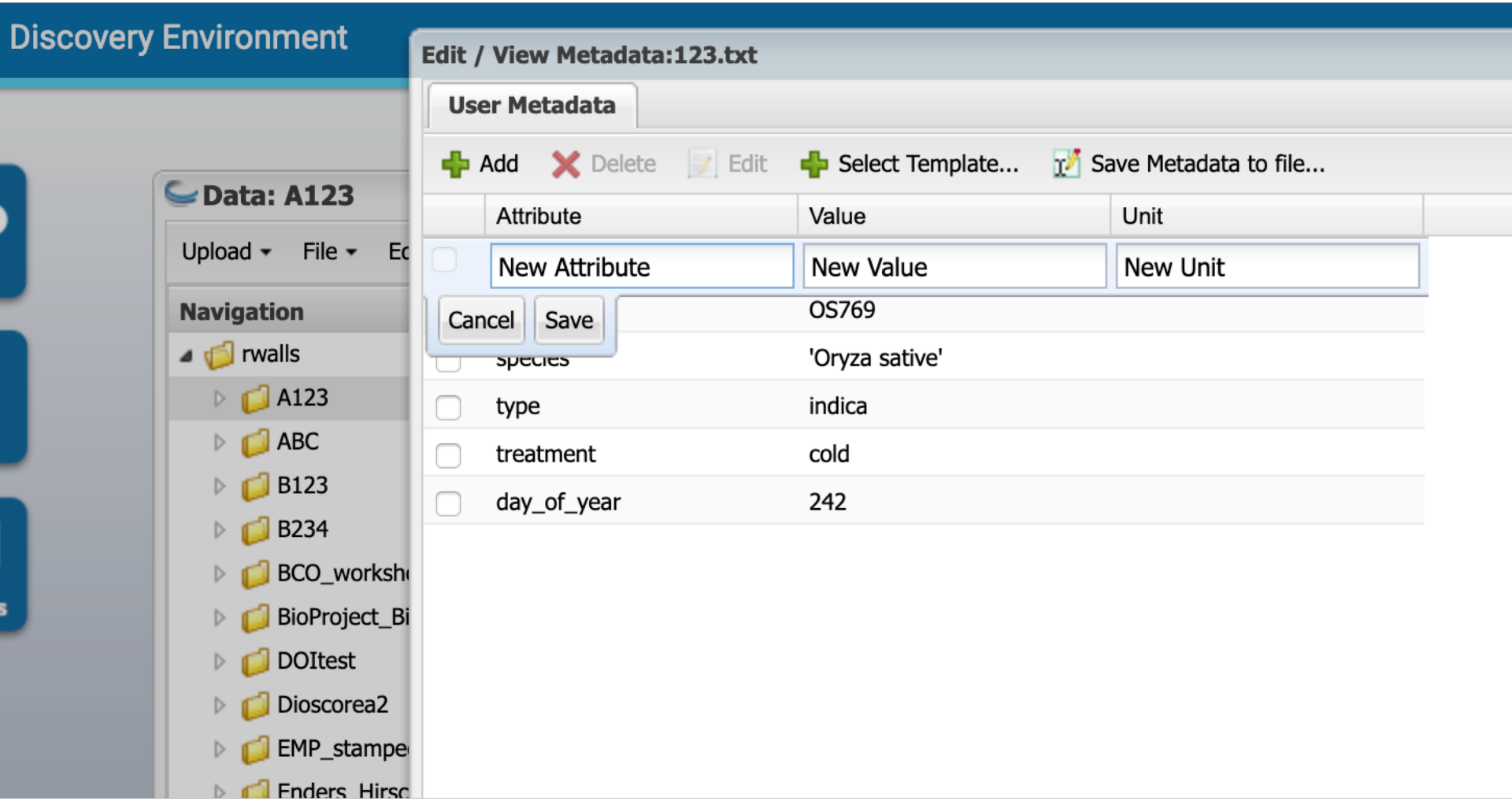
Upload, download, and organize data on CyVerse

- Data Store Guide:
  - https://cyverse-data-store-guide.readthedocs-hosted.com/en/latest/
- Discovery Environment:
  - https://de.cyverse.org/
  - https://learning.cyverse.org/projects/discovery-environment-guide/en/latest/
- Demo: upload and download via DE and iCommands
- Activity: create a folder in the DE
- Activity: upload a file via the DE and share it with a collaborator

