Research Data Storage Options

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 of digital research data. A glossary of the terms used in the summary is located at the end of this document.

UNIT	SERVICE	FREE	UNLIMITED	MUTABLE	BACKUPS	VERSION CONTROL	PII/RHI SECURE	PHI CERTIFIED SECURE	ACCESS CONTROL	SHARABLE	ARCHIVAL PRESERVATION	LONG-TERM STORAGE	PUBLIC ACCESS
CUIT	LionMail Drive	Yes	Yes	Yes	Yes	No	No	No	Yes	Yes	No	No	No
	Amazon Web Services (AWS)	Free Tier	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No
	Google Cloud Platform (GCP)	No	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	No	Yes	Yes
	Secure Data Enclave (SDE)	No	No	Yes	No	No	Yes	Yes	Yes	No	No	No	No
	Microsoft Azure	Free Tier	Yes	Yes	Yes	Yes	No	No	Yes	Yes	No	No	No
	Globus (data transfer)	Yes	N/A	N/A	No	N/A	No	No	Yes	Yes	No	N/A	No
	Columbia Data Platform	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes
	Box	Free Tier	Unlimited Tier	Yes	Yes	Yes	Yes	No	Yes	Yes	No	Yes	No
CUIMCIT	CUIMC IT Storage	No	Unlimited Tier	Yes	Yes	Yes*	Yes	Yes	Yes	Yes	No	Yes*	Yes*
	CUIMC IT FTP Server	No	No	Yes	No	No	Yes*	Yes*	Yes	Yes	No	No	No
	SharePoint Online – Office 365	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No
	Virtual Servers	No	No	Yes	Yes	N/A	Yes*	Yes*	Yes	Yes	No	No*	No
Libraries	Academic Commons	Yes	No	No	Yes	Yes*	No	No	No	Yes	Yes	Yes	Yes

UPDATED – 09.29.2022 - Roger Lefort

	Dryad	Yes	Yes	No	Yes	Yes	No	No	Yes	Yes	Yes	Yes	Yes
CUIT/Libraries	LabArchives	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No*	No
Other	PC/Mac Server (w/ options)*	No	Unlimited Tier	Yes	Yes	Yes*	No	No	Yes*	Yes	No	Yes	No

*See description below

Service Description

LionMail Drive: Available to LionMail users, it includes unlimited file storage in the cloud on LionMail Drive (Columbia's implementation of Google Drive). LionMail Drive offers many of the same features as Google Drive, with some adjustments to enhance data security at Columbia. Non LionMail users require a Gmail account to set-up Google Drive. Users can upload and store anything. Encrypted using SSL. Files are kept private until the user "invites" others to view selected files. Users can invite others to view files by entering valid email addresses for shared users. It is not intended as a permanent archive space for data, fees may be associated with the retrieval of old data files. Once a file is deleted, it cannot be recovered.

Amazon Web Services: Amazon S3 or Amazon Simple Storage Service is a service offered by Amazon Web Services (AWS) that provides object storage through a web service interface. Amazon S3 uses the same scalable storage infrastructure that Amazon.com uses to run its e-commerce network. Amazon S3 can store any type of object, which allows uses like storage for Internet applications, backups, disaster recovery, data archives, data lakes for analytics, and hybrid cloud storage. Amazon Cloud Drive is compatible with all Amazon devices. There are several templates available for customization, including version control. 5 GB is storage is free to all Amazon users. 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.

Google Cloud Platform: Google Cloud Storage is a RESTful online file storage web service for storing and accessing data on Google Cloud Platform infrastructure. The service combines the performance and scalability of Google's cloud with advanced security and sharing capabilities.

Secure Data Enclave: The Secure Data Enclave (SDE) provides Columbia researchers with a secure, remotely accessible, virtual Windows 10 desktop environment to store and collaboratively analyze PII and PHI data as an alternative to traditional "cold room"

UPDATED – 09.29.2022 - Roger Lefort

computing environments. Using a web browser, researchers can work on sensitive data and collaborate with other members of their project simultaneously. Researchers will only be able to access data explicitly placed in the virtual environment which is destroyed after use and is restricted so it can only reach other systems within the SDE. Data can only be transferred to and from the system by the designated Data Security Officer (DSO), required for each individual project.

Microsoft Azure: The Azure Storage platform is Microsoft's cloud storage solution for modern data storage scenarios. Azure Storage offers highly available, massively scalable, durable, and secure storage for a variety of data objects in the cloud. Azure Storage data objects are accessible from anywhere in the world over HTTP or HTTPS via a REST API. Azure Storage also offers client libraries for developers building applications or services with .NET, Java, Python, JavaScript, C++, and Go.

Globus (data transfer) Globus presents a secure, unified interface to identities and storage across Globus-connected sites, within the visibility and access control limits set by each site. Globus makes life easier for researchers with data on multiple systems and for system administrators who must support collaboration while maintaining secure systems. Globus is software as a service (SaaS), enabled by the cloud and built with widely-adopted industry standards. Globus works with existing systems and storage.

