HELMHOLTZ

MDC MAX-DELBRÜCK-CENTRUM FÜR MOLEKULARE MEDIZIN IN DER HELMHOLTZ-GEMEINSCHAFT

Data Storage, Backup and Access Guidelines

Max Delbrück Center for Molecular Medicine in the Helmholtz Association

A collection of guidelines for storing your data at the MDC

Current Status

Version	1.0
Date of this version:	16.06.2022
Date for original version:	30.06.2021
Reviewer:	RDM & IT
Frequency of reviews:	Annually
Created by:	Sara El-Gebali
Participating departments and committees:	RDM, IT, DPO, ISO
Approved by:	RDM, IT, DPO, ISO
Status:	Done
Confidentiality level:	Public
Available at:	

Change History

Date	Version	Created by	Description of changes
30.06.2021	1.0	Dr. Sara El-Gebali	Created
09.09.2021	1.1	Dr. Özlem Özkan	Edited
16.06.2022.	1.2.	Dr. Özlem Özkan & Dr. Inga Patarčić	Edited

Preamble

This document explains the MDC's options for data storage, backup solutions and access rules. There are different kinds of data (See the part of the definitions below), and thus, different storage options (Shares, Raw Data storage, Cluster storage, SharePoint and Source code Repository) at the MDC.

Storage and backup solutions for research data should be planned ahead in order to:

- Help prevent potential data loss and data disclosure,
- Integrate seamless workflows,
- Ensure long-term use and preservation,
- Ensure appropriate access rights for data sharing and retrieval.

Data Management Plans (DMP) also require information about data storage, backup and access rights.

Storage solutions

During your work at the MDC, you are required to store all research data using storage solutions provided by the MDC-IT and in accordance with your employment contract.

The owner of a share (the PI, group leader or project leader) is responsible for defining data access rights for members of the group and is responsible for any consequence arising from inappropriate disclosure.

For all storage options managed by the MDC-IT monitoring of activity and access is carried out monthly. Instances where files cannot be accessed by anyone anymore (for example, if the whole group leaves the MDC) will be archived and deleted. Generally, individuals are initially assigned standard quotas for data storage solutions (in general). However, **those who need more space, can request it from the central IT Helpdesk**.

MDC-Central data storage

You might be allocated more than one data share. To check which you can use, please consult insights.tools. For information on file servers offered by central IT, please refer to the guidelines available internally in German & English. For information on how to connect to directories, please refer to KB-Article on "How to Connect to Servers and Directories" (KBA00190: Which file servers are there and how do I connect to them?).

Home directory

The initial quota of the home directory is **20 GB**. This should be used mainly for storing personal documents that need to be easily accessible in order to facilitate your day-to-day tasks. You can contact the central IT Helpdesk to get an extension of this share, but it is also possible to request more space via the Self Service Portal. Please note that personal data such as emails are deleted **six months** after you have left the MDC.

Backup: A backup copy, "i.e., snapshot copy" of your home directory is created every 2 hours, and a daily version is backed up for 30 days. A full replica/copy of your home directory is created frequently during the day. After another 30 days, the content is stored in a different location.

Access: Only yourself.

Group shares

AG_groupname = **A**rbeits **G**ruppe (working group)

This designation refers to every principal investigator and research lab. Files shared between a group can be stored in the group share, which has an initial capacity of **100 GB.** Lab members and anyone else in the same group can access these file shares, but sharing **cannot** be extended to other groups or lab members.

If you are storing sensitive data in group shares, you must ensure that everyone who can access this data is authorized to do so. For this purpose, please inform and consult the PI, who is the owner of the share and thus responsible for its content.

Backup: The backup of group shares follows procedures similar to those for the home directory.

Access: All members of your group, e.g., lab team.

Project shares

PG_projectname = **P**rojekt **G**ruppe for shared storage for special project groups.

Short-term data storage is also available for project data that needs to be saved for a limited amount of time, up to a specified number of years. This can be stored under a specific designation, i.e. **PG,** the amount and duration of storage can be requested in consultation with the IT-HelpDesk.

Backup: The backup of project shares follows procedures similar to those for the home directory.

Access: Access rights are dictated by PIs and managed by the group manager application https://groupmanager.mdc-berlin.net. Access can be granted to other MDC users as specified by the PI. If the PI has left the MDC or is no longer available for some other reason, the group's most senior member will be assigned the responsibility for access and sharing rights. If an entire group departs from the MDC without prior agreements on data handling with the IT, the MDC-IT will

proceed with archival and deletion of the PG folder from production storage upon consultation with the MDC-Board.

