UCCSC2018 - Rick Jaffe - Supporting Active Research Data
UCCSC 2018 program announcement for talk:
https://uccsc.ucdavis.edu/sessions/supporting-active-research-data

[Slide 1: Title Slide]

1. Introduction, goal for this talk

Introduce myself: I am Rick Jaffe, Research Data Consultant with the Research Data Management program at UC Berkeley.

[Slide 2: 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

Goal: Today I will talk about an aspect of the research process that I, and my program, have worked hard to support over the past three and a half years. The service area, which we call Active Research Data Management or simply, Active Data -- I'll define this in a moment -- presents many challenges to UC Berkeley researchers and constitutes a large portion of the RDM Program's consulting efforts. As you will hear, we have had success in some cases, but have reached an impasse in many others. We want to do better at helping more researchers clear the hurdles that they face in this area; to produce a larger, more structural impact. As we wrestle with strategies and tactics for improving our service, we recognize that it's a good moment to reach out to colleagues at other campuses -- you -- for ideas, experiences and expertise.

Here's the outline for today's talk. I hope you'll recognize some of the common issues we face – technical, administrative/cultural, and strategic. Perhaps, the diagnostic and data movement tools that we've put into use -- however rudimentary or speculative they still are -- will be valuable. Lastly, I hope we can have time for a good discussion.

2. RDM program history

[Slide 3] Begun in early 2015, Research Data Management at UC Berkeley is a partnership of the Library and Research IT. The program's goal: "to help campus researchers manage, store, share and archive their data."[1]

[1] From RDM "Modest launch communications" planning document, September 2015: https://docs.google.com/document/d/1MRFQ0pmeyA-GzASqfGzTSBBWluGEoWyYdZK_SCl-gW0 (login req)

We organized a small team of IT staff and librarians and put out a web-based RDM guide and an email address for consultation requests, and offered support across the entire research data lifecycle.

3. Define "Active Data":

[Slide 4] So what is 'active research data'?

[Use RDM Lifecycle as a guide] Shorthand for the management of data during the active phase of a research project. Roughly, all the work between, on the one hand, planning and organizing a research project, and, on the other, sharing and preserving the products of that research. That is, between finding sources of data or preparing for data collection (working with the library to gain access to data sets; compliance with web page and publisher terms of service, and data use agreements; making arrangements for sensors, telescope/microscope access, surveys and other instruments; etc.) and sharing and publication of data for use by other researchers (metadata markup, obtaining permanent URLs; choosing a data repository, etc.).

[Slide 5] More precisely: Data migration, storage and workflow surrounding the files created, collected, processed and analyzed while a study is in process.

[Slide 6] We've done nearly 400 consultations in that time: 322 by the end of 2017, when this chart was created, plus 51 more in the first quarter of 2018. Perhaps because of our foundation in IT, as well as in the Library, we've found that nearly 30% of those consultations have dealt with active data.

[Slide 7]

Cases encountered on UC Berkeley campus

- Researcher gathering text data via large-scale web-scraping - Charter school data; needs greater computer speed, network speed and bandwidth, storage. Similar: optical character recognition (OCR) of scanned text corpora; Photogrammetry of museum objects.

- Documentary journalism program - needs to safely store footage from video production shoots. Footage will be copied to local workstations for logging, transcribing and editing, but there needs to be a protection copy stored offsite.

- Microscopy facilities - seek help to alleviate data bottleneck at acquisition computer, improve support for data backup and protection (beyond the current practice of researchers taking their data on a flash drive), extend capacity of local HD. Also: ease or automate transfer to analytic tools and services.

Cases: QB3; CRL MIC facilities; plus, various labs on campus

- Lab data protection (cloud backup) and sharing workflow - Labs request help to protect against data loss due to accidental deletion (esp. easy with Sync tools), or because an external hard drive is reaching capacity, or because it is maintaining the data on a local drive with no backup, or out of fear of data ransoming. Also, because the PI needs access to data while traveling; to all lab members' work; and/or to retain access as lab members move on.

  Case(s): More than a dozen labs on campus, in fields ranging from molecular biology and freshwater ecology to materials science and chemistry.

- *Not covered here: Personally Identifiable Information and other sensitive, confidential and restricted use data. Consultations re: these types of data comprise an equally large share of our overall requests. This area has grown into a major initiative of its own, addressing topics of guidance and policy, consulting and community, and tools and services, as we attempt to build the necessary collaboration and understanding between campus offices; build expertise among researchers; and define ready-to-use or easily-implemented secure environments on various computing and storage platforms. This is a story of its own, for a different time.*

4. Common Issues

- [Slide 8] Technical:
  - Computing and storage resources
  - Data transfer methods, timing, tech skills required
  - Speed of network, network routers and network interface cards on individual machines. (Notion of building up fast internal access to DTN and from there to PRP, etc.)
- [Slide 9] Administrative/cultural:
  - Disparate research workflows created by independent labs require individual analysis, troubleshooting and, often, solutions
  - Lab staffing and turnover; uneven levels of technical expertise. Typical scenario: grad student is tasked with developing a system, or inherits a poorly-documented one.
  - We are often not deeply experienced in the discipline, or in doing the specific type of research. There is always a lot for us to learn -- which we try to use to our advantage by asking broader questions about the process, and simplifying or finding the right resource(s) for it.
- [Slide 10] Strategic:
  - What are the (unstated) underlying issues involved?
  - 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?

