CaRCC Researcher-Facing Call, 2022-10-13

Research Data Storage: A SWOT Analysis

Speakers

- **Ryan Nakashima and Bill Homan from Research Data Services at San Diego Supercomputer Center (SDSC) on the campus of UC San Diego**

Please Note:

- ***We will record this call and post shortly thereafter on CaRCC's YouTube channel.***
- ***We expect all persons on the call to adhere to CaRCC's Code of Conduct.***

Agenda

- Welcome + Introduction to CaRCC
  - Researcher-Facing Track description
  - CaRCC is your community for research computing and data professionals. Please see this brief overview; more information on activities on our Groups web pages.
  - If you have questions about CaRCC or are interested in becoming more involved, please contact:
    - rf-coordinators@carcc.org for R-F-related activities, or
    - getstarted@carcc.org or getinvolved@carcc.org for other CaRCC work
  - We expect all persons on the call to adhere to CaRCC's Code of Conduct.
- Topic of the month:
  - Presenters: Ryan Nakashima, Bill Homan, Gavin Biffar & Willy Markuske from Research Data Services at San Diego Supercomputer Center (SDSC) on the campus of UC San Diego
  - *Abstract - The quickly disappearing option of "unlimited data storage" services along with licensing shifts in major storage and sharing platforms has caused some to rethink their research storage strategies. In our October call we will examine the strengths and weaknesses, as well as how to provide these services throughout the research lifecycle. Discussion topics will include:*
    - *Research data storage gaps*
    - *Three stages of storage across the research lifecycle*
    - *SWOT analysis for data storage*
- Open Discussion
- Announcements

Announcements

-

Sign-In (Name / Affiliation /Email)

Note: please follow the suggested sign-in format so our evolving data science intelligence routines won't trip up and forget to enter you in our $1m sweepstakes.

1. Brian Westra / University of Iowa / wstra@uiowa.edu
2. Tobin Magle / Northwestern / tobin.magle@northwestern.edu
3. Anita Orendt / University of Utah / anita.orendt@utah.edu
4. Annelie Rugg / UCLA / annelierugg@ucla.edu
5. Kirk M. Anne / Rochester Institute of Technology / kirk.m.anne@rit.edu
6. Venice Bayrd / Montana State University / venice.bayrd@montana.edu
7. Lauren Michael / Minority Serving - CI Consortium / lmichael@internet2.edu
8. Robert Henschel / Indiana University (henschel@iu.edu)
9. Brian Guilfoos / Ohio Supercomputer Center / guilfoos@osc.edu
10. Jason Yalim / Arizona State University / jyalim@asu.edu
11. Sean Anderson / Wake Forest University / anderss@wfu.edu
12. Jeanine Finn / Claremont Colleges Library / jeanine.finn@claremont.edu
13. David Nordello / University of Utah /david.nordello@utah.edu
14. Joey Netterville / University of Central Florida / joey.netterville@ucf.edu
15. Calvin Cox / Harvard Medical School / calvin_cox@hms.harvard.edu
16. Horst Severini / University of Oklahoma / severini@ou.edu
17. Chris Reidy / UArizona / chrisreidy@arizona.edu
18. Clifford Kravit / UCLA School of Medicine / ckravit@mednet.ucla.edu
19. Peter Goode / Lafayette College / goodep@lafayette.edu
20. Paula Sanematsu / Harvard FASRC / paula_sanematsu@g.harvard.edu
21. Mike Austin / University of Vermont / mga@uvm.edu
22. Jim Lawson / University of Vermont / jtl@uvm.edu
23. Kevin Brandt / South Dakota State Univ. / kevin.brandt@sdstate.edu
24. Mike Edmonds / UC Santa Cruz / medmonds@ucsc.edu
25. Vessela Ensberg/ UC Davis
26. Julie Goldman / Harvard / julie_goldman@harvard.edu
27. Dan St. Pierre / University of Michigan / dstpierr@umich.edu
28. Nan McKenna / Stanford University / nmckenna@stanford.edu
29. Randy Melen / Stanford University / randy@stanford.edu
30. Bill Homan / SDSC at UC San Diego / homan@sdsc.edu
31. Ryan Nakashima / SDSC at UC San Diego / ryan@sdsc.edu
32. Gavin Biffar / SDSC at UC San Diego / gbiffar@sdsc.edu
33. Cyd Burrows-Schilling / University of California San Diego / cburrows@ucsd.edu
34. Jessica Pierce / Harvard Medical School / jessica_pierce@hms.harvard.edu
35. Carolyn Ellis / UCSD / carolynEllis@ucsd.edu
36. Peter Wan / Georgia Tech / peter.wan@oit.gatech.edu
37. Georgia Stuart / UT Dallas / georgia.stuart@utdallas.edu
38. Jens Mueller/ Miami University / muellej@miamioh.edu
39. Paul van der Mark / Florida State University / pvandermark@fsu.edu