Raw data storage

RD_groupname: RD = **R**aw **D**ata, which can be a parent or a nested directory

In some cases, raw data produced during research might require specific storage solutions. For instance, data directly obtained from instruments (such as sequencers) can be stored in a specified storage area. The PI can request such storage for raw data from the central IT by the PI. The initial quota is **2TB**.

Backup: The backup of raw data shares follows procedures similar to those for the home directory.

Access: Access rights are regulated similarly to project shares and can be managed by the group manager application (https://groupmanager.mdc-berlin.net).

Max Cluster

To use the Max Cluster and store data there, training is required. Please get in touch with the HPC team by creating a ticket, and refer to the document Data Storage on Max Cluster. Cluster storage follows the same conventions of structuring and naming described above, i.e. home directory, AG_groupname, PG_projectname, RD_groupname. Group folders and project folders (AG_X & PG_X) are assigned an initial default quota of **10TB**.

Backup: Data stored on the Max Cluster is automatically backed up twice per day; older file-versions are kept for one month.

Access: Access rights are regulated in a way similar to folders on the server (see details above) and can be managed by the group manager's application.

There are two parts to Max Cluster: Fast storage and Restricted storage.

Fast storage on Max Cluster

This storage is for data to be processed on the HPC cluster (per user, group or project).

Restricted storage on Max Cluster

This restricted storage is dedicated to the processing of sensitive data.

Local storage on the group level computing systems maintained by the Bioinformatics Platform

Some MDC groups have their own computing systems and local storage options that the Bioinformatics Platform maintains in collaboration with the MDC-IT. This storage is directly attached to the servers/computing systems used at the group level. The capacity is specific to each system.

The owner of a share (the PI, group leader or project leader) is responsible for assigning data access rights and any consequences of potential data disclosure. For technical questions, please contact bimsb_itsupport@mdc-berlin.de.

Backup: Data stored on these servers (/home and any other shares) are backed up at least once a day (during the night) within a 30-day history.

Access: Access rights are regulated in a way similar to the MDC-Central data storage access rights and can be managed by the PI (or a project leader where that is the case). The storage can be used locally or shared over the network (NFS / CIFS).

SharePoint

For collaborative work, one can use SharePoint. SharePoint provides a space where files in diverse formats, such as .docx, .xlsx, .pdf, can be uploaded and edited. SharePoint lists can be created and configured freely for the purpose of centrally collecting data in one place. In addition, SharePoint offers a range of features to facilitate collaborative work, such as calendars, tasks and more. These features can be shared with groups or individuals, and edited directly via SharePoint. After a brief examination of the users' needs and requirements, a tailored SharePoint space can be offered by the MDC-IT group. In some cases they might recommend other tools.

Backup: A full backup of the database is carried out once per day and an incremental backup every four hours. The data is kept on the database server for seven days, followed by a complete backup by another storage system as a "snapshot", where it is kept for a period of 60 days.

Access: SharePoint provides a complex permission system which the users can configure. Please contact the central IT Helpdesk if you need any support. Subsites within SharePoint are analogous to closed rooms with more finely tuned authorization and access rights. Generally, SharePoint sites/subsites can be accessed by MDC employees and MDC external users with guest accounts.

NextCloud

NextCloud is MDC's cloud solution. It provides data space for collaborative work with documents within and outside the MDC. The cloud solution is reachable with MDC credentials via nextcloud.mdc-berlin.de. You have a **50GB** initial quota. To increase your storage quota, please contact the central IT Helpdesk.

Backup: There are two different rules for different systems: The file system is backed-up every 2 hours, and the database is backed up daily

Access: Initially, only the shareowner has access rights. But the data owner can share the data with other users.

RSpace - Electronic Lab Notebook (ELN)

MDC offers electronic lab notebooks to all researchers. ELNs help researchers document their lab work and easily share research data. They contribute to ensuring the reproducibility of research. MDC has a subscription to the RSpace electronic lab notebook tool. There are two different RSpace systems installed in MDC local servers: PROD and TEST.

- **The production (PROD) system** is the ELN you should use to record your daily lab activities. Please use this system for your lab records.
- **The test (TEST) system** should be used **for test purposes only. In the test system,** you can either try new structures for your electronic lab notebook and observe the results of these structures, then adapt them to your actual system (PROD) later; you can also use this system for educational purposes to train new researchers unfamiliar with RSpace.

Since IT tests the new RSpace updates on the test server, the system may **not** be stable and it is not appropriate to store important files there, especially if they are the only copy of data that you have. Please do **not** record your daily tasks on the test system.