5. Tools & approaches

- [Slide 11] Storage services:
    - Box and Google Drive: campus-supported, unlimited
    - UC Berkeley-hosted storage, backup
    - Cloud storage
    - CASS (UCLA), SDSC
- [Slide 12] 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/), grive2
    - Unix/Linux shell commands
    - Various third party tools, such as Expandrive
- [Slide 13] Diagnostic tools used:
    - [Slide 14] Active Research Data Storage Guidance Grid (link)
    - [Slide 15] Workflow diagram (paper form): RDM Lab data workflow documentation sheet
    - [Slide 16] Photo: completed sheets - *skipped*
    - [Slide 17] Photo: completed sheet, close

6. Successes & not-so

- [Slide 18] Success: Drive via command-line
    - Documentary journalism program - This case turned out to be an easy one. Media cards on professional cameras have a storage capacity of 128GB. Transfer time to Google Drive ~ 2-1/2 hours. Easy to upload at least one card overnight. Even if the production team returned from a week in the field with 8-10 media cards full of footage, the data could be uploaded without a backlog forming.

      Technique:
      • Transfer to Google Drive using rclone, each upload fired off manually. The rclone copy command is simple enough to teach to someone familiar with the command line. In this case, we may have introduced RClone Browser, a graphical user interface tool, to make the process even less tech-y.
    - 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!

- [Slide 19] Success: “Cloud” resources
    - 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 JetSt
    - 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)
- [Slide 20] Not so successful...yet: Microscopy facility
    - 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.]

- [Slide 21] Not so successful...yet: Lab data protection (cloud backup) and sharing workflows - This is often hard, too.

  Approaches:
  • Use of departmental (“special purpose”) accounts for Box and Drive.
  • 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.
  • Profile data, develop strategy for zipping leaf nodes as appropriate, depending upon the data needs for computational analysis.
    - • File naming, folder structure, versioning, metadata

      Downsides: Often too much work, effort. Managing data is not the primary driver for the lab, so the lab disappears from the conversation.

- 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?

7. New tools & approaches to try

- [Slide 22] New tools & approaches to try
  - Diagnostic tools:
    Unix shell tools (easier in a Jupyter Notebook?)
    DEMO of: Jupyter Notebook to assess data profile on a local computer prior to data transfer - very early draft (https://github.com/rjaffe/rdm_datatransfer)
    [Use excerpted notebook, running locally, for demo.
    >> Have Jupyter notebook running in advance.]
  - 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.

  - [Slide 23] 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
  - [Slide 24] Architectural experiments:
    - Staging server inserted into workflow
    - Eventually: augment network to Data Transfer Node?

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

8. Discussion

[Slide 25] Ask audience if these case studies sound familiar at their institutions.
What problems are they encountering?
What tools and methods are they using?

[Slide 26] Also:

- *Strategies for broader responses (i.e., 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?*

[Slide 27] Thanks! + Contact info

EXTRAS:

- **SIDEBAR: Overall data transfer** *(Leave for discussion, if it comes up)*

  Globus, with Drive (and anticipated Box) connectors, is promising, but too expensive for campus.

  Meanwhile, we’ve had success with rclone. Rclone is command-line only, which limits its audience. RClone Browser works, but is not well supported so we have refrained from recommending it. *[This episode provides insight into our understanding of the role of IT in supporting research.]* For Box, FileZilla has been useful.

- Education:
  - Consulting: Essentially, one-on-one support in response to a direct request from the researcher. We help analyze the problem, provide guidance and/or share best practices, and learn the subtleties of ‘real-world’ research workflows. We use the knowledge and experience gained to improve our training and other support efforts, and to shape advocacy for campus support of tools, services and processes.

    Our team is made up of folks with different backgrounds and experience, disciplinary and otherwise. We try to include additional team members, or subject librarians, in a consultation, both because consultations often call for multiple sets of expertise and to facilitate cross-training as a means to expand the pool and its expertise.
  - Training: Provided to:
    - Labs, at the request of a PI or lab manager
    - Classes, at the request of the instructor (usually more general in nature)
    - Self-organized groups of students (by discipline or area), at the students’ request.

    Often, the connection with the researcher is made via a reference by a librarian. If not, we try to include the appropriate subject liaison librarian.
  - Architectural experiments (infrequent; some resources for projects of this type): Proof-of-concept efforts to design systems that can provide the capabilities

identified through consultations and follow-on meetings. First step towards advocacy for a supported service.

- Working groups (supported self-help; aspirational): Researchers meet with peers, supported by RDM staff, to discuss their needs, share solutions across the group in a many-to-many (researchers to researchers) and one- or few-to-many (RDM staff to researchers) manner. This is envisioned as a way to spread knowledge and practices, and to scale support resources.