Datasets on Rosalind

To avoid users wasting their storage quota on large public datasets, we have set aside an area in /mnt/lustre/datasets to store shared data. Typically these datasets will be owned and curated by either the Rosalind administrators or by a single user who will have write access to the directory and shared with other users. Everyone will have read access to the datasets under the datasets/PUBLIC directory. Other directories will be readable to a specific group of users and the dataset curator will need to approve new requests for access.

If you would like a public dataset to be made available, or you would like to share a dataset with a group of users please ask your Rosalind administrator. Note that dataset directories are only writeable by their curator. If you want a shared space to which multiple users have write access, you will need a group working directory (again, ask your Rosalind administrator).

Dataset Metadata

Currently we only have a few datasets, but as the number grows we plan to implement a command line search tool to help users find the datasets they need and to automate the process of generating a wiki page for each dataset. For this reason, we ask dataset curators to adhere to the following guidelines:

- All datasets should be given a UUID. You can use the uuidgen program on the command line to generate one for you. This UUID refers to the dataset and should be shared by all versions of that dataset.
- By convention, shared datasets live in: /mnt/lustre/datasets/<dataset_name>/<dataset_version>
- Dataset directories owned by <curator_username>:<accessgroup> (or owned by root if the systems administrators are curating the dataset)
- Dataset directories should be writeable only by their curator and readable only by their group.
- If a dataset is public, it should live in the /mnt/lustre/datasets/PUBLIC subdirectory and have group ‘clusterusers'. If the dataset is restricted, a group should be created specifically to allow access to it. The group name should have the format: data_<dataset>
- To allow the future development of a dataset search tool, all datasets should contain a METADATA.json file with some minimal metadata in the following format (feel free to suggest format changes or additional fields to the schema) :

```
{
  'ROSID' :        'UUID',
  'VERSION' :      '1.0.0',
  'NAME':          'datasetname',
  'DESCRIPTION' :  'A brief description of the dataset',
  'KEYWORDS' :     ['a', 'list', 'of', 'search', 'terms'],
  'CURATOR' :       'k123456',  # Cluster user responsible for the dataset.
If restricted, this person must grant new access requests.
```

```
  'ACCESS_GROUP' :       'restricted'     # or public
  'URL' :          'optional source URL',
  'CREATORS' :      [ # people who created the data, own the IP
```

```
                    {'NAME':'blah', EMAIL: 'blah@example.com'},
                    {'NAME':'foo', EMAIL: 'foo@example.com'}
                  ],
   'CITATIONS' :   [
                     {'PMID': 'pubmed id','DOI': 'digital object
identifier', 'CITATION': 'Full citation string'}
                  ],
   'CREATED' :     '', # date of dataset creation by dataset creator
   'UPLOADED' :    '', # date this data was first uploaded onto the KCL
systems
   'MODIFIED' : '', # date this data was last modified
   'DELETION_DATE': '', # If applicable, date on which the dataset must be
removed from our system
}
```

For restricted datasets a subdirectory called LICENCES should contain copies of any relevant licensing or data use / material transfer agreements.

If you would like to provide additional, unstructured information about a dataset, please include a README file in the top level dataset directory.

Dataset Versioning

Updating or changing shared datasets must be done so as to avoid disrupting ongoing analyses. For smaller datasets, curators may opt to keep multiple versions. For larger datasets where keeping multiple copies is unfeasible, curators should speak to their Rosalind systems administrator to work out a versioning system that won't take up too much space.