Data Commons Workshop

October 23, 2020

We will begin momentarily...

Principles of Engagement

Housekeeping/Logistics

- The Zoom link will be live for 30 minutes before and after the workshop to allow participants to get settled and network
- If possible, please set your name display to “First Last (Institution)”
- Only presenters and moderators will use the screen share feature
- The session will be automatically recorded, no independent recordings of the sessions are permitted
- Participants should mute microphones when not actively speaking
- Moderators will verbally queue a 1 minute warning during presentations
- **Participation is highly encouraged, so please use the “Raise your hand” feature to indicate that you would like to speak or post questions and comments in the Zoom chat**
- Products of the workshop (presentations, notes, and the recording) may be shared with internal communities of interest within participant institutions to further explore the topic

Collaborative Practices

- Using welcoming and inclusive language
- Being respectful of differing viewpoints and experiences
- Focusing on what is best for the community
- Showing empathy towards other community members

Welcome & Introductions

Workshop Hosts

Mercè Crosas, PhD
University Research Data Management Officer, Harvard University Information Technology
Chief Data Science and Technology Officer, Institute for Quantitative Social Science

Stuart Snydman
Managing Director of Library Technology Services, Harvard University Information Technology

Scott Yockel
University Research Computing Officer, Harvard University Information Technology
Director, FAS Research Computing

Welcoming Remarks By

Martha Whitehead
Vice President for the Harvard Library and University Librarian
Roy E. Larsen Librarian for the Faculty of Arts and Sciences

Data Commons for Institutions

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:** Methods and standard tools for analysis, synthesis, simulation, visualization, sharing, integration, and management.
- **Data Systems Architecture:** A fast, scalable fault-tolerant infrastructure for real-time analyses and data transfer; controlled data access.
- **Data Repositories:** Well curated, validated, open access repositories required to ensure that the results of scientific research are available.

[NSF CI2030: Future Advanced Cyberinfrastructure document, 2018]

… to Data Commons

"… a data commons brings together (or co-locates) **data with cloud computing** infrastructure and **commonly used** software services, tools & applications for **managing, analyzing and sharing data** to create an **interoperable** resource for a research community."

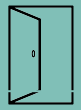
[Robert Grossman, on the NIH Data Commons Consortium initiative]

The Problem

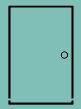
1. Researcher A doesn't know about data from Researcher B; **not easy to collaborate**, especially using sensitive data with DUAs
2. Research lifecycle steps 1, 2, & 3 are **not connected** and **duplicative**; not easy to manage data throughout lifecycle and publish research output

Researcher A

1. Data Collection



Open data (gov, cities)



DUA

Private, sensitive data (companies, hospitals)

Data collected for research (experiments, observations)

2. Active Research

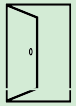
Research computing, software, methods workflows

3. Data Publication

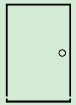
Data repository

Researcher B

1. Data Collection



Open data (gov, cities)

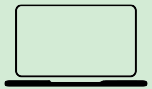


DUA

Private, sensitive data (companies, hospitals)

Data collected for research (experiments, observations)

