CaRCC Strategy and Policy-Facing Call, 2025-03-05

Title/Topic: The Data Problem

Speakers: Curt Hillegas, Princeton University, and Scott Yockel, Harvard University and NESE
Abstract: The rapid expansion of data science and artificial intelligence has ushered in an era of unprecedented data generation, yet it has also brought to light significant challenges in managing, processing, and interpreting this wealth of information. In this discussion, we'll feature speakers from Princeton, Harvard, and NESE to explore "the data problem" in research computing followed by open discussion. They'll share insights on the challenges they're tackling, the lessons they've uncovered, and their visions for the future—both where they're headed and where they aspire to go.

Table of Contents:

Please Note:

- ***We will record this call*** *and post shortly thereafter on CaRCC's YouTube channel, however if you would like to make a comment off the record, let us know and we can pause the recording.*
- ***We expect all persons on the call to adhere to CaRCC's Code of Conduct.***

Announcements

- Welcome to the Strategy and Policy-Facing Track of the CaRCC People Network!
  - Join the People Network email lists. Join the CaRCC Slack workspace.
- CaRCC is your community for research computing and data professionals. Please see this brief CaRCC overview and more information on Working Groups and Interest groups.
- If you have questions about CaRCC or are interested in becoming more involved, please contact:
  - spf-coordinators@carcc.org for Strategy and Policy-Facing-related activities, or
  - getstarted@carcc.org or getinvolved@carcc.org for other CaRCC work

Sign-In (Name / Affiliation /Email)

NOTE: This document is publicly accessible. Please keep that in mind when entering information.

1. Amit Amritkar / Penn State / amit@psu.edu
2. Curt Hillegas / Princeton University / curt@princeton.edu
3. Patrick Schmitz / Semper Cogito / patrick@sempercogito.com
4. Shane Moeykens / UMaine / shane.moeykens@maine.edu
5. Scott Yockel / Harvard University
6. Jaynal Pervez/ UT Health San Antonio
7. Jason Key / Harvard Medical School / key@hkl.hms.harvard.edu
8. Kathryn Kelley / CASC / kelley@casc.org
9. Kirk M. Anne / / kirk.m.anne@gmail.com
10. Grigory Shamov / University of Manitoba / grigory.shamov@umanitoba.ca
11. Matt Gregas/Boston College/gregas@bc.edu
12. John Heimaster / Ohio State / heimaster.1@osu.edu

13. Chris Reidy / University of Arizona / chrisreidy@arizona.edu
14. Scott Delinger / Altadel Consulting Ltd. / scott@altadel.com
15. Bill Barnett / UMass Chan Medical School / william.barnett@umassmed.edu
16. Becky Ortinez / University of Utah / becky.ortinez@utah.edu
17. Raminder Singh / Harvard University/ r_singh@g.harvard.edu
18. Barry Farmer / University of Kentucky / barry.farmer@uky.edu
19. Kristin Lepping / Rutgers University / klepping@rutgers.edu
20. Maureen Donlin / Saint Louis University / maureen.donlin@health.slu.edu
21. Melissa Cragin / Rice University / mcragin@rice.edu
22. Jan Day / Amazon Web Services / janday@amazon.com
23. Ramazan Aygun / Kennesaw State University / raygun@kennesaw.edu
24. Tom Lewis / University of Washington / tomlewis@uw.edu
25. Mirakyl Drake / University of Nevada, Las Vegas / mirakyl.drake@unlv.edu
26. Liam Forbes / University of Alaska Fairbanks / loforbes@alaska.edu
27. Mitch Rosen / University of Virginia / rosen@virginia.edu
28. Susan Ivey / NC State/ slivey@ncsu.edu
29. Rob Bjornson / Yale / robert.bjornson@yale.edu
30. Jens Mueller / Miami University / muellej@miamioh.edu
31. Matthew West / University of North Carolina, Charlotte / mwest53@charlotte.edu
32. Marina Kraeva / Iowa State University / kraeva@iastate.edu
33. John Costa / University of Saskatchewan, Canada / john.costa@usask.ca
34. Amy Nurnberger / Massachusetts Institute of Technology / nurnberg@mit.edu
35. Wei Yin / Columbia / wy2288@columbia.edu
36. Lev Gorenstein / Globus - University of Chicago / lev@globus.org
37. Christina Drummond, University of North Texas / OA Book Usage Data Trust effort / christina.drummond@unt.edu / christina.drummond@oabookusage.org
38. Sean Smith / Rice University / mrsmith@rice.edu
39. Daniel Widyono / University of Pennsylvania / widyono@seas.upenn.edu
40. Kim Wilkins / University of Nevada, Reno / kwilkins@unr.edu
41. Justin Booth / Mich State / boothj@msu.edu
42. Exequiel Punzalan / Rutgers University / ep523@rutgers.edu
43. Wirawan Purwanto / Old Dominion University / wpurwant@odu.edu
44. Bob Freeman / Harvard University / robert_freeman@harvard.edu
45. Michael Strickler / Yale University / michael.strickler@yale.edu
46. Ian Kaufman / UC San Diego / ikaufman@ucsd.edu
47. Hadrian Djohari / Case Western Reserve University / hadrian.djohari@case.edu
48. Jessica Eaton / Columbia University / je2429@columbia.edu
49. Andy Anderson / Amherst College / aanderson@amherst.edu
50. David Reddy / University of South Carolina / davidreddy@sc.edu
51. Frank Pari / Boston College / parif@bc.edu
52. Seraphim McGann / Yale / seraphim.mcgann@yale.edu
53.

