Data-Facing & Systems-Facing Joint Call, 2023-02-16 @ 1p ET/ 12p CT/ 11a MT/ 10a PT

Topic: Towards a Data Commons at Harvard

Topic Overview:
For over two years, a collaborative team from divisions and schools across Harvard University has been pursuing the idea of a "Digital Scholarship Ecosystem" to support the lifecycle of research at Harvard. The Harvard Data Commons (HDC) vision is to improve the researcher experience by automating the flow of research data from research computing environments to management, publication, discovery, and preservation environments. As a result, data integrity, provenance, and reproducibility of research data will be improved in order to meet sponsor requirements.

In building this ecosystem we aim to address following challenges faced by researchers: managing the organization of and access to data throughout the research lifecycle, facilitating the reproducibility of data, facilitating the sharing of active research data across groups within Harvard and with peers at other institutions, minimizing the risk of falling out of compliance with data policies (e.g., federal, grant sponsors, and institutional). Furthermore, we aim to work towards reducing duplication in the procurement of datasets, which results in high costs to Harvard's overall research portfolio and reduces potential collaboration. We also see practical outcomes that a data commons will help us achieve, primarily in improving research data management, access, and discovery of Harvard research, and research collaboration, sharing, and dissemination. In this talk, we will present HDC's progress toward its three objectives and our vision for the future.

Discussion Hosted by:

- Mahmood Shad, Associate Director of RSE, University Research Computing, Harvard University
- Emre Keskin, University Research Data Officer, Harvard University
- Scott Yockel, University Research Computing Officer, Harvard University

Slides & Links:

- CaRCC_HDC_Presentation.pdf

YouTube Link To Meeting Recording: https://youtu.be/4eVNl33kkvE

CaRCC YouTube Channel: https://www.youtube.com/carcc

CaRCC YouTube Channel Systems-Facing Playlist:
https://www.youtube.com/playlist?list=PLV-SC0CHLTwehApeMg_fJ5VbYndETh3yT

CaRCC YouTube Channel Data-Facing Playlist:
https://www.youtube.com/playlist?list=PLV-SC0CHLTwdrPXN8zdres6aovVAoEJIq

CaRCC Code of Conduct: https://carcc.org/about/carcc-code-of-conduct/

Join the CaRCC Slack Workspace:
https://join.slack.com/t/carcc/shared_invite/zt-9lhv4cfm-wy7RpbNI7V9qLJcvtzq8hA

Zoom Meeting Room:

https://utah.zoom.us/my/carcc?pwd=TjFuR3VVM2d5eE5zWnEvWWxDTFBCUT09
Meeting ID: 824 051 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

NOTE: **We will be recording today's meeting via Zoom**.
So if there are questions anyone wants to ask but do so anonymously please feel free to use the direct message feature in the chat to send Betsy Hillery or Brian Haymore your question and we will present the question.

Sign-In (Name / Affiliation / Email): (on call 130)

1. Brian Haymore / University of Utah / brian.haymore@utah.edu
2. Betsy Hillery / Purdue University / eahillery@purdue.edu
3. Jason Wells / Harvard University / jrwells@seas.harvard.edu
4. Venice Bayrd / Montana State University / venice.bayrd@montana.edu

