CDI Monthly Meeting 20180912

CDI Monthly Meeting - 20180912

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.

Link to Google Calendar Event

Meeting Recording

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

During the call, you can ask and up-vote questions at slido.com, event code #3514.

Agenda (in Eastern time)

11:00a Scientist's Challenge - Outcrop and field data

11:05a Welcome - Leslie Hsu, CDI Coordinator [PDF]

11:10a CDI Announcements [PDF]

11:20a **STEP-UP to Data Management**, Sue Kemp, USGS

11:30a **CDI Funded Project Report: Empowering decision-makers: A dynamic web interface for running Bayesian networks**, Erika Lentz, USGS

11:45a **USGS Thesaurus: what it is, how you can use it, and how you can improve it**, Peter Schweitzer, USGS

12:30p Adjourn

Abstracts

STEP-UP to Data Management, Sue Kemp, USGS

Last school year, FRESC was assigned a STEP-UP student to remotely work on a legacy data management challenge. Sue Kemp will describe how Gabe Reizzis assisted FRESC in the daunting task of sunsetting the Sagemap website, which includes capturing, cataloging, and maintaining public access to decades of NBS and USGS information.

Empowering decision-makers: Developing a dynamic web interface for running Bayesian networks, Erika Lentz, USGS

We present some of the early lessons learned through the ongoing conversion of a probabilistic modeling framework from proprietary to freely available open-source software, including 1) processing, storage, and hosting; 2) staffing requirements and needs; and 3) the development process. The intent of the resulting product, which will continue development in FY18 and FY19, is to create a portable interactive web-interface to serve as a prototype to demonstrate how interdisciplinary USGS science and models can be transformed into an approachable format for decision-makers.

Erika Lentz is a Research Geologist with the U.S. Geological Survey stationed at the Woods Hole Coastal and Marine Science Center. She received her PhD in Geology from the University of Rhode Island in 2010, and from 2012 to 2014 was a USGS Mendenhall Research Fellow. Her research focuses on coastal change and the processes that drive it over a range of spatial (barrier island to regional) and temporal (storms to sea level rise) scales in both natural and built environments, and is also interested in the meaningful communication of scientific information to support decision-making.

USGS Thesaurus: what it is, how you can use it, and how you can improve it, Peter Schweitzer, USGS

The USGS Thesaurus provides a means of categorizing a wide variety of information resources in a way that does not require detailed knowledge of the structure and function of the organizations within the USGS. Its specific purpose is to provide topical categorization at a general level that enables technical non-specialists to identify information resources relevant to problems of interest to them.

This presentation will explain the structure of the thesaurus, how you can use it in your work, and how you may participate in refining and further developing it.

Peter Schweitzer works on scientific information management at USGS in Reston VA. With a formal background in geology and oceanography and practic al experience in software development, he serves as a bridge between the scientific research and information technology communities. Currently, his work focuses on providing usable scientific data to the public. The largest collection of information he manages is at MRData.usgs.gov and includes geological, geochemical, geophysical, and mineral resource data produced by the Mineral Resources Program. Widely known for producing software that parses and validates geospatial metadata, he led a small group to develop controlled vocabularies and cataloging software to provide topical interfaces for scientific information on the web. He holds a Ph.D. in oceanography from the Woods Hole Oceanographic Institution and Massachusetts Institute of Technology.

Presentations

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

Highlights

1. Outcrop and field data scientist's challenge: https://my.usgs.gov/confluence/x/Rh8YJ
2. CDI RFP Wiki page: https://my.usgs.gov/confluence/display/cdi/2019+Proposals
3. DataCamp Python course: https://www.datacamp.com/courses/intro-to-python-for-data-science
4. Sign up for the Python course at: http://goo.gl/HNgp16
5. SpatioTemporal Feature Registry IdeaScale: https://esipfed.ideascale.com/a/campaign-home/23576
6. STEP-UP NPR article link: https://www.kqed.org/science/1922125/students-with-autism-excel-in-working-with-data-helping-scientists
7. From skemp : https://goo.gl/forms/1GduegaBdrvaMoT32 STEP-UP submission form
8. USGS Thesaurus: https://www2.usgs.gov/science/about/
9. USGS Thesaurus to-do:
    a. Correct, refine, and extend Thesaurus concepts
    b. Create cross-walks to other controlled vocabularies
    c. Build more web services and application interfaces
    d. Help other people use this resource effectively
10. Contact Peter Schweitzer, pschweitzer@usgs.gov, if you are interested in working on the USGS Thesaurus - we would like to include diverse scientific perspectives.

Q&A

SpatioTemporal Feature Registry

- Q from Fran Lightsom: Are all these places on land?
- A from Sky Bristol: Here's a notebook with the basic process we ran to incorporate "OBIS Areas". https://github.com/usgs-bis/sfr/blob/master/OBIS%20Areas.ipynb
    - Those are partly derived from MarineRegions.org with some processing to make the geometry more useful for analytical purposes.
    - Those areas are all in a place name index we built. I'll share the API on that as soon as it's public, and we'd love to have a conversation about it.

