UCCSC 2018

Supporting Active Research Data

Rick Jaffe, UC Berkeley

Research IT
Advancing Research@Berkeley

See presentation outline (draft):
https://docs.google.com/document/d/1u5jEFsl5Zomeo01bd6VaIr0srXkosrXX-ao1tzZD_TQ/edit?usp=sharing

Outline for today

1. Introduction, goal for this talk
2. RDM program history
3. Definition of ‘Active Data’
4. Common issues
5. Tools & approaches
6. Successes and not-so
7. New tools & approaches to try
8. Discussion

Research Data Management (RDM) at UC Berkeley

- Library + Research IT
- Goal: “To help campus researchers manage, share, store and archive their data”
- Service (soft) launch: September 2015

Research IT
Advancing Research@Berkeley

RDM Lifecycle

Active Data

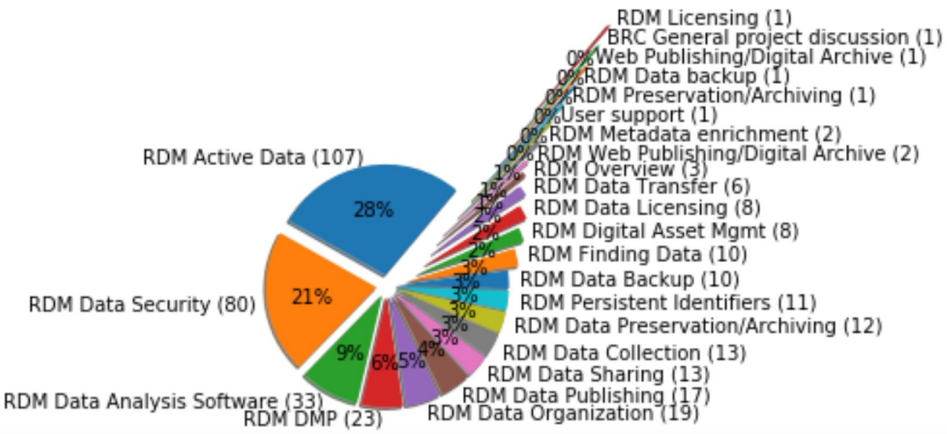
Includes:

Data migration, storage, workflow (analysis pipelines, lab workflows, etc.) surrounding the files created, collected, processed and analyzed while a study is in progress.
Topics extend to use of Box, Drive; prevention of data loss; accessibility to compute

Research IT
Advancing Research@Berkeley

400 consultations to researchers in 100 different departments or units

Active Data Consultations



Research IT
Advancing Research@Berkeley

Data for January 1, 2015 - December 31, 2017. 322 total consultations in that period.

Sample cases

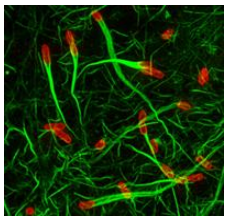
1. Researcher gathering text data via web-scraping
2. Documentary journalism program
3. Microscopy facility
4. Lab data protection (back-up) and sharing workflow

Research IT
Advancing Research@Berkeley

- Not covered here: Personally Identifiable Information and other sensitive, confidential and restricted use data. (With active data, comprises 50% of our caseload.)

Common issues: technical

- Computing and storage resources
- Data transfer methods, timing, tech skills required
- Speed of network, network routers and network interface cards on individual machines



Common issues: administrative/cultural

- Independent labs, disparate workflows
- Lab staffing and turnover; uneven levels of tech expertise
- We are often not deeply experienced in the discipline, or in doing the specific type of research.

Research IT
Advancing Research@Berkeley

- *Are there ways to treat disparate workflows as a general problem?*
- *Are there shared support models or collaborative self-help models?*
- *We try to use our lack of experience to our advantage by asking broader questions about the process, and simplifying or finding the right resource(s) for it.*

Common issues: strategic

- What are the (unstated) underlying issues?
- What resources are available on campus and elsewhere?
- Can we put together and support a work-around?
- When to advocate for funding for equitable + sustainable tools, services and support resources?