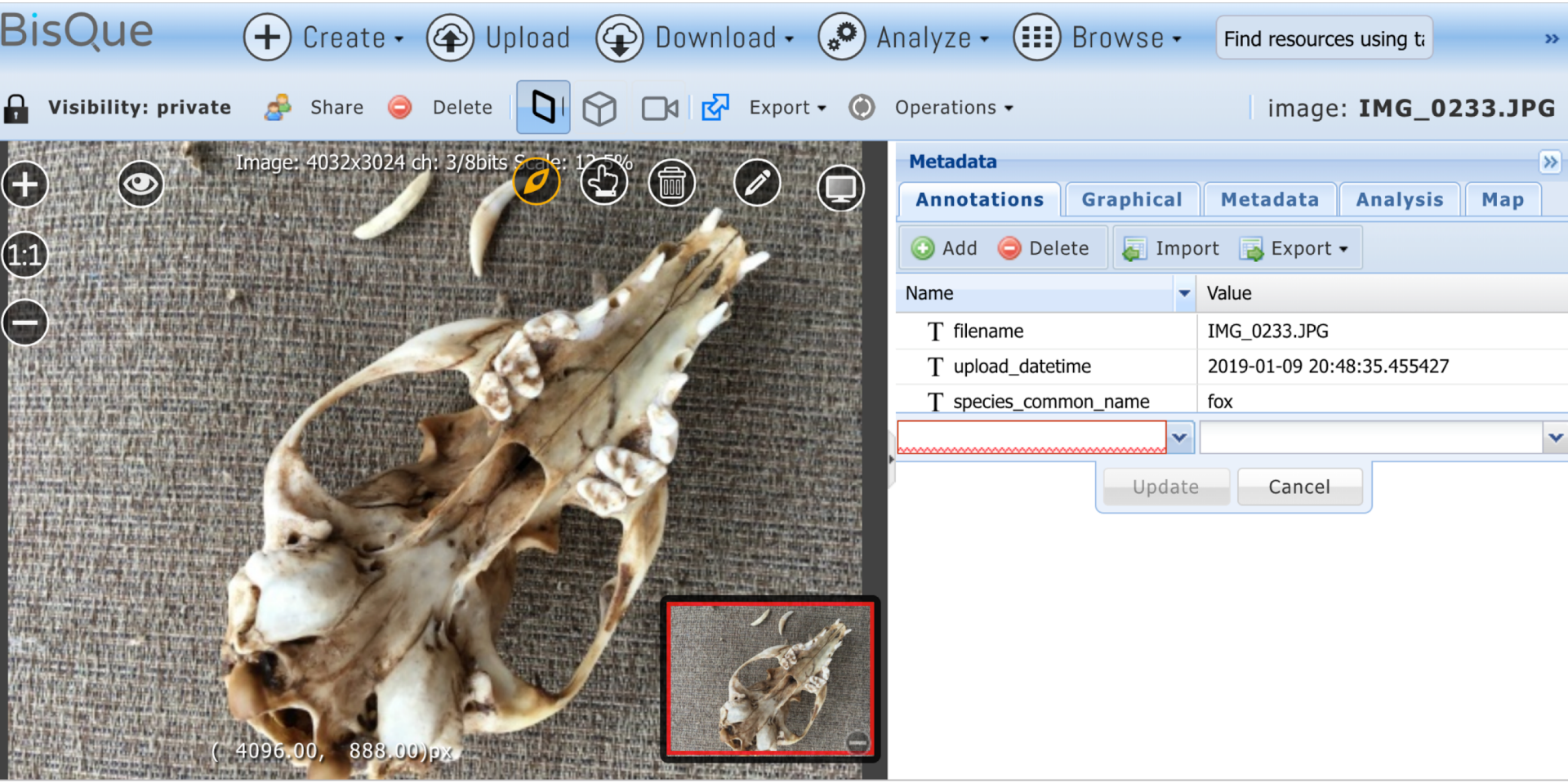
Add and edit CyVerse metadata -- three ways

- CyVerse metadata systems:
  - DE, iRODS, Bisque

DE metadata


Discovery Environment
Edit / View Metadata:123.txt
User Metadata
Add
Delete
Edit
Select Template...
Save Metadata to file...
Attribute
Value
Unit
New Attribute
New Value
New Unit
Cancel
Save
OS769
'Oryza sative'
type
indica
treatment
cold
day_of_year
242
Data: A123
Upload
File
Navigation
rwalls
A123
ABC
B123
B234
BioProject_Bi
DOItest
Dioscorea2
EMP_stampe

