*HPC Data MANAGEMENT*

User guide

Version *1.2*

*6/14/2017*

**Version History**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Version Number** | **Implemented**  **By** | **Revision**  **Date** | **Approved**  **By** | **Approval**  **Date** | **Description of Change** |
| 1.0 | Prasad Konka | 2/9/2016 |  |  | Initial Draft |
| 1.0 | Zhengwu Lu | 2/24/2016 |  |  | Comments |
| 1.0 | George Zaki | 2/25/2016 |  |  | Comments |
| 1.0 | Prasad Konka | 2/25/2016 |  |  | Updates |
| 1.0 | George Zaki | 2/25/2016 |  |  | Updates |
| 1.0 | Prasad Konka | 5/31/2016 |  |  | Comments from Zhengwu |
| 1.0 | George Zaki | 7/21/2016 |  |  | Added the curl description |
| 1.0 | Prasad Konka | 8/18/2016 |  |  | Prerequisites |
| 1.0 | Prasad Konka | 9/28/2016 |  |  | Integrated search |
| 1.0 | Zhengwu Lu | 11/7/2016 |  |  | Adding hierarchical search context |
| 1.0 | George Zaki | 11/22/2016 |  |  | Edited the hierarchical search documentation and curl commands |
| 1.0 | Zhengwu Lu | 11/23/2016 |  |  | Review and revise hierarchical search note on collections search and overall minor editing |
| 1.0 | Zhengwu Lu | 12/07/2016 |  |  | Revise hierarchical search default and expected implementation behavior |
| 1.0 | Zhengwu Lu | 12/12/2016 |  |  | Remove simple search as it is included in the general compound search implementation |
| 1.0 | Prasad Konka | 12/19/2016 |  |  | Updated server URL |
| 1.1 | Zhengwu Lu/George Zaki | 6/9/2017 |  |  | Add Web GUI section, add/update for Git migration |
| 1.2 | Zhengwu Lu | 6/14/2017 |  |  | Add/revise for Web GUI 1.2 release |
| 1.2 | George Zaki | 6/27/2017 |  |  | Added documentation to command line utilities |
| 1.2 | Prasad Konka | 6/27/2017 |  |  | Globus account setup |

TABLE OF CONTENTS

[1. Purpose 5](#_Toc486324465)

[2. Introduction 5](#_Toc486324466)

[2.1 What is HPC DME ? 5](#_Toc486324467)

[2.2 Intended Users 5](#_Toc486324468)

[2.3 Accessing HPC DME 5](#_Toc486324469)

[3 Pre-Requisites 6](#_Toc486324470)

[3.1 Account with HPC DME 6](#_Toc486324471)

[3.2 Globus Account for Asynchronous data transfers 6](#_Toc486324472)

[3.3 Pre-Requisites for HPC Client 7](#_Toc486324473)

[3.3.1 Java 7](#_Toc486324474)

[3.3.2 Get HPC Client 7](#_Toc486324475)

[4 HPC DME Overwiew 7](#_Toc486324476)

[4.1 Business Rules and characteristics of Collections 8](#_Toc486324477)

[4.2 Business Rules and characteristics of Data objects 9](#_Toc486324478)

[4.3 Metadata 9](#_Toc486324479)

[4.4 Data transfer 10](#_Toc486324480)

[4.5 User Authentication 11](#_Toc486324481)

[5 HPC DME API OverView 11](#_Toc486324482)

[6 HPC DME client Overwiew 12](#_Toc486324483)

[6.1 HPC DME Batch Client 12](#_Toc486324484)

[6.1.1 Set HPC batch client properties 13](#_Toc486324485)

[6.1.2 Running HPC Batch Client 14](#_Toc486324486)

[6.1.3 Error handling 27](#_Toc486324487)

[6.2 Accessing HPC DME API with CURL 28](#_Toc486324488)

[6.2.1 Setting up the environment 28](#_Toc486324489)

[6.2.2 Executing CURL commands 28](#_Toc486324490)

[6.3 Accessing HPC DME API using Cmd line utilities 30](#_Toc486324491)

[6.3.1 One time setup 31](#_Toc486324492)

[6.3.2 Generate a Data management API token 31](#_Toc486324493)

[6.3.3 Generate a Globus token 31](#_Toc486324494)

[6.3.4 Register a collection 31](#_Toc486324495)

[6.3.5 Register a dataobject 32](#_Toc486324496)

[6.3.5.1 Synchronously (From file system) 32](#_Toc486324497)

[6.3.5.2 ASynchronously (From Globus) 32](#_Toc486324498)

[6.3.6 Register a Directory: 32](#_Toc486324499)

[6.3.6.1 Synchronously (from file system) 32](#_Toc486324500)

[6.3.6.2 Asynchronously (From Globus) 32](#_Toc486324501)

[6.3.7 Get a collection metadata 33](#_Toc486324502)

[6.3.8 get a dataobject metadata 33](#_Toc486324503)

[6.3.9 Download a dataobject synchronously 33](#_Toc486324504)

[7 HPC DME Web Interface Overview 33](#_Toc486324505)

[7.1 Dashboard and login/logout 34](#_Toc486324506)

[7.1.1 Log In 34](#_Toc486324507)

[7.1.2 Log Out 34](#_Toc486324508)

[7.1.3 Dashboard 34](#_Toc486324509)

[7.2 Search functions 35](#_Toc486324510)

[7.2.1 Google Style Search 35](#_Toc486324511)

[7.2.2 Basic Search 35](#_Toc486324512)

[7.2.3 Advanced Search 36](#_Toc486324513)

[7.2.4 Saved Search 36](#_Toc486324514)

[7.2.5 Bookmark Function 37](#_Toc486324515)

[7.3 BROWSE functions 38](#_Toc486324516)

[7.4 SHARING functions 40](#_Toc486324517)

[7.5 Update metadata in non-batch mode 41](#_Toc486324518)

[7.6 Download Collection/file from archive 42](#_Toc486324519)

[7.6.1 Share a Globus End Point with HPC DME Service Account 42](#_Toc486324520)

[7.6.2 Download a Collection of Data Files 44](#_Toc486324521)

[7.6.3 Download a Single Data File 45](#_Toc486324522)

[7.7 Notifications and reports 46](#_Toc486324523)

[7.7.1 Notifications Management 46](#_Toc486324524)

[7.7.2 Reports Management 47](#_Toc486324525)

[7.8 Group administrator functions 48](#_Toc486324526)

[7.8.1 User 48](#_Toc486324527)

[7.8.2 Group 49](#_Toc486324528)

[7.8.3 Delete a Data Object and Associated Metadata 51](#_Toc486324529)

[8 APPENDIX A – sample collection input 53](#_Toc486324530)

[9 APPENDIX B – sample data file input 53](#_Toc486324531)

[10 Appendix c – sample permissions input 53](#_Toc486324532)

# Purpose

The user guide describes the functional capabilities of batch utility to register collections (i.e., datasets, projects, or folders) and data objects (i.e. files or folders) with HPC data management environment. The user guide further details the prerequisites needed for setting up the HPC data management environment (DME)**.**

# Introduction

## What is HPC DME ?

The HPC DME, High Performance Computing Data Management Environment, is a highly adaptable and an open ended data storage environment supporting storage and management of huge amounts of data, produced from high performance computing systems. HPC DME provides capabilities for storing, managing, transferring and sharing huge amounts of data across different systems securely and efficiently.

Users can store data objects for a long term on HPC DME, share and transfer their data such that they do not have to redistribute or maintain copies of the data on other systems by eliminating the data integrity issues. HPC DME stores and associates user defined metadata to any registered data at different levels of data life cycle, enabling the environment not only to help identify the data but also enhancing the search capabilities and to be able to attach a value factor to each dataset.

## Intended Users

The HPC DME has been built to cater to the data storage and data management needs of NCI cancer community. Any user with a valid HPC DME user account can run its client interface. Note: HPC DME supports only NCI account holders at this point of time.

## Accessing HPC DME

HPC DME can be accessed through its Service API or through its client interface. HPC DME Service APIs are developed based on REST standard. These APIs can be accessed by any interface that is compatible with REST standard. For example, these Service APIs can be accessed through Java, Perl, Ruby, Curl, JavaScript languages programmatically or through UI clients like SOAP UI, REST Client.

The following is the HPC DME Service API URL:

[https://hpcdmeapi.nci.nih.gov:8080/<Resource Name>](https://hpcdmeapi.nci.nih.gov:8080/%3cResource%20Name%3e)

HPC DME can also be accessed through its interactive client interface. The following are pre-requisites to access client interface.

# Pre-Requisites

## Account with HPC DME

You will first need to register with HPC DME for an account. Please consult HPC DME administrator for an account at [HPC\_DME\_Admin@nih.gov](mailto:HPC_DME_Admin@nih.gov) after obtaining your group or DOC approval of using the HPC DME. If you are a HPC DME administrator, please see HPC Server API specification for user registration details.

## Globus Account for Asynchronous data transfers

HPC DME supports transferring data from a Globus endpoint to HPC DME Archive storage. If you are working at FNLCR and you want to register data hosted at ISILON asynchronously, you would need to request for MOAB account. Please contact System Administrator or send an email to Frederick Helpdesk [NCI-FHelpdesk@mail.nih.gov](mailto:NCI-FHelpdesk@mail.nih.gov) for account access.

However, using MOAB and Globus is not necessarily a prerequisite to deposit data files (objects) into the Cleversafe object store or any other storage medium like isilon storage system. Generally speaking, Using Globus and the GridFTP transfer is recommended for transferring large data objects (greater than or equal to 5 GB) asynchronously.

If you are planning to use any other Globus Endpoints at NIH that you have access to, you should link your Globus identity with NIH login. Once they are linked you could use NIH IdP to get authenticated and get access to NIH Globus Endpoint. To link the accounts:

1. Login into [www.globus.org](http://www.globus.org) with your Globus Id
2. Click on “Account” button which is at the top right corner. Click on “Add Linked Identity”
3. Select “National Institutes of Health” from the IdP list box. And click on Continue button.
4. Now you will be redirected to NIH iTrust Website. Enter your NI credentials and submit.
5. If you are asked to enter your Globus Credentials, enter those and click continue.
6. Now Globus and NIH accounts are linked. FYI – Globus would not track or keep NIH credentials. It would use NIH IdP to authenticate your NIH credentials.

## Pre-Requisites for HPC Client

### Java

Download and install Java 1.8.x

<https://www.java.com/en/download/>

Please visit the following link for installation instructions

<https://www.java.com/en/download/help/index_installing.xml>

After successful installation of Java, update PATH with Java executable folder (Ex: c:/jdk1.8\_13/bin)

<https://java.com/en/download/help/path.xml>

Open cmd/shell window and type “java –version” to make sure it displays the version information you installed.

### Get HPC Client

Clone HPC DME Utils into your local machine.

Git clone <https://github.com/CBIIT/HPC_DME_APIs>

Navigate to HPC\_DME\_APIs/utils folder. Open README.md and follow the instructions

# HPC DME Overwiew

The HPC data management environment provides a high-reliability storage model for underlying collections including a collection registration system, and an API for transfer of large data objects with no-loss of data. The data object registration associates a label with a given managed data file or folder and captures extensible metadata for the managed data object.

The HPC data management environment provides a number of application programming interfaces (APIs) to operate and interact with it. At a high level, there are two important components in HPC DME. 1) Metadata management: HPC DME by default integrates with iRODS iCAT instance to manage metadata and its security for both collections and data objects. 2) Data transfer: HPC DME by default uses Globus to perform asynchronous data transfer between Globus endpoints. HPC DME pluggable architecture allows both these implementations to be replaced with alternatives easily while keeping its APIs unchanged. The basic features of HPC DME is to help users in registering and uploading their data to the HPC DME archive storage and managing it. HPC DME archive storage can be a permanent storage for the users’ data and be used as a platform to search, manage and transfer the data to other storage systems and also to share with other collaborators or users. Each data object is stored along with its required and user defined metadata associated with it. The associated metadata can be used as search criteria to identify dataset(s).

The HPC data management implementation provides users flexibility to define collections and data objects. Collections in HPC DME can be Projects, Datasets, etc. and data objects are the data files or folders. Data object is defined as an individual file or a set of files in a folder. Collections can be of different types for example, a dataset can be linked to a sequencing project, imaging study or an analysis project etc., or a project can be an entity (umbrella project) indicating a scientific ‘study’.

## Business Rules and characteristics of Collections

* Each collection can be referred with a logical path (“collection\_path”) which is unique across HPC DME.
* Irrespective of the logical path, each collection is associated with UUID metadata attribute to uniquely identify itself.
* A collection can be any category of business domain entities such as: Project, Dataset, Folder, Samples, Runs. More types can be configured in the system easily through metadata configuration. Please consult your HPC DME system administrator.
* Users can create/register multiple collections with HPC DME.
* Each collection has its own required set of metadata which needs to be submitted at the time of registration. This required metadata can be configured by a system administrator with input from each DOC’s group administrator.
* Users have an option of adding new metadata variables to the required metadata associated with each collection.
* A collection can contain one or more data objects, or child collections. This is achieved using the logical path of a given collection.

## Business Rules and characteristics of Data objects

* A data object can be a single data file or a folder compressed or uncompressed.
* Each data object can be referred with a logical path (“object\_path”) which is unique across HPC DME.
* A parent collection should be registered first with HPC DME before registering a data object.
* Irrespective of the logical path, each data object is associated with UUID metadata attribute to uniquely identify itself.
* Each data object has associated required metadata which is required to be submitted at the time of data object registration.
* Users have an option of adding new metadata variables to the required metadata associated with each data object.

## Metadata

Metadata is defined as the data about the data. It is the information describing the actual data such as the date and origin of creation, its contents, its condition, its processing, or its associations to other objects. Metadata is employed to make data searches faster, more specific and also enable and promote data sharing among scientists.

HPC DME collects metadata for each collection and data object, registered and stored in a database along with the associations. HPC DME collects two kinds of metadata related to a collection or a data object, namely, administrative and center/division specific. The administrative metadata is the required set of information which needs to be submitted at the time of registration with HPC DME. New metadata variables can be added to both administrative and center/division specific metadata sets after a user obtains proper authorizations and permissions. The metadata can also be updated by authorized users through API or command line client.

## Data transfer

HPC DME can register data objects from Globus endpoint source or from a file system. If the transfer source location is Globus, HPC DME requires user to share transfer data with the HPC DME service account. This can be done by going to [www.globus.org](http://www.globus.org) and select the folder to share. Globus do not support sharing individual files. If you have a single file to register, please create a folder for it and share the folder. The data transfer from Globus source location is done asynchronously.



