CDI Monthly Meeting 20200610

June 10, 2020: CDI Funded Projects: ScienceBase for Disaster Risk Reduction, Coupling Hydrologic Models and Data Services, and Wildlife Disease Data

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.

Meeting Recording and Slides

These are the publicly available slides. Log in to view all content. If you would like to become a member of CDI, join at https://listserv.usgs.gov/mailman/listinfo/cdi-all.

Agenda (in Eastern time)

11:00 am Welcome and Opening Announcements

11:15 am Working Group Announcements

11:25 am **Extending ScienceBase for Disaster Risk Reduction** - Joe Bard, USGS

11:45 am **Coupling Hydrologic Models with Data Services in an Interoperable Modeling Framework** - Rich McDonald, USGS

12:05 pm **Transforming Biosurveillance by Standardizing and Serving 40 Years of Wildlife Disease Data** - Neil Baertlein, USGS

12:30 pm *Adjourn*

Highlights

1. **See blog post on this meeting here.**
2. Create, update, and share your staff profile, ORCid links, or other professional page with CDI. We hope to learn about overlapping interests, geographic groupings, and more. Use this form to participate.
3. If you maintain or use APIs, please contribute information about it so we can compile and share links back to Science Information Services (SIS). Participate here.
4. See more about the USGS Model Catalog here.
5. Code for coupling hydrologic models is available on GitLab for USGS employees.
6. Explore WHISpers.

Notes

1. **Welcome and Opening Announcements**
    a. Leslie Hsu
        i. Create, update, and share your staff profile, ORCid links, or other professional page with CDI. We hope to learn about overlapping interests, geographic groupings, and more. Use this form to participate.
        ii. If you maintain or use APIs, please contribute information about it so we can compile and share links back to Science Information Services (SIS). Participate here.
    b. Kevin Gallagher
        i. Kevin discussed EarthMAP and CDI Funded Projects. USGS employees can find more on EarthMAP here.
            1. See slides for diagram.
            2. CDI works together to further USGS goals, especially by connecting USGS science priorities with grassroots ideas from practitioners. Community input and lessons learned that are given back to the community contribute further to finetuning projects and furthering USGS priorities. Today's project presentations exemplify the goals of the EarthMAP framework.
        ii. Kevin also talked about the forthcoming USGS Model Catalog. More information here.

1. The goals of the model catalog are to increase discoverability and use of models and link models to relevant information. A USGS Leader's Blog is forthcoming on this project.

c. Tim Quinn
   i. The EarthMAP webinar series on integrated modeling began since the last monthly meeting. USGS employees can find more here. The next webinar will be June 23rd.

2. **Working Group Announcements**
   a. See slides for full details.
   b. All group pages: Collaboration Areas
   c. Risk
      i. June 18th: Presentations of FY19 funded projects.
      ii. August 11-13: Risk annual meeting, which will be held online.
      iii. [9:26 AM] Ludwig, Kristin A
      June 18, 2020 @ 1p ET - Risk Community of Practice Monthly Meeting Final project presentations from the FY19 Risk RFP awardees! June topics include risk related to landslides, coastal change, invasive species, and contaminants. Please see the link below to join on 6/18 and/or sign up for our the Risk Community of Practice to receive future announcements. Sign up here.
   d. Software Development
      i. June 25: Serverless & AWS
   e. Semantic Web
      i. June 11: Discussion of a Forbes article on the Semantic Zoo.
   f. Metadata Reviewers
      i. July 6: Continue discussion of metadata for software & code with Eric Martinez.
   g. Tech Stack
      i. June 11: ESIP infrastructure and response to support the community during the pandemic.
      ii. Last meeting explored CUAHSI HydroShare.
      iii. Tech Dive Webinar Wiki
   h. Data Management
      i. July 13: Collections Management at USGS with Lindsay Powers and Brian Buczowski
   i. Open Innovation
      i. June 18: Tackling the Paperwork Reduction Act with Jeff Parillo and James Sayer.
      ii. **Subscribe to the OI Listserv** to receive OI Community meeting invites
      iii. Check out the **Guide to the Paperwork Reduction Act** in preparation for the next Ignite OI Forum on Thursday, June 18 at 2 PM ET (past meeting videos available on **Stream Channel**).
      iv. **FedGeoDay** is this Thursday and Friday online (Registration free for GOV / MIL / EDU / NGO)
      v. Check out **NASA's Summer of Citizen Science Workshops (#NASACitSci2020)**
      vi. Join the Federal Crowdsourcing and Citizen Science (FedCCS) Community Listserv for meetings and recordings, by sending an email to: **FCPCCS-subscribe-request@listserv.gsa.gov**
      vii. Nicole Herman-Mercer (USGS) - Indigenous Observation Network **"Data Quality from a Community-Based, Water-Quality Monitoring Project in the Yukon River Basin"**
      viii. Citizen Science Association (CSA) Data and Metadata Working Group - **Data Quality Resource Compendium for Citizen and Community Science**
   j. eDNA
      i. First meeting will be next month - a getting-to-know-you meeting. Please sign up for the email list for updates.
   k. Usability
      i. July 15: resource review
      ii. Town Hall Meeting for June 17 is cancelled due to speaker conflict.
   l. Data Visualization
      i. Sophie Hou, Alicia Rhoades, Amy Puls, Ellen Bechtel, and Dionne Zoanni are re-starting the data visualization group.
      ii. July 2: 1.5 hour meeting for kickoff.
      iii. Sign up for the listserv here.