Keep these in mind for the discussion later

Tools & approaches

Storage services

- Box and Google Drive:
  Campus-supported, unlimited
- UC Berkeley-hosted storage, backup
- Cloud storage
- CASS (UCLA), SDSC

Tools & approaches

Data movement tools

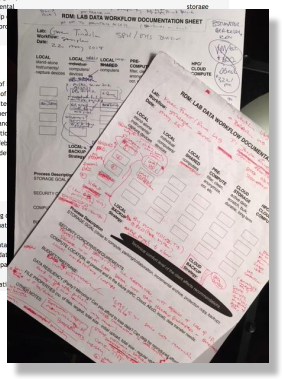
- FTP clients (FileZilla, lftp, etc.) (https://filezilla-project.org/; http://lftp.yar.ru/)
- Globus, Globus Personal Connect (https://www.globus.org/, https://www.globus.org/globus-connect-personal)
- Box and Google Drive-provided tools (Sync, Drive, Backup and Sync, etc.)
- Rclone (https://rclone.org/)
- Various third party tools, such as Expandrive

Tools & approaches

Diagnostic tools:

- Active Data Storage guide
- Lab data workflow documentation sheet

Title	Description	Best suited for	Not well suited for	Data protection levels	Cost	Connection methods
IST Performance Storage	The IST Storage and Backup team provides highly available and highly scalable systems that are offered in two billing tiers: Performance and Utility. Managed storage for high-performance applications. Hosted in UCB data center (with UCSD backup available). No direct web access.	Storing data during data preparation and high performance computation Large group file sharing, retaining group or departmental ownership Off-site pr copies		PL0, PL1, PL2	Recurring subscription	Block storage, Mountable file-based storage
IST Utility Storage	The IST Storage Team offers two data storage options depending on what type of access is needed: Storage Area Network (SAN) and Network Attached Storage (NAS). These are both highly available and highly scalable systems that are offered in two billing tiers: Performance and Utility. No direct web access. Managed storage for less I/O intensive needs or less frequently accessed materials. Hosted in UCB data center (with UCSD backup available).	Storage of volumes of with limited requiremer performan computati shares Web content de Backups				
bDrive	bDrive is a collaborative authoring platform where you can store files and collaborate with collaborators. It is an enterprise version of Google Drive, which means that it is used under an agreement approved by UC Regents. UC does not accept the vendor's	Gathering source ma and documenta Parking da later prepa analysis; Collaborati				



Active Research Data Storage Guidance Grid (detail)

Title	Description	Best suited for	Not well suited for	Data protection levels	Cost	Connection methods
IST Performance Storage	The IST Storage and Backup team provides highly available and highly scalable systems that are offered in two billing tiers: Performance and Utility. Managed storage for high-performance applications. Hosted in UCB data center (with UCSD backup available). No direct web access.	Storing data during data preparation and high performance computation Large group file sharing, retaining group or departmental ownership of files Off-site protection copies		PL0, PL1, PL2	Recurring subscription	Block storage, Mountable file-based storage
IST Utility Storage	The IST Storage Team offers two data storage options depending on what type of access is needed: Storage Area Network (SAN) and Network Attached Storage (NAS). These are both highly available and highly scalable systems that are offered in two billing tiers: Performance and Utility. No direct web access. Managed storage for less I/O intensive needs or less frequently accessed materials. Hosted in UCB data center (with UCSD backup available).	Storage of large volumes of data with limited I/O requirements Low performance computation File shares Web content delivery Backups	High performance applications	PL0, PL1, PL2	Recurring subscription	Block storage, Mountable file-based storage
bDrive	bDrive is a collaborative authoring platform where you can store files and collaborate with collaborators. It is an enterprise version of Google Drive, which means that it is used under an agreement approved by UC Regents. UC does not accept the vendor's requirement that we waive their	Gathering data, source materials, and documentation; Parking data for later preparation, analysis; Collaboration;	Backup	PL0, PL1	Free to UC users	API, Syncing app, Web browser

(link)

RDM: LAB DATA WORKFLOW DOCUMENTATION SHEET

Lab:
Workflow:
Date:

LOCAL stand-alone instruments/ capture devices	**LOCAL** individual computers/ devices	**LOCAL SHARED** computers	**PRE-COMPUTE** filter, clean, anonymize, ocr, etc.	**CLOUD STORAGE** active data: scratch, bulk, durable, long-term	**HPC/ CLOUD COMPUTE**

LOCAL BACKUP Strategy:

CLOUD BACKUP Strategy:

Technical comfort level of the client affects recommendations.