Max attendee count: 59

Notes from the Call

Slides

Campus Research Computing Consortium overview:

Session Notes (anyone can contribute)

- Curt's presentation on Tiger Data
    - Princeton Research Data Server. Host/repo for post-research/post project data.
    - Led to recognition of much greater need for active research data.
    - Move from data storage culture to data *management* culture.
    - 75 PB of active storage, 127 PB of tape for backup.
    - This is the website describing TigerData (not the User Interface): https://tigerdata.princeton.edu/
    - Work is based upon MediaFlux
- Scott's presentation on Harvard Research Data Connect
    - Like at Princeton, collaboration across units.
    - 65 PB of CEPH storage under the hood
    - 130 PB of tape storage.
    - The CEPH storage is presented to storage manager.
    -

Questions

- [ ] How does Princeton handle staging data to tape?
    - Curt: Work in progress on automated workflow and lifecycle management.
- [ ] Is data on tiger data equally available from all computing environments? HPC, lab servers, personal laptops, etc? If so, how is identity managed?
- [ ] What kind of user facing data management or movement tools do you have with Tiger Data for self management?
    - Web as well as Command line; Have learned that Globus is the fastest way to move data. MediaFlux has tools, but generally not user-facing. Expect that these UIs will continue to evolve as research needs change.
- [ ] What kind of timeframes were associated with your respective solutions
    - Curt: Started planning in 2018. Covid actually helped because the culture of (hard to schedule) in person meetings shifted, and it became easier to meet and make progress. "It all takes a lot longer than you expect…" Budget (equipment and people) was much higher than expected.

Chat Comments

Questions from chat:

- Per your comment about web interface development, what do you have for user-facing management and movement tools?
    - +1; I'm very curious about what the UX/UI looks like for users / researchers, and how much does this incentivize active use of these tools & resources?
    - Curt: in development, using agile, so still evolving. Currently have read-only support, but working towards ability to annotate metadata, etc. Everything is project-based, so role/group management is key.
        - Question about OSS? Curt: Yes, although not open while in development.

    - Scott: Harvard Library has a project “Reimagining Discovery” that is also looking at this.
- @Scott M Yockel between Data Management and Data Disposal, are you seeing any increase in interest in (Data) Impact Analysis i.e. a need from scholars / funders to report on the usage/impact of the digital objects in the repos?
    - Scott: Most of our “Data Disposal” requests are tied to DUAs as a requirement. Data Access + Citation impact would be a nice metric to track. We don’t currently have these combined.
- Are there multi-institutional DMP tools, or does every institution develop those separately?

Connection Details

1st Wednesdays Time: 9am PST / 10am MST / 11pm CST / Noon EST
Call coordinator: Amit Amritkar (Penn State)

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

Meeting ID: 824 0051 8198
Password: 31415926

One tap mobile
+13462487799,,8240518198#,,#,31415926# US (Houston)
+16699006833,,8240518198#,,#,31415926# US (San Jose)
Dial by your location
+1 346 248 7799 US (Houston)
+1 669 900 6833 US (San Jose)
Join by Skype for Business: https://utah.zoom.us/skype/8240518198

Cursor Garden ⤵ *“Don't leave your cursor in a seedy alleyway”*