2. Active Research

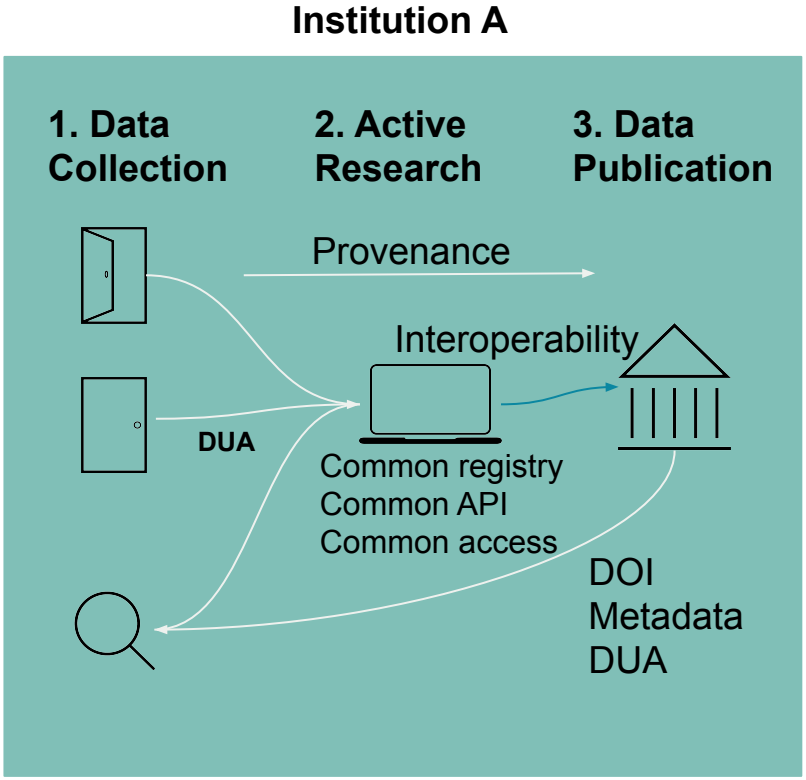


Research computing, software, methods workflows

3. Data Publication

Data repository

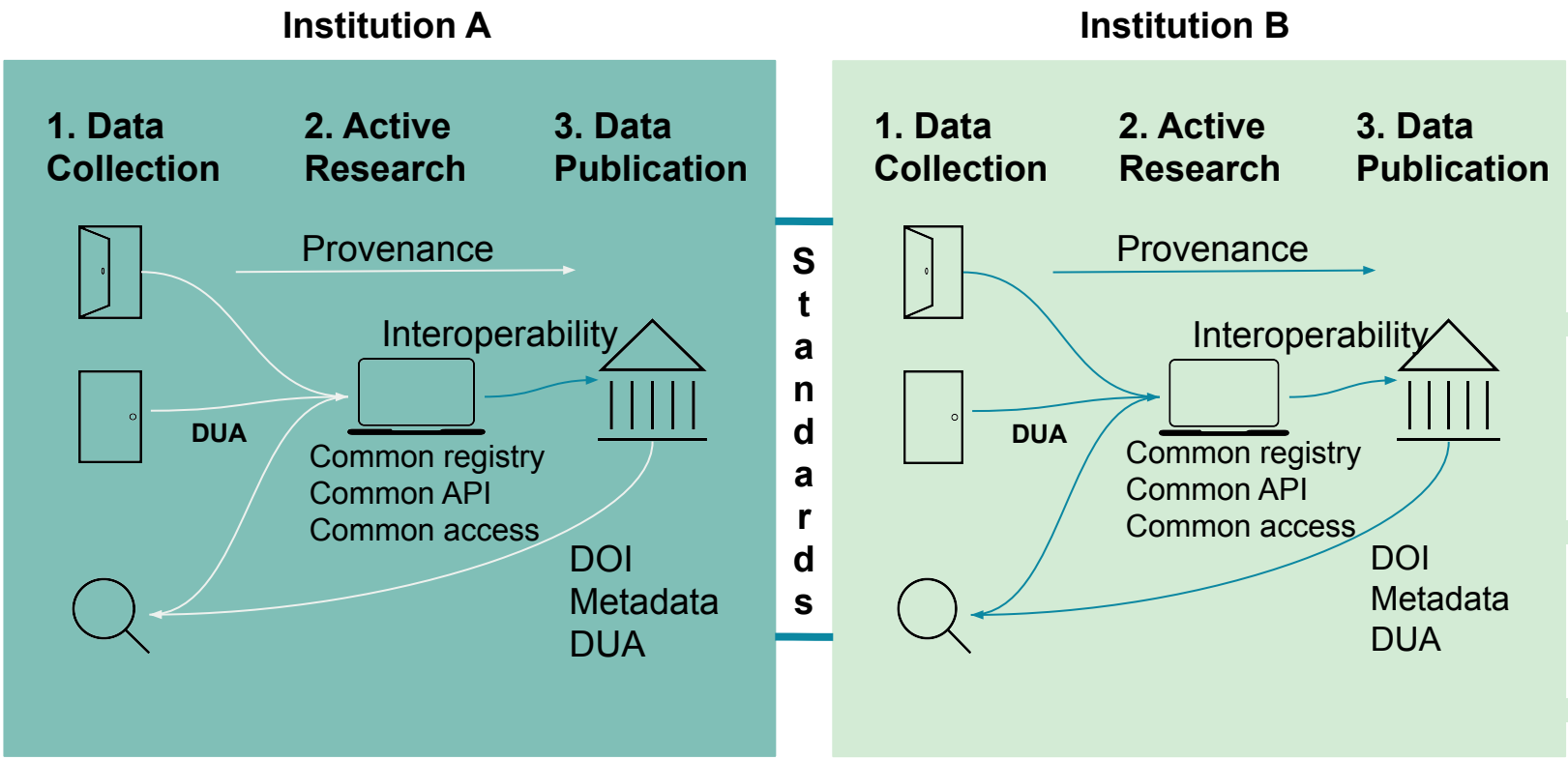
The Solution for an Institution



- **Common registry** with **metadata**, so Researcher A can find and access data from Researcher B
- **Common API**, so research tools can access data and **interoperate** between computing and repository
- **DUA and controlled access** tracked and shared throughout lifecycle
- **Provenance** tracked throughout the lifecycle to produce reproducible research outputs

The Global Solution

Agree on standards so Data Commons A can talk to Data Commons B



Workshop Objective

Hear from you:

- What does a data commons mean to you?
- What technologies, practices, or standards are you involved in that could be part of a data commons solution?
- What is your vision for next steps related to implementing a data commons for your institution?

