CDI Monthly Meeting 20200408

April 8, 2020: CDI Projects - Climate scenarios toolbox, Cloud computing at streamgages, Integrating eDNA and non-indigenous species 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

Recordings and slides are available to CDI Members approximately 24 hours after the completion of the meeting.

These are the public slides. Log in as a CDI member to view ALL of the meeting resources, including recording.
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 - Virtual work and collaboration

11:20 am Collaboration Area Announcements

11:30 am **Open-source and open-workflow Climate Scenarios Toolbox for adaptation planning** - Aparna Bamzai, USGS

11:45 am **Develop Cloud Computing Capability at Streamgages using Amazon Web Services GreenGrass IoT Framework for Camera Image Velocity Gaging** - Frank Engel, USGS

12:00 pm **Establishing standards and integrating environmental DNA (eDNA) data into the USGS Nonindigenous Aquatic Species database** - Jason Ferrante, USGS

12:30 pm *Adjourn*

Highlights

1. Remote meetings resource: https://about.gitlab.com/company/culture/all-remote/meetings/
    a. Since the last call, we have added more security to our Zoom meetings, adding a password to Zoom calls.
    b. There has been an increase in virtual meeting attendees in the last couple weeks.
2. ESIP Collaboration Areas Highlights Webinar on April 22: https://www.esipfed.org/webinars
3. To join any of the CDI collaboration areas, see https://my.usgs.gov/confluence/x/JaapJg

Notes

1. **Kevin Gallagher**
    a. The CDI is more important than ever in maintaining connection and communication.
    b. Virtual collaboration
        i. Do you have a "virtual water cooler"? Microsoft Teams and the CDI wiki are possible places for these kinds of conversations.
        ii. Share notes and highlights after virtual meetings so others can benefit from your activity. CDI collaboration areas are great for these kinds of notes.
        iii. Share your tips, tricks and ideas for working virtually with the CDI.
2. **Tim Quinn**
    a. CDI is "collaboration on a massive scale", and very important in this time.
    b. EarthMAP
        i. Feedback from CDI has been passed onto the EarthMAP team and allowed the team to identify what aspects of EarthMAP are most exciting and most confusing.