5. Kyle Hutson / Kansas State University / kylehutson@ksu.edu
6. Nicholas Polys / Virginia Tech / npolys@vt.edu
7. Rory Macneil, Research Space, rmacneil@researchspace.com
8. Jim Leous / Penn State Office of the Associate CIO for Research / leous@psu.edu
9. Yogesh Kale /University of Illinois/ ykale@uic.edu
10. Timothy Middelkoop / Internet2 / tmiddelkoop@internet2.edu
11. Deb McCaffrey / ASU / dmccaff4@asu.edu
12. Mary Murphy/ University of Wisconsin/murphy22@wisc.edu
13. Benjamin Tovar / University of Notre Dame / btovar@nd.edu
14. Zach Weidner / Purdue University / zweidner@purdue.edu
15. Chris Phillips / Purdue University / phill219@purdue.edu
16. Amit Amritkar / Penn State / amit@psu.edu
17. Maurice Dalton / Yale University- Data Intensive Social Sciences Institute / maurice.dalton@yale.edu
18. Vas Vasiliadis / University of Chicago / vas@uchicago.edu
19. Megan Chenoweth / ICPSR
20. Grigory Shamov / University of Manitoba / Grigory.Shamov@umanitoba.ca
21. Ramazan Aygun / Kennesaw State University / raygun@kennesaw.edu
22. James Kelly / Chapman University / jakelly@chapman.edu
23. Jared Baker / NCAR / jbaker@ucar.edu
24. Joshua Allen / NCSA - UIUC / jallen17@illinois.edu
25. Jeff Dusenberry / UMass Boston / jeff.dusenberry@umb.edu
26. Thomas Langford / Yale Center for Research Computing / thomas.langford@yale.edu
27. Tobin Magle /Northwestern U/tobin.magle@northwestern.edu
28. Glen Rundblom / University of Illinois / rundblom@illinois.edu
29. Lior Atar / NYU / lior@nyu.edu
30. Sandy Shew / Ohio State University / shew.1@osu.edu
31. Douglas Dodson / Penn State / douglas.dodson@psu.edu
32. Doug Johnson / OSC / djohnson@osc.edu
33. Vessela Ensberg / UC Davis
34. Flannon Jackson / NYU HPC / flannon@nyu.edu
35. James Cuff / MIT / jcuff@mit.edu
36. Katherine Holcomb/UVA/kah3f@virginia.edu
37. Chris Reidy / UArizona / chrisreidy@arizona.edu
38. Dylan Ruediger/Ithaka S+R/dylan.ruediger@ithaka.org
39. Sarah Marchese / Harvard Medical School, IT / sarah_hauserman@hms.harvard.edu
40. Jim Lawson / University of Vermont / jtl@uvm.edu
41. Jeremy Matteson / University of Missouri / mattesonj@missouri.edu
42. Rob Bjornson/Yale Center for Research Computing/robert.bjornson@yale.edu
43. Arman Pazouki/Purdue/apazouki@purdue.edu
44. Ana Trisovic / Harvard / anatrisovic@g.harvard.edu
45. Neil Bright / NCAR / ncbright@ucar.edu
46. Adrian Ho, University of Chicago Library, adrianho@uchicago.edu
47. Michael Laurentius / University of Toronto / michael.laurentius@utoronto.ca

48. Sean Cleveland / University of Hawaii/ seanbc@hawaii.edu
49. Claire Mizumoto / University of California San Diego / claire@ucsd.edu
50. Biru Zhou/ McGill University / biru.zhou@mcgill.ca
51. Michael Weiner / Georgia Tech / mweiner3@gatech.edu
52. John Blaas / NCAR / jblaas@ucar.edu
53. Will Shanks / NCAR / shanks@ucar.edu
54. Barbara Esty / Yale University Library / barbara.esty@yale.edu
55. Bala Desinghu/Rutgers University/bala.desinghu@rutgers.edu
56. Ken Taylor / University of Illinois Urbana-Champaign / ktaylo@illinois.edu
57. Lev Gorenstein / Purdue University / lev@purdue.edu
58. J. Ray Scott / Carnegie Mellon / ray@cmu.edu
59. Lora Leligdon / Dartmouth /lora.c.leligdon@dartmouth.edu
60. Paula Lackie / Carleton College / plackie@carleton.edu
61. Katia Bulekova / Boston University / ktrn@bu.edu
62. Zhiyu Li / U of Utah / zhiyu.li@utah.edu
63. Justin Booth / Mich State University / boothj@msu.edu
64. Matt Smith / University of British Columbia / matthew.smith@ubc.ca
65. Anita Orendt / University of Utah / anita.orendt@utah.edu
66. Vittorio Bucchieri / Harvard John A. Paulson School Of Engineering And Applied Sciences / vbucchieri@seas.harvard.edu
67. Megan Potterbusch / Harvard Kennedy School / mpotterbusch@hks.harvard.edu
68. Ceilyn Boyd, Harvard Library, ceilyn_boyd@harvard.edu
69. Wendy Mann / George Mason University / wmann@gmu.edu
70. Brad Spitzbart / University of Oklahoma / bspitzbart@ou.edu
71. Susan Ivey / NC State University / slivey@ncsu.edu
72. Sam Weekly / Purdue University / sweekly@purdue.edu
73. Stevan Earl / Arizona State University / stevan.earl@asu.edu
74. Ben Sabath / Harvard University / sabath@g.harvard.edu
75. Trina Marmarelli / Reed College / marmaret@reed.edu
76. Alastair Neil / George Mason University / aneil2@gmu.edu
77. Julie Goldman / Harvard Library / julie_goldman@harvard.edu
78. Susan Tussy, University of Chicago, stussy@uchicago.edu
79. Don Brower / University of Notre Dame / dbrower@nd.edu
80. Cyd Burrows-Schilling / University of California San Diego / cburrows@ucsd.edu
81. Dana Brunson / Internet2 / dbrunson@internet2.edu
82. Ash Hingham ashahing@mit.edu
83. Jenett Tillotson / NCAR / jtillots@ucar.edu
84.