HPC DME Server API integrates with Globus Connect API to provide seamless data transfer, tracking and reporting capabilities. Upon data object registration, HPC API initiates data transfer with Globus Connect Server asynchronously. HPC API gathers data transfer details including status asynchronously and keep the information within HPC DME database for reporting purposes. Data objects can be a single file or a folder where the whole folder and its contents including subfolders are transferred to the archive system.

Note: Globus is deprecating the use of Endpoint names. It encourages using Endpoint Id. You can retrieve endpoint Id by clicking on “Endpoints” tab link in the Globus portal (after login) and click on the endpoint name. You should see “UUID” if you scroll down in the “Overview” section. Use UUID value for “fileContainerId”.

HPC DME can also register data objects from a local file system without Globus dependency. This operation integrates with S3 API to synchronously transfer data from local file system to HPC DME archive. The current approach is to stream data files (objects) from local/designated file directory to the application server, and then stream that to the Cleversafe archive without concern if the service account has access right to those objects to be deposited.

## User Authentication

After enrolling with HPC, the users can login into HPC DME using their user id. The HPC DME validates the user passwords through the NCI LDAP system.

# HPC DME API OverView

HPC DME Service API is a RESTful interface. This interface is communicated over HTTPS with the same HTTP verbs (GET, POST, PUT, DELETE) that web browsers use to retrieve web pages and to send data to remote servers. HPC DME Service API can be categorized into the following sections. Please see HPC DME Server API specifications for details on each of these functions.

* User
  + [Enroll user into HPC DME](#_Enroll_User) (Create)
* Create Group
  + Add or remove user(s) from a group
* Collection
  + [Register Collection with its metadata into HPC DME](#_Register_Project)
  + Find Collection by path
  + [Find Collection by metadata](#_Find_Project_by_1)
* Data Object
  + [Register data object along with metadata into HPC DME](#_Register_Dataset) (Create and Update)
  + Find Data Object by path

[Find Data Object by metadata](#_Find_Dataset_by_2)

* Permissions
  + Assign permissions on HPC DME collections or data objects
* Download
  + Download data object/file to another Globus Endpoint

# HPC DME client Overwiew

HPC DME APIs are developed on REST standard, so any client interface can be used to access these APIs confirming REST standards. HPC DME provides an interactive client interface to process bulk data to register collections, data objects and assign permissions. If you are planning to use Globus for data transfer, please refer to section 3.2 on how to setup your Globus account.

## 6.1 HPC DME Batch Client

The HPC Batch Client is an interactive tool where users can enter commands to initiate supported functions. Current version of HPC client supports the following functions:

1. Registering collections:

Register collections and their associated metadata

1. Registering Data objects:

Register data objects and their associated metadata

1. Update Permissions:

Update users permissions on accessing HPC DME metadata

1. Search Collections:

Search for collections based on search criteria and write results into a csv/json/txt file

1. Search Data objects:

Search for data objects based on search criteria and write results into a csv/json/txt file

1. Batch registration of local files

Recursively register all files from a given local folder with HPC DME. This register each file with basic metadata

1. Batch registration of files from a Globus Endpoint

Recursively register all files from a given Globus endpoint location with HPC DME. This register each file with basic metadata

HPC Client is integrated with HPC Server REST interfaces to perform the functions mentioned above. It takes CSV (comma separated value) file as input for the commands it supports. HPC client parses given CSV file input and invokes server APIs in a batch manner to perform requested commands.

HPC Client commands are secured with valid authentication. To run protected commands, a user would need to provide HPC credentials to get authenticated. All client commands do not support multi-thread processing of input requests at this time.

### Set HPC batch client properties

Before running a HPC client, you would need to update HPC client properties file – hpc.properties.

The following are the properties:

|  |  |  |
| --- | --- | --- |
| Property name | Description | Default value |
| hpc.server.url | HPC Server API url. This is where REST interfaces are running on the server | https://hpcdmeapi.nci.nih.gov |
| hpc.user | UserId running the client |  |
| hpc.default.globus.endpoint | Globus endpoint to transfer from |  |
| hpc.collection.service | Collection service REST resource name. You don’t need to change this value unless it is changed on the server. | collection |
| hpc.dataobject.service | Data object REST resource name. You don’t need to change this value unless it is changed on the server. | dataObject |
| hpc.error-log.dir | Path of the error log file to be written | . |
| hpc.ssl.keystore.path | Path of the keystore used for 2-way SSL connectivity with HPC Server. If the value is missing, one way SSL connection is used. | hpc-client/keystore/keystore-uat.jks |
| hpc.ssl.keystore.password | Password for the keystore. You don’t have to change this value. If the value is missing, one way SSL connection is used. | hpc-server-store-pwd |
| hpc.login.credentials | Location of the file with user credentials (user name and password). Credentials should be in format of <userId:password> |  |
| hpc.job.thread.count | Number of concurrent threads to process batch input file. Setting up large number of threads may not necessarily improve processing time. It depends on various factors like client machine processing speed, network bandwidth, server response. | 5 |
| globus.nexus.url | Globus Nexus URL. You don’t have to change this property value. | nexus.api.globusonline.org |
| globus.url | Globus URL for asynchronous data registration. You don’t have to change this property value. | www.globusonline.org |
| hpc.login.token | HPC USER login token file location. You don’t have to change this property value. | tokens/hpcdme-auth.txt |
| hpc.globus.login.token | HPC USER globus login token file location. You don’t have to change this property value. | tokens/globus-auth.txt |

### Running HPC Batch Client

HPC client is a command line tool that can be run using Java. Open cmd window and navigate to the folder where HPC client is extracted to. HPC Client can be run in interactive mode or batch mode.

#### Interactive Mode:

Go to <HPC\_DME\_APIs>/utils folder and type the following command.

If you are using Unix:

export *HPC\_DM\_UTILS=<Path to HPC\_DM\_UTILS>*

*Java -Dhpc.client.properties=$HPC\_DM\_UTILS/hpcdme.properties -DHPC\_DM\_UTILS=$HPC\_DM\_UTILS -jar hpc-client/hpc-cli-1.1.0.jar*

If you are using Windows:

SET *HPC\_DM\_UTILS=<Path to HPC\_DM\_UTILS>*

*Java -Dhpc.client.properties=%HPC\_DM\_UTILS%/hpcdme.properties -DHPC\_DM\_UTILS=%HPC\_DM\_UTILS% -jar hpc-client/hpc-cli-1.1.0.jar*

You will see hpc-cli command line interface as below.



Type “help” to see all the commands supported by hpc-cli.



#### Batch Mode

The HPC client supports running multiple HPC Client commands in a batch mode.

*Java -Dhpc.client.properties=$HPC\_DM\_UTILS/hpcdme.properties -DHPC\_DM\_UTILS=$HPC\_DM\_UTILS -jar hpc-client/hpc-cli-1.1.0.jar --cmdfile <Path to HPC commands>*

For example, you may put the following commands into a text file and give that file as --cmdfile input. HPC Client will run these commands in the sequence they are given in the file.

putCollections --source samples/Batch\_Upload\_Input\_File\_Collections.csv

putDatafiles --source samples/Batch\_Upload\_Input\_File\_Object.csv

putPermissions --source samples/Batch\_Permissions.csv

You could also run any operating system commands as part of the file by using “!”. When you are using batch mode, you should set “hpc.login.credentials” property value in hpc.properties. This is to bypass login prompt during batch execution. Please keep credentials file in a safe location so that only you have access to it.

#### Collection Types

HPC DME supports user defined collection types. Some of the collections can be Project, Sample, Study, Dataset. Collections types should be defined in the system policy file to enforce validation rules. To create new collection types, please submit the following information to HPC DME System Administrator.

Collection type name and its metadata in the following format:

{

"attribute":"<Name of the attribute>",

"mandatory":<true/false>,

"defaultValue":"<Value>",

"collectionTypes":<Collection Type name>,

"validValues":[

"<Value1>",

"<Value2>"

],

"ruleEnabled":true,

"DOC":"<DOC-NAME>"

}

#### Register collections

The HPC client supports registering multiple collections with HPC DME. A collection can be a project, sample, run, dataset or folder. New collection types can be added on HPC DME as needed by a system administrator as described above. The following is the command to register collections.

putCollections --source <input file path>

Input file for collections registration would need to provide all required metadata along with any user defined metadata. The following are the sample required metadata attributes. These attributes can be configured on HPC DME as needed by a system administrator.

The following table shows sample required attributes for a Dataset.

|  |  |  |
| --- | --- | --- |
|  | **Attribute** | **Definition** |
|  | collection\_path | Logical path of the collection. Collection path is organized in a hierarchical manner where a base parent folder represents a home folder for a division. For example, FNL\_SF\_Archive is the base folder for the sequencing facility at Frederick National Labs. So the path of the collection would be /FNL\_SF\_Archive/<collection name> |
|  | collection\_type | Collection type name (Valid values are Project, Dataset, Folder). |
|  | name | Name for the dataset of files as provided by the depositor |
|  | description | Description of dataset |
|  | source\_lab\_pi | PI of the lab of the depositor at the time of deposit |
|  | lab\_branch | Lab or Branch or Program the PI belongs to |
|  | pi\_doc | Division, Organization, Center the PI belongs to |
|  | original\_date\_created | Date the dataset was created originally |
|  | data\_creator | Person or Organization lead who created the data |
|  | phi\_content | Presence of Protected Health Information in the datasets deposited via HPC DME.  Valid values are (PHI Present, PHI Not Present, Not Specified). If no value is given, default value is “Not Specified”. |
|  | pii\_content | Presence of Personally Identifiable Information in the datasets deposited via HPC DME. Valid values are (PII Present, PII Not Present, Not Specified). If no value is given, default value is “Not Specified”. |
|  | data\_encryption\_status | If the data is encrypted or not. Valid values are (Encrypted,  Not Encrypted, Not Specified). If no value is given, default value is “Not Specified”. |
|  | data\_compression\_status | If the data is compressed or not. Valid values are (Compressed,  Not Compressed, Not Specified) If no value is given, default value is “Not Specified”. |
|  | funding\_organization | Organization Funding the generation of Data |
|  | comments | General text for internal use and reference. Optional |

The following table shows sample required attributes for a Project. Please see APPENDIX A for sample batch input file.

|  |  |  |
| --- | --- | --- |
|  | **Metadata Variable** | **Definition** |
|  | collection\_path | Logical path of the collection |
|  | collection\_type | Collection type name (Default valid values are Project, Dataset, Folder) |
|  | name | Name for the Project |
|  | description | Description of project |
|  | source\_lab\_pi | PI of the lab of the depositor at the time of deposit |
|  | lab\_branch | Lab or Branch or Program the PI belongs to |
|  | pi\_doc | Division, Organization, Center the PI belongs to |
|  | original\_date\_created | Date the dataset was created originally |
|  | project\_type | Valid values are "Unspecified", "Umbrella Project", "Sequencing", "Analysis". If not value is given, “Unspecified” value is used. |
|  | Internal\_project\_id | Internal Project Id to track |
|  | comments | General text for internal use and reference. Optional |

Collections registration input file can have all collection types in a single file or they can be in separate files. If you are combining all collections into a single input file, attributes not relevant for a particular collection should have empty value.

Running the batch command would give an output as follows.



#### Registering Data Objects

The HPC client supports registering multiple data files with HPC DME. Parent collections needs to be registered first before registering data files. The following is the command to register data files.

putDatafiles --source <input file path>

Input file for data file or folder registration would need to provide all required metadata along with any user defined metadata. The following are the default required metadata attributes. These attributes can be configured on HPC DME as needed by a system administrator. HPC DME supports registering data objects from a local file system (synchronously) or from a Globus endpoint (Asynchronously). Based on the given request input, HPC DME client API initiates appropriate request.

The following table shows sample required attributes for a Data file. Please see APPENDIX B for sample batch input file.

|  |  |  |
| --- | --- | --- |
|  | **Metadata Variable** | **Definition** |
|  | object\_path | Logical path of the object. Object path is organized in a hierarchical manner with the combination of parent collection path and object path. For example, /FNL\_SF\_Archive/<collection name>/<object\_name>. When the data object is registered with HPC DME and the data is transferred using Globus, logical path of the object is used to created physical path on the file system. This path could be referring to an individual file or a folder. When submitting data registration request, you could keep file/folder name same as it is at the source or change it. If the object path represents an existing folder or file, it will be overwritten by the new request. If the object path represents a folder, the data object is placed in the folder. |
|  | name | Name for the file as provided by the depositor |
|  | description | Extensible description of File |
|  | source\_lab\_pi | PI of the lab of the depositor at the time of deposit |
|  | fileId | Data Source Globus Path |
|  | fileContainerId | Data Source Globus Endpoint. Use UUID of the endpoint. |
|  | lab\_branch | Lab or Branch or Program the PI belongs to |
|  | pi\_doc | Division, Organization, Center the PI belongs to |
|  | original\_date\_created | Date the File was created originally |
|  | data\_creator | Person or Organization lead who created the data |
|  | phi\_content | Presence of Protected Health Information in the datasets deposited via HPC DME.  Valid values are (PHI Present, PHI Not Present, Not Specified). If no value is given, default value is “Not Specified”. |
|  | pii\_content | Presence of Personally Identifiable Information in the datasets deposited via HPC DME. Valid values are (PII Present, PII Not Present, Not Specified). If no value is given, default value is “Not Specified”. |
|  | data\_encryption\_status | If the data is encrypted or not. Valid values are (Encrypted,  Not Encrypted, Not Specified). If no value is given, default value is “Not Specified”. |
|  | data\_compression\_status | If the file is compressed or not. Valid values are (Compressed,  Not Compressed, Not Specified). If no value is given, default value is “Not Specified”. |
|  | funding\_organization | Organization Funding the generation of Data |
|  | comments | General text for internal use and reference. Optional |

Running the batch command would give an output as follows.



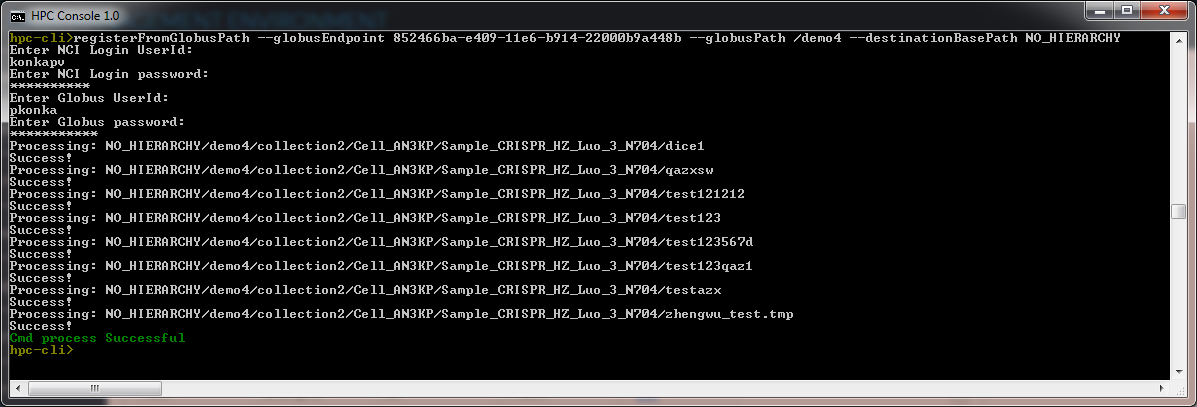
If the source of the data object(s) is a Globus endpoint, the expected operational behavior is that all data objects (files or folders) will be transferred via Globus Connect from the indicated “fileId” (associated with the “fileContainerId”) to the system configured Archive Area.

