2021 Meeting Notes - Soil ontology and informatics - ESIP Cluster

Cluster wiki page: https://wiki.esipfed.org/Soil_Ontologies_and_Informatics
Mailing list: http://lists.esipfed.org/mailman/listinfo/esip-soil-informatics
Cluster proposal: ESIP Soil Ontology and Informatics Clustering proposal
YouTube channel: https://www.youtube.com/playlist?list=PL8X9E6I5_i8iJW_fzrdsUBR6cBs3LT5Tz
Current soil data landscape: Current soil data landscape

Call in information [defunct]

Roll call here

https://docs.google.com/spreadsheets/d/1k5-3S1VTPbbeVEIXsyvQbEKG2cRQnvE2jfW__zN_pe8/edit?usp=sharing - If you are new (this year), please scroll down and add a brief introduction.

Table of Contents

Upcoming agenda and notes

Previous Agenda and Notes

13 January 2021 1700 UTC

1. How are we going to do this? How do we get the information we are wanting to model into a more ...malleable form? (not in a report or between your ears)
2. How do we want to construct/iterate on an ontology?
    a. Draft/revise
        i. One person drafts an ontology and sends it around for revisions
            1. Git?
    b. Committee construction
        i. Jamboard or similar and small breakout rooms: https://jamboard.google.com/d/1IKw68k-CTQVecbIL2fgCps3rC1UXT_fB64v562XR0rs/edit?usp=sharing
    c. ????
3. Other future topics this year

NOTES
Spreadsheet with terms/use-case/competency-questions: Soil properties and competency

- What terms?
    - Bulk density
    - Organic carbon
    - Ksat
- Use case examples
    - Link to specific questions
- Competency questions?
    - Discriminate between sieved vs unsieved and oven-dried vs field-moist bulk density?
    - How was the volume determined for bulk density?
    - What was the sample protocol and is it comparable?
    - When and where and how deep sample was the sample taken?
    - Relevant land cover classifications or ground type?
    - Frozen/permafrost
- Tools for brainstorming?
    - Miro-board (https://miro.com/): the free account will support a small number of concurrent users (have tried up to 4-ish concurrent users) collaborating on a shared miro-board (=white board), has commenting functions. (shoot an email to bwee@massiveconnections.com if you have questions)
    - Jamboard or similar and small breakout rooms: https://jamboard.google.com/d/1IKw68k-CTQVecbIL2fgCps3rC1UXT_fB64v562XR0rs/edit?usp=sharing
    -

ENVO example:
http://www.ontobee.org/ontology/ENVO?iri=http://purl.obolibrary.org/obo/ENVO_00000134
Modeling aspects of methods
CI/detection
Scales of derived data products

OTHER NOTES

1. How do we continue with creating the Ontology?
    a. Are there a list of properties or concepts similar to bulk density that we can iterate through
        i. A list of terms and a list of brainstorming tools that people feel comfortable with
            1. Word map then git repository?
            2. Table?
2. How many terms do we want to start with
    a. For GLOSIS we had huge property lists
    b. The idea is that we have several different areas/groups of terms

        i. People can jump in and out as they have time
        ii. Can go where they feel most comfortable
    c. Bulk density and carbon fraction could be good starting points
        i. KSAT - hydrology terms (Tanya)
            1. Has reach across several fields
    d. 20-50 related terms is a good place to start (Pier)
        i. Might not need more than the three suggested (bulk density, carbon fraction, and KSAT) because of their richness
            1. For example bulk density has 10 or more other terms (sieved or unsieved, etc.)
            2. Extends in different ways (methodology, etc.)
            3. What stuff around bulk density is important
                a. The soil itself or methodology
                    i. Methodology as it is important to interpret data
3. Competency questions
    a. With this ontology we should be able to achieve ??
        i. Small hierarchy of equipment, of methodology itself, etc.
            1. Allow other people to fill in the broader definitions
    b. Define with these questions to aid with the development of the ontology
        i. Overall people agree with Pier
4. Next Step
    a. How about the sensitivity of different methods/ontology on modeling aspects?
        i. Overall people liked this idea but it is a large rabbit hole to jump down right now
        ii. At some point there would be ranges of certainty (the minimum we can capture)
            1. Scales for modeling
                a. Can lead to different uncertainties
                b. Think about uncertainties as we think about the methods
        iii. Do not necessarily have to build the ontology for models
            1. Provide metadata
                a. This instance is on this scale
                b. Modelers can use this
                c. If you use consistent metadata then it will help modelers solve their own competency questions
            2. As long as we include things like detection limits, whether it was a grid, etc. this will satisfy = just associate the specific metadata included for these specific instances (Pier)
5. Competency questions
    a. Definition - the ontology should be competent of answering the decided query
        i. Want the ontology to be capable of achieving X
    b. Being about to distinguish between sieved and unsieved and between Oven dried and field dried, moist or dense