40. Trina Marmarelli / Reed College / marmaret@reed.edu
41. Jacobus Kats / UC Riverside / bart.kats@ucr.edu
42. Alisa Kang /Georgetown University /alisa.kang@georgetown.edu
43. Brad Spitzbart / University of Oklahoma / bspitzbart@ou.edu
44. Alex Pacheco / New Jersey Institute of Technology / alex.pacheco@njit.edu
45. Gladys Andino / University of Virginia / gka6a@virginia.edu
46. Sara Gonzales / Northwestern University / sara.gonzales2@northwestern.edu
47. Patrick Schmitz / Semper Cogito / patrick@sempercogito.com
48. Michael Weiner / Georgia Tech / mweiner3@gatech.edu
49. Michael Laurentius / University of Toronto / michael.laurentius@utoronto.ca
50. Moira Downey / NC State / mcdowney@ncsu.edu
51. Cal Frye / Case Western Reserve University / cxf244@case.edu
52. Hunter Hagewood / Vanderbilt / hunter@accre.vanderbilt.edu
53. Kathleen Chappell / Harvard Medical School / kathleen_chappell@hms.harvard.edu
54. Luca Belletti / Digital Commons / l.belletti@elsevier.com
55. Ashley Stauffer / Penn State/ als81@psu.edu
56. Michael Benedetto/ American Museum of Natural History
57. Jacalyn Huband / University of Virginia / jmh5ad@virginia.edu
58. Summer Wang / Ohio Supercomputer Center / xwang@osc.edu
59. Aaron Weeden / Shodor Education Foundation / aweeden@shodor.org
60. John Brussolo / University of Michigan - Michigan Medicine / jsbrusso@med.umich.edu
61. Mary Olson, Oracle Education & Research
62. Brad Thornton / Texas A&M / bthornton@tamu.edu
63. Timothy Middelkoop / Internet2 / tmiddelkoop@internet2.edu
64. Henry Neeman / University of Oklahoma / hneeman@ou.edu
65. Dana Brunson / Internet2 / dbrunson@internet2.edu
66. Martin Cuma / U of Utah / m.cuma@utah.edu
67. Mark Piercy/ Stanford / mpiercy@stanford.edu
68. Ken Lutz/ UC Berkeley/ lutz@berkeley.edu
69. Bob Freeman / Harvard Business School / rfreeman@hbs.edu
70. Evan Linde / Oklahoma State University / elinde@okstate.edu
71. Biru Zhou / McGill University / biru.zhou@mcgill.c
72. Yogesh Kale / University of Illinios / ykale@uic.edu
73. Coltran Hophan-Nichols / Montana State University / coltran@montana.edu
74. Matthew Lacy / Texas A&M University / matt.lacy@tamu.edu
75. Jeff D’Ambrogia / LBNL / jeffd@lbl.gov

Max attendee count 97

Notes from the call

Slides

https://docs.google.com/presentation/d/1fykvlfzZjNxkv5_OKDWLne8JePn7SOPP0GhI4onkwUY/

Notes

From Justin:
Licensing shifts from Box/Google are forcing Universities to think more carefully about data storage strategy

- "Unlimited" storage was appealing
- Easy to share with collaborators
- Easy to collaborate for proposals/papers
- Policies haven't kept up with how we manage data storage services

Research storage considerations

- New information security challenges
- New security requirements
- New data sharing requirements
- Can you do compute with data on storage?
- Account management -IAM, storage limits, deprovisioning

Questions for the group

- How are you getting stakeholder input?
- Do you have a research data strategy?
    - Ashley (PSU) - If there is a strategy at PSU, I'm not aware of it...but all the concerns you mention are relevant for us, Justin