Process Description

STORAGE GOAL (close to compute, parking/collaboration, incremental archive, protection copy, backup):

SECURITY CONCERNS/REQUIREMENTS:

COMPUTE LOCATION, at present and in the future (HPC, Cloud, AEoD, local), data transfer needs:

BUDGET/TIMEFRAME:

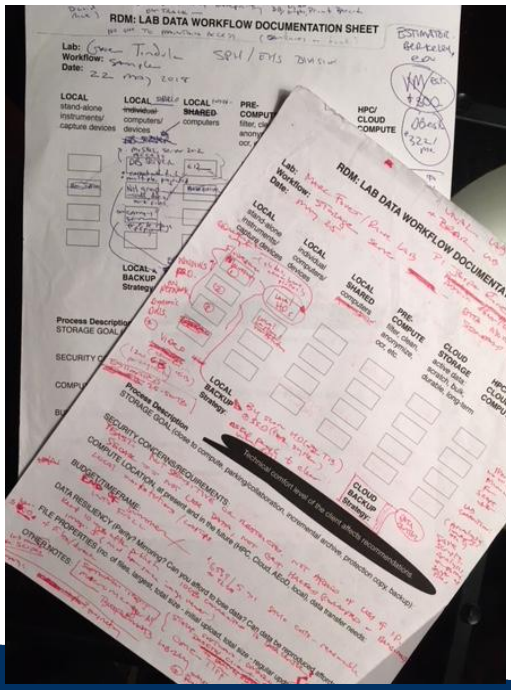
DATA RESILIENCY (Parity? Mirroring? Can you afford to lose data? Can data be reproduced affordably?):

FILE PROPERTIES (no. of files, largest, total size - initial upload, total size - regular update):

OTHER NOTES:

rev. 5-27-2016a

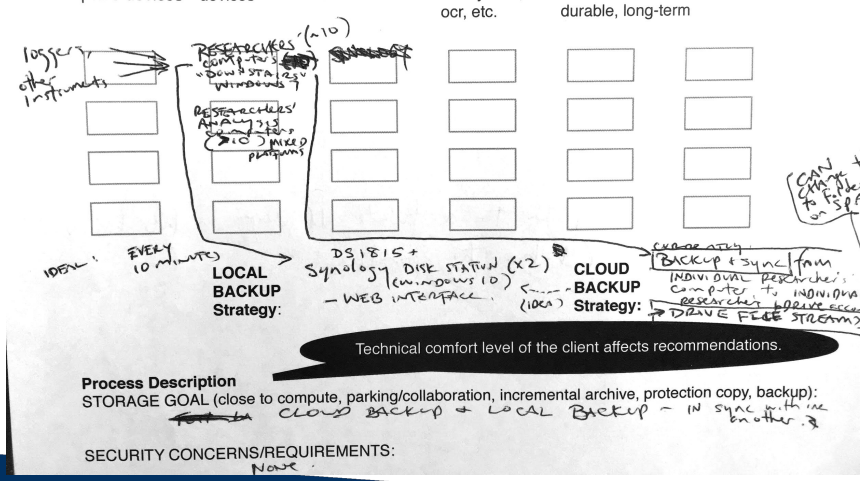
Research IT
Advancing Research@Berkeley



Research IT
Advancing Research@Berkeley

Workflow: LOCAL + CLOUD BACKUP
Date: 8/10/2018

LOCAL stand-alone instruments/ capture devices	**LOCAL** individual computers/ devices	**LOCAL SHARED** computers	**PRE-COMPUTE** filter, clean, anonymize, ocr, etc.	**CLOUD STORAGE** active data: scratch, bulk, durable, long-term	**HPC/ CLOUD COMPUTE**



Technical comfort level of the client affects recommendations.

Process Description
STORAGE GOAL (close to compute, parking/collaboration, incremental archive, protection copy, backup):
~~Parking~~ CLOUD BACKUP + LOCAL BACKUP - IN SYNC with one another.

SECURITY CONCERNS/REQUIREMENTS:
None

Success: Drive via command-line

Journalism case

- Unlimited storage on Google Drive
- Media cards on professional cameras have a storage capacity of 128GB.
- Transfer time to Google Drive ~ 2-1/2 hours.
- Easy to upload one or more cards overnight

>> rclone, Google Drive (bDrive); Associate Producer role

This case succeeded in large part because of the existence of an Associate Producer role within the video production unit. The Associate Associate Producer was responsible for handling all footage, and, as the position has its technical elements, capable of managing the file management, encryption, upload and internal instruction/training processes. The workflow successfully survived a change in Associate Producer, which pleased us greatly!