#### Registering Data Objects from a Globus location

The HPC client supports registering all data files from a Globus location path. This is a convenient command to register all data files from a Globus location. This command recursively goes through each folder and register all files along with folder structure. This process will generate “name” and “modified\_date” metadata along with system generated metadata. This command expects that there is no data hierarchy and metadata validation rules in place for the user DOC.

registerFromGlobusPath --globusEndpoint <Globus Endpoint UUID> --globusPath <Source Path on Globus> --destinationBasePath <Destination Base Path>

Running the batch command would ask you to enter your NIH credentials and Globus credentials. NIH credentials are to authenticate with HPC DME API and Globus credentials are to query Globus location for data file list.



#### Registering Data Objects from a local file path

The HPC client supports registering all data files from a local file path. This is a convenient command to register all data files from a given path on your file system. This command recursively goes through each folder and register all files along with folder structure. This process will generate “name” and “modified\_date” metadata along with system generated metadata. This command expects that there is no data hierarchy and metadata validation rules in place for the user DOC.

registerFromFilePath --filePath <Local file path> -- filePathBaseName <Base path in the file path> --destinationBasePath <Destination Base Path>

filePathBaseName: This is the relative base path for your data registration. For example, if your files path is /opt/hpc/folder1/folder2/folder3 and you want your data file relative path to start with /folder1, set filePathBaseName as folder1. The logical file path will be destinationBasePath/folder1/folder2/folder3.

Running the batch command would ask you to enter your NIH credentials to authenticate with HPC DME API.



#### Update permissions

When a user registers a collection or data file with HPC DME, that user assumes “owner” role of the entity. Owner or system admin can assign permissions on these entities to other users or groups. HPC Client provides a convenient command to update permissions on multiple entities in a batch manner. The following is the command to assign permission on collections or data objects to a user or group.

putPermissions --source <input file path>

The following table shows default required attributes for input file. Please see APPENDIX C for sample batch input file.

|  |  |
| --- | --- |
| **Attribute** | **Description** |
| object\_path | Logical path of the entity (Collection or Data file) |
| UserId | HPC DME User Id to assign permission |
| GroupId | HPC DME Group Id to assign permission |
| Permission | Permission value.  Valid values are “write”, “own”, “read”, “none”  Own – Owner of the entity. Can assign permissions to other users  Read – Read only permission  Write – Read and Write permission  None – No permission |

Running the batch command would give an output as follows.



#### Search Collections by metadata

Using the following command, users can search for collections based on a given search criteria and write resulted search records into a file.

getCollections --criteria '<Criteria>' --format <csv|json|txt> --outputfile <file full path> --details <yes|no>

**Criteria**: HPC DME metadata search API criteria is a json string with the combination of search attribute name, search attribute value and the operator to search with. The format of the criteria is as below:

{"a":"<attribute name>","v":"<attribute value>","o":"<Operator>"}

Example: {"a":"description","v":"%prasad%","o":"LIKE"}

Search cmd support searching on multiple criteria conditions. The following is the format:

{"a":"<attribute name>","v":"<attribute value>","o":"<Operator>"}**&&**{"a":"<attribute name>","v":"<attribute value>","o":"<Operator>"}**&&**{"a":"<attribute name>","v":"<attribute value>","o":"<Operator>"}

All the given criteria conditions are applied to the search with AND operator.

**format:** Search cmd takes optional “format” argument to specify format of the output file. Valid options are “csv”, “json” and “txt”. If this argument is not given, output is written in “txt” format. For “csv” format, collections “selfMetadataEntries” are written to the output file. “parentMetadataEntries” will not be written into the output file for “csv” format.

**outputfile**: Search cmd takes optional “outputfile” argument to specify output file full path. If path is not given, default path is “hpc.error-log.dir” + “getcollections\_records”+<Current date> + <type>.

**details**: Search cmd takes optional “details” argument to specify if output should include collection details or not. If not specified, default value is “yes”. If “no” value is given, output file written with collection path values only.

#### Search Collection by path

Using the following command, users can search for collections based on a logical path and write resulted search records into a file.

getCollection --path '<path>' --format <csv|json|txt> --outputfile <file full path>

**path:** Logical path of the collection

**format:** As explained in the section above

**outputfile:** As explained in the section above

#### Search Data objects

Using this command, users can search for data files based on a given search criteria and write resulted search records into a file.

getDatafiles --criteria '<Criteria>' --format <csv|json|txt> --outputfile <file full path> --details <yes|no>

**Criteria**: HPC DME metadata search API criteria is a json string with the combination of search attribute name, search attribute value and the operator to search with. The format of the criteria is as below:

{"a":"<attribute name>","v":"<attribute value>","o":"<Operator>"}

Example:

{"a":"description","v":"%prasad%","o":"LIKE"}

Search cmd support searching on multiple criteria conditions. The following is the format:

{"a":"<attribute name>","v":"<attribute value>","o":"<Operator>"}**&&**{"a":"<attribute name>","v":"<attribute value>","o":"<Operator>"}**&&**{"a":"<attribute name>","v":"<attribute value>","o":"<Operator>"}

All the given criteria conditions are applied to the search with AND operator.

**format:** Search cmd takes optional “format” argument to specify format of the output file. Valid options are “csv”, “json” and “txt”. If this argument is not given, output is written in “txt” format. For “csv” format, collections “selfMetadataEntries” are written to the output file. “parentMetadataEntries” will not be written into the output file for “csv” format.

**outputfile**: Search cmd takes optional “outputfile” argument to specify output file full path. If path is not given, default path is “hpc.error-log.dir” + “getDatafiles\_records”+<Current date> + <type>.

**details**: Search cmd takes optional “details” argument to specify if output should include collection details or not. If not specified, default value is “yes”. If “no” value is given, output file written with data files path values only.

#### Search Data file by path

Using the following command, users can search for data file based on a given logical path and write resulted search record into a file.

getDatafile --path '<path>' --format <csv|json|txt> --outputfile <file full path>

**path:** Logical path of the data file

**format:** As explained in the section above

**outputfile:** As explained in the section above

#### Perform hierarchical searches

The default hierarchical search APIs is based on the fact that every collection and dataObject inherits the metadata of all its parent collections. However, there should be no assumption made on the uniqueness of metadata attribute variables across a given hierarchy.

Assume a registered file in the following path: /Coll\_A/Coll\_B/Coll\_C/File\_D, it is possible that the same metadata variable “attribute\_X” be used for “Coll\_C” and “File\_D”. In this case, the value assigned to this attribute at the lower level (i.e., File\_D) will overwrite the value assigned to this attribute at the parent level (i.e., Coll\_C) unless otherwise specified as explained next.

Default hierarchical search using compound query (please review the HPC\_server\_API document) will search the data object at all levels (>=1) if no level value is indicated by the users. The lowest level in the hierarchy is given the value ‘1’. During the hierarchical search, users can also indicate the optional level they are basing the search on. For example, the dataObject level will equal ‘1’, the parent collection level will equal ‘2’, the grandparent collection level will equal ‘3’, etc.

Example 1: Given the path of File\_D shown above, assume you attached metadata attributes as the following:

* Coll\_A:  “X” = “1” (i.e., Metadata attribute name “X” set to a value of “1”)
* Coll\_B: “Y” = “2”
* Coll\_C: “Z” = “3”
* File\_D: “W” = “4”

Since there are unique metadata variables, there would be no difference made whether or not the users enter the level parameter values in entering the search criteria for the attribute “X”, “Y”, “Z”, or “W” because it will search at all levels.

Example 2: Assume you attached Metadata attributes as the following:

* Coll\_A: “X” = “1”
* Coll\_B: “Y” = “2”
* Coll\_C: “date\_created” = “11-11-2016”
* File\_D: “date\_created” = “11-20-2016”

Since “date created” may be assigned different values for the parent collection “Coll\_C” and the child “File\_D”, it is critical for the users to indicate at what level they are performing the search. However, if no level is indicated in the entered search criteria, default search will always search at all levels (collections and data files). If the user wants to search for a dataObject whose parent collection was created at a specific date, the compound query should include a search for the attribute “date\_created” with the value “11-11-2016” at level 2, with the level operator “EQUAL”. Therefore, the level operator specifies where to apply the search criteria.

Note that the default search level for data files is at all collection and data object levels. However, if a search is specifically indicated for collections, the default search is at all collection levels. In other words, while searching for collection, the search criteria will be executed at all collection levels (except data object level 1). The rationale is that there is nothing harmful to show a broader collections result set regardless if any data objects have been registered under certain collection path.

### Error handling

HPC client interface catches any errors returned by HPC Server API and write them into error log file. The error log is written as <record sequence number> <Error message>.

Those records not processed by HPC client due to any error are written into a separate file “errorRecords<timestamp>.txt” for each command so that only these records can be corrected to go through the HPC batch client again.

## 6.2 Accessing HPC DME API with CURL

Curl is a command-line tool for transferring data using various protocols. It can be used to interact with the HPC DME REST API. If you are using Windows environment, you will need to install tools like CygWin to be able to run Unix-like commands. Please visit <https://www.cygwin.com/> to download. Here are examples on how to use the curl tool to submit REST request to the HPC DME server.

### Setting up the environment

First make sure you have curl in your environment path. Open a terminal and type:  
$curl

You should get the following line printed on stdout:

curl: try 'curl --help' or 'curl --manual' for more information

If you do not have curl on your path, check your Linux distribution to install it.

### Executing CURL commands

In the examples shown below, any string between the two characters ‘<’ and ‘>’ is a mandatory string that should be replace with an actual value without the enclosing characters ‘<’ and ‘>’.   
  
For example, <server:port> should be replaced with the HPC DME server name and port number (e.g., <https://fr-s-hpcdm-uat-p.ncifcrf.gov:7738)>

In these examples, I dump the returned response header to a file using the -D optional url parameter. For example -D <curl-response-header-file>. Similarly, I dump the response message as a json file using the -o flag. For example: -o <curl-response-message.json>

#### Registering a new user

#### curl -H "Content-Type: application/json" -d @<user-info.json> -X PUT <server>/user -H "Accept: application/json" -D <response-header> -o <response-message.json> The input file “user-info.json” should be filled with the new user’s attributes as mentioned in the HPC\_SERVER\_API.

#### Registering a collection

#### curl -H "Content-Type: application/json" -d @<attributes-file.json> -X PUT <server>/collection/<collection-path> -H "Accept: application/json" -D <response-header> -o <response-message.json> The “attributes-file.json” should include all the mandatory attributes associated with a collection type.

#### Registering a data object

#### As mentioned in the HPC\_SERVER\_API, there are two methods to register data objects: asynchronously using Globus, or synchronously from the file system.

#### Registering a data object from a Globus endpoint

#### curl -F dataObjectRegistration=@<attributes-file.json>;type=application/json -X PUT <server>/dataObject/<dataObject-path> -H "Accept: application/json" -D <response-header> -o <response-message.json> Note that the “attributes-file.json” should include the “fileContainerID” and the“fileId” as mentioned in the HPC\_SERVER\_API.

#### Registering a data object from the file system

#### curl -F "dataObjectRegistration=@<attributes-file.json>;type=application/json" -F "dataObject=@<dataObject-file>;type=application/octet-stream" -X PUT <server:port>/ /dataObject/<dataObject-path>

#### Note that the attributes-file.json should NOT include the “fileContainerID” nor the “fileId” as mentioned in the HPC\_SERVER\_API.

#### Search for collection

Curl -X POST <server>/collection/query/compound –H "Content-Type: application/json" -d @<compound-query.json> -H "Accept: application/json" -D <response-header> -o <response-message.json>

The “compound-query.json” file should contain the attribute to be searched, and the comparison operation as mentioned in the HPC\_SERVER\_API document.

#### Search for data Object

Curl -X POST <server>/dataObject/query/compound –H "Content-Type: application/json" -d @<compound-query.json> -H "Accept: application/json" -D <response-header> -o <response-message.json>

The “compound-query.json” file should contain the attribute to be searched, and the comparison operation as mentioned in the HPC\_SERVER\_API document.

#### Assign Permissions

#### curl -H "Content-Type: application/json" -d @<permission-file.json> -X POST <server>/acl -H "Accept: application/json" -D <response-header> -o <response-message.json>

#### Download data object to a Globus endpoint

#### curl -H "Content-Type: application/json" -d @<destination-description.json> -X POST <server>/dataObject/<object-path> /download --config $HPC\_DM\_TEST/utils/config -H "Accept: application/json" -D <response-header> -o <response-message> The “destination-description.json” file should contain the fileContainerID” and the “fileId” as described in the HPC\_SERVER\_API

#### Download data object to a file system

#### curl -H "Content-Type: application/json" -d @<empty.json> -X POST <server>/dataObject/<object-path> /download -H "Accept: application/json" -D <response-header> -o <response-message> In this call, the “empty.json” file should contain just the two characters “{}”. The data object will be downloaded with the name “response-message”

## 6.3 Accessing HPC DME API using Cmd line utilities

Some of the curl command mentioned in the [curl section](file:///C:\Users\luz6\AppData\Local\Temp\HPC_User_Guide-1.docx#_Acessing_HPC_DME) have been wrapped with bash functions to make them easily accessible from command line. These functions are available as part of the HPC DM API git repository:

<https://github.com/CBIIT/HPC_DME_APIs/tree/master/utils>

Here are examples on how to use some of the utilities:

## One time setup

To get the latest version of the utilities from git hub, execute the command:

git clone <https://github.com/CBIIT/HPC_DME_APIs.git>

Follow the /path/to/HPC\_DME\_APIs/utils/README.md file for the one time setup instructions of the utilities. All utilities start with the prefix “dm\_”.

## Generate a Data management API token

Run the following command to generate a DM API token (instead of putting your password every time):

dm\_generate\_token

Note that token expires every two hours.

## Generate a Globus token

If you would like to register directories directly from Globus, first you need to edit your Globus information as mentioned in the /path/to/HPC\_DME\_APIs/utils/README.md file. Then run the following command to generate the Globus token:

dm\_globus\_generate\_token

Note that you need to create a Globus account and link it to your NIH account as explained [here](#_Registering_Data_Objects).

