**High Performance Computing (HPC) Data Management Environment (DME) Getting Started: DOC Group Administrators**

**Purpose**

This document is intended to serve as a guide for getting started on using the High Performance Computing (HPC) Data Management Environment (DME) for the National Cancer Institute (NCI). Specifically, this guide is intended to help the reader understand the duties of a DOC Group Administrator, should the reader be designated such role in facilitating a DOC’s utilization of the HPC DME.

**General Duties**

A DOC Group Administrator (Group Admin) is expected to be a knowledgeable user of the HPC DME, serving as a technical leader and guide in his/her respective DOC. Therefore, he/she plays a critical role in the adoption and ongoing utilization of the HPC DME by others who belong to the same DOC.

One of the responsibilities of a Group Admin is creation and/or maintenance of the metadata policies for their DOC. The metadata policies are rules and conventions that must be followed by HPC DME users to put or update content in the DME. Within the DME, one DOC may be designated one or more base paths, and for each base path, there may be one metadata policies set in effect. The set is divided into 3 parts. The first is the **data hierarchy specification**. It defines the permitted structure and organization of Collections under the applicable base path. The second part of the metadata policies set applied to one base path is the **Collections metadata validation rules**. It contains definitions of the expected metadata accompanying Collections and sets rules such as which metadata attributes are mandatory and what are acceptable values for a metadata attribute. The third part of the metadata policies set associated with one base path is the **Data Objects metadata validation rules**. This is the counterpart to the Collections metadata validation rules, but it is within the scope of Data Objects instead of Collections.

It is strongly recommended that every DOC utilizing the DME has comprehensive metadata policies sets for all its base paths in the DME. With well-devised metadata policies sets in effect, the DME users of a DOC shall find that Data Objects they register in the DME shall be easier to index, search, and organize, and treatment of Data Objects can be more consistent and uniform as they all conform to the same rules concerning metadata.

Currently, the mechanism for creating or updating metadata policies sets is a manual procedure that an HPC DME System Administrator (Sys Admin) executes. Each metadata policies set is composed as 3 JavaScript Object Notation (JSON) documents for the 3 parts of a policies set as described previously: **data hierarchy specification**, **Collections metadata validation rules**, and **Data Objects metadata validation rules**. If a DOC Group Admin is comfortable working with JSON, he/she could, without the aid of a Sys Admin, write or edit these JSON documents. Any text editor application could be used for this purpose. However, it is outside the scope of this document to offer detailed instructions on how to write or edit JSON using any text editor application or other type of software product. Refer to the figures in the next section are examples of the 3 parts of a metadata policies set expressed in the expected JSON form and applicable to one base path.

Since metadata of interest is widely variable across different organizations or work teams, an HPC DME Sys Admin would not be able to offer definitive guidance on what should be in any DOC’s metadata policies set. The Sys Admin could assist with the mechanics of creating or updating a policies set, but ultimately, a Group Admin is responsible for producing the content in his/her DOC’s metadata policies set.

Another responsibility of an HPC DME Group Administrator is administering users and groups. Tasks related to this responsibility can be accomplished via the HPC DME web GUI client. For details about that, please refer to the [HPC DATA MANAGEMENT USER GUIDE](https://github.com/CBIIT/HPC_DME_APIs/raw/master/doc/guides/HPC_User_Guide.docx), and read the section about using the web GUI.

**Example JSON Documents**

Figure 1: Example of Data Hierarchy Specification

|  |
| --- |
| {  "collectionType": "Project",  "isDataObjectContainer": true,  "subCollections": [  {  "collectionType": "Dataset",  "isDataObjectContainer": true  },  {  "collectionType": "Run",  "isDataObjectContainer": false  },  {  "collectionType": "Folder",  "isDataObjectContainer": true  },  {  "collectionType": "No\_metadata\_folder",  "isDataObjectContainer": true  }  ]  } |