Success: “Cloud” resources

Web-scraping case

- XSEDE: Jetstream (compute, storage)
- Campus Champion
- Python scripts running in Jupyter notebook to gather data, access Box API
- Unlimited storage in Box

>> OCR, Photogrammetry, high performance computing

Research IT
Advancing Research@Berkeley

Move web-scraping scripts to Jetstream to take advantage of network speed, multithreaded capacity, bandwidth, etc.). Push scraped data by script to campus-supported Box. Similar cases: OCR, photogrammetry. These make use of campus HPC infrastructure (Savio), Box/Google Drive, Jupyter notebooks. Even using Kubernetes to replicate notebook set-up so undergraduate research team can do the work.

Jetstream is such a valuable resource on the compute side of research; how can we do something similar for data?

Links to XSEDE and JetStream, noting Jetstream’s associated storage (max 500GB): https://www.xsede.org/ ; https://jetstream-cloud.org/ ; Jetstream storage volumes
Link to Savio home page (http://research-it.berkeley.edu/services/high-performance-computing)

Not so successful…yet

Microscopy facility

- High storage costs, management time required.
- Facilities might support a system admin, but labs struggle

>> “I support three scopes in three different buildings”
>> “I’m a biologist, but I spend my time moving files to Box”
>> Output directories containing > 30,000 files

Research IT
Advancing Research@Berkeley

This has proven to be a hard nut to crack due in part due to storage costs and management time/capability required. Some facilities can support a system admin, but the labs struggle to develop and maintain the expertise to install and run the system.

“I’m responsible for three microscopes in three different buildings. Having scripts on each machine running in the background to push data somewhere via Globus presents too many potential headaches for me to take on.”

Box FTP can only display only 20,000 files at a time, so any FTP command to test the data before writing (newer than, larger than, etc.) fails. [Perhaps checksum would fail, too.]

Not so successful…yet

Lab data protection and sharing workflows

- Departmental accounts (“SPAs”) for Box, bDrive
- Consider strategy for zipping/tarring files
- File naming, folder organization, versioning, metadata

>> Lab culture / turnover
>> Too much work; labs disappear from conversation

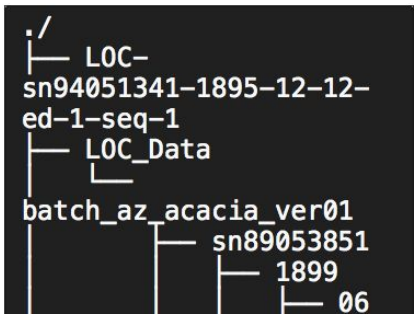
Research IT
Advancing Research@Berkeley

Another example: 3TBs shared with a colleague at a different institution, or needing to be moved from the researcher’s previous university.
Profile data: How many files? How often are they created or new ones come in? How large individually? In total? Other copies?
Data sensitivity.
Immediately get a copy *somewhere* (to Box, Drive, an external HD...) if there is only one copy of the original data sitting on a local machine!
Managing data is not the primary driver for the lab, but fear of data loss motivates them.

New tools & approaches to try

Diagnostic tools:

- Unix shell tools (easier in a Jupyter notebook?) DEMO



Research IT
Advancing Research@Berkeley

This notebook (Shell_play-excerpt) is available to be cloned at:
https://github.com/rjaffe/rdm_datatransfer/
The Library of Congress (LOC) data used in this demo is available at:
https://goo.gl/XECaJK
(Full URL:
https://drive.google.com/file/d/1I0P5J-LM1DNJD6u6A7CmeIH_DtB1v3-5/view?usp=sharing)
To use the LOC data with this notebook on your computer, download the LOC_Data.zip archive to the 'test data' folder included in the notebook code and unzip the archive.

New tools & approaches to try

Knowledge sharing:

- Carpentries- or bootcamp-style training
- Embed undergrads in research units to manage workflows
- Part-time hire of data manager as consultant
- Lab manager working groups

New tools & approaches to try

Architectural experiments:

- Staging server inserted into workflow
- Eventually: augment network to Data Transfer Node?

Research IT
Advancing Research@Berkeley

Infrequent; some resources for projects of this type. Proof-of-concept efforts to design systems that can provide the capabilities

Discussion

Do these case studies sound familiar at your institutions?
What problems are you encountering?
What tools and methods are you using?

Discussion (cont.)

Specific strategies for broader support responses

- How to reach more researchers at one time?
- Any experiences, and tips for, organizing across labs?
- Experience with “embedding” in a lab? Or bringing on a research project’s data manager as a partner/resource to share expertise with other researchers?

Thank you!

UC Berkeley Research Data Management

researchdata@berkeley.edu

https://researchdata.berkeley.edu

Research IT
Advancing Research@Berkeley