Except for technical storage limit, there is no quota per group. The storage limit can be extended based on needs. You can upload files up to **50MB** directly into the RSpace ELN. Bigger files must be uploaded to the IT-provided storage systems and linked to the lab notebook. For this purpose, the RSpace ELN has a samba share integration (to link big data). **Sensitive data must stay on "Restricted storage on Max Cluster" or another secure type of storage designed for sensitive data and should only be linked in the RSpace ELN.**

To start using MDC RSpace, please contact the Research Data Management Department or write an email to eln@mdc-berlin.de. For technical questions, please create an IT ticket.

Backup: The RSpace server is backed up once per day. The integrations and samba shares have their own backup rules.

Access: Access rights can be defined by the PI of the group.

Bitbucket - Source code repository

The MDC offers a source code versioning system, "Bitbucket, " which is similar to Gitlab and Github. Bitbucket also allows teams to plan projects, collaborate on code, test and deploy software on target servers and work seamlessly and collaboratively with the help of pull requests and reviewing capabilities (git.mdc-berlin.de). Bitbucket uses GIT under the hood to store files and their versions, create branches and commit changes before pushing them to the central GIT server.

Every MDC employee and research group can get one or more Bitbucket GIT repositories. A repository can handle an unlimited number of files and up to several GB of disk space.

Backup: The repository metadata is stored in a relational database, and files are stored on the filesystem. Both are backed up once a day.

Access: Bitbucket has fine-grained access control of repositories and branches. Research group members can get access rights that are project-specific and project-wide or specific for certain repositories. Access rights can be defined per user and on the group basis.

The research group PI or the project owner is responsible for the repositories that belong to the project.

Other storage solutions

Wherever possible, please ensure that your data is stored in facilities provided by central IT. When absolutely necessary and the use of external devices is unavoidable, please consider the following:

- Do I have an approval from a DPO (data protection officer) and ISO (infomation security officer)? Copying, editing or sharing of specific types of data on external devices or sensitive research data require approval from DPO and ISO.
- Who has access to your data?
- Is your data sensitive?
- How often do you perform backups?
- Which tools do you use for automated backups?
- Where do you store your backups?
- How many backups do you have?
- More information on backup strategies is available on the UK Data Archives website.

Furthermore, please note the following requirements:

- The data is backed up regularly
- Storage must be reliable
- Access rights and permissions are controlled
- Data should be encrypted when it is needed (e.g. when travelling); see Data Organization Guidelines, section Quality control
- The deletion of files has to be done securely; if you need support on safe disposal, you can contact the IT by creating a ticket.
- Approval from the information security officer and the data protection officer for sensitive research data

It is advisable to apply general guidelines for backup and follow the **3-2-1** concept:

- Create 3 **copies** of your data (your production data and 2 backup copies)
- Stored on 2 different **media,** e.g. hard drives, CDs and thumb drives
- Stored at 1 **offsite**/different location for disaster recovery

External hard drives and removable media

Please note that external hard drives are subject to physical degradation and long-term wear and tear; therefore, it is advisable to;

- Maintain a copy of data on a different (MDC-owned) device
- Backup should be performed regularly
- Check data integrity regularly
- Limit the usage to the active data and do not use it for long term archival purposes

Furthermore, deleting of files and reformatting a hard drive will not prevent a possible recovery of data that has previously been on that drive. Therefore, it is mandatory to securely erase files

according to instructions provided by the Information Security Office; alternatively, the device can be handed into central IT (please contact the IT-helpdesk). When using external hard drives to backup personal devices, data must always be encrypted and a copy of it stored on the MDC's central facilities.

Laptops and PCs

Devices that are not controlled or overseen by the central IT group will not gain authorization to access the scientific production environment. The owner/user of the device is responsible for ensuring that there are secure backups and security measures in place to prevent unlawful access.

Group-owned servers

If your group maintains a server of their own, central IT can offer backup solutions**.** The frequency and the choice of data to be backed up must be agreed on and indicated in the backup strategy arranged between the PI and the central IT server operator responsible for your server. Central IT maintains virtual servers, and regular backups are carried out nightly.

Abbreviations

DMP: Data Management Plan

DPO: Data Protection Office

HPC (Cluster): High-Performance Computing

ISO: Information Security Office

PROD (Server): Production

Definitions

Primary: Raw Data, e.g., measurements, recordings from instruments or observations, data from collaborations. This includes data in raw format, tables, databases, etc

Secondary: Analysis, e.g., results of calculations, data from collaborations, including algorithms and measurements, revisions

Final: Published data

Metadata: Documentation of protocols, contextual information, including machine settings

Code: Software or algorithms

Sensitive research data: According to Article 4(13), (14) and (15), Article 9 and Recitals (51) to (56) of the GDPR, *the following personal data is considered "sensitive" and is subject to specific processing conditions:*

- personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs;
- trade-union membership;
- genetic data, biometric data processed solely to identify a human being;
- health-related data;
- data concerning a person's sex life or sexual orientation.