## Register a collection

To register a collection, use the command:

dm\_register\_collection <collection-description.json> <destination-path>

The git repository contains a sample file for collection-description.json in the directory: /path/to/utils/templates. Feel free to edit the sample file to add more metadata to your collection.

For example, your command can be:

dm\_register\_collection /path/to/HPC\_DME\_APIs/utils/templates/collection-metadata.json /<top-level-directory>/my-collection

## Register a dataobject

### Synchronously (From file system)

To register a data object synchronously, use the command:

dm\_register\_dataobject <dataobject-description.json> <destination-path> <source-file>

The git repository contains a sample file for dataobject-description.json in the directory: /path/to/utils/templates.

For example, your command can be (after registering a collection above):

$dm\_register\_dataobject  /path/to/HPC\_DME\_APIs/utils/templates/dataobject-sync-metadata.json /<collection-path>/my-dataobject /path/to/dataobject-file

### ASynchronously (From Globus)

To register a data object asynchronously, use the command:

dm\_register\_dataobject <dataobject-description.json> <destination-path>

The dataobject-descripiton file should contain a valid Globus endpoint and path for your source dataObject. Also, the system account should have a read permission to the Globus endpoint as mentioned [here](file:///C:\Users\luz6\AppData\Local\Temp\HPC_User_Guide-1.docx#_Data_transfer). The repository contains a sample file for asynchronous dataobject-description.json as shown in this example:

dm\_register\_dataobject  /path/to/HPC\_DME\_APIs/utils/templates/dataobject-async-metadata.json /<collection-path>/my-dataobject

## Register a Directory:

With the command line utilities you can recursively register all files within a directory and its subdirectories to the archive. By running the next two commands, basic metadata information (e.g., filename, size, etc.) will be created for every file and pushed to the archive. The user is expected to append these metadata when more information is available.

### Synchronously (from file system)

To register a directory from your file system, execute the command:  
  
dm\_register\_directory <source-directory-path-on-file-system> <destination-path-on-archive>

### Asynchronously (From Globus)

To register a directory that exists on a Globus endpoint, first generate a Globus token as shown [here](#_Generate_a_Globus), then execute the command:  
  
dm\_register\_globus\_directory <source-directory-path-globus-endpoint> <destination-path-on-archive> [globus-endpoint]

Note that by default, the Globus endpoint you defined in your hpcdme.properties file as mentioned in the README.md is used. If you would like to use a different Globus endpoint, you can add the optional endpoint UUID at the end of the command.

## Get a collection metadata

To get all the metadata associated with a collection, execute the command:

dm\_get\_collection <collection-path> [response-message] [response-header]

If the [response-message] is omitted, the metadata will be printed to stdout.

## get a dataobject metadata

To get all the metadata associated with a dataObject, execute the command:

dm\_get\_dataobject <dataobject-path> [response-message] [response-header]

If the [response-message] is omitted, the metadata will be printed to stdout.

## Download a dataobject synchronously

To download a dataObject file from the archive to your current file system, execute the command:

dm\_download\_dataobject\_sync <dataobject-path-on-arhive> <destination-path-on-local-file-system>

# HPC DME Web Interface Overview

HPC DME web graphic user interface is developed on the REST APIs to exemplify how end users may utilize the web browsers to fulfil common use case scenarios in scientific data/metadata archiving and data management activities. Please note that what has been presented here is just a reference implementation. Labs, groups or functional units are welcome to implement their own user interface using the REST APIs.

HPC DME UAT Web GUI link for user acceptance testing:

<https://fr-s-hpcdm-uat-p.ncifcrf.gov/>

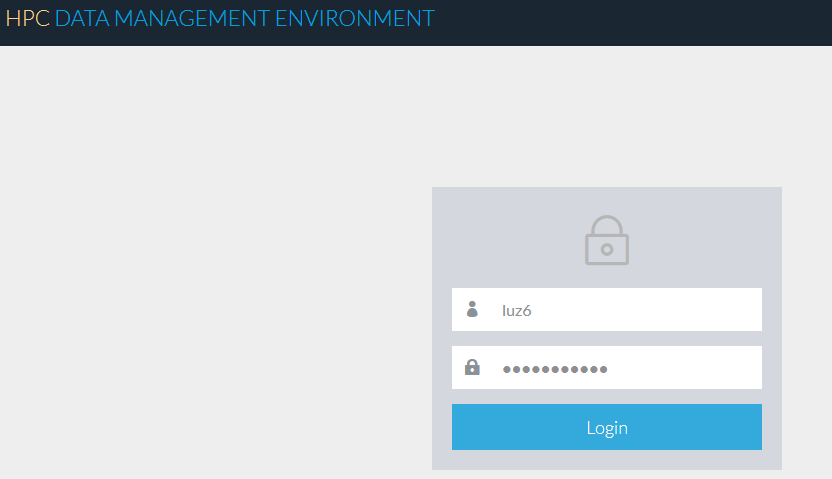
HPC DME Production Web GUI link for production use:

<https://hpcdmeweb.nci.nih.gov>

## Dashboard and login/logout

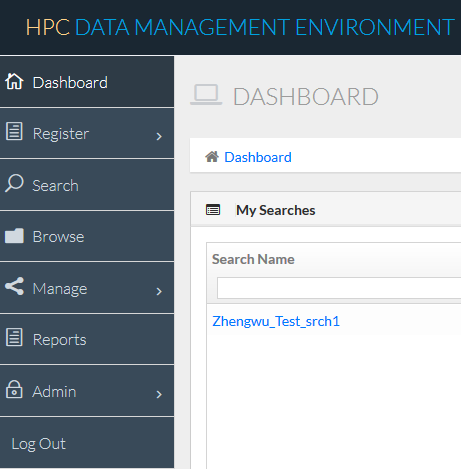
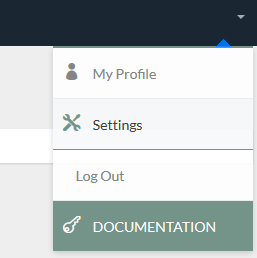
### 7.1.1 Log In

Users are authenticated using the NIH ActiveDirectory – Only NIH AD user IDs are stored at the HPC DME database for tracking authorization. A subsequent iTrust integration will align HPC DME with NIH Secure identity solutions.



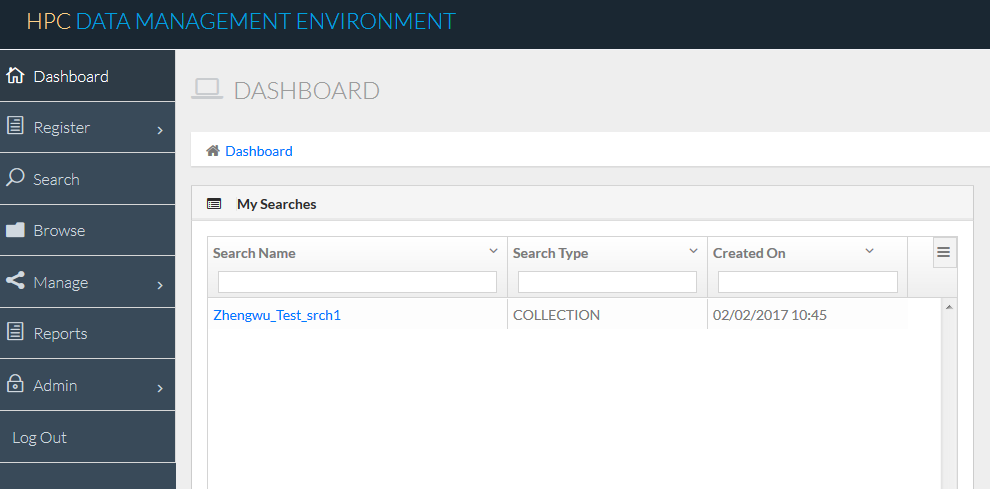
### Log Out

A user may exit out by clicking “Log Out” menu item from left navigational panel or top right corner drop-down menu.

### Dashboard

Default Dashboard shows saved searches and user subscribed events/reports in a listing format for “My Searches” and “My Notifications”, respectively. Clicking any of the listing grid headers will sort the listings in order.



## Search functions

The HPC DME GUI has multiple pages enabling users to enter search criteria and execute targeted searches to identify interested data files/collections. The following summarizes the search capabilities:

1. All text searches are supported case insensitive, i.e., a last name search of “ample”, “AMplE”, or “amPLE” will retrieve any and all users, whose last name in the database is “ample”
2. Wild card “\_” represents any one character and “%” indicates 0, 1 or more characters
3. For group administration only, all records in the domain User or Group tables will be listed if no search criteria are entered for any of the searchable text fields

HPC DME API/GUI client supports the following functions:

### Google Style Search

Users may leave the default entry as “Any” for the “Attribute Name” dropdown, but enter any text at the “Attribute Vale” box for a google style search, either for Collection or Data Object search type. Note: S/he needs to select an operator from the “Operator” dropdown.

### Basic Search

Users may perform a typical [hierarchic](https://www.google.com/search?q=hierarchical&spell=1&sa=X&ved=0ahUKEwif36GIgNPTAhXLQyYKHYdzApgQvwUIJCgA&biw=1280&bih=560) search by entering a level attribute per their DOC organization of collections structure. This will help drill down the list of “Attribute Name” by showing only corresponding attribute names for the level selected earlier. A User may add additional search filter criteria via clicking the button”Add Criteria” and entering his/her search “Level”, “Attribute Name”, “Operator” and “Attribute Value”. Multiple search parameters will be combined with “and” to further nail down the list of search results for either search type “Collection” or “Data Object”.

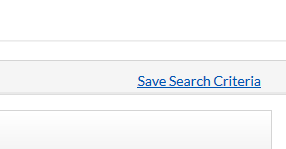
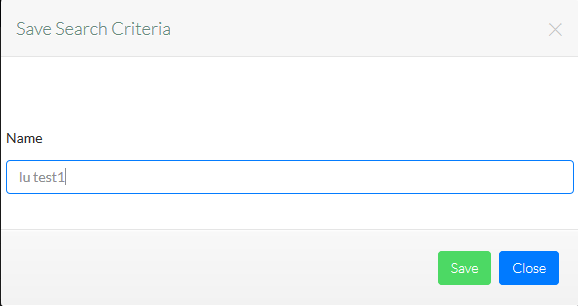
### Advanced Search

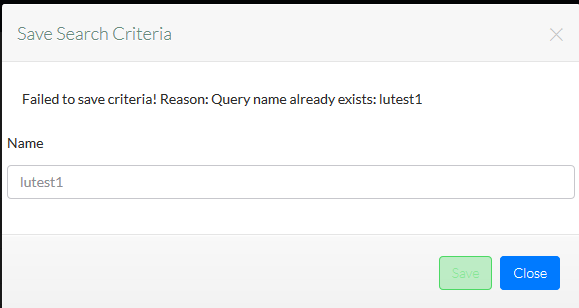
Users may perform an advanced search by typing the previously entered search ID at the basic search section such as “A1”, “A2”, “A3” but entering his/her own desired filtering logic operator “AND”, “OR”. Users may also use parentheses to indicate the filtering order/sequence. For instance, the expression “(A1 OR A2) and (A3 or A4)“ will return any search results meeting both criteria “(A1 or A2)” and “(A3 or A4)”, where the filtering inside the parentheses will be assessed first.

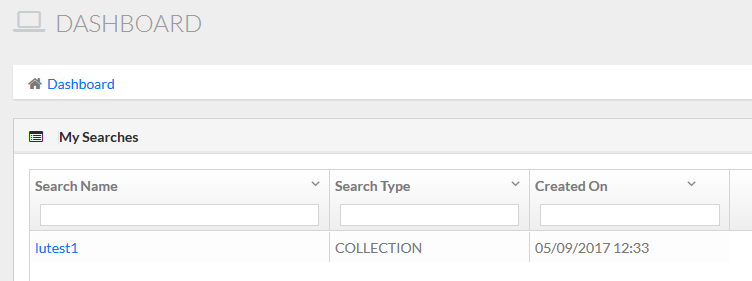
### Saved Search

Users may save an executed search by clicking the hyperlink “Save Search Criteria” and then typing a name for the search for late retrieval. Users may get a prompt message indicate “Query name already exists” if they are creating the same search name. Any saved searches will be displayed at the home page “Dashboard My Searches, where the Search Name, Search Type and Created On (Date/Time) stamps are listed for information purpose.

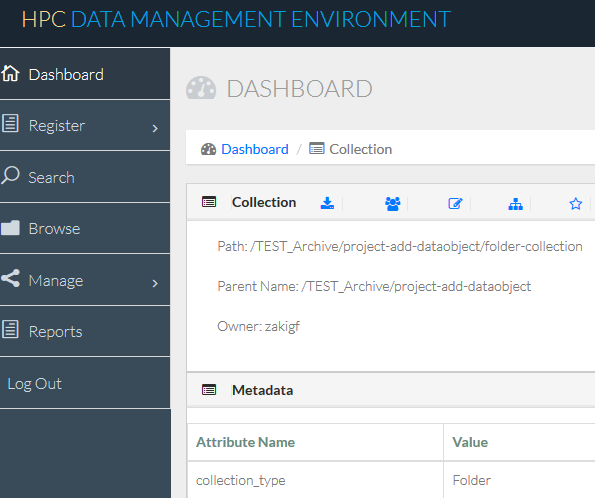
The Save Search feature may be enhanced further to include capability of editing(deleting) and sharing among users.

