2014-09-25 HIPAA - NSF-ACI - Funders' view of Data Management: CASC Meeting debrief

From the invitation to this meeting:

Please join us Thurs 25 Sept for a Research IT Reading Group on multiple topics highlighted at this month's meeting of the Coalition for Academic Scientific Computation (CASC). [...] In addition to some overview and impressions of the CASC meeting, Research IT's David Greenbaum and Chris Hoffman will facilitate discussion of:

- HIPAA: bringing Box into HIPAA alignment (at Indiana University)
- NSF's Advanced Cyberinfrastructure program
- Data Management: what's next from NSF and DOE

Please review the following articles / slide decks prior to the meeting:

HIPAA
Bringing Box into HIPAA Alignment

NSF ACI
Overview of NSF's Advanced Cyberinfrastructure (ACI) Program
Vision/Strategy for Software in Science, Engineering, Education

Data Management
DOE on Digital Data Management

Optional reading / viewing context:
Agenda for CASC 2014 (papers posted here)
ACI site (focus on programs and initiatives)
WWSSPE site (Sustainable Software)
DOE Public Access Plan

Attending

Larry Conrad, CIO
Bill Allison, IST-API
David Greenbaum, Research IT
Chris Hoffman, Research IT
Aron Roberts, Research IT
Rick Jaffe, Research IT
Aaron Culich, Research IT
Quinn Dombrowski, Research IT
Camille Villa, Research IT
Patrick Schmitz, Research IT
Steve Masover, Research IT
Ray Lee, Research IT
John Lowe, Research IT
James McCarthy, Space Sciences Lab
Scott Peterson, Morrison Library
Harrison Decker, Data Lab
Tim Dennis, Data Lab
Dav Clark, D-Lab
Neil Maxwell, Research Admin and Compliance

Notes
Now formally a member of Coalition for Academic Scientific Computation
Been around for 20+ years
Started with pioneers across higher ed, figuring out how to design, build, make available high performance computing facilities
Has evolved, not only looking at HPC but broad array of issues around supporting research using IT in higher ed
Strong connections with important actors in federal government
Intersection of research community, IT support community
Presentations from domain-based communities doing cyberinfrastructure work (e.g. astrophysics, urban planning)
Rich set of presentations from representatives of federal government (including director of NSF)
Slides aren’t yet up, but they will be

Domain-based cyberinfrastructure
Ian Foster, Computation Institute at UChicago
Talked about “long tail” of science
“Big science” has substantial resources for computation, data management support
Much larger group of people with less resources, harder to do the same kind of work
Shifting Globus to be more of a cloud-based platform— supporting research labs w/o access to large facilities
Pasteur’s quadrant approach: pure basic research, use-inspired basic research, applied research
tranSMART foundation: data-sharing analytics platform for medical schools, drug companies and others trying to share large amounts of complex data back and forth; strong industry connection
New NIH project: Microbiome Cloud Project; supporting microbial genomics data
Sloan Digital Sky Survey: astrophysics at John Hopkins university
Partnership with Microsoft and Jim Gray
20-30 years of extensive data gathering
Urban Center for Computation and Data: organized cities/data/democracies project
Working on bringing together tools and infrastructure support for people doing work on cities, impact of urban development, housing development
(Charlie Catlett, Argonne)
EarthCube (geo side), NEON (national ecological observatory network) — building earth observatory with massive sensors, federally funded airplanes, complete picture of earth environment
Trying to bring together shared resources to help in particular sciences
Question of which of these to pay attention to, how, how to learn things that could be applicable

Recently-funded NSF Projects from cyberinfrastructure — not domain-based
Chameleon and CloudLab
Moving further on what cloud computing will be like for research
2 $10M projects
CloudLab: coming out of Clemson, want to make it as easy to turn on a cloud as it is to turn on a VM
What does it look like to harvest resources from multiple HPC labs across country?
Clemson has put a lot of IT energy around research computing
Other cloud one: Chameleon, coming out of Texas Advanced Computing Center (TACC)
Similar research agendas, different approaches to problem
What one might do in a cross-cutting way

Federal government: head of advanced cyberinfrastructure initiative, biological, geological, social/behavior sciences representatives
Giving overview of where NSF is
Deeper understanding of what’s going on in advanced cyberinfrastructure
What it does on its own, cross-cutting, NSF-wide
Trying to link up cross-cutting cyberinfrastructure work with individual scientific areas
Need to know more what’s happening here; in RIT, have put focus on relationships with Mellon Foundation
People on campus have secured funding for networking, but need to help bring in funding that leads towards development of cyberinfrastructure
NSF grants are challenging to write, but they’re worthwhile
NSF skills development program — used to augment consulting staff
Have funded close to 130 NSF CID initiatives — have gotten funding for increasing bandwidth coming to campus
Putting focus on HPC, data, investment in sensors of all kinds
Didn’t focus on sensors in RAE benchmarking
Should do more coordinating conversation about what’s happening with use of sensors
Head of bio directorate of NSF, geo directorate
Impressed at level of connection back to computation side, comprehensiveness of vision
Cyberinfrastructure group seen as facilitating across different programs; recommendations of funding within other programs
Less discussion around social sciences; more focus on where they’re going on data management
Representative of Office of Science and Tech Policy from White House
New effort coming out of OSTP to spearhead across multiple federal agencies — government approach to what’s needed from computation side to support future research
Another federal government report coming out to do some synchronizing
Unclear how far they were on that effort
Presentation from faculty from National Academies; not federal agencies but play important role, distinguished faculty, reports and efforts on various topics
Overdue report that started in 2013 to help map out what the future directions should be for NSF, infrastructure, etc.
Organizer for CASC asked them to hurry up, report will shape what NSF does in terms of investment
Spirit of trying to get feedback from this group about what’s working, what’s important

