**Hutchinson data storage policy creation**

As one of the groups moving from Hutch IT to the Clinical School Computing Service (CSCS), we need your support to deliver the migration project. This purpose of this document is to help you make good decisions regarding what to do with your data.

It is important that you determine what to do with your data because investing time in effective data management will produce the following benefits:

* Reduced monthly storage costs for your group, less data equals storage.
* Reduced transfer time for moving the data to its new Service.
* Reduced risk during the data transfer process, (moving a high number of large data volumes across networks increases technical complexity)
* Data will be easier to manage / improved search function

CSCS recommend that you prioritise establishing some data storage policies within your group, (if you do not have these already), as these will help you make better decisions regarding data management going forward.

At a minimum, you should consider 4 types of data: Personal Data, Live Data, Cold data and Computational Data; (depending on your needs, you may actually have other additional data types).

You should consider the University [guidelines](https://help.uis.cam.ac.uk/service/security/data-sec-classes) when creating policies around data storage. The suggestions below are appropriate for Level 0-2 data as classified by the University. Level 3 data is subject to different controls and rules and may need fleshing out to meet your group’s needs.

Acceptable storage and security measures for Level 2 data also include being geographically dispersed, regular backups and ability to retrieve historical data.

After classifying and establishing your policies, following the guidance below there will be a need to prepare your data, ready for migration. This will involve

* moving/classifying your data into these different types,
* appropriately labelling/indexing files/folder,
* compressing larger datasets or .tar small sets of files,
* data housekeeping and removal of un-needed files and data.

**Personal Data**

People often store department data to cloud or other personal storage. This is not recommended practice and should be avoided, as this can create problems; e.g. should colleagues need access to that data when you are absent due to sickness or holidays, or should that person move on to another job, as valuable data may end up going with them. The group should have a clear policy on what is personal data and where it should be stored.

We suggest that personal data be moved from computers, USBs or external storage drives to one of the below solutions to have secure and backed up access to the data to mitigate against the risk of data loss.

|  |  |
| --- | --- |
| **Service** | **Storage Per User** |
| CSCS Home Drive | 50GB |
| University OneDrive | 5TB |
| GDrive | 5TB |

[OneDrive storage](https://help.uis.cam.ac.uk/onedrive) is provided by the University to all Staff and Students, as is [GDrive storage](https://help.uis.cam.ac.uk/service/collaboration/workspace). CSCS have produced a guide on how to migrate files to OneDrive which can be found [here](https://confluence.medschl.cam.ac.uk/x/PaLXBQ). The CSCS Home Drive is provided as part of the £5.37 p/m user account charge; which will be provided to you as a minimum standard for new CSCS users, this also includes services such as email and help desk access, full details [here.](https://confluence.medschl.cam.ac.uk/x/vQhEAQ)

**Live Data**

Live Data is be best defined as data being collected for on-going projects/research which may need to be accessed by multiple members of a team, whilst needing to be securely stored and backed up.

This group data should be stored centrally so that it is protected against loss, multiple people can access it and it survives any staffing changes. Any data that you classify as live data will be migrated by CSCS and stored using the University Information Services (UIS) Institutional File Store (IFS) Service.

Hutchinson groups moving to the Clinical School will be provided with an allocation of IFS storage by CSCS which can be used for data storage to support the migration project. However, once the project has been completed and you fully migrate to using CSCS Services, appropriate monthly charging would then apply.

Some examples of live data:

* Image files
* Word, Excel, PDF etc, office files
* Data that just gets processed on a laptop or desktop
* This should reside on IFS
* Data files generated by instruments such as microscopes, particle sizers, flow cytometry etc

Please note that the maximum volume size for any large data set moving to IFS storage should be 65TB, but in reality, we will need your volumes to be smaller than this to allow capacity for future expansion.

Before transfer, **folders or directories containing large numbers of small files should be stored in a .tar archive**. Datasets containing large number of small files have been found to slow down data transfer speeds and increase the risk of transfer failure.

Long file names or directory paths also have the potential to affect data transfers and therefore we recommend that you limit the character length for files to no more than 25-35 characters.

A good overview of best practices for managing research data can be [found here](https://libguides.princeton.edu/c.php?g=102546&p=930626).

Once Group Drive data on [\\minerva\centre](file:///C:\minerva\centre) has been migrated to CSCS infrastructure, ANY custom permissions on folders with disabled Inheritance (aka different permissions from the top-level security group of the share) will be lost during data migration. **Please can you therefore ensure that any personal data or level 2/3 data that needs to be protected is identified/labelled appropriately**. It will be your responsibility to highlight this to CSCS and to agree the best method of transfer.

