PATH-RELATED CHANGES BETWEEN C-SUB & JADE

This document describes a list of data locations/data types, how they change between the two clusters, accommodations we have implemented, and places where users need to change their workflow (commands issued and paths mentioned in scripts). We believe that we have succeeded at minimizing changes.

A place for everything, and everything in its place

Users do not need to use all of these locations. PIs and their assistants should be aware of the options so they can create the best work flows for their research.

One of our design goals was to **provide a place** for the standard types of data people might need to be able to store. We provide a variety of locations so users can place different types of data in appropriate places. Appropriateness is related to data size, desired data access speeds, and access controls needed to meet the NIST 800.171 standard.

JADE supports researchers working on sensitive data from several providers, such as CMS and dbGaP. This requires JADE to **standardize procedures** to be able to control access and audit the success or failure to protect data. Consistent conventions allow scripts and tools to enforce correct ownership and permissions as well as to audit access. (Each access of a CUI-containing file will be logged noting the time, user, type of access, etc. These logs will be reviewed by IT@JH cybersecurity teams. This is required by JHURA (Research Administration) for them to approve new and renewed dbGaP DUAs.)

JADE uses **separate file systems** for some of these location types, which in addition to providing access control permissions, allows us to serve those file systems from a collection of file servers with different capacity, speed, and cost attributes. As usage grows, new file systems can be created and additional file servers acquired without creating a multitude of exceptions from the conventions. This design allows PIs to be able to choose between faster and pricier SSD or slower and cheaper storage. (We currently don't have spare SSD space but can acquire it if there is enough interest.)

Location Index:

Conventions:

In this document, path elements enclosed in angle brackets are variables to demonstrate the general scheme/convention.

<comm> (short for community) is a unique string, i.e. "cms" or "dbgap" or "sysadmin"
<comm_letter> is the first char of a <community>, i.e. "c" or "d" or "s"
<dua> is a number, e.g. 55548
<cdua> is the concatenation of <comm_letter> and <dua>, a single string, e.g. c55548
<username> is an account name, e.g. c-jxu123-55548
<username> = <comm_letter>-<jhedid>-<dua>

HOME DIRECTORY

/users/<community>/<cdua>/<username>/

Data Type Description:
Each user has a home directory to store their primary files. This location is stored on a file server (named dcs06-jade) equipped with fast solid-state disks. That file server has only limited space, and its disk space costs more than it does on our bulk file servers, which utilize large numbers of spinning hard drives. Consumption of space is controlled by per-user disk quotas, as users cannot be allowed to fill up /users/ file system, as that would prevent operations by the entire cluster.

User Action Needed on JADE:
Users do not need to do anything, as we have created symbolic links which allow the C-SUB path to work on JADE.

Details of Paths On Each Cluster:
The path to it changes

C-SUB
/users/<dua>/<username>
JADE
/users/<community>/<cdua>/<username>

Example:
/users/55548/c-jxu123-55548
becomes
/users/cms/c55548/c-jxu123-55548

Backwards Compatibility:
We have created symbolic links in /users/ which allow either the old path or the new path to work.
/users/55548 → cms/c55548

PER-DUA-GROUP-SHARED (FAST)

/users/<community>/<cdua>/shared/

Data Type Description:
We created in the C-SUB a group-writable shared directory at the same level as user home directories, so users could share files with their group without opening their home directories to the whole group or having to create and maintain ACLs. This space might also be used for code for pipelines used by SLURM jobs, which you would want to be able to access frequently and quickly from compute nodes. Files in this location are controlled by the same per-user disk quota as mentioned above.

User Action Needed on JADE:
Users do not need to make any changes.

Details of Paths On Each Cluster:
The path to it changes

C-SUB
/users/<dua>/shared
JADE
/users/<community>/<cdua>/shared

Backwards Compatibility:
The symbolic links described earlier (/users/<dua> → /users/cms/<cdua>) also handle the shared directory.