Empowering Decision makers: A dynamic web interface for running Bayesian networks

- What are some considerations if you want to figure out if your research question would benefit from Bayesian network modeling?
    - Can be used as a complement to more specific deterministic models. They can be built to ask about specific management actions. If there are some variables you know really well, and other variables you know less well, the Bayesian networks help. They can point to where you need to know more.
- Are there difficulties explaining probabilities and model details to your stakeholders?
    - Some people really get probability, like the insurance industry. But we want to make sure the probabilistic outcomes are framed in a helpful way, that is part of this tool development. Does the way we are presenting outcomes resonate with the stakeholders? What else can we say that would help address stakeholders understand probabilities to answer their questions?
- Do you have plans to go back to the user group that you created the user stories with to see how well your bayesian network assists them with their decisions?
    - Yes, that a strength of this project. We have a group of stakeholders ready to continue to give us feedback as this develops.
- Have you considered using the geoplatform.gov for hosting your public facing app? https://www.geoplatform.gov/ Email: servicedesk@geoplatform.gov
    - Tell me more!
- Re geoplatform: I have been told that we are required to use Cloud Hosting Solutions? Not sure if this is a USGS WMA policy or if it is USGS-wide?
    - A: Re:geoplatform: The requirement to use Cloud Hosting Solutions (CHS) is USGS-wide. For anything cloud related, contact CHS and they can help you find a solution.

USGS Thesaurus: what it is, how you can use it, and how you can improve it

- I get confused about thesauri, vocabularies, and ontologies, as a user, do I need to know the difference?

    - The more general idea is Knowledge organization systems. There are flat lists, authority files, and thesauri. Ontologies are more complicated than the thesaurus we have, ontologies have more complicated ways to describe relationships. We were counciled to focus on formal thesauri because that was the level needed to accomplish the browse function of theUSGS website, which was used for about 10 years. We used it to accomplish browse functions in apps. Compromise in complexity in structure and information. Constraint that it would work in a web environment.
- The thesaurus seems like it would integrate in some way to the taxonomy required to move on to the USGS Drupal hosting of websites. Has this happened at all?
    - It was connected to the USGS website for awhile. But currently they don't need it anymore, the current USGS website taxonomy is more for web navigation. We can look for differences in the two systems and see if they can inform the thesaurus.
- Is there an API to make use of this, or link it into things like ScienceBase? (Rather than downloading and re-hosting)
    - The web services could be used for that. We haven't linked directly to ScienceBase yet. ScienceBase has some ways of dealing with controlled vocabularies, so that would be a fruitful conversation to have with their team.
- How would we link in more detailed disciplinary vocabs? Or should we?
    - That is an interesting philosophical and technical question. We have a lithology vocabulary that is much more detailed than the USGS Thesaurus. When you get into the fine details of scientific investigations, you need to make more distinctions than when you are just trying to point people to the correct resources. The level of detail needed for different tasks. Where does the other vocabulary plug in? This is related to the concept of cross-walking vocabularies that I mentioned earlier - this requires more investigation!
- What relational database is this data stored in?
    - We've used a different databases: PostgreSQL, and we also have a SQLite database.
- How does the USGS Biocomplexity Thesaurus fit (or not) into this project?
    - It uses a different data structure. For USGS Thesaurus, every term only has one broader term, that gives a simple tree structure. Biocomplexity Thesaurus has a different structure which is more complicated to show on the web. Better for direct search interfaces. There is room for more integration, though it is a bit complex. It also has a lot more terms (~4400 vs 900 in the thesaurus).
- Can you describe the background engine that analyzes and makes use of the vocabulary, and making a relevant and correct inference?
    - I am categorizing resources (like publications or datasets) has a card catalog entry in another database. There is thin metadata - citation and keywords. That is an efficient way to bring up all resources related to a keyword and those with keywords that are narrower than that. Simple inference to help people find information pertinent to a topic.
- How often is the data catalog updated. If someone enters new data but not yet in Thesaurus catalog, how long would it be before the various data labels would function?
    - Depends on the catalog. I separate data structure of the vocabulary and the catalog using the vocabulary. The catalog using the vocabulary might have a copy of our vocabulary. Contact me by email if you have more detailed questions.
- If the thesaurus has a 'use for' recommendation, should scientists follow the thesaurus, or are we 'allowed' to use the non recommended term?
    - It depends on the context. In a web interface, the 'use for' is fair game. In a metadata record, you'd be better off using the 'use for' term.
- Sounds like simple searching/indexing, a typical DBMS query function. Why not use some AI and ML to maximize the value of the thesaurus. Like Google Search!
    - Maybe it could be done. :-) I like to understand how things are actually working, so this is conceptually simpler than some of the technologies are out there now. Those technologies may be able to do the same function.

Attendees

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