**Cold Data**

Older, inactive or unstructured data should be stored in a less expensive service: cold storage. Older data typically does not need to be used or accessed on a regular basis, but you will need to retain a secure copy of it. Typically for ethics purposes or as a requirement of your grant funding.

This data will be stored using UIS’s [Research Cold Store](https://www.hpc.cam.ac.uk/research-data-storage-services/research-cold-store) (RCS) as it is a cost-effective way of storing large amounts of data at a competitive price, with enterprise level protection and backup for the storage. RCS is also suitable for data sets that have been received but are not being processed in the immediate future.

Cold storage data is procured based on a 1-5 year contract, with a minimum of 5TB of data to be stored, costs below[[1]](#footnote-1):

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Service** | **Price (£ per TB)** |  |  |  |
|  | **Per month** | **Per year** | **Per 3 years** | **Per 5 years** |
| Research Cold Store (RCS) | £2.56 | £30.72 | £92.16 | £153.60 |

Any data that is moved to RCS will be fully accessible to you once you have migrated over to using CSCS Services. To do this, you will simply need a Raven based account to access the store via and SFTP connection (we can provide guides on how to access via SFTP from devices).

All data stored here is backed up in 2 locations and all located within the UK. Further details on the service can be found on the UIS HPC website [here.](https://www.hpc.cam.ac.uk/research-data-storage-services/research-cold-store)

Please note that in order to move your data to RCS:

* Data needs to be compressed into 1-2TB files
* This must be using standard compression tools; tar, zip, bzip, pbzip etc
* RCS projects are accessible from the UIS HPC cluster, this means data can be pulled onto RDS for processing (see section below on this UIS service)

Additional to the above categories it would be suggested that the department define policies on usage for the below data types.

**Computational Data**

Data classified as bioinformatic, genomic or very large data sets. Including large reference data sets that get downloaded (and can be retrieved again) such as UK Biobank. This would typically be stored or processed by a dedicated server (that the group owns and you may need CSCS to host) or a cluster.

This data would require fast storage that is accessible from the central server/cluster. Typically, this is not backed up and the storage is only used as a scratch space for data that is being processed.

Department to define their policy on usage and providers to be used for these services. As part of this work, things to consider are:

* How much of the data you have is HPC/bioinformatic data that requires a cluster for processing?
* Can you get the raw data again? If no, how much do you need to backup?
* Do you remove intermediate files or compress data as part of your workflow?
* Data for processing goes to RDS, copies of raw data, results or data you need a backup of goes to RCS

CSCS do not offer a dedicated HPC service as this would duplicate a service already provided by University Information Services (UIS). Therefore, any critical Hutch IT non-standard services such as HPC will remain in operation but will not be migrated over to CSCS. Groups will be provided with remote access to these resources as an interim solution. As part of any pre-migration project engagement, CSCS will facilitate initial engagement with the UIS HPC team, to review your requirements and discuss potential services that can be offered to you at UoC.

**Additional HPC Guidance**

* Users that require the ability to process data with standard Bioinformatic tools will need an HPC account
* When you apply for an account the lead PI for the group is allocated 200,000 core hours per quarter
* Core hours are then shared with the group and additional computation time is chargeable.

**PID (Patient Identifiable Data)**

Department to define their policy on usage and providers to be used for these services.

CSCS operate the [SDHS (Secure Data Hosting Service)](https://cscs.medschl.cam.ac.uk/server-services/secure-data-hosting-service/) which is a certified Safe Haven and can be used to store PID, alongside services offered by the NHS.

**External Storage – Hard drives, USB, NAS Drives etc.**Department to define their policy on usage. CSCS recommend the department plan out the moving of data and elimination of existing external storage plus policies regarding procurement and future use of external devices for data storage wherever possible to protect against data loss.

**Summary: What CSCS need from you prior to any migration**

Classify your data according to the categories we have just outlined.

* How much data is cold?
* How much data is live?
* How much data do you need to access using HPC (or a dedicated processing server)?
* Do you have data lifecycle management as part of your processes for computational data?

We would expect you to take some time to:

* Complete some data cleansing prior to migration
* Chose the right storage option/s (based on the options above)

1. Information on pricing from <https://www.hpc.cam.ac.uk/research-data-storage-services/price-list> [↑](#footnote-ref-1)