2014-01-30 Data management in science domains

Announcement
We'll be joined by staff who are well-versed in current data issues in the sciences: Kathy Durkin (from the Molecular Graphics and Computation Facility in Chemistry, [http://glab.cchem.berkeley.edu](http://glab.cchem.berkeley.edu)) and Donna Hendrix (of the QB3 Institute, [http://qb3.berkeley.edu](http://qb3.berkeley.edu)). Kathy and Donna will talk about some of their current practices and the data management needs they are coming across in their respective domains.

We're also inviting a task force of the IT Managers' Forum (ITMF) that is looking at storage needs with a particular emphasis on research and scholarship.

The readings below provide some context for the discussions but are slightly broader in scope.

**Faculty/Researcher Survey on Data Curation (2013) by Greg Janée**
A survey of UCSB faculty about their data management (data curation) needs. The section on "Help Needed" is especially interesting and could help us think about the kinds of help people at UCB need.

**DMVitals: A Data Management Assessment Recommendations Tool (2012) by Andrew Sallans, Sherry Lake**
Slide deck describing methods used by a team at U. Virginia to assess needs for their campus and individual scholars

Optional readings:

**Data Sharing by Scientists: Practices and Perceptions** (2011) by Carol Tenopir, Suzie Allard, Kimberly Douglass, Arsev Umur Aydinoglu, Lei Wu, Eleanor Read, Maribeth Manoff, and Mike Frame (PLoS One. 2011; 6(6): e21101) [Long article, more about issues related to data sharing, but very interesting]

And just for kicks, we'll include the DataONE primer on research data management

Notes
Santa Barbara survey - what kind of help do you need re: data management activities?
- Cluster of needs around storage (backups, archives, repositories, etc.)
- Where do you store it? In long, short term?
- Other clusters around getting more information, metadata, funding
- Responsibility tri-plots: who do scholars think is responsible for data management?
- Scholars think they're responsible for curation, but other parties are involved
- Local department vs external repository vs campus (library)
- Where do scholars turn to for collaboration?
- Humanities and social sciences were largely "not me", sciences clustered around "me" (local department, external repository)

Questions:
- Do researchers understand what data management is, that is the broader aspects beyond storage and backup? What terms/perspectives are gaining traction -- data management, curation, lifecycle?
- Who is responsible for data curation/management?
- What are unique aspects of data that make it difficult to curate/manage?
- What are biggest challenges to campus in data lifecycle? (pre-award, post-award, active award? plan, create, archive, store)

Kathy Durkin:
- Chemistry: don't know what we're doing, who to turn to
- Ad-hoc data management
  - 10-20 computational clusters, in data center, 1-2 PIs own each, 1 large shared cluster
  - Individually owned clusters - hardcore computational chemists, they don't have labs beyond computers
  - Everyone else - experimental chemists, but they do some sort of computer modeling on shared cluster
  - Kathy - manages large shared cluster, manages indirectly the individual cluster (supervises individual sysadmins)
  - Three categories, different data types:
    1) molecular/atomic simulations - user generated data, 1 TB/user/year, lifecycle in immediate access sense is about a year (lifecycle of postdoc, grad student), after that it would get archived
    2) bulk simulations - user generated data, same storage as #1
    3) informatics - external databases, less individual-user generated data, chemistry doesn't do a lot of this, except for overlap w/ QB3 "chemistry is about anything with atoms in it, which is everything."
    "I never throw anything out. I have a folder called 'eternity'."
  - By the time a year has passed, scientists have distilled data down to what they need, data moves into archive
  - During the year of active use, nightly backups w/ 1 month of rollbacks (quietly have ~6 months of rollbacks)
  - Question of how much RAM to buy is "how much will fit?"
  - Nodes have 128 GB of RAM, 3-5 TB of scratch space, and that's not enough
  - Lots of calculations don't need it, but enough calculations do
  - #1-2: dynamic, large data, lots of I/O (more RAM means less I/O)
  - #3: black box, public DB - usually using in public domain, not hosting them; less demanding for storage needs, likely to be a big growth area
  - Data sharing is all ad-hoc, internal among researchers
  - No formal architecture in place for data sharing, no architecture for any lifecycle aspect; just save everything
  - All curation is at the user level
  - Standards for metadata are still developing
  - With supplemental data for publications, get a sense of who's generous and who's not
  - Some constraints on what journals allow/require for data
- #1 - non-proprietary formats, lots of structured text files
- #2 - lots of proprietary formats (e.g. Matlab simulations)

Concerns around proprietary formats - tied to a program, continuing to license it, will it be able to read legacy files?