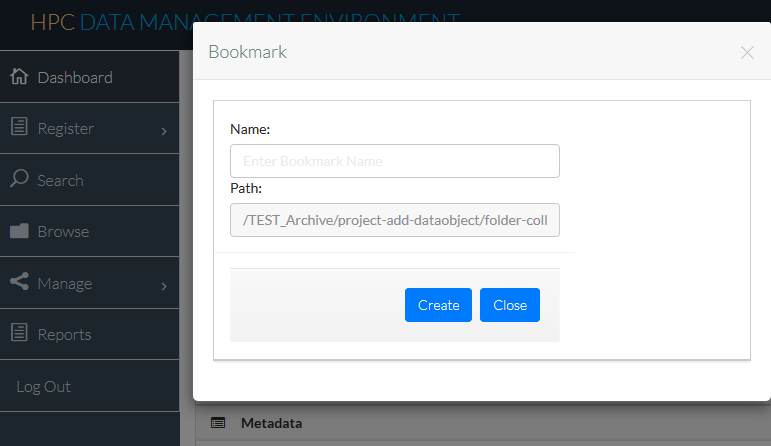
 

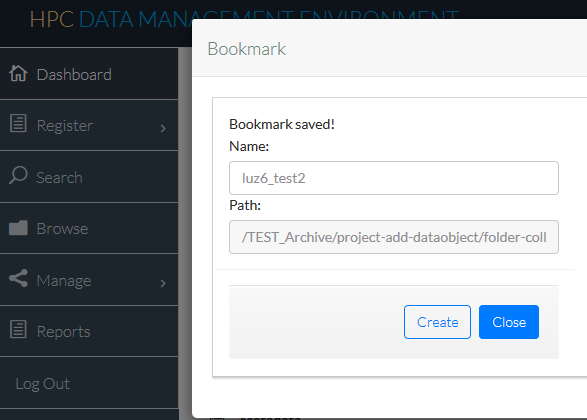




### Bookmark Function

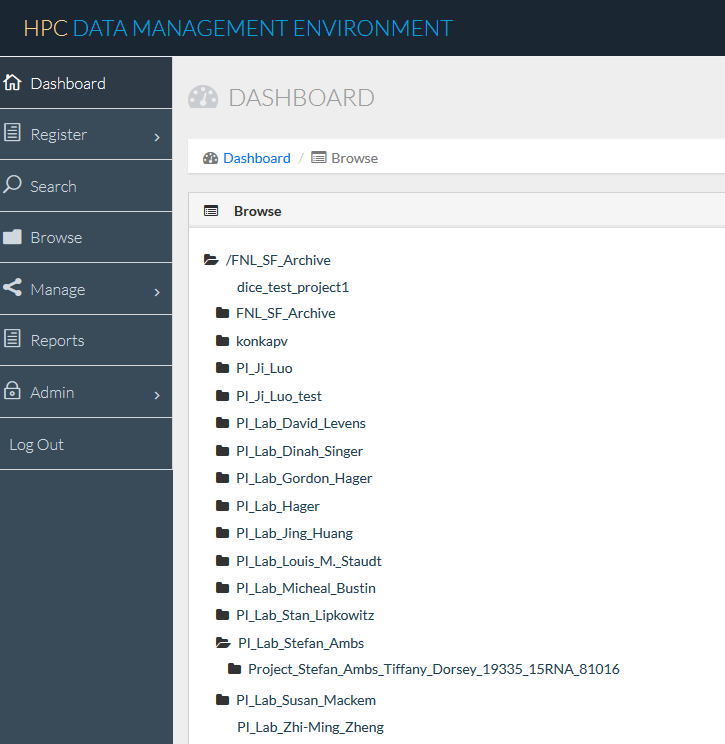
Users may bookmark any type of search results and enter a bookmark name for easy later retrieval or use the browse feature to drill down the specific branch tree per their own authorizations designated. To perform such function, a user needs to select the “Star” icon at the details page of a collection or data file. S/he enters a name for his/her collection or data file path, and then enter Create to have the bookmark saved and tracked in the database. Note: Bookmark name is not allowed to have empty space between words but may be concatenated with “\_” or “-“



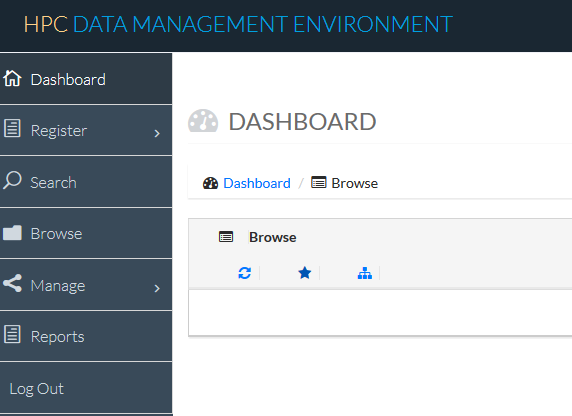


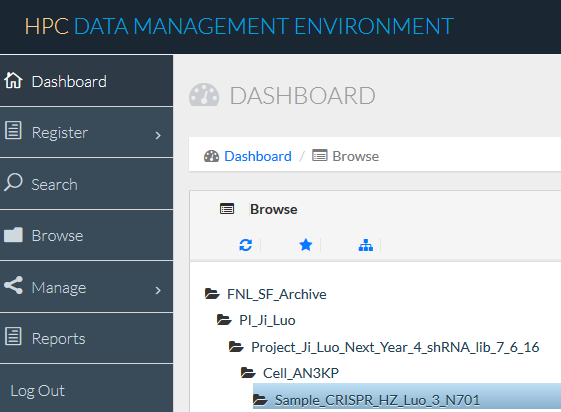
## BROWSE functions

The Browse menu offers a user to navigate their DOC data/metadata hierarchies through a tree-branch expandable user interface and allows users to drill down through the data organizational layers unique to their own DOC policy.

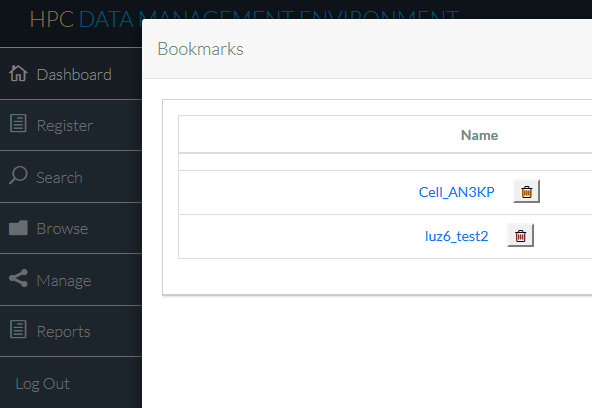


If a user does not have a root access to his/her own DOC or other groups’ base path but wishes to browse through navigating children collections (directories) s/he has access to, s/he may use the Bookmark function to bookmark the collections URL first via Search menu, select the Bookmark icon in the Browse menu as indicated below, identify the bookmark to drill into from the listing table, and then click the Name for the intended bookmark.





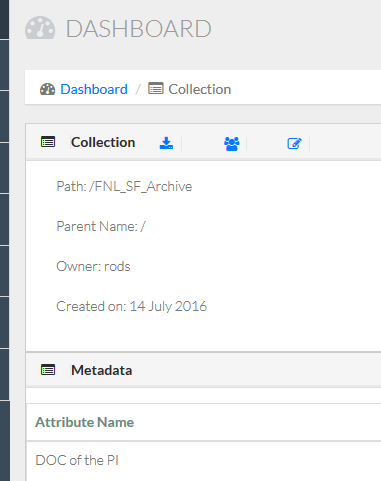
Please note that saved bookmarks may be deleted from the Browse menu if they are no longer required by clicking the “Delete”icon at the right side of the bookmark name as indicated below..



## SHARING functions

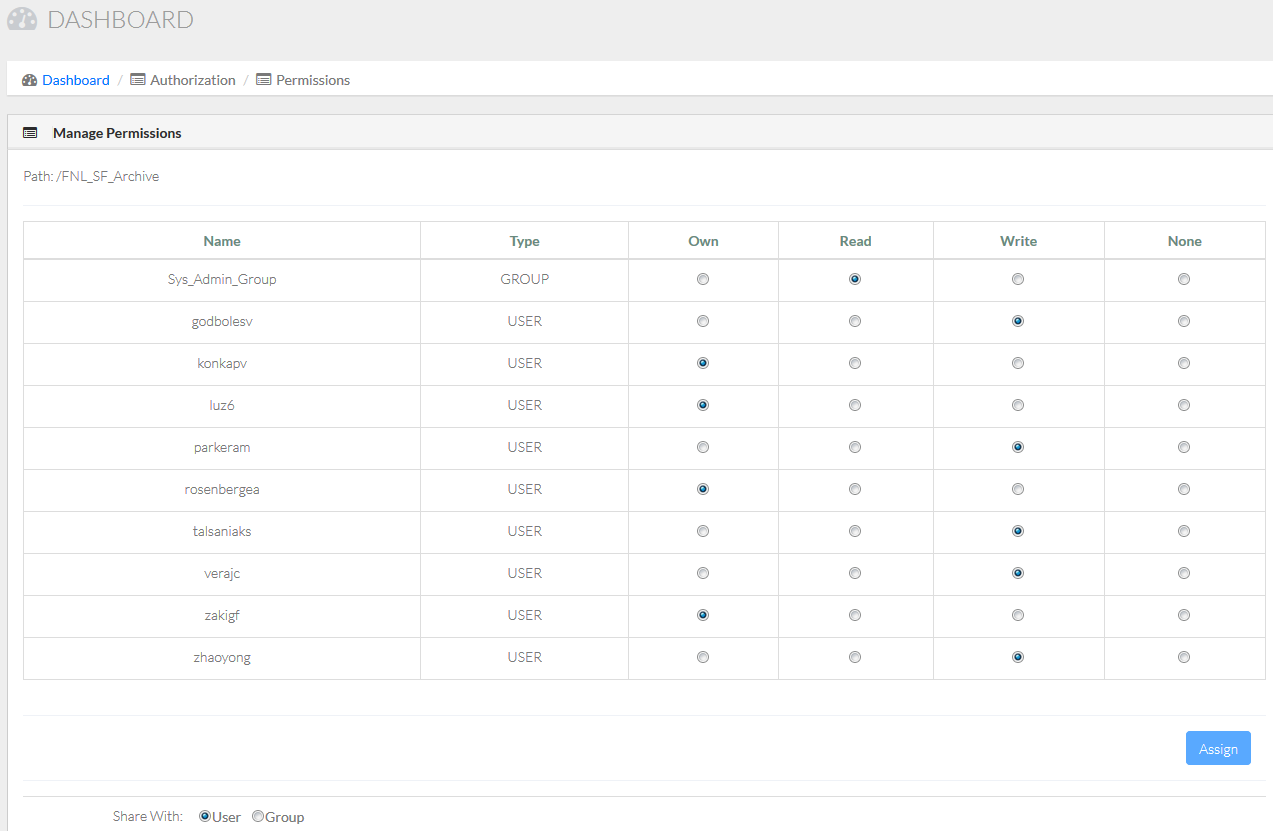
The sharing function is a critical part of the HPC scientific data management and collaboration at NCI. The emphasis and workflow relies on the fact that each data owner/producer wants to decide how they share their data. Check out the HPC DME Web GUI -Download, Sharing and Update Metadata icons at below snapshot.

***HPC DME Web GUI -Download, Sharing and Update Metadata icons***



A data owner may assign who (either via an individual user account or group membership) can own, write, read or NONE certain collections or data files through the GUI Browse or Search menu items. The middle icon in the above figure along the “Collection” indicates that the login user may be authorized to assign access or to have read permissions on the path specified. If a user has read permissions only, s/he will not be able to select/unselect corresponding radios of permissions.

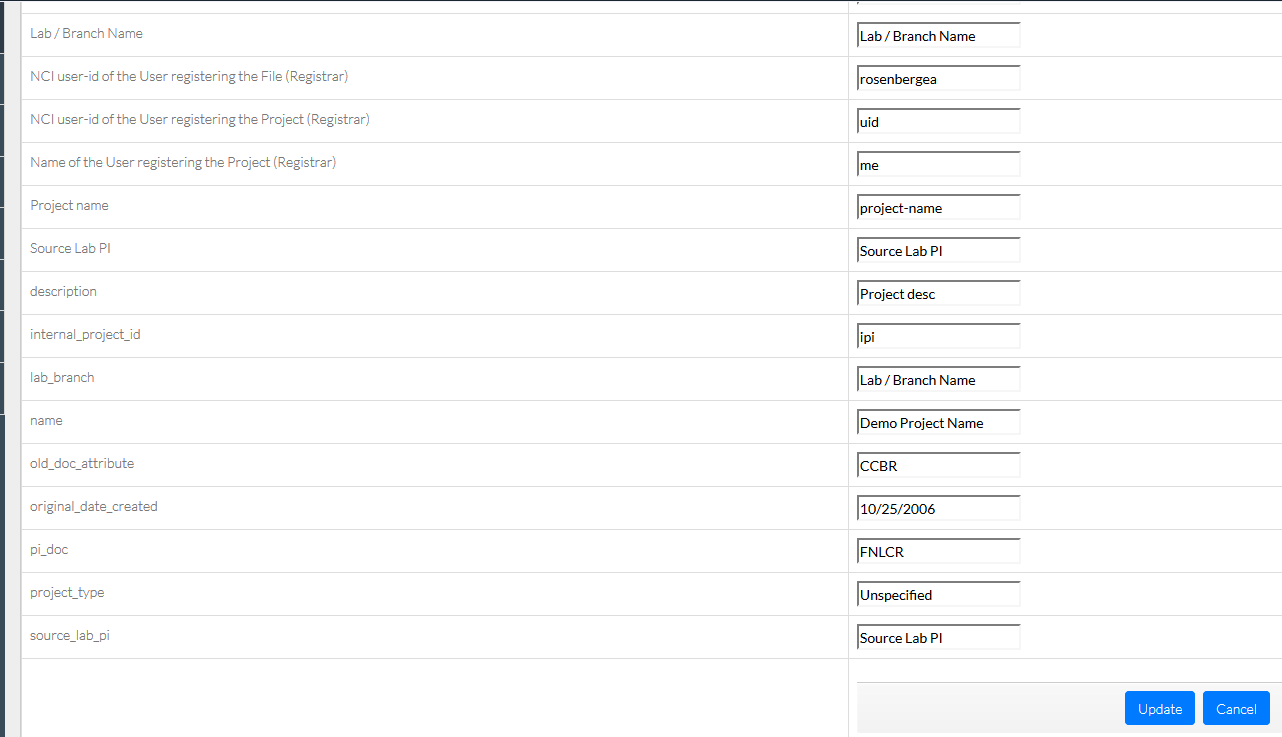
Clicking that icon takes the user into the below “Manage Permissions” figure of showing all the existing users and/or groups with either of the following privileges: Own, Read, Write, and None. From here, the user may assign a user or group either of the 4 permissions to the indicated collection path or data file (object).



## Update metadata in non-batch mode

The web GUI allows any users, through either individual user account’s authorization or group membership, to be able to make update to metadata attribute values. The right icon at previous figure HPC DME Web GUI -Download, Sharing and Update Metadata icons along the “Collection” indicates that the login user is authorized to update custom metadata attributes on the path specified.

Clicking that icon takes the user into the below “Update Metadata” page showing all enabled text box entries of existing custom metadata attributes for value update. From here, the user may enter any new values for their interested metadata attributes and click the “Update” button at bottom right to commit the transaction/update on the backend database. Please note that certain system metadata attributes such as CheckSum is for display purpose only Nobody will have permission to perform update on these system generated metadata attributes.