Figure 2: Example of Collections Metadata Validation Rules

|  |
| --- |
| {  "metadataValidationRules": [  {  "attribute": "collection\_type",  "mandatory": true,  "validValues": [  "Project",  "PI\_Lab",  "Flowcell",  "Sample",  "Folder"  ],  "ruleEnabled": true  },  {  "attribute": "pi\_name",  "mandatory": true,  "collectionTypes": [  "PI\_Lab"  ],  "ruleEnabled": true  },  {  "attribute": "project\_id\_CSAS\_NAS",  "mandatory": true,  "collectionTypes": [  "Project"  ],  "ruleEnabled": true  },  {  "attribute": "project\_name",  "mandatory": false,  "collectionTypes": [  "Project"  ],  "ruleEnabled": true  },  {  "attribute": "lab\_contact",  "mandatory": false,  "collectionTypes": [  "Project"  ],  "ruleEnabled": true  },  {  "attribute": "bioinformatics\_contact",  "mandatory": false,  "collectionTypes": [  "Project"  ],  "ruleEnabled": true  },  {  "attribute": "project\_start\_date",  "mandatory": false,  "defaultValue": "System-Date",  "collectionTypes": [  "Project"  ],  "ruleEnabled": true  },  {  "attribute": "grant\_funding\_agent",  "mandatory": false,  "collectionTypes": [  "Project"  ],  "ruleEnabled": true  },  {  "attribute": "flowcell\_id",  "mandatory": true,  "collectionTypes": [  "Flowcell"  ],  "ruleEnabled": true  },  {  "attribute": "run\_name",  "mandatory": true,  "collectionTypes": [  "Flowcell"  ],  "ruleEnabled": true  },  {  "attribute": "run\_date",  "mandatory": true,  "collectionTypes": [  "Flowcell"  ],  "ruleEnabled": true  },  {  "attribute": "sequencing\_platform",  "mandatory": true,  "collectionTypes": [  "Flowcell"  ],  "ruleEnabled": true  },  {  "attribute": "sequencing\_application\_type",  "mandatory": true,  "collectionTypes": [  "Flowcell"  ],  "ruleEnabled": true  },  {  "attribute": "read\_length",  "mandatory": true,  "collectionTypes": [  "Flowcell"  ],  "ruleEnabled": true  },  {  "attribute": "pooling",  "mandatory": true,  "collectionTypes": [  "Flowcell"  ],  "ruleEnabled": true  },  {  "attribute": "sample\_id",  "mandatory": true,  "collectionTypes": [  "Sample"  ],  "ruleEnabled": true  },  {  "attribute": "sample\_name",  "mandatory": true,  "collectionTypes": [  "Sample"  ],  "ruleEnabled": true  },  {  "attribute": "initial\_sample\_concentration\_ngul",  "mandatory": false,  "collectionTypes": [  "Sample"  ],  "ruleEnabled": true  },  {  "attribute": "initial\_sample\_volume\_ul",  "mandatory": false,  "collectionTypes": [  "Sample"  ],  "ruleEnabled": true  },  {  "attribute": "sfqc\_sample\_concentration\_ngul",  "mandatory": false,  "collectionTypes": [  "Sample"  ],  "ruleEnabled": true  },  {  "attribute": "sfqc\_sample\_size",  "mandatory": false,  "collectionTypes": [  "Sample"  ],  "ruleEnabled": true  },  {  "attribute": "RIN",  "mandatory": false,  "collectionTypes": [  "Sample"  ],  "ruleEnabled": true  },  {  "attribute": "28s18s",  "mandatory": false,  "collectionTypes": [  "Sample"  ],  "ruleEnabled": true  },  {  "attribute": "library\_id",  "mandatory": false,  "collectionTypes": [  "Sample"  ],  "ruleEnabled": true  },  {  "attribute": "library\_lot",  "mandatory": false,  "collectionTypes": [  "Sample"  ],  "ruleEnabled": true  },  {  "attribute": "library\_name",  "mandatory": false,  "collectionTypes": [  "Sample"  ],  "ruleEnabled": true  },  {  "attribute": "sfqc\_library\_concentration\_nM",  "mandatory": false,  "collectionTypes": [  "Sample"  ],  "ruleEnabled": true  },  {  "attribute": "sfqc\_library\_size",  "mandatory": false,  "collectionTypes": [  "Sample"  ],  "ruleEnabled": true  },  {  "attribute": "source\_id",  "mandatory": false,  "collectionTypes": [  "Sample"  ],  "ruleEnabled": true  },  {  "attribute": "source\_name",  "mandatory": false,  "collectionTypes": [  "Sample"  ],  "ruleEnabled": true  },  {  "attribute": "source\_organism",  "mandatory": true,  "collectionTypes": [  "Sample"  ],  "ruleEnabled": true  },  {  "attribute": "source\_provider",  "mandatory": false,  "collectionTypes": [  "Sample"  ],  "ruleEnabled": true  }  ]  } |