c. What is the extraction process? Does it conserve the volume within the environment?
    i. How is the material volumetrically removed from the site to sustain it? Volume constraint? How is that maintained?
    ii. We need the process and the properties before and after sample was taken
        1. Depth of sample taken
        2. Where? When? How deep? Horizon?
    iii. Relevant land cover class, permafrost
        1. Permafrost may or not have ice
        2. Process of melting
        3. Whether or not the permafrost melts determines whether there is melting and how the structure changes.
            a. Crosswalk to cryo-semantics
        4. Example: http://purl.obolibrary.org/obo/ENVO_00000134
d. Use case
    i. Derive competency questions from use case?
    ii. Agricultural soil or other soils
        1. Active or retired
e. Ultimate goal of the ontology
    i. Methods or sample?
    ii. If I am putting in a sample, I want someone to be able to determine the relative metadata
        1. Depending on how you use it there are different levels of granularity

6. Dr. TB will send out a long message or survey to other members that couldn't make this meeting so that they can brainstorm.
    a. Jamboard!
    b. During brainstorming, list properties that are important to you
        i. Top term that gets voted up we will start with
        ii. https://docs.google.com/spreadsheets/d/11oY2TIQIVnaNYfG-nXiOfWyQ17uVlReNX4qrdOYddeg/edit?usp=sharing
7. Meeting on the 10th at 12:00PM Eastern
    a. Pier might be able to lead meeting
    b. Brandon is also a maybe
    c. Kathe can run backup with some pointers

~~27 January 2021 2300 UTC~~

ESIP WINTER MEETING => No cluster call

10 February 2021 1700 UTC

Recap
Thank you to everyone who made it onto the cluster call today! We talked about ways forward and got some feedback on next steps. Kathe T-B sketched out some possible future activities that could lead to new tools (or identification of existing tools) for data harmonization. Tanja W made a point of reminding us that we shouldn't leave the qualitative human readable data descriptions behind as we move forward. Fenny vE identified some European data sets that we might consider in addition to CSIRO and NRCS. Margaret O offered to help process EDI data holdings by scraping up the meta-data registered for those data sets.

To get us started I've created a new repository. If **you are familiar with a soil data product please help** us get started by going here and identifying how layer level samples are described in the data: https://github.com/ESIPFed/SOI_harmonization/issues/1 and identify the soil data product itself here: https://github.com/ESIPFed/SOI_harmonization/issues/2 Or you can get in touch with me (Kathe) and I'll be happy to wrangle contributions with a couple of questions. The next step will be to expand and recast the data description tables (see notes from 10-Feb https://docs.google.com/document/d/1_SIm_3e9xpni_64zdAU1DZrFEd8YclerbsihDJq4qwY/edit?usp=sharing) to capture these descriptions. I plan to duplicate the description tables from the soil_data_model_survey repository into this new SOI_harmonizaiton so no need to contribute to those tables directly. And finally I hope that the thesaurus will link to existing ontologies in the future (so INSPIRE folks, watch this space!).

If this activity gets traction we'll talk about how to streamline this process to get more off-line engagement and start working up scripts to harmonize the data products using these tables.

