Research Data Storage, Sharing and Transfer Options

COLUMBIA UNIVERSITY| READI PROGRAM

Principal investigators should establish a research data management system for their projects including procedures for storing "working data" collected during the conduct of the research. The PI should communicate these procedures to all group members. The procedures should ensure that the PI is able to access all data produced by the research group and must meet all applicable security requirements. Below is a list of options for the storage, sharing, and transfer of digital research data. A glossary of the terms used in the summary is located at the end of this document.

		Service	Cost	Capacity	Archive	Working Data	Backed-up	Access Control	PHI/PII Certified Secure	Version Control	Shareable	Public Access
Columbia University Affiliated	CUIT	LionMail Drive	Free	Unlimited	N	Y	Y	Y	N	N	Y	N
		Amazon Web Services (Enterprise Agreement with Columbia)	See below	Unlimited	N	Y	Y	Y	Y	N	Y	N
		High Performance Computing	See below	See below	N	N	N	Y	N	N	Y	N
		Virtual Servers w/ Storage	See below	See below	N	Y	Y	Y	N	N/A	Y	N
		Web Hosting	See below	See below	N	Y	Y	Y	N	Y	Y	Y
		Secure Data Enclave (SDE)	$526+/Yr*	See below	N	Y	N	Y	Y	N	N	N
		Microsoft Azure	See Below	Unlimited	N	Y	Y	Y	N	N	Y	N
		Globus	Free	N/A	N/A	N/A	N	Y	N	N/A	Y	N
		Cloud Research Computing Consulting Service	See Below	See below	N/A	Y	N/A	N/A	N	N/A	N/A	N/A
	CUIMCIT	CUIMC IT Storage	$1,172/TB/Yr	Increments in TB to UNL	Y*	Y	Y	Y	Y	Y*	Y	Y*
		CUIMC IT FTP Server	$258/MB	500 MB of storage up to 5 users*	N	Y	N	Y	Y*	N	Y	N
		SharePoint Online – Office 365	$1,004/site/Yr	5 TB/site	N	Y	Y	Y	Y	Y	Y	N
		Virtual Servers	$7,668/server/ Yr	See below	N*	Y	Y	Y	Y*	N/A	Y	N
	Libraries	Academic Commons	Free up to 10 GB/file (UNL # files), $5/GB/file after 10 GB/file	Up to 100 GB/file	Y	N	Y	N	N	N	Y	Y
	DSBIT	Data Storage	See below	See below	Y	Y	Y	Y	Y	N	Y	N
		High Performance Computing	See below	See below	N	N	N	Y	N	N	Y	N
		Virtual Servers	See below	See below	Y	Y	Y	Y	N	Y	Y	N
		Web Hosting	See below	See below	Y	Y	Y	Y	N	Y	Y	Y
		Colocation	See below	See below	N	N	N	Y	N	N	N	N
	ELN	LabArchives	Free with valid Columbia UNI	Unlimited	N*	Y	Y	Y	N	Y	Y	N

Commercial	Figshare	Free 20 GB private space; Free UNL public space	Private space: 1GB-20GB, Public Space: UNL	Y	Y	Y	Y	N	Y	Y	Y
	Dropbox	See pricing options below	2GB-UNL	Y	Y	Y	Y	N	N	Y	N
Open Source	Open Science Framework	Free	Unlimited	Y	Y	Y	Y	N	Y	Y	Y
Other	PC/Mac Server (w/ options)*	Depends on cost of machine	As much as willing to pay for	Y	Y	Y	Y*	N	Y*	N	N

*: See description below
****Note**: prices and availability may vary as noted in table above, see websites below for most up-to-date pricing and availability

CUIT

CUIT LionMail Drive: When you create new documents using LionMail Drive, you typically create online Google Docs, Spreadsheets and Slides. These programs are collaborative tools that allow users to share files and documents with multiple users. There are also web-based editors to create drawings, forms and fusion tables. These online documents are tightly integrated with other Google Apps and provide very powerful real-time collaboration features.
http://cuit.columbia.edu/lionmail-drive

CUIT Shared High-Performance Computing: Columbia's centrally-managed High-Performance Computing (HPC) resources on the Morningside Campus offers a Linux-based compute cluster. Options for entry include periodic purchase rounds, rental shares, or free access. See http://hpc.cc.columbia.edu or email rcs@columbia.edu for more information.