Figure 3: Example of Data Objects metadata validation rules

|  |
| --- |
| {  "metadataValidationRules": [  {  "attribute": "object\_name",  "mandatory": true,  "collectionTypes": [  "Sample"  ],  "ruleEnabled": true  },  {  "attribute": "file\_type",  "mandatory": false,  "collectionTypes": [  "Sample"  ],  "ruleEnabled": true  },  {  "attribute": "reference\_genome",  "mandatory": false,  "collectionTypes": [  "Sample"  ],  "ruleEnabled": true  },  {  "attribute": "reference\_annotation",  "mandatory": false,  "collectionTypes": [  "Sample"  ],  "ruleEnabled": true  },  {  "attribute": "software\_tool",  "mandatory": false,  "collectionTypes": [  "Sample"  ],  "ruleEnabled": true  },  {  "attribute": "md5\_checksum",  "mandatory": false,  "collectionTypes": [  "Sample"  ],  "ruleEnabled": true  },  {  "attribute": "phi\_content",  "mandatory": true,  "collectionTypes": [  "Sample"  ],  "defaultValue": "Unspecified",  "validValues": [  "Unspecified",  "PHI Present",  "PHI Not Present",  "Not Specified"  ],  "ruleEnabled": true  },  {  "attribute": "pii\_content",  "mandatory": true,  "collectionTypes": [  "Sample"  ],  "defaultValue": "Unspecified",  "validValues": [  "Unspecified",  "PII Present",  "PII Not Present",  "Not Specified"  ],  "ruleEnabled": true  },  {  "attribute": "data\_encryption\_status",  "mandatory": true,  "collectionTypes": [  "Sample"  ],  "defaultValue": "Unspecified",  "validValues": [  "Unspecified",  "Encrypted",  "Not Encrypted",  "Not Specified"  ],  "ruleEnabled": true  },  {  "attribute": "data\_compression\_status",  "mandatory": true,  "collectionTypes": [  "Sample"  ],  "defaultValue": "Unspecified",  "validValues": [  "Unspecified",  "Compressed",  "Not Compressed",  "Not Specified"  ],  "ruleEnabled": true  }  ]  } |

**Appendix: More on Metadata Validation Rules Expressed in JSON**

The following is a summary of the attributes in each JSON object defining a metadata validation rule element regardless of whether the context is Collections or Data Objects.

1. **attribute**: Metadata attribute name
2. **mandatory**: Flag to indicate if it is required or not. Valid values are “true”, “false”
3. **defaultValue**: Default value of the attribute if no value is given
4. **collectionType**: Collection type name applicable for this attribute rule. Only one value is allowed.
5. **validValues**: List of valid values for this attribute
6. **ruleEnabled**: Flag to indicate if this rule is enabled or not. Valid values are “true” or “false”. If value is set to “false”, this rule will not be evaluated during validation process.
7. **DOC**: Division name applicable. Any registered user account is assigned one DOC in the system. Any specific DOC dependent rules may be attached and activated.