

CDI Monthly Meeting 20161109

CDI Monthly Meeting - November 9, 2016

The Community for Data Integration (CDI) meetings are held the 2nd Wednesday of each month from 11:00 a.m. to 12:30 p.m. Eastern Time.

WebEx:

<https://usgs.webex.com/> - Under the Meeting Center tabs, search for meeting name: "Community for Data Integration."

Audio:

USGS/DOI Dial In Number: (703) 648-4848 (for USGS and DOI offices)

Toll Free Dial In Number: (855) 547-8255 (for other offices and telecommute locations)

Conference Code: 47919# (same for both numbers)

Webex Recording

Webex recordings are available to CDI Members approximately 24 hours after the completion of the meeting. Please login to view the recording. If you would like to become a member of CDI, please email cdi@usgs.gov.

Agenda (in Eastern time)

11:00a **Scientist's Challenge** - [What collaboration methods and workflows are scientific programmers using?](#) Jeremiah Lant, USGS

11:10a Welcome - Kevin Gallagher - Associate Director for Core Science Systems and Tim Quinn - Office of Enterprise Information Chief

Rich Signell: There has been some talk of an opportunity for building infrastructure to support simulation data and high performance computing. Do you guys have plans for taking advantage of that opportunity.

Kevin: You are talking about the data call from OMB that has been working its way down the chain. The goal is to identify potential infrastructure investments. At the USGS, we have a large facility maintenance backlog (e.g. issues with building 20 on the Denver Federal Center and consolidation of Menlo Park offices with NASA on the Ames facility base. Items from that backlog will be part of USGS's response to that data call, but we have also had discussions about including other types of investments in our response such as monitoring equipment (e.g. stream gages) and computing infrastructure. Tim and I are both big proponents of those types of investments. Right now, the data call is coming in and we will see how folks are responding across different mission areas and the ELT will work to categorize the different needs. I am very hopeful that one category might be science computing infrastructure and that that can be a shared investment across the Bureau. These kinds of data calls are not frequent, but they do come around every so often. I want to make sure not to oversell it because I am not sure if it is a data call solely of interest or if it might actually get legs (e.g. and infrastructure bill). That being said, we are always interested in making sure we are participating in these calls and putting relevant plans on the table.

Tim Quinn: From an IT infrastructure point of view, we are interested in exploring data center consolidation and cloud storage. We do have some real challenges in these areas especially in the movement of data. These challenges can't be solved overnight. As money becomes available, we will take advantage of it to improve our science.

Kevin: Some of the priorities that I mentioned earlier, such as climate prediction, hazard early warning systems, and early warning of the spread of wildlife diseases, all require high performance computing.

11:15a CDI Working Group Report Outs [\[PDF\]](#)

11:25a **CDI FY17 Statements of Interest Close of Voting Session** - Leslie Hsu, USGS & Madison Langseth, USGS

[FY17 RFP Survey - did you participate?](#)

11:45a **Council of Data Facilities: History and Overview** - Danie Kinkade, Woods Hole Oceanographic Institution

Abstract:

The Council of Data Facilities (CDF) was formed in 2014 as an outcome of an EarthCube End-User Domain Workshop for geoscience data facilities, in order to coordinate with the many elements of the EarthCube initiative at a time when society's expectations of data facilities are increasing in scale and scope. The Council serves the community in a coordinating and facilitating role to provide a collective voice on behalf of its members to the NSF and other foundations; endorse, and promote standards and best practices; identify and support development and utilization of shared infrastructure and services; and foster innovation through collaboration. This talk will provide an overview of the CDF, including its history, structure and function, and recent activities.

Presentation: Slides are available to CDI Members. Please login to download the slides. If you would like to become a member of CDI, please email cdi@usgs.gov.

Danie Kinkade is an Information Systems Associate at the Biological and Chemical Data Management Office (BCO-DMO) at the Woods Hole Oceanographic Institution. Her background includes a master's degree in oceanography and over 17 years of experience managing oceanographic research data in both federal and academic environments. She is currently serving as Co-chair on the Council of Data Facilities, and recently served on the EarthCube Leadership Council as an at-large member representing the oceanography community. Her current interests lie at the intersection of data management, informatics and cyberinfrastructure, where use of novel technologies and community best practices facilitate data discovery, access and analysis as part of the research process.

Q&A

Rich Signell: Are there challenges that Earthcube has that CDI can assist with?

Danie: I have limited knowledge of CDI, but we are interested in seeing what CDI is doing. The executive council is interested in engaging with CDI and open to learning about other community endeavors to leverage.

Viv Hutchison: What is the CDF's end game? Is it the registry or is there an intention to have a physical connection among these data facilities?

Danie: We are leveraging the DataONE model for the CDF registry. I think we would love to do that, but we haven't talked about it in detail. Right now we are taking baby steps.