CUIT Virtual Servers: CUIT offers a managed virtual machine environment with configurable Storage, Compute, Backup and Disaster Recovery options and will assume responsibility for management as outlined in the SLA. Configurations can be customized based on your needs and prices will reflect accordingly. See https://columbia.service-now.com and go to *Catalog,* go to *Virtual Server Hosting*.

CUIT Web Hosting: CUIT Interactive Services offer a variety options to suit your web hosting needs. See http://webservices.columbia.edu

Amazon Web Services (AWS) through CUIT: Amazon has signed a BAA with Columbia. Amazon Web Services offers a broad set of global compute, storage, database, analytics, application, and deployment services that help organizations move faster, lower IT costs, and scale applications. Costs will be comparable to AWS list price, however aggregation of all linked Columbia University AWS accounts under a single University billing account will allow for discounts that will be shared amongst all customers. Pricing plans begin at 20 GB of storage, for $10 a year. The cost of storage is the number of GB divided by 2 per year. For example, for 100 GB of storage costs $50/year, 200 GB of storage costs $100/year, and so on. Additionally, CUIT will implement an AWS Direct Connect private network peering will lead to reduced data egress rates and some additional technical capabilities. All AWS purchases should be done through CUIT. For more information regarding CUIT and AWS enterprise agreement, email aws-request@columbia.edu. For more information see cuit.columbia.edu/aws

Secure Data Enclave (SDE): The SDE is a secure Windows 10 Virtual Desktop environment for the purposes of sensitive data analysis, allowing researchers to work on data requiring special handling, restricted access, or other security related measures outlined by the data agreement. The SDE is certified by CUMC Security as HIPAA compliant, and acceptable for use with personally identifiable information (PII), protected health information (PHI), and research health information (RHI). The SDE virtually emulates a traditional "cold-room" environment, where a desktop computer was locked in a room without internet access to protect the data and allow the researcher to run analysis safely.

Columba discounts are available for groups with multiple projects. Non-standard hardware configurations may incur additional costs. https://cuit.columbia.edu/sde

Microsoft Azure: Microsoft has signed a BAA with Columbia. See https://azure.microsoft.com for services offered.

Globus: With Globus at Columbia, subscribers can move, share, publish and discover data via a single interface – whether your files live on a supercomputer, lab cluster, tape archive, public cloud or your laptop, you can manage this data from anywhere, using your existing identities, via just a web browser.

Columbia has standard Globus licensing which entitles all Columbia researchers to the benefits without charge. Additionally, CUIT and CUIMC are planning on obtaining the Globus High Assurance license for Columbia University use that will support the use of data that requires additional protection.

If you'd like help using Globus, or would like access to the advanced features of Globus, please contact globus@columbia.edu. For additional information, please visit: https://cuit.columbia.edu/research-data-transfer

Cloud Research Computing Consulting Services: For research computing in the cloud, you can contact rcs@columbia.edu if you have questions or need assistance getting set up for Cloud services, particularly AWS.

CUIMCIT

CUIMC IT Storage: HIPAA compliant. Included with storage is "drop box" type of solution called DataAnywhere that allows Medical Center users to share files outside of the Medical Center via secure links. Archive solution as long as client continues to fund the storage space. Supports version control to previous version with certain limitations.
https://secure.cumc.columbia.edu/cumcit/secure/howto/remote/index.html

CUIMC IT FTP Server: Ideal for transient storage, such as transfer of large files**.** Access to server set up by CUMC IT, does not require a UNI or MC account. Intended for temporary storage. Currently under review for PHI/PII certification.
https://secure.cumc.columbia.edu/cumcit/secure/storage.html

SharePoint Online – Office 365: HIPAA compliant. CUIMC IT offers SharePoint Online – Office 365 for Medical Center groups and departments. A SharePoint site provides an intuitive area for collaboration online, including document, calendar and list sharing with only an approved MC Domain account required. These sites are managed within your group, allowing for granular levels of access based on your needs. Cannot recover files once deleted and files remain on site as long as client continues to pay.
http://www.cumc.columbia.edu/it/getting_help/online.html

