VU Computer Science Data Storage Platforms

version v0.4, November 5, 2024

Kees Verstoep Shuai Wang

Table of contents

1 Selecting the Right Data Storage Platform for Research Projects

This document is created as a guideline for the selection of storage platform for the research projects in the Computer Science department of the Vrije Universiteit Amsterdam. This guideline makes concrete suggestions for our department based on the 2020 university RDM

policy and information provided by the UBVU (University Library) RDM team and VU/IT department. The guideline takes into account the current state of several ongoing and recently completed projects in our department, based on a department-wide survey held in 2023.

The document provides an overview of available options for data storage, explains how they are suggested for use in our department, and helps researchers choose the most suitable storage option for projects. While this document covers only data storage, the data stewards are aware that in practice, there could be much more to take into consideration. Feel free to consult the data stewards in specific cases.

Using many different storage options can lead to fragmentation and operational overheads when working with multiple data sources. Therefore, in our department we want to limit the commonly used storage options to a preferred shortlist. This document serves as a guideline for this convergence.

The online VU tool Data Storage Finder has no less than 10 storage options listed. More detailed information and metrics about these options can be found on the University Library's Data Storage guide.

The data storage options are classified as appropriate or not according to the following criteria:

Data classification (sensitivity)	low / medium / high / very high
Data sharing	not needed / with VU colleagues / with anyone
Data volume	below 500 GB / above 500 GB
Possible features	• Basic storage • Compute/HPC • Fine-grained access rights • Collaboration tools • Archiving/Publishing

Note: these classifications are per "project". The formal status, scale, type and personal participation of projects can differ significantly:

- some large projects have budgets of their own (with partial external funding);
- smaller projects might be fully supported via the "zero-cost" base resources available per project;
- in between are projects that are supported by a contribution from the budget of the PI.

Apart from the options mentioned, personal laptops and storage tied to data processing resources like HPC clusters are also important data storage resources for research. However, these should only be used to work on (copies of) the actual datasets that are stored on and shared via the selected primary storage platform.

Version management of primary and derived research data is an important topic in itself, which needs to be addressed in a project's data management plan, and which should be made explicit in operational guidelines which can also include suitable synchronization mechanisms. It is orthogonal to the selection of the storage platform.

Closely related to data storage is data *sharing* and data *archiving*. This will be addressed in future versions of the guidelines.

Another important aspect to consider is the *backup* of the data, to be able to retrieve data that was lost or that was somehow corrupted. A general rule for backup is the 3-2-1 rule and does scale depending on how backups are structured: 3 copies (1 original, two backup), on 2 different types of media, with at least 1 off-site (physically) or in separate, dedicated cloud storage.

Typically each of the storage platforms discussed below will have some mechanism to restore a previous version of a specific file from backup, but the extent to which this is possible may vary significantly. Being able to retrieve a single file version from a week ago should not be a problem, but for many projects it is necessary to be able to retrieve a *consistent snapshot* of a large number of files (see the discussion of version management above). Where requirements for retrieval of previous file versions are higher than offered by the storage platform directly, additional versioning functionality will have to be designed and implemented (e.g., daily/weekly consistent snapshots of the project data). This should be made explicit in the project's data management plan.

2 Survey results on the current use of Data Storage

In our survey of 2023, it was reported that during the development of the projects, most data was stored on researchers' own laptops. Interaction with GitHub and similar version control tools was common. Nation-wide servers such as the SURF compute clusters and the DAS compute clusters are well used in the department, but it should be noted these are not (or should not be) primary storage platforms as such.

While designing and compiling the results of the survey, we also encountered some storage services that we were not aware of. Some respondents reported using additional group-specific servers. These customised storage solutions were typically developed as part of a particular research project or to support a specific research domain. This does lead to the question if it is necessary to maintain additional services which also have a data storage role. To properly deal with these special use cases, there should be procedures in place where the data stored on

specific servers is still backed up by standard storage solutions, thus guaranteeing long term data safety.

The YODA platform has been promoted at a university level. However, it is not yet very actively used in the department. We are now building up more experience with it, but it is clear that changing storage platforms midway during an active project can be problematic.

3 Data Storage Suggestions

Here we group the storage options available in three categories regarding suitability for research in our department: **most suitable**, **suitable for specific purposes**, and **typically less suitable**; finally, there are some **legacy options** worth mentioning.

