We'll focus this week on Google Drive, which recently announced unlimited (no quota) storage for all Google Apps for Education customers, including UC Berkeley.

Discussion will include examination of issues that Google's service presents, as well as storage on Google Drive through the lens of use cases that will be described by researchers who plan to participate in the discussion. We'll also touch on the campus Box.com offering, which allows for association of custom metadata templates with stored files.

**When:** Thursday, December 4 from noon - 1pm  
**Where:** 200C Warren Hall, 2195 Hearst St (see building access instructions on parent page).  
**Event format:** The reading group is a brown bag lunch (bring your own) with a short ~20 min talk followed by ~40 min group discussion.

The presentation and discussion will be facilitated by Research IT's Chris Hoffman. Bill Allison (IST - Architecture & Platform Integration) will present on the campus's Google Drive offering.

**Please read/review the following in advance of the 12/04 meeting**

---

- [Google's announcement of unlimited storage for Apps for Education customers (30 Sep 2014)]
- [Google Drive: Flaw Costs New Client Thousands of Dollars]
- [Comparison tool being developed to help provide guidance about collaboration tools at Berkeley: https://kb.berkeley.edu/page.php?id=44390]
- [Box.com metadata tutorial]

---

**Attending:**

Aaron Culich, Research IT  
Aron Roberts, Research IT  
Camille Villa, Research IT  
Chris Hoffman, Research IT  
Glen Jackson, Research IT  
Quinn Dombrowski, Research IT  
Patrick Schmitz, Research IT  
Rick Jaffe, Research IT  
Alex Walton, IST  
Bill Allison, IST-API  
Brian Waechter, L&S IT  
Dav Clark, D-Lab, BIDS, EdEx  
Ian Crew, IST-API  
James McCarthy, SSL  
Jon Stiles, D-Lab  
Jon Skelton, IST  
Marilyn Saarni, LBNL Earth Science Division  
Nico Tripcevich, Arch Research Facility  
Neil Maxwell, Research Admin IT  
Perry Willet, CDL  
Raymond Yee, most recently w/ School of Information  
Ron Sprouse, Linguistics  
Scott Peterson, Doe Library

**Framing questions:**

- Is "unlimited storage" a game changer? What does it mean, or what does it NOT mean?
- What are the factors researchers consider when selecting a service or combination of services to use for storing their data (and...
Chris: Researchers or those who work with researchers here today will help advance our conversation about how the available services that Bill and Ian will talk about today. We'll likely look in a future reading group about security and protected data issues. Additional conversations later in the spring will include the storage team and CDL bringing us a look at their work and perspectives. There will also be a meeting later this month among all the UCs at which RDM will be a major topic, so I'll have more to report about that.

Bill: Remember back about four years to the time when the campus was looking at outsourcing email; the CalMail crisis (outage) of that time. Berkeley's publication of thought process that led to decision to go with GMail had Steve Balmer throwing chairs around the office (Bill heard...). Our decision was organized around addressing mission needs: research, teaching, public service; not administration. Looking not only at email but at larger suite of services we were likely to want to and have to support around what we now know as 'cloud services.'

Bill: We are hiring a service manager. We had people to run mail and calendar, but not bandwidth to support other cloud/collaborative tools. The service manager will beef up a very thin staffing; the hire will permit us to turn on Google Groups. Because UC is a constitutional corporation, we have to have restrictions on the services we use from Google, and aligning our requirements for terms of service with services outside what Google terms the "Core" has been legally problematic. We're working on it: the short story is that we have a short-term solution in the works that meets legal requirements. It will have to do with requesting exceptions where there's an academic need for services not currently/yet enabled for the campus as a whole.

Bill: A bucket of data is not nearly as useful as a bucket of data associated with metadata and accessible via API.

Bill: Box will remain "open" for UC business at least through June 2017, a changed 'ruling' since it looked like UCB's Box engagement was going to be deprecated due to budget pressures. We can't offer Box unlimited now, but we as a campus have a lot of extra storage space in our current allocation (~75% of 100TB) and are working to see how we're going to possibly offer Box in unlimited mode. Stay tuned, 3-4 month timeline to work this out.

Ian: Currently 50GB default. CSS can bump a personal account to 100 under their own authorization; a departmental account to 250GB. But even that is 'bumpable' with any reasonable justification.

Patrick: We've got a possible need for 170TB of data. An outlier, for now, but needs at these scales do occur in research space.

Bill: Let us know about these. Can pull in telecom, use advanced file compression technologies. Not sure how these work with Google currently, but they seem interested in working with these.