3. **EarthMAP update from Sky Bristol**
    a. Blog post (link for USGS employees)
    b. Intranet page (link for USGS employees)
    c. MS Team (link for USGS employees)
4. **Announcements from Collaboration Areas (*see slides for full details*)**
    a. Usability
        i. Town hall meeting: April 15th, Testing Usability with Users
        ii. Resource review: May 20, Usability and Building Trust
        iii. Have usability questions? Post them at: https://my.usgs.gov/confluence/x/yZCpJg Interested in being a usability tester? Sign up at: https://my.usgs.gov/confluence/x/ZMmpJg Want to stay in touch? Join Listserv via: https://listserv.usgs.gov/mailman/listinfo/cdi-usability
    b. Semantic Web
        i. Paper discussion, April 9
        ii. "Best Practices for Implementing FAIR Vocabularies and Ontologies on the Web" by Daniel Garijo and María Poveda Villalón: https://arxiv.org/pdf/2003.13084.pdf
        iii. 2020 - 2021 SWWG Meetings
    c. Metadata Reviewers
        i. Last meeting, April 6
        ii. Next meeting, May 4
        iii. Meetings of the Metadata Reviewers Community
    d. Tech Stack
        i. Next meeting, April 9, "Unidata Science Gateway" https://science-gateway.unidata.ucar.edu/ http://wiki.esipfed.org/index.php/Interoperability_and_Technology/Tech_Dive_Webinar_Series#9_April_2020:_.22Unidata_Science_Gateway.22_Julien_Chastang
    e. Data Management
        i. Next event, April 13, Upcoming changes to the Science Data Catalog with Lisa Zolly
        ii. Last event, March 9, Value Propositions with Science Gateways
    f. Software Development
        i. Next event, April 23
    g. Open Innovation
        i. April 17, COVID-19 Open Innovation Efforts: https://my.usgs.gov/confluence/display/cdi/COVID-19+Open+Innovation+Efforts
        ii. The Opportunity Project (TOP) – Earth Sprint (Problem Statement Due Friday, April 10 – email me at sophiabliu@usgs.gov if you would like to help): https://opportunity.census.gov/sprints/
        iii. TOP Earth Sprint Roundtable Notes: https://docs.google.com/document/d/1UE8cMjDL2_aJpwShHv7gn1hThrvpTQC9K5uBQ0zXadA/edit?usp=sharing
        iv. FEMA PrepTalk on "Crowdsourcing & Citizen Science as Force Multipliers for Emergency Management" by Sophia Liu: https://www.fema.gov/preptalks
        v. Citizen Science Association Webinar: https://www.citizenscience.org/events/webinars
        vi. Citizen Science Association COVID-19 Resources: https://www.citizenscience.org/covid-19
    h. Risk
        i. Next meeting, April 16, Human-Centered Design Thinking with Impact360 Alliance (part 3)
        ii. Risk Community of Practice Community Survey: https://tinyurl.com/vp3xla4
        iii. Risk page: https://listserv.usgs.gov/mailman/listinfo/cdi-risk
    i. ICEMM
        i. Annual meeting was March 17-18; ICEMM CDI website has all recordings here: Interagency Collaborative for Environmental Modeling and Monitoring
5. **Open-source and open-workflow Climate Scenarios Toolbox for adaptation planning: Aparna Bamzai-Dodson**
    a. Link to website: https://www.earthdatascience.org/cst/index.html
    b. Scenario planning - a way to consider the range of possible outcomes; 3-5 plausible divergent scenarios. Managers and scientists can use this information for adaptation strategies.
    c. The Climate Scenarios Toolbox is attempting to take the pain out of working with climate data
    d. The Toolbox is open and usable, allowing other users to contribute open code. The Toolbox hopes to do the following:
        i. lower the barrier to entry
        ii. automate common tasks
        iii. reduce the potential for errors
        iv. empower a larger user community
    e. The link above includes a getting started guide for the Toolbox.
        i. There is extra support for the National Park Service, as NPS was a partner for this project.
    f. Engaging CDI
        i. Install and use the Toolbox
        ii. Provide feedback on issues/features
        iii. Contribute to the package
6. **Develop Cloud Computing Capability at Streamgages using Amazon Web Services GreenGrass IoT Framework for Camera Image Velocity Gaging: Frank Engel**
    a. Gaging (measuring water quanitity)
        i. Sometimes we can't measure
            1. flashy regimes
            2. unsafe/unreachable
            3. indirect (post flood) methods aren't cheap
    b. How do we get past these issues?
        i. non-contact methods
            1. imagery combined with software - gets complicated; requires training; and some subjectivity is involved
            2. want to automate this process and take some of the pain out of it
    c. CHS/AWS IoT Cloud Processing Goal
        i. First required building a cloud infrastructure
        ii. Successes
            1. Auto-provisioning to the cloud
            2. MQTT Schema (in progress)
            3. Generating global actions (see something, do something)

4. Initial time-lapse video Lambdas (for SSTL)

iii. Lessons learned
1. Cloud computing knowledge takes a lot of work to acquire
2. A lot of hands in the cookie jar
    a. In the short term, it can be difficult to sort through the differing needs of stakeholders
3. Time!

7. **Establishing standards and integrating environmental DNA (eDNA) data into the USGS Nonindigenous Aquatic Species database: Jason Ferrante**
    a. eDNA is genetic material released by an organism into its environment (skin, blood, saliva, feces into surrounding air, water, soil, etc.).
    b. Why add a data layer to the NAS database specifically for eDNA?
        i. Want to combine the traditional specimen sightings and eDNA detections for a more complete distribution records to improve response time to new invasions.
    c. Aquatic invasive species data specifically are species of interest
    d. Need to establish strong community standards that will allow high-quality data that can be validated.
    e. What did we do?
        i. Experimental Standards
            1. eDNA literature review
            2. establish standard criteria regarding sampling design and collection, laboratory processing, and data analysis
        ii. Stakeholder Backing
            1. Reviewing criteria among stakeholders
            2. Input by eDNA community of practice
                a. pre-submission form to vet data before it is included
            3. Teleconferences to gain consensus (ongoing process)
            4. Produce a white paper
        iii. Integration into NAS
            1. Community standards
            2. Web submission form/template
            3. Prototype web viewer (map)
        iv. Pre-submission survey
            1. Two blocks of questions, some that will require a "yes" in order to move forward, some that will vet the data better
        v. Quick start guide for the database became a need during the feedback process
        vi. See slides or recording for mock-up of map view
        vii. Challenges
            1. Expected challenges:
                a. Getting to consensus on submission form
                b. Staying organized and keeping lines of communication open
            2. Meeting the needs of managers and researchers (getting feedback)
                a. town hall style meetings to present ideas and garner feedback
                b. if you're interested, it will be Monday, April 13 - contact West Daniel if you'd like to attend
            3. Take aways/follow up
                a. Networking is very important. Use existing infrastructures (such as CDI!); Teams is also working very well
                b. Within the CDI group, many are looking for help developing new tools which use eDNA data. Working on a manuscript that provides insight about the process