Identify together:

Technologies, processes, and standards that we can all use to build a Data Commons for our institutions, while avoiding creating a “data commons silo” for each institution.

Tools and Technologies to build a Data Commons

Agenda

FACILITATOR(S)	SESSION	TIME	PRESENTATIONS
N/A	Networking	9:30 - 10:00 AM	None - Open Session / Meet and greet
Mercè Crosas Scott Yockel Stuart Snydman	N/A	10:00 - 10:10 AM	Welcome and Workshop Introduction
Mercè Crosas	Technologies and Practices to build a Data Commons	10:10 - 10:15 AM	Vivien Bonazzi, Deloitte
		10:15 - 10:20 AM	Tim Clark, University of Virginia
		10:20 - 10:25 AM	Carole Goble & Bill Ayres, University of Manchester
		10:25 - 10:30 AM	Ilya Baldin & Jonathan Crabtree, University of North Carolina
		10:30 - 10:35 AM	--------
		10:35 - 10:40 AM	Ian Foster, Globus, University of Chicago
		10:40 - 11:00 AM	Q&A / Session Discussion
		11:00 - 11:10 AM	**BREAK**

Break

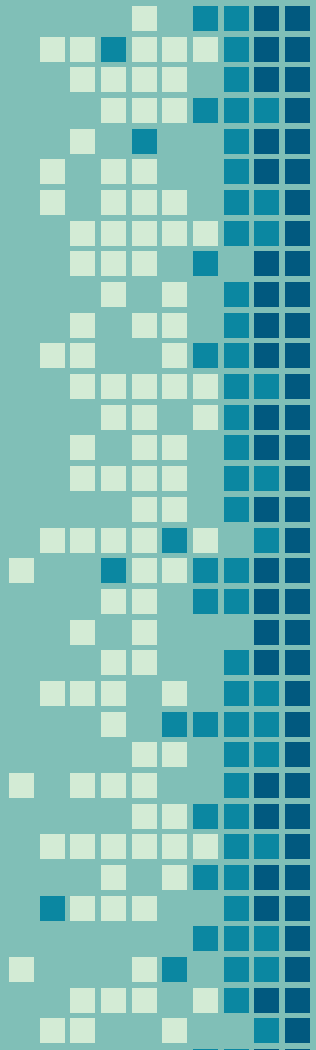
Data Commons from the Library perspective

Agenda (continued)

FACILITATOR(S)	SESSION	TIME	PRESENTATIONS
Stuart Snydman	Data Commons from the Library perspective	11:10 - 11:15 AM	Philipp Conzett, UiT The Arctic University of Norway
		11:15 - 11:20 AM	Jon Stroop, Wind Cowles & Curt Hillegas, Princeton University
		11:20 - 11:25 AM	Heather Yager & Amy Nurnberg, MIT
		11:25 - 11:30 AM	Wolfram Horstmann, University of Göttingen
		11:30 - 11:35 AM	Tim McGeary - Duke University
		11:35 - 11:40 AM	Erin Foster, University of California, Berkeley
		11:40 - 12:00 PM	Q&A /Session Discussion
		12:00 - 12:10 PM	**BREAK**

Break

Data Commons from the Research Computing perspective



Agenda (continued)

FACILITATOR(S)	SESSION	TIME	PRESENTATIONS
Scott Yockel	Data Commons from the Research Computing perspective	12:10 - 12:15 PM	Jim Wilgenbusch, University of Minnesota
		12:15 - 12:20 PM	Greg Madden, The University Corporation for Atmospheric Research
		12:20 - 12:25 PM	Ruth Marinshaw & Tom Cramer, Stanford University
		12:25 - 12:35 PM	Q&A / Session Discussion
Mercè Crosas Scott Yockel Stuart Snydman	N/A	12:35 - 1:00 PM	General Discussion, Conclusions, and Next Steps
None	Networking	1:00 - 1:30PM	None

Wrap-up

Thank you!

Workshop notes and participant contact list can be found in the Workshop's Google folder:

https://tinyurl.com/yyh5uc34