## Download Collection/file from archive

## Share a Globus End Point with HPC DME Service Account

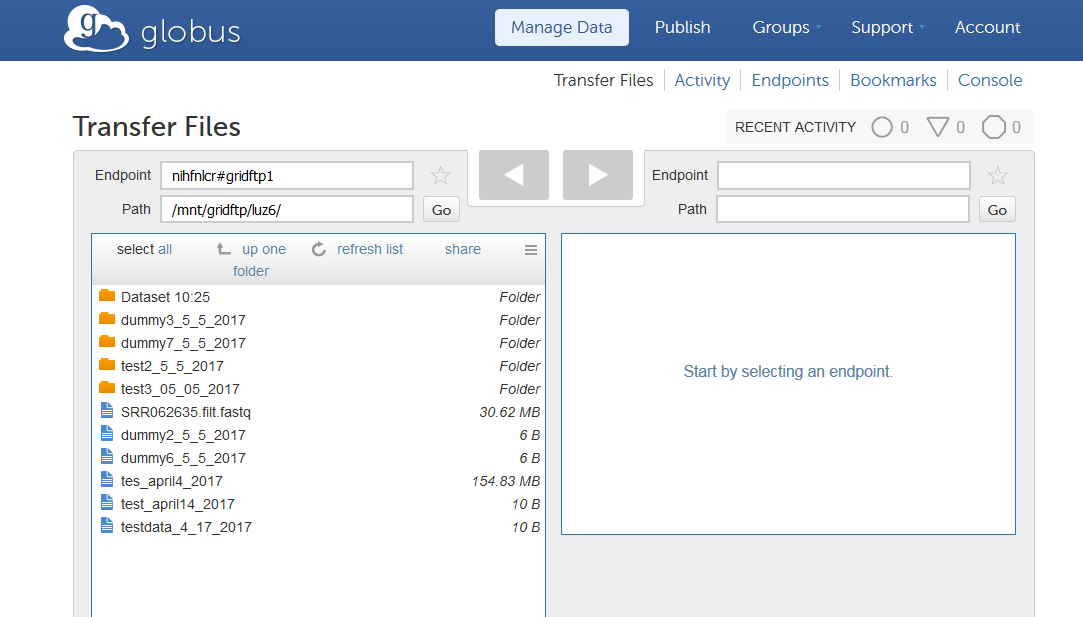
When a user uploads a large data file into archive from an existing Globus end point, the actual data transfer may be performed asynchronously via s system service account. The user needs to grant the service account read access to their source end point before the upload request is submitted. Similarly, a user needs to grant the service account write access to their target Globus end point before a collection or file download request is submitted. When a user downloads a large data file from an Archive into the target Globus end point, the actual data transfer may be performed either asynchronously or synchronously via Globus s system service account.

***HPC DME System Service Accounts Summary Information:***

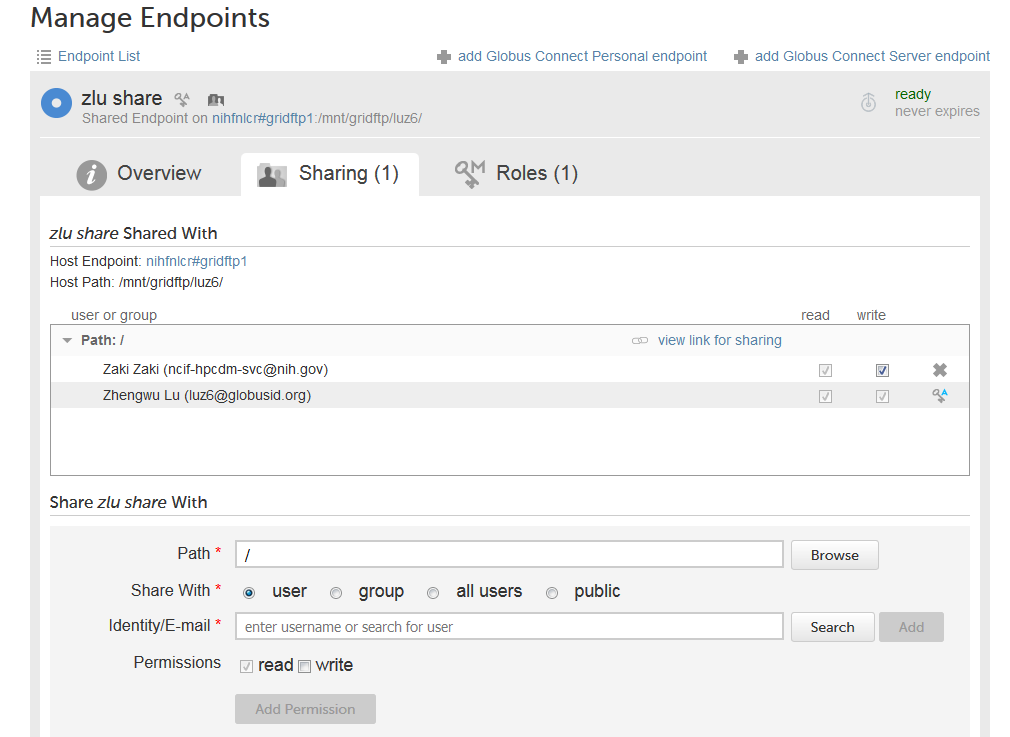
|  |  |  |  |
| --- | --- | --- | --- |
| Service Account | POC Name | POC email | Applicable HPC DME Tier (s) |
| ncif-hpcdm-svc | Zaki Zaki | George.zaki@nih.gov | Development/UAT/Training |
| ncifhpcdmsvcp | Tiru Paalepu | Tiru.paalepu@nih.gov | Production only |

The following summarizes the key procedures for a user to share a file or collection(folder/directory) with the HPC DME service account:

1. Create a Globus share of an existing directory, where you want to share with the Service Account through Globus “Transfer File” menu.



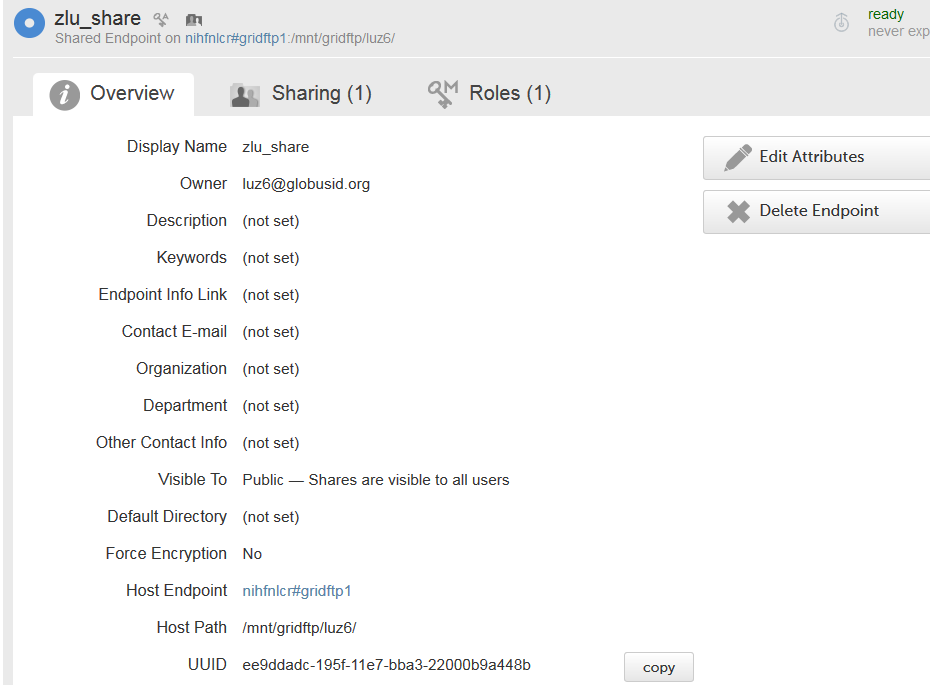
1. Share with HPC DME Service account through Globus “Manage Endpoints” Sharing tab following the creation of the share.

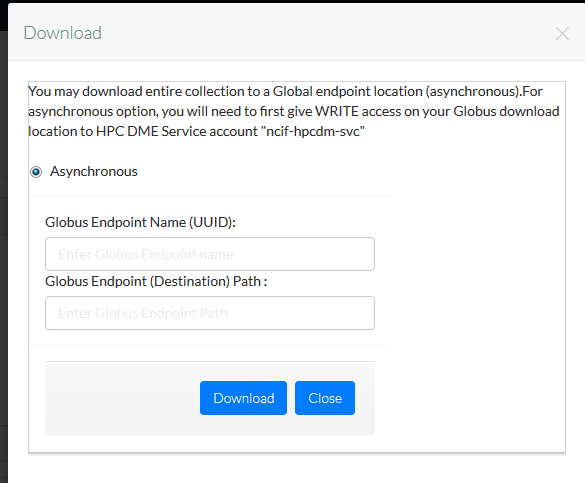


1. For “Path”, you may leave as is to share the original directory or enter/browse to a subdirectory. Unfortunately, Globus does not allow sharing a file or files directly.
2. For “Share With”, you want to check the radio button “user”
3. For “Identity/E-mail”, you want to enter or search the corresponding HPC DME service account email depending on whether you are running the solution at UAT or Production
4. For “Permissions”, you will select the default “read” or check the “write” permission: Uploading a file/collection to an archive requires granting read access to the source end point (s) while downloading from Archive requires granting write access to the service account on the target end point (s).

## Download a Collection of Data Files

The web GUI allows any users, through either individual user account’s authorization or group membership, to be able to download a collection of files (via Globus asynchronous transfer) or a single file (via either synchronous or asynchronous transfer). When performing a Globus transfer, a user may copy the UUID from the target endpoint Globus “Overview” tab UUID field and paste that as the input value for “Globus Endpoint Name (UUID)” at the Download dialog. S/he needs to manually enter the entry for “Globus Endpoint (Destination) Path” at the prompt dialog. The collection of files will be transferred to the target Globus endpoint with a directory name created using what is entered. The left icon at Figure HPC DME Web GUI -Download, Sharing and Update Metadata icons along the “Collection” indicates that the login user is authorized to download the selected collection (of data files) to a Globus target end point.

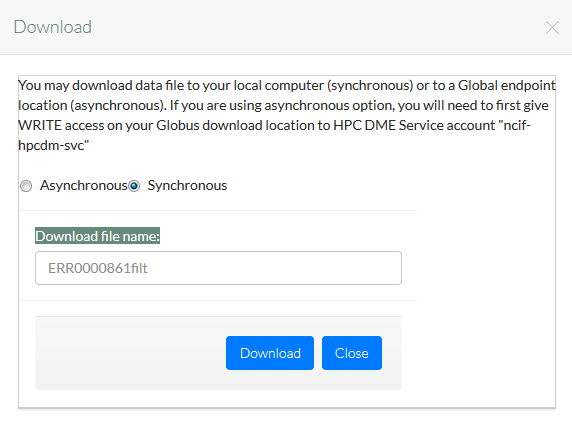


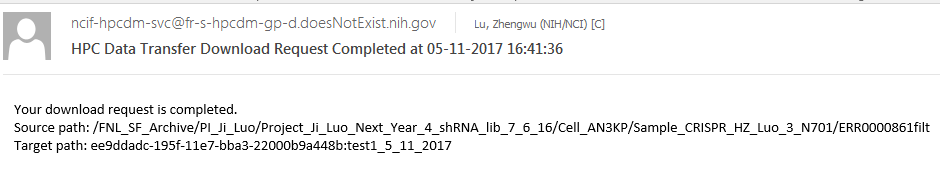


## Download a Single Data File

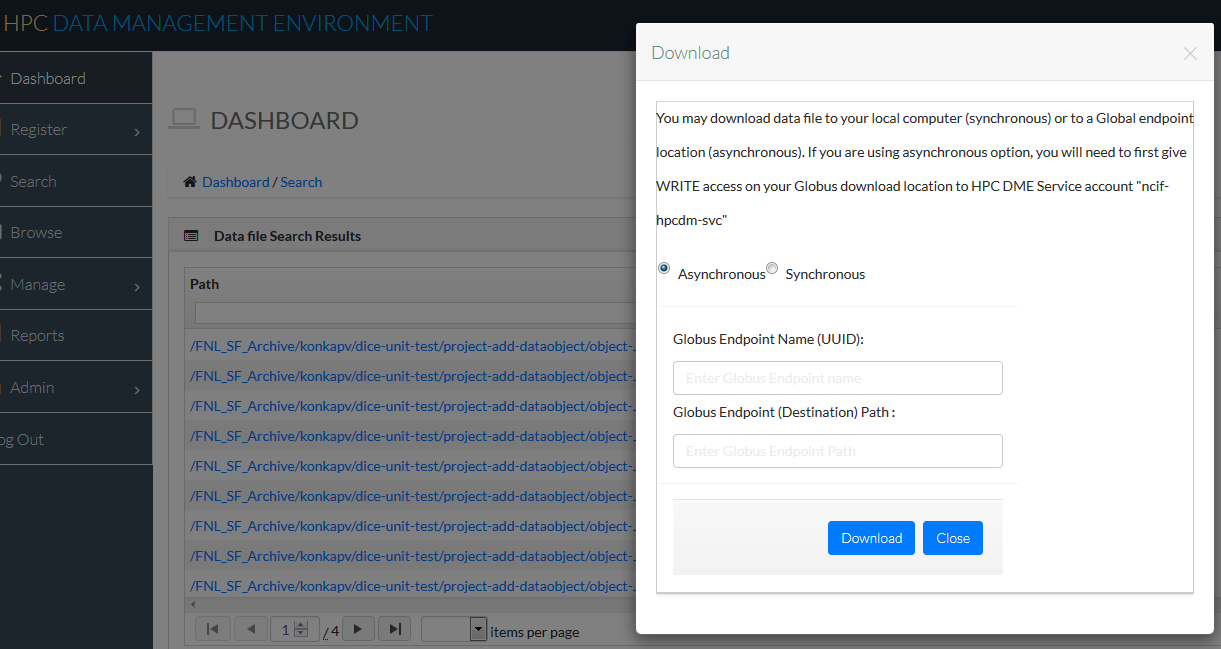
To download a single file, you want to navigate through the HPC DME Search menu and check Search Type “Data Object”. After executing a search, you click the “Download” icon on the adjacent file to initiate the download action. Subsequently, you decide how to download the data file: Synchronous download to local share or asynchronous to a Global endpoint location.