3 presentations about different data management national/international consortia
Pragmatic work coming out of U of Illinois NCSA, Research Data Service
Basic things he’d like to see supported on his campus
Permanent identifier to store and make accessible data sets
Next level of evolution of federal requirements on data management
What’s the next step of what we want people to do on data reuse and preservation?
Mandate vs guidance, respecting different communities
Department of Energy has published next level of guidance

Vint Cerf — father of the internet, one of developers of TCP/IP
Future of the internet; went around the room and asked for opinion
Talked about kind of training needed for grad students now for doing data science
A lot of developers don’t know anything about hardware

CASC lobbies, in touch with what’s going on in DC
Making the case that this is important, need for ongoing investment

Has HIPAA working group
Med schools used to have their own IT, now they’re merging, have to deal with new kinds of data
Bill from Indiana University - HIPAA alignment
There’s not compliance with HIPAA, it’s not a standard
They try to become compliant with other standards, 800-53 NIST standard for securing data
Policy side, what standards have been used, how they’ve been modified to fit local environment
Part of Internet2 Box community
Demand for HIPAA-aligned Box service; had to create own BAA (Business Associates Agreement) with Box to make this work
Were able to demonstrate that Box is sufficiently secure
Have to work on one piece of authentication system
All this is about “can you survive an audit?”
At Box conference a few weeks ago, Box is clearly paying attention to this; differentiator is a service on campus
There are templates out there like Indiana to offer secure data storage services

Report by Ruth Merinshaw at Stanford, needed secure data for med school
Have been burned by disastrous security breaches, happened in med school and hospital
Creating very robust system not just for storing data securely, but built in tools for things like only being able to connect to service if you can demonstrate encrypted disk storage
3-year agreement, a huge amount of money going into this
Separate medical Box BAA with Box — very expensive, unlimited storage, only 3500 users
Med schools and hospitals have “more money than God” to do this kind of thing

TACC — different approach; doing 2-factor authentication on everything
Built a module that can be plugged into everything
2-factor authentication: just to shorten conversation with lawyers and compliance people
“That’s good, check.”

Need to be paying attention to software and applications
Should be treating software like infrastructure, thinking about the lifecycle
Working towards Sustainable Software for Science: Practice and Experiences
Workshop filling gap in old conference structure
Software with 30-year lifecycle, but we deal with 3-4 year grant cycle
How to support tools that get developed

Strategic trends affecting HPC
Departure from Moore’s Law
Can turn crank one more time in traditional architecture
Big data
Offshore microelectronics — one company in the US that does it, that’s an issue (used to be core industry)
Fab lines become more expensive with each generation; only half-dozen around the world that can produce high-end chips
Need best synergy between best people working in big data and best people working on big computation
Power limitations
Industry is “sucking talent” from universities

In data and data management, what NSF is doing
Provides a framework, deploy in phases, learn from one phase to the next, focus on publications initially, integration, working with communities
NSF will post approved plan (for data management) within a year
FAQs, guidance will evolve
Technical pilots, changes to procedures
Waiver process for changes to 12-mnth embargo
Will retain existing requirements and practices to extent possible
Contrast: DOE — ahead of things
In July, published public access plan
Lays out high-level principles about open access
Statement on digital data management, set of requirements that go into effect October 1 (Office of Science directorate)
Will put them in place for all of DOE w/in a year
Will affect labs that get DOE funding
Won’t apply to people using time on federal systems, but for getting grants
What to do if people can’t meet requirements, what do sharing and preserving mean
Have to have a letter coming in support of your claims of where you’ll put your data
Significant thing: allowance for publication/data costs in budget proposal
Moving towards model of declaring 10-year costs on a 3-year grant; can pay upfront

This from DOE, what NSF is going to do is really important
If they roll out new series of even light-touch requirements, all PIs need to figure out how to respond
Importance of more coordination with the library around research data management

How to keep staff in higher ed, how to fund training around this
Need to carve out resources for development within OCIO to take long look at funding opportunities, how can we leverage this to broaden our service offerings

What led to membership in CASC?
Up until a year ago, LBNL ran advanced computing for campus, they were CASC member
When we decided to get involved directly, made sense to become member and start attending
They said it was definitely worthwhile
We pay for it out of BRC budget, consider it important aspect of work

BIDS — logical place for engaging grant writing activity
Concerned about interface between privacy and openness in data, mobile device based data collection
Mobile data connection — if people want that, they go to industry
“It’s dumb trying to do that in academia right now. I’m trying to do that in academia.”

Need to have another conversation with David Trinkle to coordinate efforts
Trying to track landscape of what’s happening here
Universities providing a pre-filter on grants, given limits on how many proposals a campus can bring forward

Discussion about how many people can be sent per campus
Free to go — dues go towards meeting, but only supposed to send 2 people
Working group “beyond hardware”
Some discussion around virtual participation

Some people on the ground, trying to create science as service platforms
Representative from National Data Service — platform for bringing together community
Research Data Alliance — international organization, self-organizing group focused on data sharing, with working groups
Renci — iRODS platform/consortium
Supporting iPlant collaborative (users contributing information about plants on an hourly basis)

Securely supporting research data will be a major initiative of RDM program development
Best practices for de-identification (it’ll never be perfect)
Bill: Happy to pull together a discussion about FERPA/PHI and Box
FERPA — fewer issues than PHI
Don’t have HIPAA compliant box today