Bisque metadata

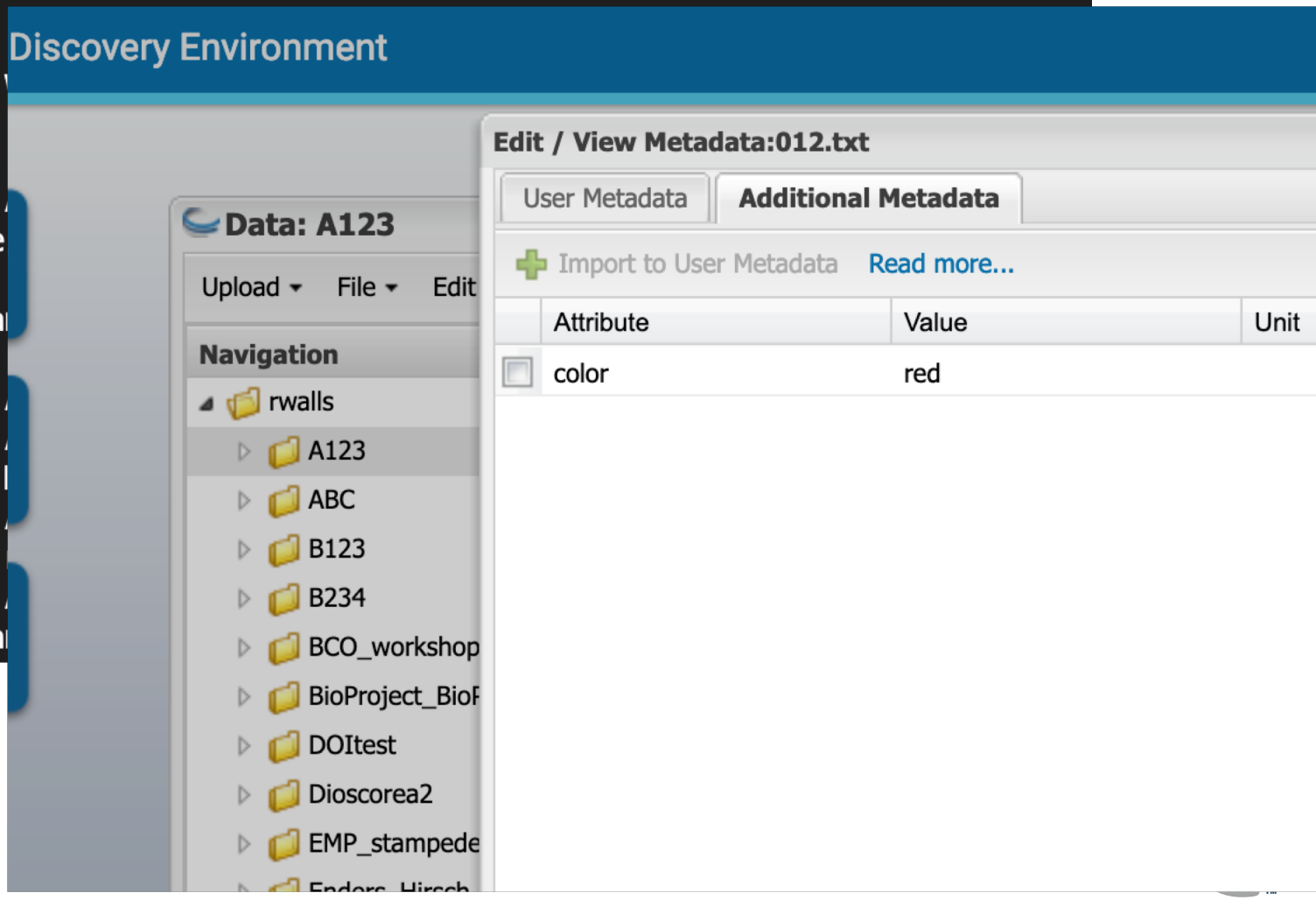


iRODS metadata

```
[rwalls-cyverse:~ rwalls-iplant$ imeta -h
Usage: imeta [-vVhz] [command]
 -v verbose
 -V Very verbose
 -z Zonename  work
 -h This help
Commands are:
 add -d|C|R|u Name
 adda -d|C|R|u Name

 addw -d Name AttNa

 rm  -d|C|R|u Name
 rmw -d|C|R|u Name
 rmi -d|C|R|u Name
 mod -d|C|R|u Name
      (modify AVU;
 set -d|C|R|u Name
 ls  -[l]d|C|R|u Na
```



Add and edit CyVerse metadata -- three ways

- CyVerse metadata systems:
  - DE, iRODS, Bisque
- Demo:
  - Add iRODS metadata with iCommands
- Activity: Add metadata in the DE
- Activity: Search based on metadata

Using and creating metadata templates

- Templates as views on metadata
  - Demo: Apply the DataCite and Dublin Core templates
- Creating a metadata template
  - Demo: Creating templates in the administrator interface

Links

- Please complete the survey!
  - Webinar evaluation survey
- FAIR webinar:
  - https://wiki.cyverse.org/wiki/display/Events/FFW%3A+Making+your+data+FAIR+with+CyVerse
- CyVerse Learning Center:
  - https://learning.cyverse.org/en/latest/