Columbia Data Platform: The Columbia Data Platform is a cloud-based solution for research data storage, discovery, analysis, collaboration and archive. Features include data storage and discovery, data analysis and exploration, collaboration and data archive (coming soon!).

Box: Box is a cloud-based content management system with collaboration, security, analytics and other features related to files and information. There is a core Box service, then add-ons for different industries and situations. Box can be used to manage, share, and collaborate on digital files. Example pricing plans below:

- Basic – Free up to 10 GB
- Pro (Personal) – $10/month for up to 100 GB
- Business Starter (Teams 3+) – $5/user/month for up to 100 GB
- Business Plus (Teams 3+) – $25/user/month for unlimited storage

CUMC IT Storage: HIPAA compliant. Included with storage is “drop box” solution. Archive solution as long as client continues to fund the storage space. Public access is in process of being implemented.

UPDATED – 09.29.2022 - Roger Lefort

CUMC 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. 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

SharePoint: CUMC IT offers SharePoint 2010 web sites 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.

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. CUMC IT sets up, installs, and regularly monitors servers. Regular backups are performed to a secure off-site location using Symantec NetBackup. Typical setups include the following with additional storage and customization available:

- Windows - 80GB of hard disk space, 4GB of RAM, on Windows Server 2008 R2
- Linux/LAMP - Linux, Apache, MySQL, and PHP

Academic Commons: Digital repository for Columbia University faculty, students, and staff and affiliates. Maintained by Center for Digital Research and Scholarship at Columbia University Libraries. 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.

Dryad: The Dryad Digital Repository is a curated resource that makes research data discoverable, freely reusable, and citable. Dryad provides a general-purpose home for a wide diversity of data types. Key features include:

- Flexible about data format, while encouraging the use and further development of research community standards.
- Fits into the manuscript submission workflow of its partner journals, making data submission easy.
- Assigns Digital Object Identifiers (DOIs) to data so that researchers can gain professional credit through data citation.
- Promotes data visibility through usage and download metrics and by allowing content to be indexed, searched and retrieved.
- Promotes data quality by employing professional curators to ensure the validity of the files and descriptive information.
- Contents are free to download and re-use under a Creative Commons Zero (CC0) license.

UPDATED – 09.29.2022 - Roger Lefort

- Contents are preserved for the long term to guarantee access to contents indefinitely.
- Open source, standards-compliant technology.

Lab Archives: Columbia University provides an Electronic Research Notebook service for researchers, instructors, and students. Electronic Research Notebooks are designed to replace paper notebooks and lab manuals to support research staff productivity and efficiency, and securely protect lab research with automatic backups and comprehensive audit trails. Our Electronic Research Notebook solution is powered by LabArchives, a cloud-based solution that enables 24/7 collaboration and sharing of content, from any device with a web browser.
LabArchives is approved for PII, RHI, and PHI. It is registered in RSAM #5644. Please note: At this time, LabArchives may be used on all CU campuses, but is approved only for use in research (excepting research studies that involve the provision of health care services for which study subjects are billed).
Electronic research notebooks also support funding agencies' Data Management Plan requirements, managing team progress remotely, and interconnecting all your lab data and image files to your observations and notes.
Columbia University has an enterprise license with LabArchives for both the Professional Edition and Classroom Edition of their electronic notebook solution.

Advanced Research Computing Services (ARCS) Data Storage: storage services for a number of applications, ranging from desktop file storage to high-performance computing applications. An Isilon clustered file system provides 1 PB of high-speed, redundant storage for our compute clusters and user data. A secondary Isilon clustered file system provides daily replication of valuable data to a secondary site as well as additional iSCSI Ethernet SAN storage for infrastructure support. Multiple Linux-based file servers provide storage for specific applications. A storage area network (SAN) provides 7.2 TB of reliable storage to a pool of database servers and backend storage for server virtualization. A large, scalable tape robot and pair of backup servers provided automated backups of all relevant storage to tape for long-term backup.

Service	Description	$/TB/Year
Home	Home-class storage	$1,500
Data	Data-class storage	$1,900
Scratch	Scratch-class storage	$800
Archive	Archive-class storage	$425

http://systemsbiology.columbia.edu/data-storage
http://systemsbiology.columbia.edu/advanced-research-computing-services

UPDATED – 09.29.2022 - Roger Lefort

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.

Glossary of Terms

- Mutable – Refers to a database structure in which data can be changed. Any data changes made simply overwrite and replace the previous record. This means that previous iterations of data are lost unless there is a system of back-ups and transaction logs that track changes.

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

- Version Control – Also known as revision control, source control, or source code management) version control is a class of systems responsible for managing changes to computer programs, documents, large web sites, or other collections of information.

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

- Archival Preservation – All data can be saved for long-term (permanent) storage.

- 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.

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

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

UPDATED – 09.29.2022 - Roger Lefort