CDI Monthly Meeting 20160810

CDI Monthly Meeting - August 10, 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:** Leah Morgan, USGS, Data release for a diverse group of geochemistry labs

11:10a **Welcome** - Cheryl Morris - Director, Core Science Analysis, Synthesis, and Libraries

Related Links

- Virtual Training on Reviewing Metadata, 8/11 12-1:30p ET - Register and more info at https://my.usgs.gov/confluence/x/_46YI
- Prepare for the CDI FY17 RFP - View past funded projects: https://www2.usgs.gov/cdi/products-publications.html
- CDI FY17 In-Person Annual Workshop: Contribute and vote on ideas: https://usgs1.uservoice.com/forums/398538-annual-meeting-fy17-ideas
- FY17 RFP Collaboration Forum

11:15a **Public Access/Open Access at USGS: A Story of a Science Data and Publishing Evolution -** Viv Hutchison, USGS

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.

Abstract:

The USGS is undergoing a change in the way we handle science data associated with scholarly conclusions that will have huge impacts on our bureau's contribution to science as a whole. We are striving to better ensure that the science data we use in our scholarly conclusions are accessible, understandable, and available globally to other researchers, our partners, and the public. We are beginning to make publications available using more modern methods such that sharing and free access is the norm, and the connections between pubs and data are prominent. There are cultural changes happening in science on a global level, and USGS is making remarkable progress in this new shift in the way science is done. Let's explore how we got here and where we might go!

11:45a **Supporting Academic Data Science: The Data-Driven Discovery Initiative at Moore** - Carly Strasser, Moore Foundation

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.

Abstract:

Today academic researchers are faced with a huge array of new tools and techniques for handling large data sets and engaging in computationally intensive research. While the research community recognizes the need academics to understand and use these new tools and techniques, there is a critical shortage of practitioners. Science may be data-rich, but will remain discovery-poor without the institutional commitment, people-power and technology needed to mine data and reveal hidden breakthroughs. The Data-Driven Discovery initiative is an effort within the foundation's Science Program focused on promoting the both the researchers and the practices required for the modern era of data-driven research, with a broader vision of accelerating scientific discovery.

Presentation Q/A

Question from Chat: What is the difference between R Markdown and jupyter?

Carly: Others might have a better explanation, but one difference is you can use Jupyter with a lot of different languages other than just R. You can put the Jupyter notebooks on GitHub and allow people to interact with them. I'm not sure how much of that functionality is available through R.

Rich Signell: It is interesting that it took the Moore Foundation and Sloan Foundation to fund something that is as useful as the Jupyter project. How do you guys interact or don't interact with other groups such as Earthcube, that are sort of playing in this same arena?

Carly: NSF and Federal funders are doing a better job of supporting infrastructure projects like Jupyter Notebook. The original creator of IPython Notebooks, Fernando Perez, is a researcher with Berkeley and has not been treated well by Berkeley in terms of getting jobs and support and it's kind of stunning the someone who has created something that is so important to the research community, has no status or stature within the academic world because it is not discovery-based project that he is working on. NSF recently announced some funding for software sustainability-type projects is funding infrastructure-based projects, so I am optimistic that they are seeing the importance of this. In terms of our interactions, my boss, Chris Mentzel, spends a lot of time with the Big Data Hubs people and the NIH BD2K project. Those organizations are definitely thinking about this space. My interactions are pretty minimal but I do spend a lot of time with organizations like AGU, thinking about how do we promote the data science practices and techniques within the AGU community, which of course is funded by NSF. I don't interact at all with Earthcube, but I did interact with DataONE back in the day. There is some overlap between what those two groups are thinking about. I am optimistic that the Federal government is getting on board with this space. One of the things that we think about at our Foundation is "What should we be funding that is hard to get funding for from traditional sources?" So that is kind of our criteria for giving money is "Would they be able to get this project funded through some other group?" And if the answer is yes, then we may not feel that we should support it.

Rich: When you give money to people, like Fernando Perez, does that help them get tenure? Have they been more successful in the future?

Carly: Yes, we have seen really positive effects from the funding that we have given. Universities love pleasing funding agencies. One investigator was hired on as non-tenure track associate and he was just given the opportunity for a tenure-track position.

Rich: We spend a lot of time talking about in this group discovering data. How do you find these datasets that you might do scientific discovery with beyond just Google searching. Do you guys play in that arena or have tools that you support that help with this issue?