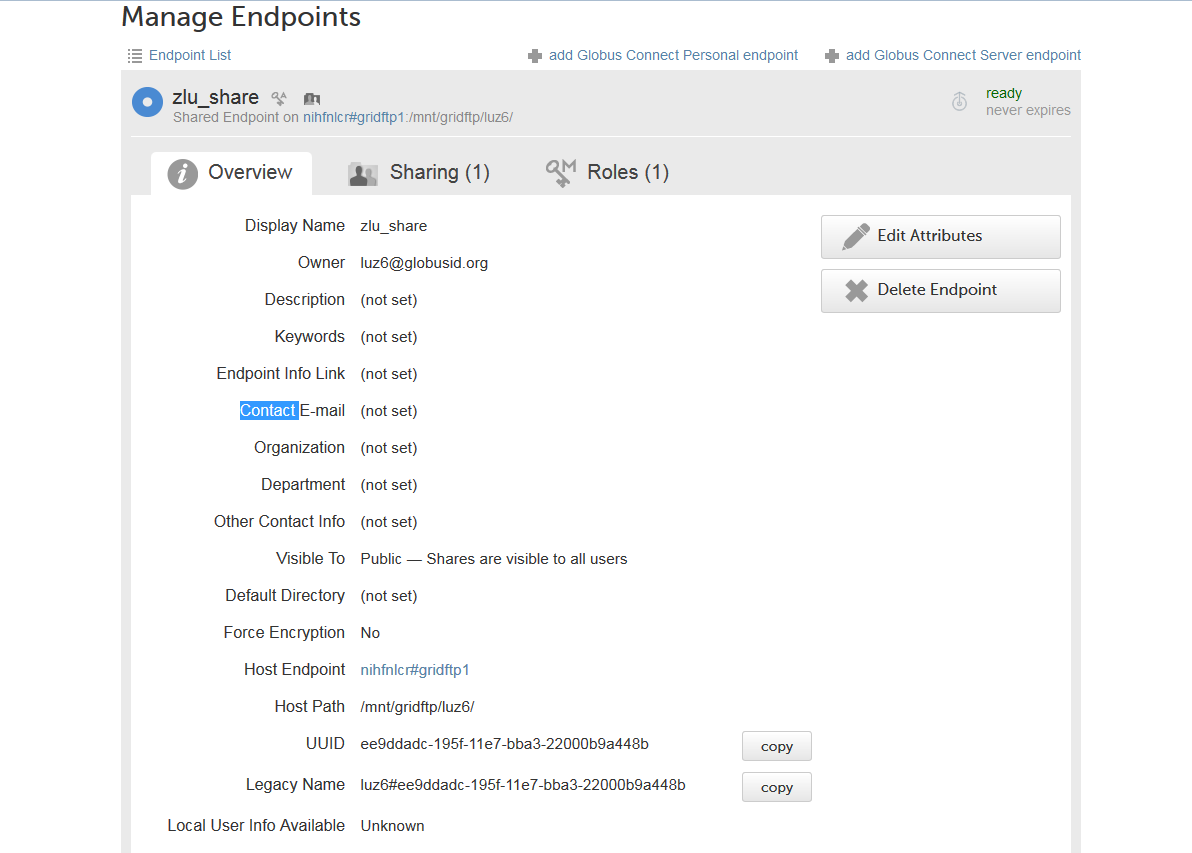
For synchronous download(s), a user has the option of adopting the existing file name or typing over to use a different name. Once the file name is entered, clicking download will transfer the data object from Archive into your local drive/share using the determined file name. If s/he has subscribed to the event type “Data Transfer Download completed”, the user will be informed by email after the download is either completed or runs into failure with relevant contextual information of the collection downloaded.





If a user uses asynchronous option, s/he will need to first give WRITE access on their Globus download location to HPC DME Service account beforehand. The user may copy the UUID from the target endpoint Globus “Overview” tab UUID field and paste that as the input value for “Globus Endpoint Name (UUID)” at the Download dialog. S/he needs to manually enter the entry for “Globus Endpoint (Destination) Path” at the prompt dialog. The single file will be transferred to the target Globus endpoint with the file name created using what is entered. If s/he has subscribed to the event type “Data Transfer Download completed”, the user will be informed by email after the download is either completed or runs into failure with relevant contextual information of the file downloaded.

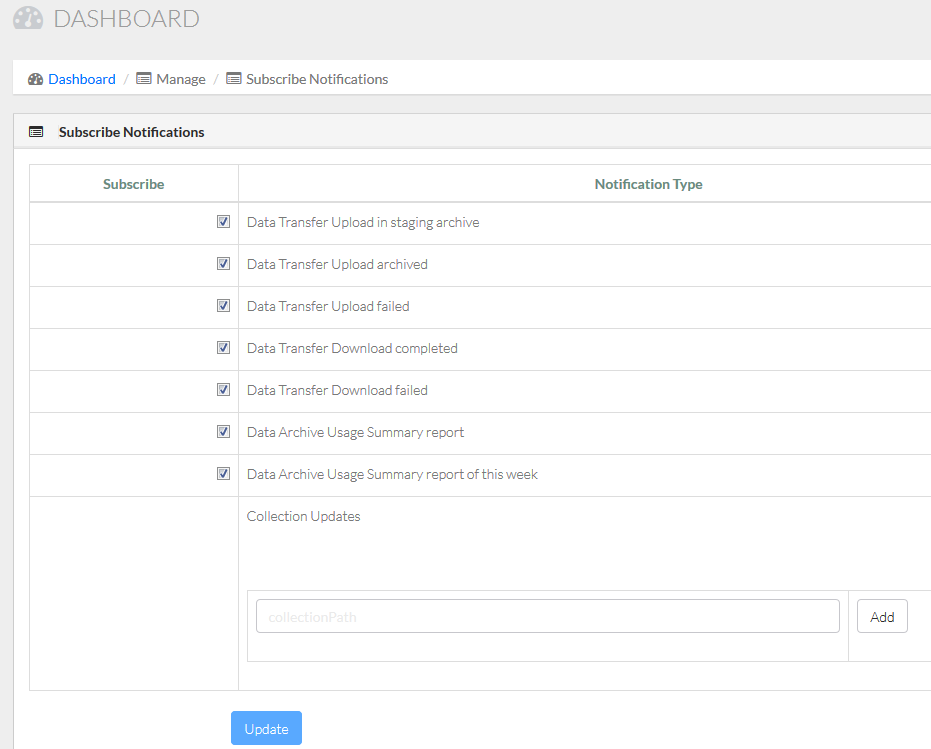
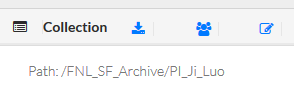




## Notifications and reports

### Notifications Management

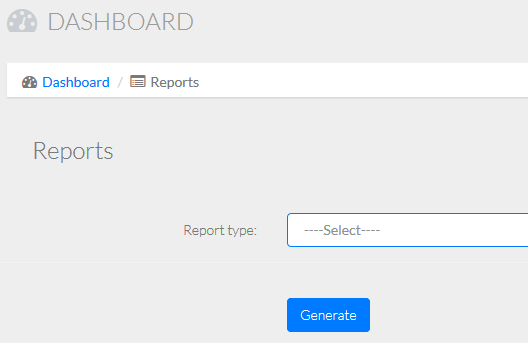
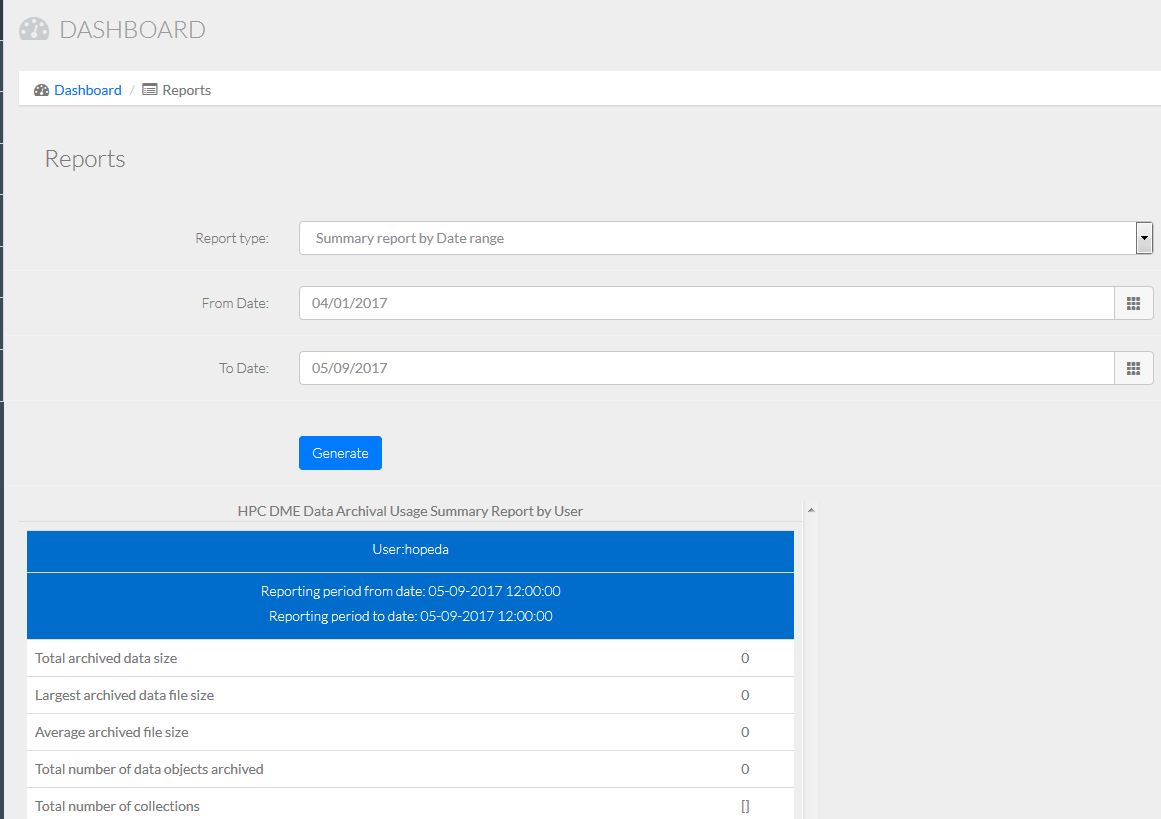
A user manages his/her own notifications by subscribing/unsubscribing to the different event types related to data transfer (upload/download), reporting, or data/metadata updating occurring on their interested collection (s). If a user subscribes to a parent collection for notification, any updates occurring on the parent and all its children collections will be informed due to the inheritance feature implemented. To subscribe/unsubscribe to an event type, a user needs to check/uncheck the adjacent checkbox. To subscribe to a collection update of your interest, you want to browse, navigate to the collection details page, copy what is after the “Path:” into the clipboard, and then paste this path into the “collectionPath” text box. You finish the transaction by clicking Add and then Update buttons on the “Subscribe Notifications” page. To unsubscribe, you want to uncheck those collection paths no longer interested, then click Update to commit

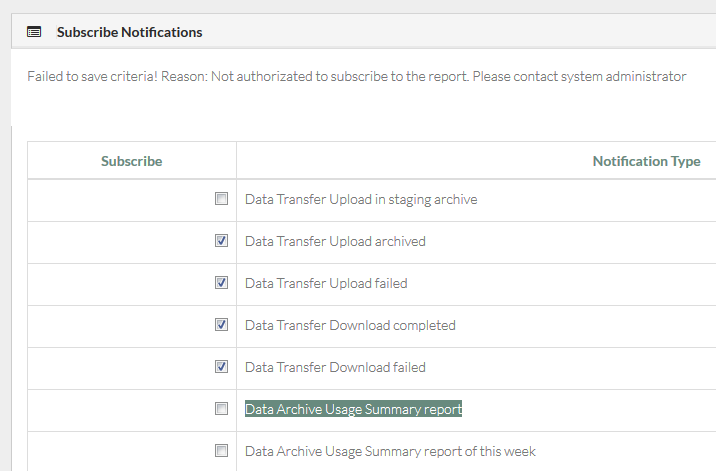
 

### Reports Management

A user may access existing reports through Report menu and pick a report from the following Report types: Summary Report; Summary Report by Date range; User Report; and User Report by Date range. A report dialog will open up prompting for a user selection or entry if a User Report is initiated or for a date range entry if either Data range report is in action (Summary or User Report). You present the result by clicking the Generate button on the report dialog page. Additional reports may be requested and custom developed/deployed in future.

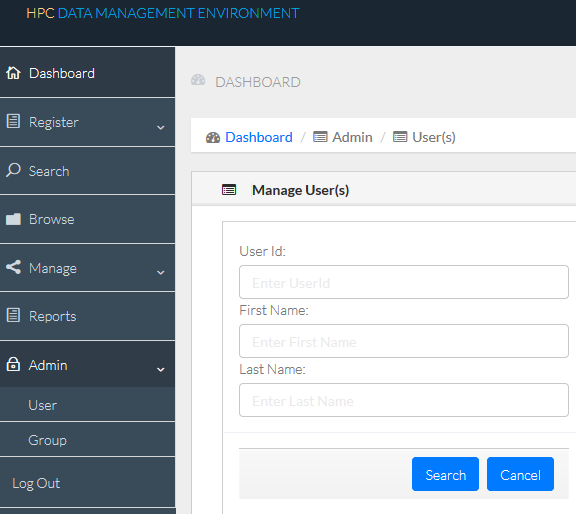
Note: Two reports “Data Archive Usage Summary report” and “Data Archive Usage Summary report of this week” are developed for system administrators only, for which non-system administrator users will get error prompt if they try to subscribe to them.



## Group administrator functions

The web GUI allows a group administrator to perform basic user/group account management functions such that each group has their own administrator to manage common account issues and has full control on how their data/metadata needs to be annotated, shared and controlled in a consistent manner. This will also alleviate the burden for a system administrator without much domain knowledge about each DOC(group)’s business operations.

