LabCAS User Guide

LabCAS ("Laboratory Catalog and Archive Services") is a web-enabled environment that allows users to publish, share, search and download a wide variety of biomedical datasets. This page contains a collection of documents targeted to end-users of the LabCAS portal and services.

Data Publishing

At present, data can be published into LabCAS in two possible ways: by using the LabCAS User Interface to upload and trigger ingestion of data; or by uploading data directly to the LabCAS server. Before publishing, data should be "curated" i.e. organized in a meaningful directory structure, and provided with enough metadata to make them useful to other users.

Data Structure

Data in LabCAS is organized according to the following logical hierarchy:

- **Collections**: broad sets of related data from the same study, the same analysis, or the same project.
- **Datasets**: different sets of related files within the same collection - for example, the different patients of a clinical study, or the different labs participating in a study.
- **Files**: all the files in a given dataset - for example, all the images for a single patient, or all the data supplied by a single lab.

Examples:

- Collection=RNA Sequencing, Dataset=ERR164773, Files=ERR164773_1.fastq, ERR164773_2.fastq, gene.counts
- Collection=University of Colorado Lung Images, Dataset=UCHSC_1467, Files=22021 P06-32 B4 Bronchus Intermedium x40 D1.jpg

Before publishing data into LabCAS, you should organize them in a meaningful directory structure conforming to the model above.

Metadata

In LabCAS, metadata can be associated to data at all 3 levels: collections, datasets, and files. When publishing, it is recommended that the following metadata fields be supplied to make the data searchable and more useful (mandatory fields are <u>underlined</u>). When using the LabCAS UI to publish data, some of these fields are automatically populated based on the user login, or the value of other supplied fields. Draft LabCAS metadata CDEs provide detailed data element information.

Collection

- <u>CollectionName</u>: short name for high level data collection (example: MD Anderson Lung Images)
- <u>CollectionDescription</u>: a few sentences describing the high level data collection (ex: Lung images for clinical studies conducted during 2001-2010.)
- <u>OwnerPrincipal</u> (ex: uid=amos,dc=edrn,dc=jpl,dc=nasa,dc=gov)
- <u>Consortium</u> (ex: MCL)
- <u>Discipline</u>: (ex: RNA Sequencing, Pathology, Oncology, etc.)
- <u>LeadPI</u> (ex: Chris Amos)
- LeadPIId
- *QAState* (ex: Public)
- <u>Organ</u> (ex: Lung, Pancreas, etc.)
- OrganId (ex: 3)
- <u>Institution</u> (ex: Dartmouth, MD Anderson Cancer Research Center)
- InstitutionId
- ProtocolName
- ProtocolId
- <u>CollaborationGroup</u>

Dataset

- <u>DatasetName</u>: short name for this dataset (ex: ERR318895)
- <u>DatasetDescription</u>: a few sentences describing this dataset (ex: Data for patient X)
- PubMedID
- <u>Species</u>
- Instrument
- SpecimenType
- DatasetURL (ex: http://someotherwebsite/patients?id=X)

File

- <u>FileName</u>: parsed from file system
- <u>FileLocation</u>: parsed from file system
- FileType
- <u>FileSize</u>: parsed from file system
- FileDescription
- ProcessingLevel

Any other custom metadata can be supplied and stored as well, at all levels of the data hierarchy.

Publishing via LabCAS UI

The LabCAS web portal (http://labcas.jpl.nasa.gov) provides a web-enabled workflow for uploading data to the server, and trigger publication to the LabCAS archive. The user is guided through the process of selecting the files from their own desktop, and populating the required and recommended metadata fields. Because uploading data through the browser is not very efficient, this process is recommended for datasets that are not very large (up to a few GB in size).

Uploading Data to the LabCAS Server

Very large datasets should be transferred to the LabCAS server via some means other than a web browser. At this time, LabCAS offers two ways to upload data directly: SFTP and WebDAV. Both protocols are supported on a variety of platforms including Linux, Mac OSX, and Windows. Before uploading data, please make sure to:

- Organize your data in a directory structure corresponding to one collection and one or more datasets, as described above
- For each dataset, provide as much metadata as possible in a file colocated with the dataset, and named <dataset>.cfg (see example file)

Then proceed through one of the two methods below.

SFTP

- Generate a public/private RSA key pair with the following command, or use an existing pair if you have one:
    - ssh -keygen -t rsa
- Send your public key (id_rsa.pub) to the LabCAS team so they can enter it in the list of allowed user keys
- Simply use the SFTP client of your choice to connect to the LabCAS server (with no username or password, since authentication is provided by the key), navigate to your assigned root directory, create new directories for the Collection and Dataset to upload, then transfer the data (all files in the current local directory, including the metadata file <dataset>.cfg):
    - sftp sftpxfer@zipper.jpl.nasa.gov
    - cd data/<user directory>
    - mkdir <collection directory>
    - cd <collection directory>
    - mkdir <dataset directory>
    - cd <dataset directory>
    - put *

WebDav data upload

- Contact the LabCAS team to obtain a WebDAV username and password, which will be stored in the LabCAS LDAP database
- Use one of the supported clients on your desktop to upload the data (see below). The server URL is: https://labcas-dev-dav.jpl.nasa.gov/staging/. After logging in with your credentials, drop your data in the location staing/<collection>/<dataset>
    - Mac OSX: macs come with a pre-installed WebDAV client. In the upper bar menu, choose Go > Connect to Server, and enter the URL above
    - Linux: we recommend using the *cadaver* client (see http://www.webdav.org/cadaver/):
    - Windows:

Note that at this time uploading data through SFTP or WebDAV will NOT automatically trigger data publishing (although it will in the near future). Please contact the LabCAS team to let them know new data is available, and to coordinate the publishing phase.