Carly: Not so much. It is an interesting space of how to find the data that you are looking for. We are promoting new publication models particularly puslishing all types of outputs including data. We get asked a lot to fund databases, but that is not a space that we dabble in. Our researchers don't really have trouble finding data. They are really tech savvy and people with data tend to reach out to them. We don't promote data discovery tools.

Leslie: How you might convince other communities to start embracing preprint world?

Carly: It is easy to forget that it is hard to get communities to embrace preprints. ASAPBIO.org has some really interesting articles abou the benefits of preprints and the value of early release of information. I think that a big component of getting people excited is showing them that it is not going to hurt their chances of publishing their work later. Publishers have started allowing people to publish preprints and still have their manuscripts accepted for publication later. Some publishers are being "bullied" into accepting papers that have been published as preprints. Many groups are saying they are encouraging preprints. Preprints can be included on proposals as outputs; however, they don't necessarily count for tenure. It is still going to require some additional work.

Question from Chat: There will be ice skating in Hades before USGS and other Fed scientists are allowed to circulate preprints. How you think that will work out long term? (FSPAC Bruce Taggart would be a good contact to answer this)

Keith Kirk (via email to CDI): The Federal Gov is disallowed from placing draft manuscripts where they are available to the pubic owing to the 2004 OMB Peer Review Bulletin which directs federal agencies to ensure their research has been peer reviewed prior to "public dissemination". So unless OMB changes their mind the USGS will not allow draft manuscript to be made available on JA web sites unless they have received USGS Bureau approval first.

12:20p **Working Group Reports**

- Citizen Science - Sophia Liu and Dave Govoni
    - Open Innovation Leaders Blog Post and Poll: http://internal.usgs.gov/thecore/leaders/?p=44278. Sophia helped cowrite this to rraise awareness on participatory approaches to science.
    - Mapping Innovation Workshops: https://sites.google.com/a/usgs.gov/usgs-mapping-innovation-series/mapping-innovation-workshops
    - Sophia has been helping with organizing the mapping innovation workshops. These are meant to be internal to USGS; however, we are also hoping to have a more open mapping innovation workshop that other non-USGS participants can get involved with.
    - **Rich:** is this ESRI focused?
    - **Sophia:** The idea is not to think about it as ESRI focused. They have been a big player, but there are a lot of other tools and technologies involved in mapping innovation workshops. We are updating the Google site with agendas, etc. that will hopefully demonstrate how this is not just ESRI focused.
- Communication - JC Nelson and Marcia McNiff
    - Holding monthly calls, go on the wiki to see how to get involved.
- Data Management - Heather Henkel and Viv Hutchison
    - We didn't have a meeting this month. We will meet again in September. Most recent activities: Cassandra is leading subteam looking at data management strategies for centers in addition to data management planning and best ways to go about that.
- Earth-Science Themes - Roland Viger
    - The Earth-Science Themes Working Group is still gestating. Although no regularly scheduled calls have occurred in the last quarter or so, there is in fact activity in and around ETWG. This is chiefly occurring in the form of collaboration with other communities and USGS Programs. Examples include the extension of NHDPlus software to create value-added derivatives for the high-resolution National Hydrography Dataset, discussions about different types of hydrological connectivity (including temporally varying ones), preliminary investigation of new methods for integrating elevation and hydrographic information, tying together of US and Canadian hydrographic and hydrologic information, and the use of soils information in water quality modeling. While much of this work is being done beyond the

usual “CDI teleconference” spheres of interaction, but we are working to attract people, their ideas, and energy back to ETWG. For now, ETWG has become a place to start asking earth science-driven questions and (hopefully) finding folks working in these areas.

- Semantic Web - Fran Lightsom
  - Reviewing metadata training is tomorrow. SWWG has been heavily involved in developing this training event.
- Tech Stack - Rich Signell
  - We have a monthly call tomorrow to hear about the Community Data Analysis Tools (CDAT) from Lawrence Livermore National Laboratory. It’s a great package that has been around for a while. See http://wiki.esipfed.org/index.php/Interoperability_and_Technology/Tech_Dive_Webinar_Series for more information.
- Connected Devices - Tim Kern and Lance Everette
  - 8/28 meeting will include Elizabeth McCarthy from the National Map Corps

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.