PER-DUA-GROUP-SHARED (BULK)

/data/<community>/<cdua>/shared/

Data Type Description:

We have added a new per-DUA group-shared data location in JADE. The one described above, nestled inside the /users/ file system, is stored on a file server (named dcs06-jade) equipped with fast solid-state disks. That file server has only limited space, and its disk space costs more than it does on our bulk file server (named dcs10-jade) which is equipped with spinning hard drives. Dcs10-jade has much more storage capacity than dcs06-jade.

This is a per-DUA *file system* built on dcs10-jade, so it needs to be sized by the PI appropriately for both current and future needs. No per-user disk quotas are planned, although they could be created if requested.

Users and groups can choose to put files in either location. This location opens the door to storing larger amounts of data in a cost-effective manner. For example, a research group might have some reference data set or a saved set of benchmarks which need to be accessed by group members, but which is not Controlled Unclassified Information (CUI) and doesn't need to be accessed at the fastest speed.

User Action Needed on JADE:
Users do not need to make any changes. They can use the directory, or not.

Details of Paths On Each Cluster:
The path to it is:

C-SUB
(none)
JADE
/data/cms/<cdua>/shared

DUA-SPECIFIC DATA FROM CMS (CUI)

/data/<community>/<cdua>/cui/

Data Type Description:
Users sign Data Usage Agreements in order to work with CMS data sets containing CUI. This location stores those files (whether raw or processed to be in necessary formats/forms). When DUA's expire, this data is supposed to be deleted. It is hoped that files stored here are mainly statice (read-mainly). They are owned by a "data steward" account named <comm_letter>-steward1-<dua>, e.g. c-steward1-57285. That account can be used by one or more people designated by the PI to maintain these files. These files are stored on file server dcs10-jade, equipped with spinning hard drives, so it can hold multiple TBs. This is a file system, so it needs to be sized by the PI appropriately for both current and future needs. No per-user disk quotas are planned, although they could be created if requested.

User Action Needed on JADE:
Users do not need to do anything, as we have created symbolic links which allow the C-SUB path to work on JADE.

Details of Paths On Each Cluster:
The path to it changes

C-SUB
/cms01/data/dua/<dua>/
JADE
/data/cms/<cdua>/cui/

Example:
/cms01/data/dua/55548/
becomes
/data/cms/c55548/cui/

Backwards Compatibility:
We have created symbolic links which allow either the old path or the new path to work.
/cms01/data/dua/55548 → /data/cms/c55548/cui/

DUA-SPECIFIC DATA FROM CMS (CUI Intermediate)

/data/<community>/<cdua>/cui-intermediate/

Data Type Description:
This location is meant for CUI-containing files which are regularly modified or generated by the group. When DUA's expire, this data is supposed to be deleted. Ownership of its subdirectories is up to the PI. If a PI wants each user to have a directory owned by those users, we can create the directories and set their ownership & permissions. These files are stored on file server dcs10-jade, equipped with spinning hard drives, so it can hold multiple TBs. This is a file system, so it needs to be sized by the PI appropriately for both current and future needs. No per-user disk quotas are planned, although they could be created if requested.

User Action Needed on JADE:
Users only need to make changes if they want to start using this location as part of their workflow.

Details of Paths On Each Cluster:
The path to it changes

C-SUB
(no standardized location)
JADE
/data/cms/<cdua>/cui-intermediate/

Example:
/cms01/data/dua/57285/users
becomes
/data/cms/c55548/cui-intermediate/users

Backwards Compatibility:
Where this kind of data has been stored in the C-SUB, we can create symbolic links which allow either the old path or the new path to work.

/cms01/data/dua/57258/users → /data/cms/c57285/cui-intermediate/users

CROSS-DUA GROUP-SHARED

/data/crossdua/<community>/

Data Type Description:
If a PI manages a research group whose members work one or more DUAs, the research group may have material that they want everyone in those DUA groups to be able to see, such as tool sets that the group uses. These files should not contain CUI, as that material is legally only able to be seen by signers of each agreement. (DUA-specific material should be shared in the file system /data/<community>/<cdua>/shared/, mentioned earlier.)

User Action Needed on JADE:
None, unless they want to request that such a file system be created.

Details of Paths On Each Cluster:
The core of this location is the leading path elements: /data/crossdua/<community>/
PIs can specify names to use below that level. We can imagine strings which are: the PI's surname, a particular research project, the name of a lab or organization, ...

