Data Commons Projects

The Data Commons (DC) is growing and evolving, with components spread throughout the CyVerse CI. Our development approach is to build the pieces that will provide the most benefit to our users as quickly as possible — and remain flexible and responsive to future needs. If you have suggestions or feedback on our projects, please log in to the CyVerse wiki and add your comments below.

This page is updated approximately once per quarter.

Data Commons Website - datacommons.cyverse.org and dc.cyverse.org

First available: Q3 2016

The Data Commons (DC) website is the main access point for data in the DC. The DC contains the subsets of CyVerse public data that have been shared with the public by community members (*Community Released)* and the data that have has been assigned permanent identifiers (PIDs) and are permanent and stable (*CyVerse Curated Data*). All of these data are considered "anonymous" data, because you do not need a CyVerse username and password to access them. DC data are stored in the Data Store under /iplant/home/shared/commons_repo/curated. All datasets in the Data Commons have minimal required metadata, based on the DataCite and Dublin Core schemas, as well as a ReadMe file that explains the contents of the dataset. The Data Commons also house CyVerse tutorials and example data.

You can follow development of our public data page on the GitHub repo and the Data Commons JIRA project (must be logged in with your CyVerse account to view JIRA).

New features for 08/08/2017 release: 2.2

New landing pages for CyVerse Curated Datasets

- Improved display of metadata
- Citation display and download
- Display data license
- Links to analysis platform/tool now available as metadata
- See also https://pods.iplantcollaborative.org/jira/browse/DC-100

New features for 01/09/2017 release: 2.1

New home page for data commons at datacommons.cyverse.org and dc.cyverse.org

- Links out to Community Released and CyVerse Curated data
- Links out to wiki pages on how to publish data through the DC (SRA, CR, CC)

New features for 09/16/2016 release: 2.0

- CyVerse branding
- App rewritten with Django + AngularJS
- Refactored to use DE Terrain API for file metadata instead of using `python-irods` to query iRODS directly
- Direct downloads (when supported) are offloaded to the CyVerse anon-files service. This frees up the DC server to continue handling web requests and not be responsible for managing download streams from iRODS. Direct download is supported for files up to 2GB.
- Improved file preview support. All text-like files can be previewed up to 8kB. For larger files, the first 8kB is shown. Currently supported previews include files that Terrain reports content-type `text/*`, e.g., `text/csv`, `text/plain`, etc.
- Prototype landing pages that display combined template metadata and iRODS AVUs
    - Metadata displayed for any object that has metadata
- Link to data in the DE

Features under development

- Landing pages for Community Released datasets (Oct. 2017 release)
- File display based on info or mime type
- Improved QA tests

Planned features

- Search from DC portal - waiting for completion of CyVerse search refactoring.
- Different renderers for different types of metadata views

- Community branding

Permanent Identifiers

First available: Q3 2015

CyVerse decided to start issuing PIDs when we realized that there were no canonical repositories for many of the data types being generated through our CI. CyVerse users wanted a way to share their published data where it would be easily accessible for new analyses within CyVerse, so what better place than our own Data Store. We have a contract with the California Digital Library to issue DOIs and ARKs through their EZID s ystem. See Requesting a Permanent Identifier in the Data Commons for more information.

Existing features

- DataCite metadata template for DOIs.
- Dublic Core metadata template for Community Released data. New requests are required to use this. Will work with users to add it to exi sting folder after Data Store reorganization.
- Pipeline for requesting DOIs through the DE. Using this pipeline, scientists organize their dataset into a single folder, apply the appropriate metadata and add a ReadMe file, and then request the DOI. At that stage, the folder is moved automatically to a staging folder and the PID curators can review the request using our administrative interface (Belphegor). The process is described in DOI Creation SOP for Curators.
- Ability to have multiples of the same metadata attribute (e.g., multiple contributors) as part of a template

Planned features

- **EZID is transferring their DOI service to DataCite, so we will need to adjust development to fit the new service.**
- Ability to link metadata fields (e.g., link contributorType to Contributor) - in API, needs UI
- Pipeline for generating ARKs using API
- Ability for curators to update metadata using the EZID API via Belphegor
- Science-specific metadata templates for images and sequence data
- Preparation of PID datasets through the Projects Interface

Projects Interface

First available: TBD

The Projects Interface (PI) is intended to support management of the data, people, tools, and analyses that are part of a scientific research project. Although the PI will be housed within the DE, it should managing projects that use any component of CyVerse CI, through the use of shared APIs. Atmosphere currently has its own interface for creating and managing projects, which will remain. However, by integrating APIs from the DE, users will be able access information about Atmosphere projects from with in Data Commons PI.