Patrick: E.g., Google as endpoint for Globus Online. ScienceDMZ integration also worth a look.

Aaron: EECS is interested in Google Cloud platform and what Berkeley wants to do there. Eric Brewer (at EECS and VP of Google Infrastructure)

Chris: Limitations that you (Ian) have seen with Google Drive and Box?

Ian: Both Drive and Box have strengths and weaknesses. No perfect/right answer. I am thinking now that Google's strength is not in Drive but in Apps offering, collaborative / real-time work. Box is strong in space of EFSS (Enterprise File Sync and Share) re: handling existing documents. Word docs, spreadsheets, whatnot. Box supports more file formats, preview of larger file formats -- that kind of feature exceeds Google's. Box supports metadata (cf. reading list for today's meeting), whereas Google currently has only a description field for each file. Metadata templates enabled for an enterprise (not individual); we've got several templates in UCB's pipeline: EXIF, IPTC, Dublin Core, CDWA Lite, and VRA Core.

Ian: Google does not offer a level of permission that allows someone to edit and upload yet be forbidden to delete docs and modify the folder structure. This can be a serious problem when coupled with Drive Sync -- local deletion propagated to Drive, deletes for everyone; and restoral (requires involved manual process, contacting Google) restores files out of folder structure. More finely grained permissions are available in Box.

Alex W: There's a non-obvious setting on Google Driver permissions: owner reserves right to change who can see doc. This prevents deletion of commonly-held copies. Only local deletion.

Chris: Let's hear from some researchers present...

Raymond Yee: Dropbox Pro, $100/year. On limited budget, I liked that there's a fixed cost, predictable, does not need monitoring. Programmability around Dropbox storage is attractive (API). Computing over 33,000 photos in personal Flickr account. Use AWS for computation, so am looking to minimize pulling data out of AWS (costly), but okay to push data in for computation (no charge). Not on campus, I have a slow connection (home); so would like to move data not through home connection, but among services with fat pipes.

Chris: Mark Ingles and I talked about Physics and Chemistry, whose faculty use Dropbox. Mark explained that a staff member leaving a group was linked to the PI's Dropbox account. The staff member was subjected to a phishing attack, caused loss of tens of thousands of files. Now talking about moving from Dropbox to Box -- but that's a hard sell given inertia of Dropbox use. Catastrophic loss does happen, though, and we need to think about protecting from these (very costly and disruptive) instances.

Dav: EdEx data. Processing, archiving, making data available. Need data to be on a server that does computational work. Have been using Box for manual kinds of work, but in institutionalizing (off my laptop) -- looking at benefits/costs of providing these files via the a web server (in place once institutionalized processing -- off personal laptop), via Amazon S3, via Box. You could have it on Box but you need it someplace else as well (for computation).
Patrick: Maybe Box copy to avoid egress charges from S3?

Dav: Yes, that's a factor.

Ian: A lot of these services seem to be fast enough to to pull files down onto a server for computation

Aaron: How Drive integrates with processing platforms is critical. Also, as Raymond said, predictable costs.

Marilyn S: Use case is that we must prove to Federal Government that we archive data that is meant to be available permanently. Giant data sets in climate change, the field in which I work. I don't hear that being addressed, but it's a problem that needs to be dealt with.

Dav: Regents own the data that needs to be archived, and they assert that ownership in order to fulfill legal requirements imposed by the Federal Gov't. So maybe the question is how do we serve researchers, the UC (Regents/institution), and address the Federal (and long-term) requirements.

Chris: We'll be talking with the IST Storage team about addressing this at a campus level.

Marilyn: At a UCOP level is where it needs to be created.

Perry W: That's why I'm here (CDL)

Ian: On Google Drive, institutional ownership of folders can automate institutional ownership of files created in the folder.

Nico: Did a project using apps and Google Earth Engine project w/ Google. Google has uploaded a lot of USGS data. A lot more going on at Google in terms of public data and computation over it than at Box.

Perry: Pay once store for 10 years is in works.

Bill: Hope to keep quotaless Google Apps for Edu for alumni.

Bill: Note that in serious cases contact bConnected team for restoral.

Bill: We have associated with Dropbox ad hoc berkeley.edu accounts -- 64 TB in 21,665 accounts / 540 $100/year; 70 are business accounts; our marching orders are to do fewer things, but in talking to Dropbox I've told them that if they pull in a lot of Berkeley customers because they have the service that's needed, there may be opportunities to talk further....