Virtual Servers: Ideal for running applications or programs. Infrastructure powered by VMWare. Content remains on server as long as client continues to pay for services. Individual servers may become PHI/PII certified. CUIMC IT sets up, installs, and regularly monitors servers. Regular backups are performed. Backups to tape are stored in a secure off-site location. Typical setups include the following with additional storage and customization available: Windows - 80GB of hard disk space, 4GB of RAM, on Windows Server 2012 R2 Linux/LAMP - Linux, Apache, MySQL, and PHP
https://secure.cumc.columbia.edu/cumcit/secure/storage.html

Libraries

Academic Commons: Digital repository for Columbia University faculty, students, and staff and affiliates. Any digital content can be uploaded and is freely available to the public. A URL is given to each document uploaded so that it is citable.
http://academiccommons.columbia.edu/

DSBIT

Department of Systems Biology Information Technology (DSBIT) website:
https://systemsbiology.columbia.edu/dsbit

Department of Systems Biology Information Technology (DSBIT) High Performance Computing: Department of Systems Biology Information Technology (DSBIT) maintains several high-performance computing systems, including multiple high-performance compute clusters as well as high-memory systems. In 2013 we installed a new cluster with 6,336 CPU-cores and 73,728 CUDA-cores (GPU). It has a maximum performance of 212 TFlops, almost 9 times the performance of its predecessor. The system is on the Top500 list of supercomputers worldwide.
Additionally, DSBIT has two high-memory systems with 1 TB of system memory each, and a pool of computational servers for compilation, debugging, and job control. The DSBIT high-performance computing platform is available for researchers at Columbia University and elsewhere who conduct data-intensive research. Pricing for CPU time is determined on a CPU-per-hour basis.
https://systemsbiology.columbia.edu/high-performance-computing

Department of Systems Biology Information Technology (DSBIT) Data Storage: Storage services for a number of applications, ranging from desktop file storage to high-performance computing applications. A HIPAA compliant system, the 5 PB enterprise-grade storage system provides high-speed, redundant storage that is tightly integrated with the HPC cluster to support big data analyses. A disk-to-disk replication storage system is used for long-term backup. And a Scalar I2k 300 slot tape robot is used for off-site disaster recovery (DR) backup.
See website below for more information:
http://systemsbiology.columbia.edu/data-storage

Department of Systems Biology Information Technology (DSBIT) Hosting Services: DSBIT offers web and database hosting, server virtualization, data center colocation and desktop support.
For Information: https://systemsbiology.columbia.edu/colocation-and-server-hosting

Electronic Lab Notebook (ELN)

LabArchives: LabArchives has signed a BAA with Columbia. This service is provided by CUIT and the Libraries, in collaboration with EVPR, and is free to instructors and researchers with a valid Columbia UNI. Data within LabArchives is never deleted, but researchers should still consider an additional service for archiving their data.

LabArchives may be used on all CU campuses, but is not approved for clinical activities. LabArchives may be used in research (other than research studies involving the provision of health care services for which study subjects are billed).

Under Columbia's enterprise license agreement, Columbia users have access to both the Professional Edition and the Classroom Edition of LabArchives. Key features include:

- Many customizable options
- Secure, cloud-based program, with the ability to access anywhere, including on mobile devices
- Real-time 24/7 collaboration, sharing and feedback between professors, TA's, and students
- Automated backup, audit trail, and version control
- Unlimited storage and number of notebooks
- Classroom edition available for those teaching courses with a laboratory component

To learn more please visit: https://labnotebooks.columbia.edu/

Commercial*

Figshare: A cloud based system for securely managing research data. Any type of data can be uploaded with options for sharing with select people or making information publically available, discoverable, and citable. Many file formats are capable of being visualized within Figshare's website, without users requiring specific software. Data is backed up in multiple institutions around the world, DOIs provided by DataCite, content is hosted on Amazon Web Services, which provides virtually limitless file storage and fast upload and download times. Fulfills public access requirements for many funders and publishers. General features for FREE accounts include (https://figshare.com/features):