Discussion Notes & Q&A:

- Towards a Data Commons at Harvard
  - Mahmood Shad, Associate Director of RSE | University Research Computing | Harvard University

- Emre Keskin, Harvard University Research Data Officer
- Scott Yockel, University Research Computing Officer

- Addressing a gap between data management and storage management–people-facing and systems-facing; University Research Computing–beyond HPC
- What challenges do you face in sharing and publishing research data?
- What technology or tooling would you like to see in place to reduce the friction in our research process?
- The problem:
    - Lack of data centric tools throughout the research lifecycle
    - Lack of interoperability between tools
    - Planning, storing, analysis, management, disposal tools
    - Ease the organization of large data
    - Facilitating the reproducibility of data
    - Sharing of active data internally and externally
    - Supporting data retention compliance
    - Tracking the provenance of data
    - Providing the metrics of data access and use
- From data centric Systems
    - *NSF should support foundational cyberinfrastructure research for data science with a focus on frameworks and tools for data science, data centric systems architectures, and data repositories*
        - Data Science:
        - Data Systems Architecture: A fast scalable
        - Data Repositories: WEll curated validates open access repositories required to ensure that the results of scientific research are available.
    - NSF CI3040: Future Advanced Cyberinfrastructure document, 2018
- … to a data Commons?
    - "... A data commons brings together (or co-locates) data with cloud computing infrastructure and commonly used software services, tools & applications for managing and …
- Harvard Data Commons Vision
    - Seamless data science experience to researchers.
    - Will improve the research experience by automating the flow of research data and derivative scholarship to management, publication, and preservation environments. As a result, data integrity, provenance, dissemination, and reproducibility of research data will be improved.
- Current State
    - Graphic about RC tools and Storage
    - Examples: RSpace, Jupyter, R-Studio, Globus, Starfish
    - Data Acquisition flows to Data Analysis which is in Storage Tier 1 or 2, and those flow to tier 3 for publications data sharing and preservation
- Library Systems
    - Create Acquire, Discover, Deliver, Manage, Preserve
    - Discover has DASH, Dataverse,