Feel free to reply here if I missed anything. In this recap. As always, detailed notes are in the meeting document https://docs.google.com/document/d/1_SIm_3e9xpni_64zdAU1DZrFEd8YclerbsihDJq4qwY/edit?usp=sharing thanks to our fantastic undergrad scrib this week Lillian Haren! Gary B-C hit record early in the meeting and that can be found here: https://transcripts.gotomeeting.com/#/s/6a14f15fd75e72e0ca6549b99292ae9621a3ad067ab7734845af46ddd7a5c6c7

Our next meeting is **24-Feb at 2300UTC** and we'll recap where things sit with this activity.

-Kathe

Agenda
Lead: Kathe Todd-Brown

- Introductions and roll call
- 5 minutes to read/digest summary below
- Feedback for plan

- Action items for next meeting

Below is a summary of what I'm (Kathe) hearing the group talk around. Please feel free to contribute/improve on any of these topics:

- Problem space
    - Data providers describing their data need highly detailed and precise language (generally big labs, but definite use-case for small PI)
    - Data harmonization need linkages to more general terms to bridge the gap in method documentation in data rescue/recovery
- Possible use cases - merge soil data
    - Align major vocabularies in CSIRO, NRCS, GSP, INSPIRE, LTER, and other major labs.
        i. GACS Core has Global Agricultural Concept Space core concepts
    - Align individual-PI soil data registered on repositories.
- Playing Go-Fish
    - Pick a data 'type' as a group
        i. Space-time location of soil sample
        ii. Bulk density
        iii. Organic carbon fraction and K-sat
    - Pick a data set as individuals
        i. Major Lab data products: NRCS, CSIRO, LUCAS, BioSoil others (global: ISRIC/WoSIS/SOTER?)
        ii. Community data products: ISCN, ISRAD and others
        iii. Individual researcher data: Top hits on DataOne or personal favorites
    - Describe and compare the data set on call (or off-line contributions)
        i. Data model: https://github.com/ESIPFed/soil_data_model_survey/blob/master/data/data_structure.csv
            - Add column description
            - TODO: how to describe long data table
        ii. Meta data (needs control vocabulary added to it?): https://github.com/ESIPFed/soil_data_model_survey/blob/master/data/data_meta.csv
            - Add type for keeping in human descriptions
            - Needs control vocabulary strategy
        iii. Thesaurus: https://github.com/ESIPFed/soil_data_model_survey/blob/master/data/thesaurus.csv
- Map the data models
    - Can we match the data with an existing ontology or do we need to modify something?
    - Add these to the Thesaurus
- Merge the data scripts

    - Can we design general scripts to merge the data descriptions (ala SoDaH and SOC-DRaH)?
- Timelines
    - Feb-April: Go-Fish and Map for 3 data types
    - Summer: Scripted data merge based on maps

Notes -

Apology: Megan Wong (CeRDI) - happy to contribute user cases from our DB as we can (as helpful). DB underpins various platforms (famer group, catchment level focus) for e.g. https://data.soilcrc.com.au/map/about . Will touch base offline w Kathe to see how we can contribute to gitHub docs

Meeting Notes -

- **Group read notes summary above and discussed**
    - **Gary Recorded**
- GSP-GLOSIS/INSPIRE: Different in their models
    - trying to make making as easy as possible
    - have the same codelist basis
    - INSPIRE links: INSPIRE website / knowledge base: https://inspire.ec.europa.eu/ INSPIRE Community Forum / Thematic Clusters: https://themes.jrc.ec.europa.eu/
    - Good practises, implementation examples: https://inspire.ec.europa.eu/inspire-tools
    - INSPIRE Soil: https://inspire.ec.europa.eu/Themes/127/2892
- Other vocabularies that we should look at in South American, African, or Chinese? (Kathe)
    - Look at what was published, we think it is good to keep what we have because we have professionals in the field in the group
    - Vocabulary in French from Christine Le Bas: discussed vocabularies in different languages other than english
    - Gary in the chat brought up LTER which includes controlled vocabulary and AGROVOC (Agronomy Ontology) which includes farming/agricultural vocabulary.
    - Biosoil - Forest/Biosoil Data that Christine Le Bas worked on.
- Keep focus on soil data
    - Not just listing existing vocabularies and merging them
        - Fall down a rabbit hole of which concepts connect and which don't
    - Focus on looking at datasets and seeing what we need to describe them.
    - Proposing Go Fish! Method that has been used in Kathe's lab discussed above
        - Kathe will draft up project management
- Talking about the tables in the summary
    - The group is talking about better description, only original provided definition in the metadata that was taken from the original data
        - Want flushed out descriptions in the data structure table
    - Methods need to be better determined
    - Long vs. Wide data tables, how to describe both?