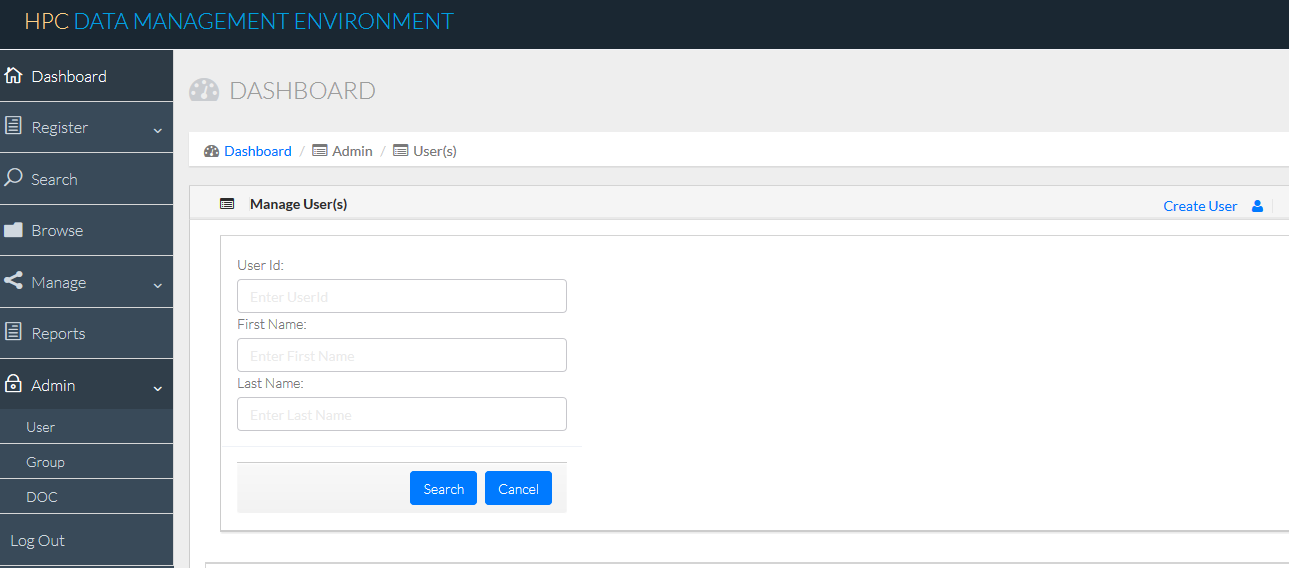


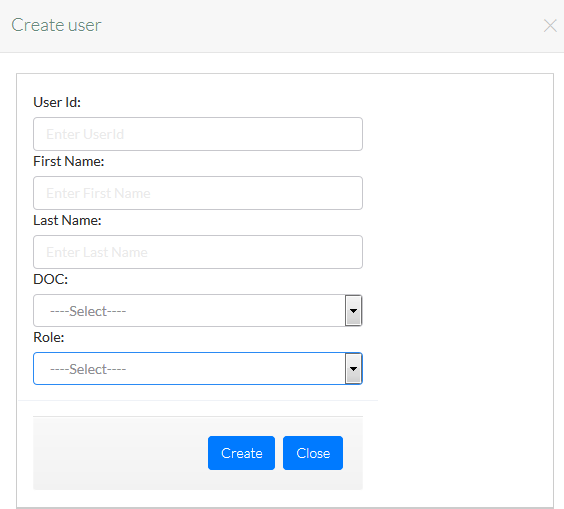
### User

A Group Administrator accesses the Admin User menu to add a new user for his/her group.

The Create user action can be triggered by clicking the hyperlink “Create User” on Admin Manage User(s) page. On the following user entry form, the administrator wants to ensure that:

1. NIH login user ID is entered for the User Id text entry field
2. The user’s actual first name is entered for First Name text entry
3. The user’s actual last name is entered for Last Name text entry
4. The actual DOC (group) value will be pre-populated on the DOC dropdown using the DOC the administrator is associated with
5. For the Role dropdown, either USER or “GROUP\_ADMINISTRATOR” may be selected or entered depending on the actual need





An administrator can search existing users using the Search button. All common search features/capabilities are supported.

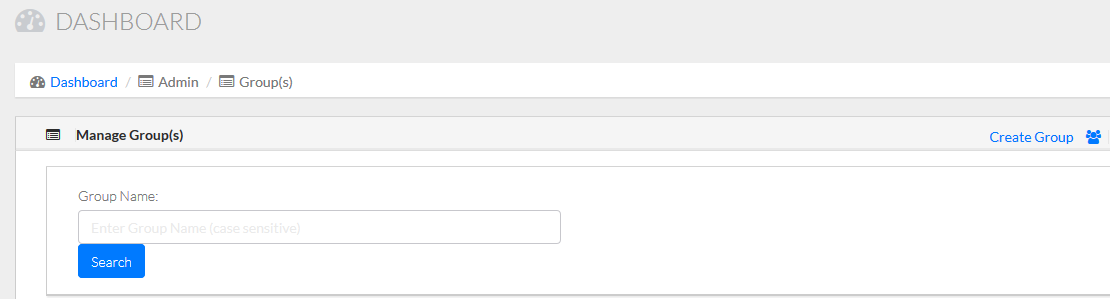
### Group

A Group Administrator accesses the Admin Group menu to add a new user group for his/her own DOC (Lab/Group/Operational Unit) business operational needs. The action of Create group

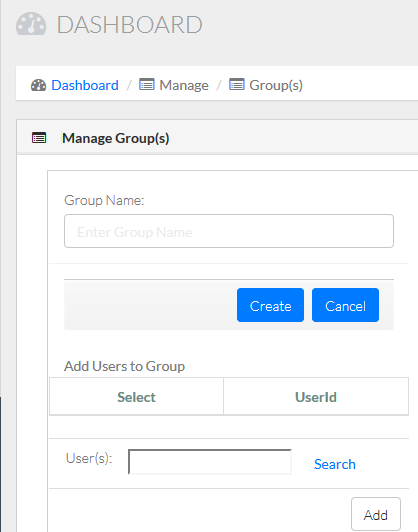
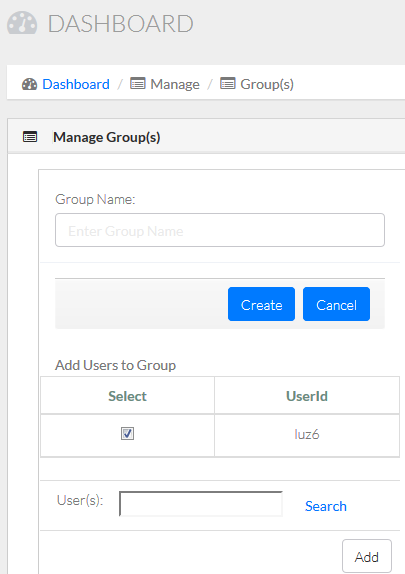
can be triggered by clicking the hyperlink “Create Group” on Admin Manage Group(s) page. On the following group entry form, the administrator wants to ensure that a meaningful user group name is determined and entered. Please note that the group name is not allowed to have empty space between words but may be concatenated with “\_” or “-“. A user group name of “Dummy test” with a pace in between “Dummy” and “test” will not be accepted; However, name entry of “Dummy\_test” will be accepted and transacted. Moreover, a user group name will be unique and will not be allowed to be edited once it is created. Therefore, the Update function will be solely for changing membership associated with the group.

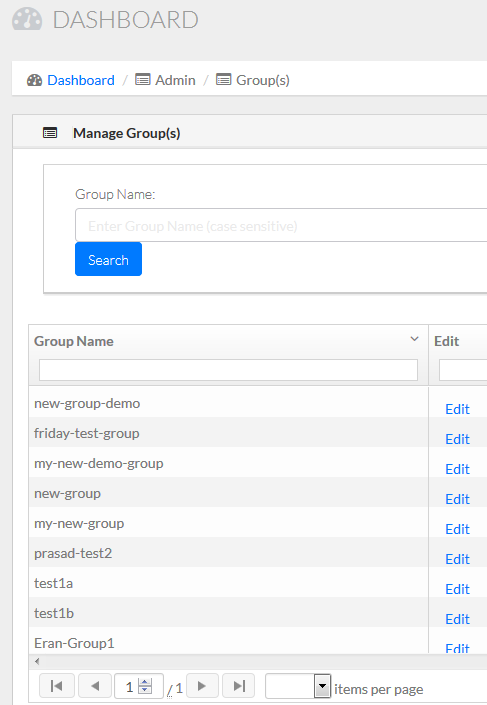
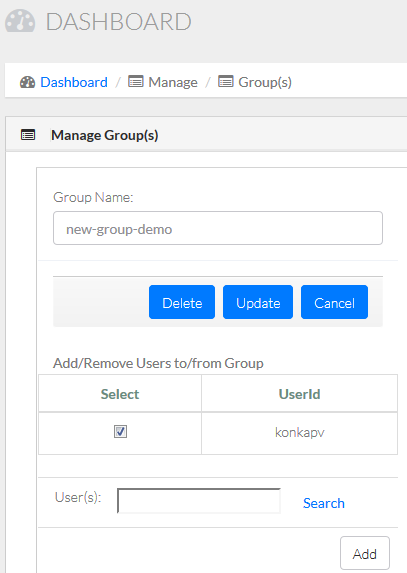
To maximize the benefits of adopting this feature, creation of a user group needs to be focused on the common authorizations desired to be in place on a group level. It will be advisable to create a user group, grant read access to this group, and then add individual users to the group if an administrator wants to share read access to some external staff/collaborators.

1. When creating a user group, an administrator may add the new group without associating a user member. However, s/he may add a user or users to the new group and create the group in one shot.
2. To add/remove a user or users to an existing user group, the administrator will search and locate the Group to edit first. When editing the Group selected, s/he will search and assign all interested users first to the grid list under “***Add/Remove Users to/from Group***”. Lastly, the Administrator will check/uncheck the checkbox adjacent to the user(s) to be associated/removed from the Group. A click on the Update button then will commit the transaction to update the group membership eventually.



Similarly, an administrator can search existing user groups using the Search button. All common search features/capabilities are supported.

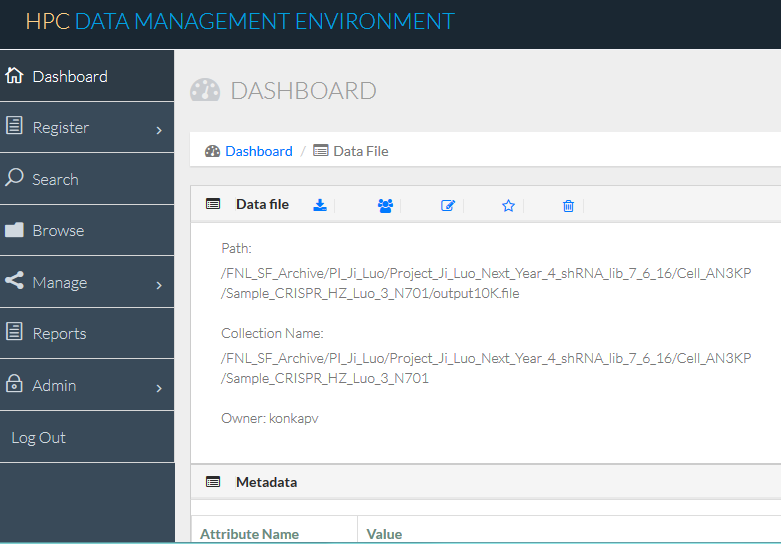
 

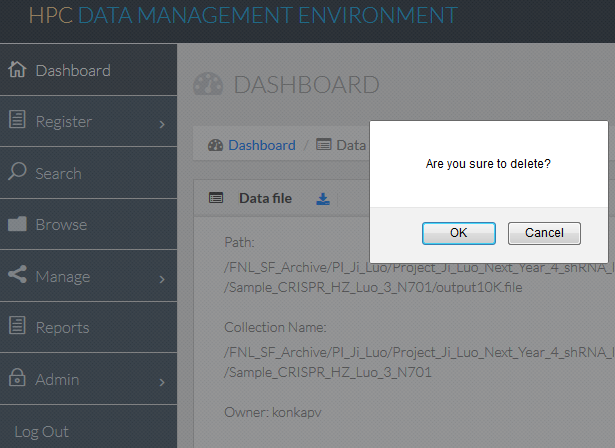
 

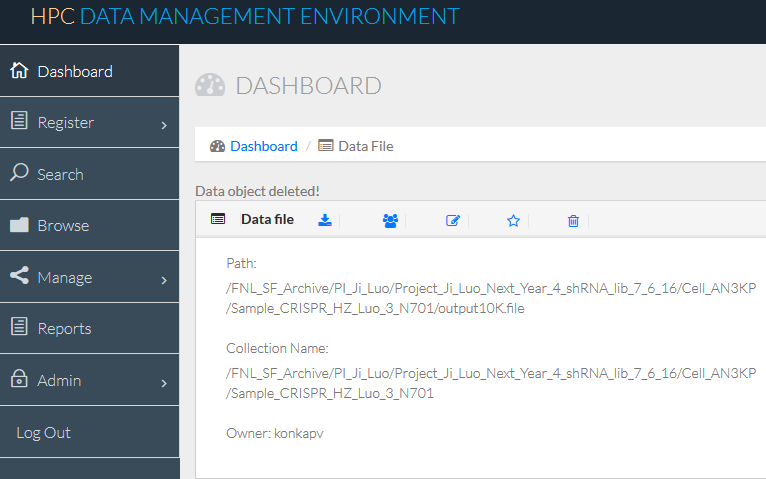
### Delete a Data Object and Associated Metadata

Sometimes though occurring not frequently, a user may end up uploading the wrong data files into the Archive. There would be a need to delete these data files and associated metadata while the integrity and such usage is exercised to the minimum extent. At this point, we will give such delete privilege to the Group Administrator role only. When a group admin selects this delete, the data object from Archive and all the associated metadata from iRODS backend database will be deleted physically. Therefore, a group administrator is advised to perform such delete operation only when s/he is absolutely certain that such data files are entered in by mistake and need to be deleted permanently.

The action of Create group can be triggered by clicking the most right“Delete” icon on the top of [Dashboard](https://fr-s-hpcdm-gp-d.ncifcrf.gov/dashboard)/Data File page for the selected data object to be deleted permanently. On the following confirmation dialog, the group administrator will select “OK” to proceed or “Cancel” to back out. After confirming the operation, the Web GUI performs the transaction via the HPC DME API and indicates that data object deleted with the collection (directory), file path, and owner presented on the screen.







# APPENDIX A – sample collection input

[https://github.com/CBIIT/HPC\_DME\_APIs/tree/master/utils/hpc-client/samples/Batch\_Upload\_Input\_File\_Collections.csv](https://ncisvn.nci.nih.gov/svn/HPC_Data_Management/branches/hpc-prototype-dev/src/hpc/hpc-irods-client/samples/Batch_Upload_Input_File_Collections.csv)

# APPENDIX B – sample data file input

[https://github.com/CBIIT/HPC\_DME\_APIs/tree/master/utils/hpc-client/samples/[Batch\_Upload\_Input\_File\_Object\_Cleversafe.csv](https://github.com/CBIIT/HPC_DME_APIs/blob/master/utils/hpc-client/samples/Batch_Upload_Input_File_Object_Cleversafe.csv)](https://ncisvn.nci.nih.gov/svn/HPC_Data_Management/branches/hpc-prototype-dev/src/hpc/hpc-irods-client/samples/Batch_Upload_Input_File_Object.csv)

[https://github.com/CBIIT/HPC\_DME\_APIs/tree/master/utils/hpc-client/samples/[Batch\_Upload\_Input\_File\_Object\_Globus.csv](https://github.com/CBIIT/HPC_DME_APIs/blob/master/utils/hpc-client/samples/Batch_Upload_Input_File_Object_Cleversafe.csv)](https://ncisvn.nci.nih.gov/svn/HPC_Data_Management/branches/hpc-prototype-dev/src/hpc/hpc-irods-client/samples/Batch_Upload_Input_File_Object.csv)

# Appendix c – sample permissions input

[https://github.com/CBIIT/HPC\_DME\_APIs/tree/master/utils/hpc-client/samples/Batch\_Permissions.csv](https://ncisvn.nci.nih.gov/svn/HPC_Data_Management/branches/hpc-prototype-dev/src/hpc/hpc-irods-client/samples/Batch_Permissions.csv)