3. **Extending ScienceBase for Disaster Risk Reduction** - Joe Bard, USGS
   a. The Kilauea volcano eruption in 2018 underlines the need for near real-time data updates.
      i. Bard helped create lava flow update maps that would inform decision-making.
      ii. Previous methods for sharing latest data updates was by attaching GIS data to an email - a flawed method.
   b. When you upload GIS data to ScienceBase, web services are automatically created.
      i. Web services are a type of software that facilitates computer to computer interaction over a network. You don't need to download data to access it; and it can be easily accessed problematically.
      ii. Data updates can be automatically propagated through web services, to avoid versioning issues.
      iii. Use of ScienceBase during the Kilauea volcano crisis met unforeseen issues around reliability related to hosting on the USGS server and many simultaneous connections.
   c. This project explores a cloud-based instance of Geoserver on the AWS S3 platform wherein the user can publish geospatial services to this cloud-based server. This method is more resilient to simultaneous connections and takes into account load-balancing and auto-scaling.
      i. Opens the possibility of dedicated Geoserver instances based on a team's needs
      ii. On ScienceBase beta, there is a function to publish data directly to S3.
      iii. The related Python tool is available on Gitlab. Makes downloading data form the internet and posting on a ScienceBase item easy.
      iv. Ex: pulling in data from ASH3D and adding to an SB item.
   d. Next steps
      i. Finalize cloud hosting service deployment and configuration settings.
      ii. Check load balancing and quantify performance.
      iii. Explore setting up multiple Geoserver instances in the cloud.
      iv. Evaluate the load balancing technologies (e.g. Cloudfront).
      v. Ensure all workflows are possible using SB Python library.

4. **Coupling Hydrologic Models with Data Services in an Interoperable Modeling Framework** - Rich McDonald, USGS
   a. Why?
      i. Integrated modeling is an important component of USGS priority plans.
      ii. Goal is to use an existing and mature modeling framework to test a modeling sandbox.
   b. Modeling framework

        i. Frameworks are founded on the idea of component models. Model components encapsulate a set of related functions into a usable form.
        ii. Going through a BMI means that no matter what the underlying language is, the model component it can be made available as a Python component.
    c. To test the CSDMS modeling framework, the team took the PRMS modeling system and broke it down into its 4 reservoirs (surface, soil, groundwater, and streamflow) and wrapped them in a BMI. They then re-coupled them back together. Expectation is that we could couple PRMS with other models.
    d. See recording for demonstration of the tool.
        i. Note the model run-time interaction
        ii. Data services example
        iii. PRMS is in Fortran and we're running it in Python
        iv. Code is available on Gitlab.
    e. Challenges and Takeways
        i. It takes effort to wrap a a model with BMI.
        ii. New coupling opportunities possible with MODFLOW and WEBMOD.
5. **Transforming Biosurveillance by Standardizing and Serving 40 Years of Wildlife Disease Data** - Neil Baertlein, USGS
    a. Over 70% of emerging infectious diseases originate in wildlife.
    b. The National Wildlife Heath Center (NWHC) has been dedicated to wildlife health since 1975.
        i. Biosurveillance the NWHC has been involved in includes: lead poisoning, West Nile Virus, Avian influenza, white-nose syndrome, and SARS-CoV2.
    c. NWHC has become a major data repository for wildlife health data.
        i. WHISPers and LIMS (laboratory information management system)
            1. WHISPers is a portal for biosurveillance data. Events are lab verified and the portal allows collaboration with various state and federal partners, and some international partners, such as Canada.
    d. Problem
        i. Need to leverage data to inform public, scientists, and decision makers
            1. Data is not FAIR (findable, accessible, interoperable, and reusable)
            2. There are nearly 200 datasets in use
            3. Data is not easy to find
            4. Data exists in various file formats
            5. Limited or no documentations
    e. 5 step process to make data FAIR
        i. Definition: creating a definition. Created a template in which we capture users responsible for data, what the file type is, where they're stored. A data dictionary was also created.
        ii. Classification: provide meaning and context for data. Classifies relationships with other datasets, other databases, and identifies inconsistencies in data.
        iii. Prioritization: identify high-priority datasets. High-priority datasets are ones that we need to continue to use down the road or are high-impact. Non-priority datasets can be archived.
        iv. Cleansing: Next step for high-priority datasets. Includes fixing data errors and standardizing data.
        v. Migrating: map and migrate the cleansed data.
    f. How?
        i. Dedicated staff - hired 2 student service contractors
        ii. Conducted interviews with lab techs, scientists, and PIs
        iii. Documented datasets
        iv. Organized documentation
        v. Began cleansing data
    g. Where We are
        i. 130 datasets ready for archiving and cleansing
    h. Challenges
        i. Training of staff
        ii. Work is labor intensive
        iii. No documentation available for some datasets
        iv. Databases build with limited knowledge of database design
        v. Variation between lab and individuals
    i. Takeaways
        i. Staff have been great to work with
        ii. Data collectors need to think through data collection
            1. Be intentional with data collection process - is it FAIR? Are my methods standardized? How is the data collected now and how will it be collected in the future?
        iii. Documentation is important
            1. Documenting the process and management of data collection and compilation.
        iv. Data migration needs dedicated resources
    j. Next Steps
        i. Finish those 200 datasets, focusing on a few migrating a few key datasets first.