3.1 Most suitable

Research Drive	• SURF supported • sharing data inside and outside VU • suitable for potentially sensitive data
YODA	• SURF supported • sharing data inside and outside VU • suitable for potentially sensitive data
SciStor	• VU/IT supported • low latency high throughput access at VU • good for quickly synchronizing many files • the department has pre-paid 40 TB SciStor capacity, so many projects can benefit from this without project-specific funding • BUT: only possible to share data within VU

Here are a few important differences between YODA and Research Drive:

- YODA has a basic classification that is higher than Research Drive (Research Drive can hold data with classification High, but needs additional measures);

- Research Drive enables fine-grained access permissions: you can give collaborators access to only one folder and not the entire project (this is not possible with YODA, and quite useful when e.g. working with students or external project partners);

- YODA includes a vault for archiving while ResearchDrive does not;
- YODA enables adding metadata as you go; Research Drive does not.

3.2 Suitable only for specific purposes

Google Drive VU	• useful for quick ad hoc sharing, including online editing shared contents • can store datasets with high data volume (many TB) without privacy concerns, for which no project budget is available
Microsoft Sharepoint (Teams)	• useful for quick ad hoc sharing, including online editing shared contents • BUT: Use only if Microsoft-based document sharing is the best match for a small scale project. Sharepoint is a relatively new option supported by VU/IT, but the functionality is quite diverse and partially overlaps with the Google option, just in a Microsoft flavor.
Open Science Framework	• specifically suitable when collaborating with external partners in Open Science contexts, where widely sharing data, analysis techniques and results are the primary goal.
Encrypted portable storage (SSD)	• useful for regular laptop backups • BUT: Do not use this to store research datasets, given the risks of losing the only physical copy, and the challenges to share the data
Github, GitLab, Bitbucket	• very useful for software development and sharing • might be used for small scale data sharing, e.g., as part of a publication whose focus is on a developed algorithm or system • BUT: less suitable for large scale data sharing, given the limits on file size and data volume

3.3 Typically less suitable

If possible, best avoid the next options since they usually have little benefits compared to the suitable options above, but do have some downsides, and we would like to avoid fragmentation and interoperability problems for projects.

OneDrive	– Tied to personal VU-relation of owner, so not suitable for long term project storage
SURFdrive	– Basically a more limited precursor of ResearchDrive, so use that instead.
Custom solutions	– E.g., group/project-specific RAID systems. Typically there are hidden costs maintaining the setup. If there are very specific reasons why none of the standard solutions seem to match, best talk with the RDM advisors before introducing a custom solution.

3.4 Legacy storage options

In the past, some more storage options were available. They are not mentioned in the online storage selector, and their use should be limited as much as possible. Preferably they should be phased out, by migrating them to preferable options listed above. New projects should not be started using these "legacy" storage options.

Legacy Home folder, or "H drive"	• The old home folders are still accessible when logging in to the SSH stepstone, or when accessing the BAZIS cluster. However, capacity is limited, costs are high, and it is not possible to share this data directly with other users, at the VU or elsewhere. It is currently (early 2024) actively being phased out, with VU/IT suggesting to use OneDrive instead. For our department, SciStor will actually be a more suitable alternative since it preserves almost all the existing functionality.
Legacy Group/Projects folder	• The old group/projects folders are also still available via the SSH stepstone, but capacity is limited, and costs per gigabyte are high. This data should also be migrated, and no projects should be added anymore.

Dropbox	• Dropbox has useful syncing options between various devices, and can be used for ad hoc file sharing, but the costs are relatively high. Dropbox is not a VU/IT supported storage platform and it should not be used for research datasets.

More guidelines on data archiving will be provided in future guidelines.

4 Detailed overviews

The descriptions in this overview are mostly copied from the VU Data Storage Finder which is provided by the RDM support department of the VU Library; they are included here for easy reference.

4.1 Research Drive

Research Drive enables you to easily store and share files with other users, inside and outside the VU. Please note: storing data with a 'high' classification in Research Drive is permissible, but may require taking extra security measures. Please get in touch with the Research Data Management Support Desk via rdm@vu.nl if you want to store high-classification data in Research Drive.

Research Drive details	
Max. storage size	Terabytes of data
Costs	• below 500 GB: free of charge • 500 GB to 2 TB: €200 per year • above 2 TB: €200 + €250 for every TB above 2 TB
Sharing and collaboration	With anyone
Location	SURFsara (NL)
File recovery	Via versioning, Deleted files or file restore request
Data classification	Medium, but with additional steps: High

4.2 YODA

YODA is a platform that supports research data management throughout the entire research cycle: from safe and easy storage and sharing of data during the research process, to sharing of data within research groups and projects and, ultimately, to research data archiving and publication.