- Upload files up to 5GB
- 20 GB of free private space – choose when to make it public
- Unlimited public space
- DOI for work – extra citations for work
- Upload any file format – able to preview files in browser
- Accessible anywhere – cloud and web-based allows for access anywhere
- Desktop uploader – drag and drop files to upload
- Use the figshare API – automate research workflows
- Collaborative spaces – control access to files with collaborators
- Private link sharing – quickly share large files with a link
- Reserve a DOI – for those needing a DOI (for publications) but do not want to release data
- Collections – group content together to showcase research

For other pricing options, customers are encouraged to contact Figshare. http://figshare.com/

Dropbox: Files can be uploaded to Dropbox, then can be edited from any location and shared. Everything is private until user chooses to share with other parties. Files are secured with256-bit AES encryption and two-step verification. Three different pricing plans are available. Dropbox uses "refer a friend" benefits to acquire additional space.
Basic- Free up to 2GB
Pro- $9.99/month up to 100 GB
Business- Additional administration features and version control for 5 or more users for $15/user/month for unlimited space.
https://www.dropbox.com/

*Please note these commercial options are not Columbia University endorsed

Open-Source*

The Open Science Framework (osf.io): provides free and open source project management support for researchers across the entire research lifecycle. As a collaboration tool, researchers can work on projects privately with collaborators and make parts of their projects public, or make all of the project publicly accessible for broader dissemination. As a workflow system, researchers can connect the many services they use to streamline their process and increase efficiency - including Zotero, Mendeley, OpenSeasame, JASP Stats, R, and storage options like AmazonS3, Box, Dataverse, Dropbox, figshare, Git Hub, GoogleDrive, and OwnCloud. More connections coming soon to services like DMPTool, bitbucket, OneDrive, Dryad and Fedora/Hydra . As a flexible repository, researchers can store, share, and archive their research data, code, protocols, and materials from across a wide landscape of popular tools:

- Structured projects: Access files, data, code, and protocols in one centralized location and easily build custom organization for project
- Controlled access: Control which parts of a project are public or private, making it easy to collaborate and share with the community or just your team. Private sharing links, anonymized sharing, and persistent public urls make working & sharing anything connected to OSF project simple.
- Enhanced workflow: Automate version control, get persistent identifiers and DOIs for projects and materials, preregister research, and connect to third party services directly to the OSF using the OSF API (developer.osf.io)
- Extend Research: Automatically create a preprint or meeting abstract. Easily manage multi-institutional projects.

*Please note this option is not Columbia University endorsed

Other

PC/Mac Server: PIs and researchers may choose to set up their own private server housed within their research laboratory. Users who choose this option need to contact CUIT to set up a static IP address for the machine. The PI is responsible for the maintenance of the server, performing back-ups, and access control. Special considerations need to be considered for PHI/PII and other sensitive information, including keeping the server in a locked facility and ensuring properly functioning firewalls. For large amounts of data storage (several TB) can purchase networked attached storage (NAS) or other devices which can be on its own, or connected to server. Can elect to have an FTP, see below.
http://www.wikihow.com/Build-a-Fileserver
https://www.apple.com/mac-mini/server/

FTP Server: Uses a client-server design. FTPs are often secured using SSL/TLS. Many FTP options are available, including free services. PIs should exercise caution when choosing a server to transfer their research data, because they can be vulnerable to hacking.
http://www.slideshare.net/mwGSU11/choosing-an-ftp-client-8294642
http://www.mediacollege.com/internet/ftp/clients.html

Glossary

Access Control- PI has the ability to control who can view, alter, upload, and download content. Access is secured with user name and password.

Archive- All data can be saved for long-term (permanent) storage.

Backed-up- Data is regularly backed-up automatically with the PI involvement.

PHI/PII Certified- Certified by CUIT/CUMCIT for being the highest level of security possible for PHI and/or PII information.

Public Access- Options to make data publicly available to fulfill funders and/or publishers requirements.

Shareable- Able to share certain data, as decided by PI, with collaborators from all over the world.

UNL- Unlimited

Version Control- Protocol set in place for versioning of data files being used by multiple users.

Working Data- Data produced that is in preparation for publications, grant submissions, presentations, etc. that has not been formally published.