A key technical requirement for the PI was the ability to create groups, which now exists.

Existing features

- Integration of Grouper service into DE/Data Store (Q3 2017)
- First release of Teams - Users can create public and private teams. Allows sharing of apps and tools with teams (Oct. 2017)

Planned features

- Project creation (timeline uknown)
    - Step 1:
        - Ability to create a project
        - Ability to add metadata to a project
        - Ability to manage people, data, apps, tools, analyses, and allocations within a project
    - Step 2:
        - Ability to create datasets within a project
        - Ability to publish datasets from within a project, to DCR or external repositories

Ontology-based Metadata Management (OMM)

First available: Q2 2016

A big hurdle to using adequate metadata is that scientists often have to supply different sets of metadata for different uses of the same file such as getting a DOI, submitting to NCBI, or publishing to Dryad. Often the different fields contain the same information, but with slightly different labels, meaning that researchers have to enter the same data multiple times, fuss with formatting, and often end up getting frustrated and giving up. Another hurdle for getting scientist to use metadata is the large amount of effort required for limited value in return.

The DC hopes to help overcome these problems through the use of cutting-edge metadata technologies that will make using metadata easier for scientists and more make the metadata itself more useful. We group these developments together under the heading of "Ontology-based Metadata Management or OMM". All OMM meetings start and end with a group chant, which helps to ground us and focus our energy.

Metadata views (Q4 2016 release)

The way metadata is handled in the DE has been completely revamped. The DE metadata database (DB) is now the primary source for DE metadata, in order to allow functionality beyond what is possible with iRODS. iRODS metadata can still be used via iCommands.

Templates still exist from a UI point of view, but on the back end they are treated as views on the metadata attached to an object, allowing users to only view the attributes associated with a template.

New features:

- Ability to add multiples of the same attribute

Still to do:

- Ability to link one AVU to another
- Ability to specify a linking relation between two AVUs

Ontology annotation (Oct. 2017 release)

First set of features released:

- Ability to associate a value with an ontology term IRI using term completiong.
    - Points to OLS or UAT SPRQL endpoints
    - Only available through templates.
    - Displays term label

DataOne member node (late 2018)

- Services under development to make CyVerse Curated data available and discoverable through DataONE

Data Store folder reorganization (Winter 2017/2018)

iRODS zone renaming and moving of root folders

- On hold until after site visit.
- See JIRA epic DS-114

Re-organization of Community Data

CyVerse Data Policies

- A new Data Commons User Agreement and Data Policy were released 04/2017.
- Data policy was updated 08/2017 at the request of the Science team. Users can now request up to 10 TB without executive team review.

Submission to external repositories

Submission to NCBI SRA repository

First available: Q2 2015

CyVerse users can now submit data to SRA through the DE. The process is documented on the NCBI Sequence Read Archive (SRA) Submission (Workflow Tutorial) wiki page.

Submission to NCBI WGS repository

First available: Q2 2017, in beta testing

Whole Genome Shotgun (WGS) projects are genome assemblies of incomplete genomes or incomplete chromosomes of prokaryotes or eukaryotes that are generally being sequenced by a whole genome shotgun strategy. WGS projects may be annotated, but annotation is not required. We are adapting the CyVerse SRA submission pipeline (see below) to submit sequences to WGS.

- Similar to SRA submission features. Users must create a BioProject and BioSample and apply appropriate metadata before submission

Submission to NCBI TSA (Transcriptome Assembly)

Status: Planning

Future projects

Improved data discovery and reporting

- using ELK stack, Google Analytics
- Urgent need is to display data access stats on user portal

Data model

The DE and Agave already capture the inputs and outputs of analyses. We plan to use the Provenance Ontology (PROV-O) or some variation of it for organization and visualization of complex datasets.

Support for live streaming data

From sensors.

Metadata Thesaurus

The metadata thesaurus currently exists as spreadsheet, with mappings among MIXS, NCBI, and DataCite. Dublin Core and others are planned for the near future.

A test server for TemaTres was set up during the plant phenotyping hackathon in Feb. 2017.