Lindsay Powers: Earthcube has been actively engaged in developing the architecture plan for bringing the Geoscience community together and that architecture is designed around the Geoscientist workbench, which will provide an environment where workflows can be created and shared. That is one area where these data facilities or outcomes of the CDI projects can contribute and get connected.

12:00p **Trusted Digital Repositories (TDR): Proposed New Criteria and Process Flow** - John Faundeen, USGS, Keith Kirk, USGS, and Clara Brown, USGS

Abstract:

The Trusted Digital Repositories working group was formed at the request of Tim Quinn during the March 2016 Public Access Plan implementation meeting in Reston. The purpose of this working group is two fold: 1) to explore possible enterprise solutions for data repositories within the USGS including cloud solutions and 2) to explore and recommend an approach for certifying USGS public-facing servers providing scientific data to the public, to meet new federal requirements associated with trusted digital repositories. John Faundeen, is an archivist with EROS and is co-chair with me of the FSPAC Data Preservation Subcommittee. He is also co-chair of the TDR working group along with Clara Brown, who came to the USGS in 2016 and who is Chief, Digital Library Services, CSS. Clara is responsible for the USGS Publications Warehouse. The three of us are making this presentation together today.

The trusted digital repository working group aligned with the FSPAC digital preservation subcommittee in June as subject matter experts to inform the Data Preservation subcommittee. The results of our efforts are presented here today.

Presentation: Slides are available to CDI Members. Please login to download the slides. If you would like to become a member of CDI, please email cdi@usgs.gov.

John Faundeen has worked as the EROS Archivist at the U.S. Geological Survey (USGS) Earth Resources Observations and Science (EROS) Center since 2001. His role involves policy, oversight and guidance for the observational, cartographic, and elevation data created and maintained at EROS. John allocates most of his time to preservation and appraisal functions. The preservation activity includes environmentally managing a 20,000 square foot archive containing 100,000 rolls of analog film and thousands of magnetic tapes. Establishing an off-site archive containing several petabytes of electronic data continues to be a centerpiece of EROS's data management risk mitigation strategy. MOUs with the U.S. National Archives and Records Administration were established based upon proven data management capabilities. He has previously served as the acting USGS Records Officer on two separate occasions.

Keith Kirk is a Bureau Approving Official (BAO) in the Office of Science Quality and Integrity OSQI) in the Office of the Director. He is an active participant in a number of Bureau committees, including the USGS Data Policy committee and various Fundamental Science Practice Advisory Committee (FSPAC) subcommittees and an active member of USGS Community of Data Integration (CDI). Keith was the lead for developing the USGS Public Access Plan approved by the Whitehouse Office of Science and Technology Policy (OSTP) and Whitehouse Office of Management and Budget (OMB) in January 2016.

Clara Ruttenberg Brown has worked as the Chief of Digital Services for the USGS Libraries since November 2015. She comes to the Survey with almost twenty years of working in Federal, Academic and Public Libraries where her focus has been on anticipating

customer needs and seeking ways of improving their experience in providing unfettered access to the library's digital and electronic resources. Clara enjoys yoga, kayaking and thinking about exercising in general and if you are in Reston she enjoys planning and attending happy hours.

Q&A

John Faundeen: If there are folks out there who have an interest in joining the Trusted Digital Repository pilot, we would be interested in working with you.

Leslie Hsu: How well do you know the USGS entities that might be interested in becoming Trusted Digital Repositories?

John: We have some hints based on conversations over the last year, but I'm sure we will have some surprises.

Keith: We have already had three requests from different facilities to participate in the pilot.

Clara: I agree. We have hints, but we don't have a complete picture.

Participant: So to clarify, if we are interested in participating in the pilot, we should contact the three of you?

Keith: Yes.

Michelle: There is a link to an internal FSP page with a list of Trusted Digital Repositories. Could you explain that and the connection with this pilot effort?

Keith: That list was put together by the FSPAC data preservation subcommittee. It is on the public website and that list doesn't say "Trusted Digital Repositories." It says, "acceptable repositories where USGS scientists can put their data." That list is evolving based on scientist input. We purposefully did not call them "trusted" because there are very few federal repositories that have that seal. EROS has the seal and John can talk about that.

John: EROS went through the Data Seal of Approval certification in 2013, which was prior to the combination of the two bodies (Data Seal of Approval and World Data Systems). EROS is certified through 2017, but did not go through this current process. EROS will be going through this new process next year.

Michelle: Will the repositories on that list be required to go through this process in the future?

Keith: Many of these repositories are not in our purview. Many fall under NSF or OSTP. Some are state repositories and will not be required to achieve the status. Since this is part of an executive order and OSTP hasn't agreed on what will constitute a Trusted Digital Repository, we are moving very slowly because we don't want to have to duplicate efforts in the future if OSTP decides to go a different direction.

12:30p Adjourn

Attendees

A WebEx Participant Report is available to CDI Members. Please login to download the report. If you would like to become a member of CDI, please email cdi@usgs.gov.