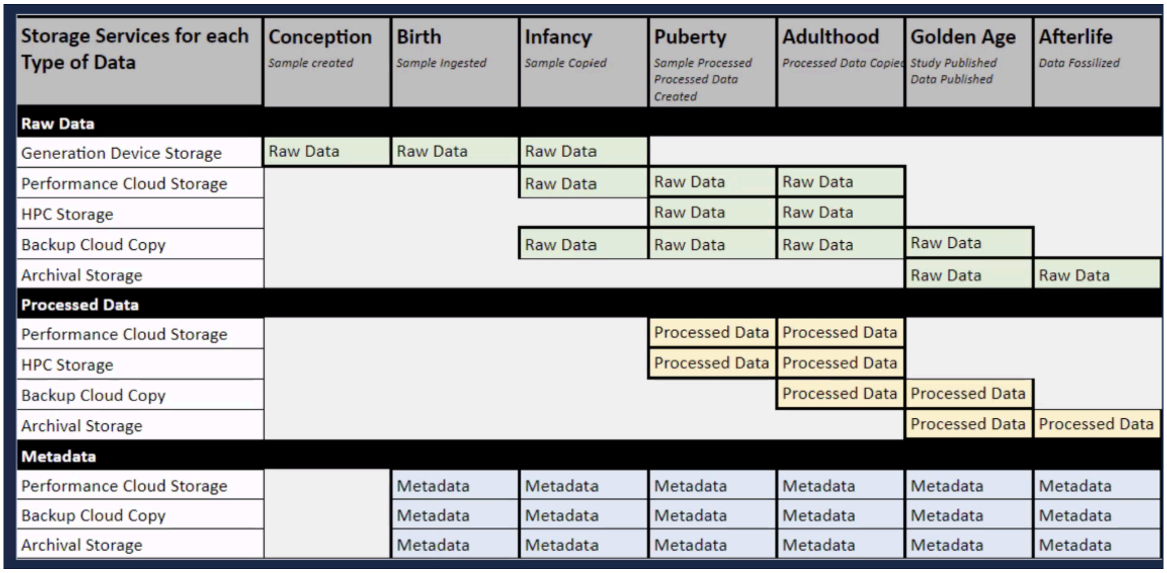
Storage Basics -
https://docs.google.com/presentation/d/1LBmYiOHc40cdIkSy7RlvXHLs5JDU6s00XQ92MiI72yY/edit?usp=sharing

Bill Homan - San Diego Supercomputer center
Research data services division

- Talk to lots of researchers in SD and elsewhere (US and international, public and private sector)
- Lots of use cases

3 stages

1. Raw data generation
2. Place data into a repository: complicated, many choices, often the focus
3. Archival:



Storage Services for each Type of Data	Conception *Sample created*	Birth *Sample Ingested*	Infancy *Sample Copied*	Puberty *Sample Processed Processed Data Created*	Adulthood *Processed Data Copie*	Golden Age *Study Published Data Published*	Afterlife *Data Fossilized*
Raw Data							
Generation Device Storage	Raw Data	Raw Data	Raw Data				
Performance Cloud Storage			Raw Data	Raw Data	Raw Data		
HPC Storage				Raw Data	Raw Data		
Backup Cloud Copy			Raw Data	Raw Data	Raw Data	Raw Data	
Archival Storage						Raw Data	Raw Data
Processed Data							
Performance Cloud Storage				Processed Data	Processed Data		
HPC Storage				Processed Data	Processed Data		
Backup Cloud Copy					Processed Data	Processed Data	
Archival Storage						Processed Data	Processed Data
Metadata							
Performance Cloud Storage		Metadata	Metadata	Metadata	Metadata	Metadata	Metadata
Backup Cloud Copy		Metadata	Metadata	Metadata	Metadata	Metadata	Metadata
Archival Storage		Metadata	Metadata	Metadata	Metadata	Metadata	Metadata

5 main Types of storage (above)

Factors

- Cost
- Lifetime
- Technical limitations - like number of files
- Security - often lumped in with redundancy/backup, prevent ransomware attacks, tension with accessibility
- Support - all systems break eventually - who fixes it? (on prem vs hosted)
- Accessibility
- Discoverability
- Development time for processing

SWOT analysis: framework to understand forces that affect research storage

- Strengths
- Weaknesses
- Opportunities
- Threats

Questions

- How much responsibility for maintaining archived data is put on Research Computing and how much is on the researchers?

