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 are 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
	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
Commercial	Figshare	Free Tier	Unlimited Tier	Yes	Yes	Yes	No	No	Yes	Yes	No	Yes	Yes
	Dropbox	Free Tier	Unlimited Tier	Yes	Yes	No	No	No	Yes	Yes	No	Yes	No
Open Source	Open Science Framework	Yes	Yes	Yes	Yes	Yes	No	No	Yes	Yes	No	Yes	Yes
Other	PC/Mac Server (w/ options*)	No	Unlimited Tier	Yes	Yes	Yes*	No	No	Yes*	Yes	No	Yes	No

*See description below

Service Description

CUIMC IT Storage: HIPPA compliant. Included with storage is "drop box" solution. Archive solution available as long as client continues to fund the storage space. Public access is in process of being implemented.
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 CUIMC 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: CUIMC 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.
http://www.cumc.columbia.edu/it/getting_help/sharepoint.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. CUIMCIT 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

https://secure.cumc.columbia.edu/cumcit/secure/storage.html

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.
http://academiccommons.columbia.edu/

Advanced Research Computing Services (ARCS) Data Storage: storage services for various 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

LionMail Drive: Columbia University currently offers Google Drive access, available only to users with a LionMail account. Collaborating users will 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.
http://www.google.com/drive/index.html

Amazon Web Services: Any digital content can be uploaded and accessed remotely. 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.
https://www.amazon.com/clouddrive/learnmore#features-section

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 publicly 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. Figshare+ offers data deposit as a one-time Data Publishing Charge (DPC) to share the datasets and materials supporting a specific publication or project.

Figshare+ Pricing					
100GB	250GB	500GB	750GB	1TB	1.25TB
$395	$395	$395	$395	$395	$395
1.5TB	1.75TB	2TB	3TB	5TB	5TB+
$745	$745	$745	$745	$745	Contact

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. Example pricing plans below:

- Basic – Free up to 2GB
- Plus (Personal) – $9.99/month for up to 2,000 GB
- Professional (Business) – $19.99/month for up to 3,000 GB
- Advanced (Teams 3+) – $20/user/month for unlimited storage

https://www.dropbox.com/

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

https://www.box.com/

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 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/CUIMCIT 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.