- DASH: central, open-access repository of research by members of the Harvard community
    - Built on DSpace (https://dspace.lyrasis.org/)
- Digital Repository Service
    - Harvard infrastructure ensuring the persistent integrity, authenticity, accessibility, and usability of University digital collections across time, technical and cultural distance
- Dataverse
    - An open source web application to share, preserve, cite, explore, and analyze research data. (https://dataverse.org/)
    - Open to all researchers, both inside and out of the Harvard community.
    - A tool for data management, advancing searching, dataset discovery and download, text mining and more to support…
- Data Commons @ Harvard
    - Executive Sponsors are the University CIO Klara Jelinkova, Martha Whitehead, University Librarian, John Shaw, Vice Provost for Research
    - The team has University Research Computing, Harvard Library, Harvard University IT, Harvard Medical School, Institute for Quantitative Social Science, Harvard Business School
    - Leadership: when you have a large group, individual people can change, but the work can still move forward
- Data Commons: MVP (minimum viable product): Objectives
    - Automating the technical publine
    - Enhancing Dataverse
- Objective 1
    - Integrate the environments and Harvard repositories
    - Current Status
        - Direct Upload exists now, (no unzip), used at multiple Dataverse installations
        - Globus Testing phase and will be available in Dataverse upcoming release
        - Remote Store
            - Working prototype exists for 5.11 (non-ingested files)
            - Implementation supporting remote file and interested files ‘experimental’ expected for upcoming Dataverse release
    - Direct Upload to S3
        - DVUloadeder is another application that can be used: https://github.com/GlobalDataverseCommunityConsortium/dataverse-uploader
    - Remote Storage
    - Globaus Transfer to Dataverse
- Was the Globus S3 Premium Connector used? Yes, we’re doing both. Can also use Multipart S3 upload. Also working on a Posix interface.
- Harvard Data Commons MVP: Objective 2
- Is the backend storage on dataverse S3? Yes.
    - Graph showing Explore Phase (personal/future self), refine Phase (small team), Produce Phase (community). The segments are meant to reinforce each other.
- Objective: Support advanced research workflows and provide packaging options that can be deposited in a repository to enable reproducibility and reuse

- Current Status
    - Available in Data verse v. 5.12
    - Documentation is available
- Harvard Data Commons MVP: Objective 3
    - Integrate Harvard Dataverse with the Digital Repository Service for lon-term presentation
        - Enhance Dataverse to explore datasets for long-term preservation
        - Enable curators
    - Integrate Dataverse with DASH to connect datasets with open access publications
- Harvard Data Commons MVP: Object 3A slide
    - Graphic showing workflow of DRS
- Harvard Data Commons MVP: Objective 3B slide
    - Shows workflow of Harvard DASH to Dataverse integration.
- Harvard Data Commons MVP: Objective 3 (again)
    - Status: DRS is implemented but, current version of DRS doesn't support versioning
    - The process is not automated
    - Status: …
- Related Initiatives
    - Research Computing and Data Community
    - Regulated Data Strategy (ReD)
    - Dataverse, NIH Other Transaction Award
    - Northeast Storage Exchange (NESE)
- Future Ecosystem
    - An ecosystem of applications, services, resources integrated by standards-based and service-oriented framework that will be populated by Library Services, University RC and school-based RC services, and Office of VP of Research support as well as researchers themselves working in partnership
- Details
    - Support the entire data lifecycle for conducting research
    - Support creation, sharing, and curation of digital content
    - Researches will include data access for findability, computation
    - Built on Service Oriented Architecture (SOA)
    - Based on loosely coupled, distributed, interoperable services and tools
    - Modes of access that any research can have (access through web browsers)
- Challenges do you face in sharinging and publishing research data?
- What technology or tooling would you like to see in places to reduce the friction in your research process?
- Harvard dataverse uses AWS S3 and has a backup at the NESE
- Need a project manager to manage a collaboration across 40 people; divide into 4 subgroups https://sites.harvard.edu/harvard-data-commons/people/;
    - Very specific statement of work
    - 2x week meetings, needed to overcome domain gaps; Aug 2021–June 2022;
    - The participants were already meeting about these issues, but separately
    - Not director-level

- Soliciting faculty input: faculty ambassadors program; working on a strat plan for research and data and that will produce feedback
- Who will lead this operationally after this project? Run for MVP. Existing groups that run components are still running them. This group just provided the glue to stitch them together. Facilitators will be needed to help guide people through these utilities.
-

Next 2 Months Presentation Plans:

- March 16th, 2023: TBD
- April 20th, 2023: TBD

Suggested future topic for a systems facing call (Feel free to add one):

-