Please contact the Data Protection Office (DPO) if you need assistance with sensitive research data.

Short-term data: Short-term data is generally categorized as **live or active data** that is in use, and could be related to a current project and/or your day-to-day activities. This includes temporary files intended for short-term use.

Long-term data: Generally speaking, long-term data is data which is infrequently accessed or used and intended for archival purposes. This includes data pertaining to a research project that has ended or is no longer active, i.e. data no longer in active service. If you need assistance with long-term preservation and archival services, please contact the IT department by creating a ticket.

Research: Research is any creative and systematically performed work with the goal of furthering knowledge (including discoveries about people, culture and society) and the use of such knowledge for new applications.

Principal investigator: PI is the most senior researcher associated with a research project and the primary individual responsible for the research project's implementation and management and the integrity of the design, conduct, and reporting. Additionally, the PI holds the responsibility for the direction and oversight of compliance.

Researchers: All members of an institution, including scientists under contract, students and support staff, as well as others with a formal affiliation at the MDC, who have access to, generate and/or manage research data. Visiting researchers or collaborators may also be expected to comply with the policy.

Research data: Research data refers to all information (independent of form or presentation) needed to support or validate the development, results, observations or findings of a research project, including contextual information. Research data include all materials that are created in the course of academic work, including digitization, records, source research, experiments, measurements, surveys and interviews. This includes methods/protocols, metadata, software and code. Research data can take on several forms: during the lifespan of a research project, data can exist as gradations of raw data, processed data (including negative and inconclusive results), shared data, published data, and open access published data, and with varying levels of access, including open data, restricted data and closed data.

Data sharing: Data sharing is the practice of making scientific data used for scholarly research available to others (e.g. researchers, institutions, the broader public), for research re-use or in knowledge transfer activities.

Data storage options at a glance

	Aim of the storage option	Initial storage quota	Back up schedule	Access rights	Collaboration and/or sharing outside the MDC	Storage of sensitive research data	Contact for help
Home Directory	Data is used for day-to-day tasks only by a single researcher.	20 GB	Every 2 Hours	The owner (individual)	No	No	IT Ticket
Group Shares	Data shared within one MDC-group only.	100 GB	Every 2 Hours	All members of the group, e.g., lab team.	No, sharing cannot be extended to other groups or lab members.	Depending on the DPO's and ISO's assessment and IT measures.	IT Ticket
Project Shares	Data storage for project data, can be shared between multiple MDC groups & only granted for a limited duration.	Based on needs	Every 2 Hours	Access rights are dictated by PIs and managed by the group manager.	No, only within MDC-groups.	Depending on the DPO's and ISO's assessment and IT measures.	IT Ticket
Local storage on group servers maintained by the Bioinfo Platform	Data shared within one MDC-group only.	Not limited	Daily: multiple snapshots per day	All members of the group, e.g., lab team.	No, only within MDC-groups.	Depending on the DPO's and ISO's assessment and IT measures.	bimsb_itsupport@mdc-berlin.de
Raw Data Storage	Raw data produced during research.	2TB	Every 2 Hours	Access rights are dictated by PIs and managed by the group manager.	No, only within MDC-groups.	Depending on the DPO's and ISO's assessment and IT measures.	IT Ticket

Fast storage on Max Cluster	Data to be processed on the HPC cluster (per user, group or projects).	10TB	Multiple snapshots per day	Share owner, MDC-group members, Project-group members.	No, data needs to be placed elsewhere for sharing with outside the MDC.	No	IT Ticket
Restricted storage on Max Cluster	A dedicated storage is available for processing of sensitive data.	10TB	Multiple snapshots per day	PI is responsible for potential breaches. Contact the IT for access rights details on sensitive data.	No, sensitive data can be stored only on the Cluster.	Yes, the storage is designed for sensitive data	IT Ticket
SharePoint	Data space for collaborative work where files can be uploaded in multiple formats, can be shared with MDC guests.	Based on needs	Every 4 hours	SharePoint sites/subsites can be accessed by MDC's employees and external users with an MDC guest account.	Yes, by providing a link, sites and subsites can be shared with external groups or individuals.	No	IT Ticket
NextCloud	Personal Synchronization & Share Tool Data space for collaborative work on documents	50GB	File system: Every 2 hours Database: Daily	Initially only the share owner. The owner can share the data with anyone.	Yes	No	IT Ticket
RSpace ELN	ELN to document research, experiments, and procedures performed in a laboratory.	Not limited	Daily	Access rights are defined by the PI of the group.	No	No	Technical Questions: IT Ticket To set up a new lab: eln@mdc-berlin.de
Bitbucket	Source code versioning system	Based on needs	Daily	Access rights are defined per user and group basis.	No	No	IT Ticket