Questions and Answers

1. Q: I'm wondering how the review and approval works for this use case. I'm guessing you are using a "provisional" release, but how do you actually make the data public? I thought that was controlled by the ScienceBase team.
    a. Joe Bard: Provisional release that can be made public.
    b. Drew Ignizio: yes that would be a matter for the PIs to work through on how the data could be released. But we can accommodate as needed in terms of item permissions, etc.
2. What kind of data QA and how documented?
    a. Joe Bard: As new data sources come in, the data is revised. A continual process of refining the data so that the long-term data is QA/QC as much as it can be.
3. This is very cool, who pays for the S3 bucket / services? Is that covered by ScienceBase?

    a. Dell Long: Right now the costs are paid by ScienceBase. As we flesh this out and create new instances of Geoservers, we will work on the cost model for those instances.
    b. Drew Ignizio: The ScienceBase team assumed the cost in this case, but this approach would allow us to break these out more so than we have been able to before.
4. Joe are you coordinating with Peter Cervelli on this project?
5. Once the BMI is built, what kind of updates are needed when the original model has new updates? Is there maintenance involved?
    a. The BMIs themselves and the interoperability layered to produce the Python components are all set up in GitHub, so that when you update your code, you get a test to see if the testing functions work, then it sort of dynamically also uploads to the Python component. Nice workflow that CSDMS has developed to keep track of everything and keep everything updated.
6. Would using the Pangeo framework make sense for this?
    a. Absolutely. if there was a data proximate workspace you could use. Could also import from CSDMS to Pangeo.
    b. CSDMS framework is built on Pangeo framework as well.
7. Interested in what data standard you're using? Or are you creating your own?

    a. Depends on what kind of data we're dealing with. Location data uses FIPS, ISO 3166. Species uses ITIS. We may need to create custom standards further down the line.
8. What are your plans for building the database in the future? I am sure you don't want to go through the cleaning process for new datasets or contributors.

    a. Focusing on building WHISPers system and getting a new LIMS system. Most of our diagnostic work is captured outside of a LIMS. We've ended up with many spreadsheets and databases. Wanting to get a more robust platform, i.e. WHISPers.
9. Following on Abigail's question, if not a single standard, how do you ensure at least a minimum interoperability with diverse data?

    a. From Neil after the meeting: A single standard for individual variables would be applied to data for migration to a unified data system. If applicable, widely recognized terminology standards can be applied across data sets (like ISO-3166, ITIS). Terminology standards unique to NWHC may include data such as NWHC contacts (e.g., Is Jennifer Davis = Jenny Davis?) or result interpretation (e.g. Pos, Positive, P, Yes, etc.).
10. In the past, WHISPERS only presents aggregated data for download. Are you changing that?

    a. Leann White: It presents aggregated data because that is what our partners are comfortable with sharing. So it likely will not change in the future.
    b. Follow up question: Then how can you call it FAIR data?
        i. Note from Leslie, Maybe team can address: on the FORCE11 FAIR site there are different levels of FAIR defined, because not all data are able to be completely open, for example "Level 3. The Data Elements themselves in the Data Objects are 'technically' also FAIR, but not fully Open Access and not Reusable without restrictions (for instance Patient data or Proprietary data)." (from: https://www.force11.org/fairprinciples)
    c. From Neil after the meeting: As LeAnn alluded to, the WHISPers architecture only captures summarized/aggregated data as this was most suitable for our collaborators (typically federal, state, and tribal entities). Currently, the home page export functionality in WHISPers aggregates this further to provide 1 record per event. It is our intention to eventually expand the export capability to include event details for multiple events, but are currently limited by available resources to develop that capability. Being that WHISPers is a 3-tiered system, there is a web services layer. We are still in active development, but an API will soon be available for WHISPers users to access the data by developing their own tools and data extractions. Instructions for its use should be coming out in the upcoming months.