    - Supercomputing center has home directory/scratch space, but it’s mostly on the researcher
    - Not all supercomputing users - guide people to best solutions
- Can researchers mint DOIs for their data at the project vs. file level?
    - How do you achieve discoverability
        - Depends on what kinds of data, who needs to discover it - very nuanced. Happy to talk with anyone about those nuances in your cases.
- Noting that U Michigan's annual research expenditure is ~$1.7B (#2 in the US), and UCSD's is ~$1.4B (#6): How do these research data management approaches translate to institutions that have annual research expenditures of a few hundred million, or a few tens of millions?
    - Similar question could be asked about how things change when there are sizeable grants for storage vs. not
    - What happens: Unfunded projects with growing amounts of data
- How are you defining 'Performance Cloud Storage'?
- Aren't many of the important functions in the "Golden Age" (discovery, access, preservation) covered by repositories? Or is this more about the challenges for providing those functions for "big data"?

Chat Comments

- (Timothy Middelkoop) Pro-tip: don’t have lunch in the kitchen while a a talk about google is going on… Unfortunately the google assistant did not have an answer for your question....
- (Cal Frye) Synology and QNAP to the rescue! (kidding, not kidding)
- (Jeanine Finn) My grad student institution just emailed all the alums who were promised "lifetime gmail"....get it down to 1 MG or they're deleting your account :)
- (John Brussolo) At U Mich, we completed our migration off Box 11 months ago. We’re just hearing about the new Google Drive storage changes - it’s early days regarding that.
- (Annelie Rugg) Regarding strategy, are there standards for research data storage that can help guide our approaches? And are those standards broadly applicable, or discipline-specivic, or country-specific?
- Regarding the table shared by Ryan and team:
    - I like the analogy!
    - Love this table. … Permission to "steal" the table for our outreach events
    - as a social scientist, I appreciate the developmental life stages as a reference
    - Same here, great comparison. If that's OK I'll use that in my storage lecture.
- (Cyd Burrows) UC San Diego has published this Research Data Storage Explorer (with thanks to Cornell University): https://researchdata.ucsd.edu/finder
    - (Jason Yalim) We have had something similar at ASU: https://researchstorage.asu.edu/
- (Jeanine Finn) For long-term storage and curation DCN's curation primers are very helpful for understanding file types and requirements for reproducibility across a number

of data types and disciplines
https://datacurationnetwork.org/outputs/data-curation-primers/

- (Alisa Kang) In our experience, most of the large data generators are not HPC users, they need powerful desktop and GPU to process data, and mounting cloud storage to the desktop doesn't work with this type of set up.
- (Jeanine Finn) Coming from library world -- these are good questions for data librarians at your institution. Discoverability is central for us. [referring to the DOI questions and discussion – see questions above]
- (Cyd Burrows) Unfunded projects with growing amounts of data
- (Lauren Michael) +1 on cloud resources being insufficient. Tobin was involved with centralized research storage at UW while I was there, and might comment further, but we definitely had some very large data analyzed or generated by our large-scale computing users, but that's also tied to our significant HTC-catered computing resources (our HPC-optimized cluster was smaller, mostly used for simulations/generation).
    - (Timothy Middelkoop) @Lauren I assume you mean the consumer/office storage cloud systems (google drive, o365, box, etc) not things like AWS S3 and other “cloud” storage systems?
    - (Lauren) I'm referring to ResearchDrive, which is UW-Madison's locally-provided, core research storage; mountable to laptops/servers, and available via command-line transfer to the large-scale computing center (and some other servers/services). Each PI 5TB, free. https://it.wisc.edu/services/researchdrive/
    - (Tobin Magle) @Lauren there were some really sophisticated data storage use cases that were integral to how Box functioned and couldn't really work on other platforms
    - (Jim Leous) That has been our experience with O365 here. Thanks for bringing that up.
- (Nan McKenna) Back to one of the original questions, I'm very curious as to whether/how anyone is getting structured feedback from researchers (vs. more anecdotal).
- (Clifford Kravit) Just for reference: https://research-it.ucsd.edu/_images/researchlifecycle_image.png
- (Alisa Kang) I hear from our library an institution can buy storage from https://datadryad.org/stash to make institution-specific repository, maybe that's an option instead of google or other cloud
- (Tobin) I don't think researchers are necessarily ready to put everything in a public repository. not to mention data that can't be public

Connection Details
Agenda
Announcements
Sign-In (Name / Affiliation /Email)

Notes from the call

- Slides
- Notes
- Questions
- Chat Comments

Connection Details

https://utah.zoom.us/my/carcc?pwd=TjFuR3VVM2d5eE5zWnEvWWxDTFBCUT09

Meeting ID: 824 051 8198

-