YODA details	
Max. storage size	Terabytes of data
Costs basic storage	• below 500 GB: free of charge • 500 GB to 2 TB: €200 per year • above 2 TB: €200 + €250 for every TB above 2 TB
Costs archiving	• below 500 GB: free of charge • 500 GB to 2TB: €25 per year • above 2TB: €25 + €25 for every TB above 2 TB NOTE: Archiving will be billed for a 10-year period
Sharing and collaboration	With anyone
Location	SURFsara (NL)
File recovery	Via versioning or file restore request
Data classification	High

4.3 SciStor

Besides basic storage, SciStor is best suited for over-the-network use with lab instruments and high-performance computing (HPC/BAZIS), either via SMB or NFS.

SciStor details	
Max. storage size	Terabytes of data
Costs	€0,10 per GB per year for reserved space (price doubles with backup for file recovery)
Sharing and collaboration	VU employees and students
Location	VU Campus (NL)
File recovery	Via file restore request or versioning
Data classification	Medium

4.4 Google Drive VU

This is the well known Google Drive, only with a VU license allowing storing more data. Besides Google Drive, most applications within the Google Workspace environment, such as Google Docs, Google Sheets and Google Slides, are also available and accessible on any device.

Google Drive VU details	
Max. storage size	Terabytes of data
Costs	None
Sharing and collaboration	With anyone
Location	Google (WORLD)
File recovery	Via versioning
Data classification	Low

4.5 Microsoft Sharepoint (Teams)

Microsoft 365 cloud storage and collaboration.

Microsoft Sharepoint/Teams details	
Max. storage size	1TB, up to 25TB
Costs	None
Sharing and collaboration	With anyone (A Microsoft account may be required.)
Location	Microsoft Cloud (EU)
File recovery	Via versioning or Recycle Bin
Data classification	Medium

4.6 Open Science Framework

The OSF is an open-source collaboration tool geared towards Open Science.

Although not primarily a storage tool, OSF can be a very useful and easy to use way to share and publish your data, project documentation, or pre-registrations.
It is possible to connect Research Drive or Dataverse storage to OSF if your data volume is larger than what OSF offers.

The OSF is developed by the Center for Open Science (COS) that strives to increase openness, integrity, and reproducibility of research.

Open Science Framework details	
Max. storage size	• Private project: 5 GB • Public project: 50 GB (Extra storage possible on request)
Costs	None
Sharing and collaboration	With anyone
Location	Google (EU)
File recovery	Via versioning or Recycle Bin
Data classification	Low

4.7 Encrypted portable storage

Encrypted USB drive/disk or USB flash drive (USB stick).
You can order a drive or disk via the Order accessories form found in the service portal.

Please note that it is highly recommended to always use encryption with portable storage, even if your data is classified as low.

Encrypted portable storage details	
Max. storage size	Depends on the device chosen
Costs	Depends on the device chosen (about EUR 100 per TB)
Sharing and collaboration	Not applicable
Location	Local, direct connection
File recovery	Can be configured
Data classification	Medium (with encryption), Low (without encryption)

4.8 OneDrive

OneDrive for Business is part of VU Microsoft 365 cloud - also referred to as Digital@VU - and is available for employees and students. Please note that this is personal storage that is linked to the existence of your account at the VU. If you disappear, your data in OneDrive also disappear.

OneDrive details	
Max. storage size	1TB
Costs	None
Sharing and collaboration	With anyone
Location	Microsoft Cloud (EU)

OneDrive details	
File recovery	Via versioning or Recycle Bin
Data classification	Medium

4.9 SURFdrive

SURFdrive is a personal cloud storage service for the Dutch education and research community, offering staff and researchers a secure and easy way to store, synchronize and share files in the SURF community cloud. Please note that SURFdrive is not available for students. However, it is possible to share files with students and others.

SURFdrive details	
Max. storage size	500GB
Costs	None
Sharing and collaboration	With anyone
Location	SURFsara (NL)
File recovery	Via versioning or Deleted files
Data classification	Medium

4.10 Custom solutions

If none of the options satisfy your requirements, a custom setup or configuration may be possible. Contact the University RDM Support Desk (rdm@vu.nl) or the department RDM contacts Shuai Wang (shuai.wang@vu.nl) and Kees Verstoep (c.verstoep@vu.nl) to explore the options.

For example, in our department there are currently two additional group-specific data storage options: The CI "rippers" and the LOD (Linked Open Data) server. The CI rippers are being used by researchers in the Computational Intelligence group. The LOD server is mostly used by researchers in the UCDS group and the former KRR group. The server has approximately 20TB of data storage. There are some valuable datasets on it that are not available via other data storage/archiving options. Also, the server provides LOD-specific tools to support a particular research domain.

Another example is the storage attached to compute facilities like DAS-5 and DAS-6 and the SURF Snellius. These storage solutions enable storing local copies of original and derived datasets for purposes of efficient high performance computing. Still, the primary copies of the source and processed datasets should typically not be kept permanently on such resources, but on the suitable storage options discussed earlier.