C-SUB
(none)
JADE
/data/crossdua/<community>/[<PI> | <project> | <lab>]

PUF NON-DUA-SPECIFIC DATA

/data/crossdua/cms/puf/

Data Type Description:
This is a instance of the cross-DUA convention described in the previous section. There are CMS-related data sets which are not restricted by a DUA. They can be viewed by everyone in the CMS community. We have created in JADE a file system which can hold such data, /data/crossdua/cms/

User Action Needed on JADE:
Users do not need to do anything, as we have created symbolic links which allow the C-SUB path to work on JADE.

Details of Paths On Each Cluster:
The path to it changes

C-SUB

/cms01/data/puf-free/

JADE

/data/crossdua/cms/puf/

Backwards Compatibility:
We have created a symbolic link which allow either the old path or the new path to work.

/cms01/data/puf-free → /data/crossdua/cms/puf

CMS DATA MODERATION RELATED LOCATIONS

INCOMING-VIA-SFTP DATA

/transfer/in/<comm>/<cdua>/<username>/

Data Type Description:
In both clusters, users create SFTP sessions (on an external computer) to the appropriate cluster login node (using TCP/IP port 22011). The SFTP server places the newly-connected user in a starting directory, which is common to all users. To be able to upload files to the cluster, users need to change directory into their personal directory.

User Action Needed on JADE:
After connecting with SFTP, users need to use a slightly different path that they cd into. When logged onto the cluster they need to use a different path.

Details of Paths On Each Cluster:

C-SUB

/cms01/incoming/<username>

JADE

/transfer/in/<community>/<cdua>/<username>

THE PROPOSED-FOR-TRANSFER-OUT-OF-CLUSTER DIRECTORY

/proposed/<comm>/<cdua>/<username>/

Data Type Description:

CMS requires that data leaving the cluster be moderated to ensure that it protects Personally Identifiable Information. HARP manages a team of data moderators, who review proposed material and then copy approved files to a location that users can access from the outside.

User Action Needed on JADE:
Users do not need to do anything, as we have created symbolic links which allow the C-SUB path to work on JADE.
Data moderators need to view proposed material using a new path, /proposed/cms/<cdua>/<username>/

Details of Paths On Each Cluster:

C-SUB

Inside of each home directory is a proposed subdirectory, i.e. ~/proposed/

JADE

Inside of each home directory is a symbolic link to the new location,
i.e. ~/proposed → /proposed/cms/<cdua>/<username>/

Backwards Compatibility:
When systems administrators migrate a CMS user to JADE, their scripts make these changes:

a. Rename the existing ~/proposed/ directory to ~/proposed.csub/
b. Create a new directory /proposed/cms/<cdua>/<username>/
c. Create a symbolic link which makes ~/proposed/ point to the new location.

THE APPROVED-FOR-TRANSFER-OUT-OF-CLUSTER DIRECTORY

/transfer/out/<community>/<cdua>/<username>

Data Type Description:
Each user has a directory for files which are ready to be extracted from the cluster using SFTP. (Because those files have been approved by the data moderation team.)

User Action Needed on JADE:
Users need to change the path that they cd into after they connect with SFTP. They need to use a different path when logged onto the cluster.
Data moderators need to copy approved proposed material (in a new location) to a different destination (also a new location).

Details of Paths On Each Cluster:

C-SUB

/cms01/outgoing/<username>

Example:

/cms01/outgoing/c-jxu123-55548

becomes

JADE

/transfer/out/<community>/<cdua>/<username>
/transfer/out/cms/c55548/c-jxu123-55548

In both clusters, users create SFTP sessions (on an external computer) to the appropriate cluster login node (using TCP/IP port 22027). The SFTP server places the newly-connected user in a starting directory, common to all users. To see and retrieve their approved files, users need to change directory into their personal directory.

C-SUB

SFTP-connected users “land” in /cms01/outgoing/

They then change directory to their <username>

JADE

SFTP-connected users “land” in /transfer/out/<community>/
They then change directory to <cdua>/<username>