What's in it for a researcher to curate data, besides mandate from funding agency?
Grad student's project is over 4-5 years, postdocs continue to collaborate after leaving
Incentive is so they themselves can figure out what the data is down the road
Have to have community standards, seen all sorts of schema, but nothing uniform
Have to structure model so that computation is finished in a timeframe meaningful to you as a person
If you had better resources, you'd make a better model - calculations always seem to take the same amount of time, regardless of technical resources available
Growth in amount of reuse of data?
Would be ideal if there was a community standard for putting data out on public server after certain period of time
No resources currently to make it available
Would be great to have a big archive
Is experimental reproducibility a pressure? - having this would reduce what problems there are
Meta-analysis: ran into problems all the time, either no community standards or people weren't doing it
Can have community standards, but need infrastructure to support it
Most people will send you their data if you ask them
Donna:
QB3 - Governor Gray Davis Institute for Science and Innovation
Mission of ensuring future of CA economy by promoting research and innovation, at interface of biology and quantitative sciences
Cooperative effort -- state, private industry, venture capital, campuses
Promoting multidisciplinary research, create innovative educational programs, foster industry partnerships
240 faculty affiliates across 3 campuses
UCSF is headquarters of QB3
Research areas: bio, chemistry, physics, engineering, others
At Berkeley, affiliation with QB3 is in addition to main academic appointment in home department
104 faculty affiliates, 5 schools and colleges, 16 academic departments
QB3 facilities serve broad community of academic/research labs
State-of-the-art instruments and services, class rooms, meeting rooms, etc.
9 core research facilities, each has a PhD level scientist director, technical expert; work with faculty director
Recharge services, used throughout campus
204 lab/20 depts charged for use in 2013
Data in CIRM/QB3 shared stem cell facility, genomics sequencing laboratory, CGRL (Computational Genomics Resource Laboratory)
Shared stem cell facility, high throughput screening facility (microscopy, images): 6 TB in each facility, avg amount of data/experiment is 500-1000 MB for SSCF, HTSF 1000-5000 MB, store for 6 mos
Genomic sequencing lab (sequence, text): raw runs between 110 GB-250 GB, raw run files are saved for 4-5 mos, capacity is ~145 TB, 112 TB in use
Some file somewhere saying what software was used to analyze what, etc.
Whether they can dredge everything up and redo analysis -- doubt it
Million dollar instrument comes with proprietary software
Some machines are driven by Windows 98, firewalled because software needed to drive it requires it
Lots of machines that can't be on the network
Put raw runs on FTP server - not very secure, everyone sees everyone's stuff
After a few months, "did you get your data? I'm deleting it."
What those people do with their data after it leaves core facility -- who knows?
Different practices in different labs, what they keep/choose to keep is highly variable
Core facility structure: if there's a point of intervention for archiving, that's the place, before it leaves
They talk among themselves about data management plans, but don't come to anyone official
Very ad hoc
Some people do go through archived data for other analysis - volume of data is growing
Whole process for depositing genomic data in public databases
Some labs are sophisticated, have servers that exceed central capacity
Kathy: metadata not embedded in file itself, image in a database that also stores the metadata
Computational genomics resource lab:
Vector cluster: 50 users, 9 compute nodes, 74 cores, 57 TB storage
Volumes: 20 GB in home directory/user, 300 GB in instrument data per lab/group, unlimited scratch, can increase volume on request, reference data sets are sizable and available on cluster
User/data management
- Not backed up (home directories are) -- just there fora couple months
User directories deleted after users leave
Scratch may be deleted if not touched for some period of time
Permanent data on quota, clean as needed
People are totally new to this: don't know Unix, scripting, don't know how to log in
Wanted place where they could learn how to do it, have access without investing in it for own lab
Get mail saying that people will just buy their own stuff, or "why can't you analyze this for us?" (set it up this way so students can learn and get jobs)
New director, adding analyze-it-for-you at small scale
Galaxy - open source web-based platform for data intensive biomedical research; platform for genomic analysis and workflow management
- Available on public server, local demand
- Encapsulates/integrates many analytical methods
- Captures workflow for duplicating or sharing analysis workflow
- Hoping to set it up here; currently running on public server
- Provides versioning
Tracks what software you use, runs, maps it all out, can take a snapshot, can copy it, give it to someone else
Trying to teach command line, but galaxy can provide UI for workflows including it
Galaxy just for CGRL right now - capacity issues
Authentication concerns, setting it up, queueing, etc.
Would probably debug it on Galaxy, run it on other servers
People using Galaxy don't want to deal with logging in
Kathy: try to avoid having to teach people Unix
Theoretical chemists who run own clusters would be okay with Unix (internal mechanism for bringing each other up to speed), but teach 5 minutes of Linux commands, mostly just double-click icons, written a middle layer that handles it all
QB3 has other approach - teaching people command line a few times per semester
What labs need:
- Archiving - recommended best practices, support/service to make it easier to manage
- Repository: turnkey methods to share data
  Good, cheap, how everyone else does it, can rely on it
Requirements from funding agency: data management plans, standards for distribution
Issues:
  Writing proposals with reasonable DMP
  Preserving old data: resolving aging URLs/hostnames/accession IDs
  Reliability of servers managed by lab staff
  Copying things out of other proposals
  Campus-level DMP would make things a lot stronger
  Will come up with simple list of questions, ask them to various labs
  How many people are unhappy about it? Or are they just accepting what they have, not thinking to ask for help?
It's like an earthquake plan - think about it on earthquake preparation day (until an earthquake happens)