Questions

1. Based on recent USGS guidance, will this call be moved off of Zoom to Teams?
    a. Leslie: We are testing external participation now and will keep the CDI informed of tech choice. Anyone interested in testing or discussing further, get in touch with me!
2. Would you be able to see if you have any "surprising" new users as a result of the tool, or do you have ideas of how to learn if you do?
    a. The package is not officially released on GitHub but we are working on it, and there will be a publication in Journal of Open Source Software. Hope we see people fork code, can incorporate user modifications back into the main branch. Hoping the user community picks this up and makes it into a bigger and better toolbox.
3. What was your process for identifying your main users and their needs?
    a. Our center is part of a USGS network meant to work with natural resource management partners to help understand climate adaptation science. We have worked quite extensively with Fish and Wildlife Service and National Park Service over 7 years on supporting their science needs. We saw inefficiencies in the workflow, and commonalities existed in the data requested, but we were starting from the ground up whenever we had to provide it. Anyone doing research across continental US can use it, so hoping to expand past initial stakeholders.
4. Is the climate reanalysis data included so that historical climate (and weather) can be downloaded too?
    a. Yes, it can do historical and future comparisons.
5. Could you say more about the Journal of Open Source Software?
    a. Open review process; way to release new tools that are allowing people access to new data. Write-ups for publication are pretty short; description of the package/software, what problem it solves, and how you are contributing (not doing something that is already done). Goes through peer review process and is released as a publication. Nice way to get the tool out to a broader community.
6. Is it possible to include a vignette that uses the software in a JOSS publication?
    a. Will find that out and get back to you.
7. Comment from sli.do: Thank you for the clear explanation of what a Lambda function is! Hearing about them everywhere these days....
    a. Thanks!
8. Is "edge computing" the same as "data proximate analysis"? Can you explain it a bit more?
    a. Edge computing is when you place some sort of computing power at the “edge” of your network. Really common examples include your smartphone, or house thermostat, which are at their core internet connected computers. They contain programs which compute metrics, perform tasks, etc. while communicating (near constantly) with a cloud server. For my use case, edge computing is putting a small computer at the bank of a stream to collect environmental observations. Harry House (CHS Lead), answers below that data proximate analysis would be a subset of edge computing.
    b. hrhouse : I am assuming the working definition of "data proximate analysis" just means you run the analysis of the data "near" where the data sits. For instance, running a cloud-based application against data that resides on-premise would not be data proximate analysis.

Edge computing, which means you are running jobs "next: to devices in-situ where the data is collected would be "data proximate analysis". In other words, edge computing is a subset of data proximate analysis.

9. Next time we have a workshop in person, could we serve raspberry pie at the data blast?
    a. Engel: I love pi(e)
10. Can a user trigger the camera remotely? Or, based on water surface elevation?
    a. Yes, the user can currently trigger the camera in 3 ways:

        1. By a schedule (this is the standard approach—capture videos on some set interval)

        2. Using an internet connection, by checking a NWIS web streamgage parameter against a user-defined threshold

        3. Using a relay kit we distribute with a typical Raspberry Pi setup

        4. We are working on using MQTT to do control and command so that the Raspberry Pi may be triggered from the cloud, or from other IoT sensors in the network.
11. How do you handle security with IoT? Are these RaspberryPIs protected since they have AWS credentials to upload the data? Is there a DMZ for video uploads?
    a. We are developing a stig for RaspberryPIs. We enforce how people set up modems in the field so it is unreachable from the outside world.
    b. We address logical and physical security in several ways. The most important thing we likely do is provide training. In our typical setups, we are using cell modems for internet connectivity. We prescribe that these cell modems be setup in a way which only lets internet traffic from “safe” IPs even gain access. We also advise that all passwords for the system be strong. Finally, we want to make sure that the other connected devices (e.g., IP cameras) are not exposed. Physical security can be trickier. People who plan to vandalize a site will do so, pretty much no matter what we do. But, we disable hardware Bluetooth and wifi at sites, so that nobody should be able to spoof a connection. We also make sure that our hardlines are well secured. For example, we ensure that the network jack for a camera is not exposed in a way that someone could (easily) gain access and connect to the local network that way.
12. Are you using raspberri Pi camera?
    a. Yes, we have included support to work with the various raspicam-complaint models out there, like the Pi NoIR v2 and others.
13. Can you elaborate a bit on Infrastructure as Code practices you're using for this IoT project? You mentioned that you had to create infrastructure first
    a. In general, we are using ThingLogix Foundry as our cloud templating framework. Foundry is touted as a “no code platform in the serverless cloud” which is supposed to enable easy solution development without the need to actually code them. However, we are doing things that the Foundry core hasn’t implemented, so it requires us to work closely with ThingLogix to overcome these needs
    b. hrhouse : The Infrastructure as code environment just means that the environment is implemented typically via Cloud Formation scripts which can be easily be replicated as needed. This is a key principle in how the CHS environment is architected and presented. That is the only way we can support the environment at scale, and is really a best practice. The sensor processing system, like all other systems in CHS, are required to adhere to this basic design principle. So the good news is we can support such systems at scale, and the customer can also scale their own work as well. The bad news is as Frank mentioned, it does take a new level of skill set to understand what needs to be done to work within that framework. The CHS program provides a base environment, but the customers also have an obligation to build out systems without those boundaries in this manner. We recognize how this limits adoption, and are working to bring on some support engineers who can work with customers to help them.
14. How many IoT RPis cameras are there. Do they provide constant video feed?
    a. There are 3 Cameras in the CHS AWS environment at this time. There are at least 20 Raspberry Pi “camera gages” in various locations around the country. Once we finalize our Raspberry Pi scripts, we will offer an upgrade path to the users of other camera cages in the network.
15. great presentation of a tool that I am looking forward to seeing being more widely available for implementation. Thankyou
    a. Thanks! This has been an amazing road so far. I’m truly excited to see the final product, and hopefully broader adoption in the USGS (as well as the Open Source IoT) communities.
16. I think we need a “Lending Lab of Low-Cost Instruments and Sensors” as an off shoot of what the USGS HIF provides but bringing together these low-cost sometimes disruptive innovations leveraging IOT and other sensors that we can maybe push out to the public for crowdsourcing or citizen science projects.
    a. This is a great idea. Ultimately, our group would like to see that our sensor “network” is mature enough that we can start to distribute it through the HIF in a way that is as “turn key” as possible.
17. I am assuming the working definition of "data proximate analysis" just means you run the analysis of the data "near" where the data sits. For instance, running a cloud-based application against data that resides on-premise would not be data proximate analysis. Edge computing, which means you are running jobs "next: to devices in-situ where the data is collected would be "data proximate analysis". In other words, edge computing is a subset of data proximate analysis.
    a. Engel: Thanks for the clarification Harry!
18. Is that RPi stig available anywhere?
    a. Currently not. I will look into how we can expand the development of the STIG, and whether/how we might share that out.
19. What happens to the video stream after processing? Is it archived somewhere?
    a. For now, all video and derivative products are being saved in an S3 bucket. There are interesting, and challenging discussions about what items should be archived in a cloud processing framework, especially in light of the fact that bandwidth is often an issue. For example, if we are doing processing of video “at the edge” and sending derivitave products to the cloud (e.g., image frames, computed values from sensors, image processing results) rather than the original video—do we need to still archive the video? How would we accomplish that in low or spotty bandwidth locations?
    b. hrhouse: You can do whatever you want with the video stream artifacts. That is up to the owner of the system.
20. Can you comment about any "disagreements" that came up on the submission form when you got input from your community?
    a. Earliest iteration was just a CSV file, and more work would be done to vet data. Idea for pre-submission survey came up as input from the community. Lots of conversations about controls, that controls were in place, making sure we had questions that vetted the assay that was being run, making sure people were taking multiple samples from the field. Wanted to be as inclusive as possible, while maintaining a high level of quality.
21. Is the eDNA data being used to validate/reinforce other species detection/occurrence data in NAS?
    a. Not specifically, but can work to the ability to do that. NAS does a lot of work to vet photos/data that come in.
22. can you talk a bit more about spatial controls? links to NHD?
    a. This data layer is not going to be linked to anything, but this is one of the types of areas that might help to inform broader understanding of species distribution. We are interested in ways to pair eDNA with covariates.
23. Have you looked at how your community standard will translate to the biological data standard: Darwin Core?
    a. Yes, they are similar. There's going to be a lot of overlap, and would like to make it overlap as much as possible.

Other comments from the chat

From Sophia Liu : I think we need a “Lending Lab of Low-Cost Instruments and Sensors” as an off shoot of what the USGS HIF provides but bringing together these low-cost sometimes disruptive innovations leveraging IOT and other sensors that we can maybe push out to the public for crowdsourcing or citizen science projects.

From Abby Benson : Jason you might consider looking at the Hydrolink Tool do link the eDNA occurrences to the NHD.

From Jake Weltzin : check out the hydro-link tool for snapping a sample location to NHD(+) network