        - Kathe will add the example of the long data table currently in harmonized tables above
        - Margaret O'Brien has code that can show us visualizations
    - Human descriptions to be added into the metadata. Have it connected to the dataset.
        - Not everything can/should be translate into machine readable at the beginning of the project
    - Everyone would like a constructed workflow on the individual action items they will be working on
        - Everyone appreciated the organization of Kathe's process
    - Not creating a ontology from scratch, trying to recruit existing ontologists
        - Hopefully we can connect/build off wosis or Inspire
        - Glosis had a hard time keeping people on project for long enough to contribute. Engagement is hard.
        - Glosis implementation, don't want to be tied to one implementation.
            - Want to be aligned with glosis
            - Mapping

24 February 2021 2300 UTC

Recap: We had a fantastic group of nine folks call in for this meeting. Kathe walked through the data descriptors on https://github.com/ESIPFed/SOI_harmonization. Mark S pointed out that a public URI should really be linked in the 'thesaurus' portion of the data descriptors (but I'm not quite sure I understood your point correctly here so please feel free to clarify/correct me). There was some back and forth on the specificity of the column names vs method descriptions that will be fun to continue to untangle in later conversations. How sampling and processing fit into these data descriptions and when we need to address them in the harmonization process is still an open issue.

Independent of these data descriptors, we are starting to collect representations of sample location, organic carbon, bulk density, and hydrologic connectivity in the Issues here: https://github.com/ESIPFed/SOI_harmonization/issues We talked fairly extensively about what should be included in the location identification and horizon designation was proposed as a possible addition. If you work with a specific data product please feel free to add your descriptions. After we have 4-6 descriptions collected we'll take a look at specific ontologies associated with individual measurements and see what needs to be extended to describe the data holdings.

Our next meeting is 10 March 2021 at 1700 UTC and will be a community presentation by Emilio Mayorga. Stay tuned for details!

Recording:
https://transcripts.gotomeeting.com/#/s/3bb110233b7b0903261df52df0f482dc45edbea5077b79d6e107bf95313db9e3

Lead: Kathe Todd-Brown

- Recap status of layer location in data sets
- Set up description tables: https://github.com/ESIPFed/SOI_harmonization

Notes:

- Need to determine what we want our focus / next steps to be
- Feedback on 3 data-table descriptors (https://github.com/ESIPFed/SOI_harmonization)
    - When looking at examples of a data point
        - The zero-depth
            - Something that isn't shared across databases
            - Is this the surface or the organic contact
            - Need to specify this
        - The dates themselves
            - Encourage geo first
                - Sorting
                - Complete communication with EU and US
        - More than one special geolocation
            - Throw them all in there
            - We want to show how each dataset represents it
                - Collect the diverse extent of the different types of structures
    - With the current approach, trying to describe aspects of a normalized database
        - Might be good for ISCN3
            - Mark is starting with a wide table and got rid of all the complexities and denormalized everything (SERGO)
                - Usually what the scientist wants
                - Semantic contents of the columns and don't have to worry about the tables
                - Focused on measurements
                    - There are existing ontologies for site and latitude
                    - Thinks we should focus on the nitty gritty
        - Variable location
            - Wide and long tables
            - Don't know if this is the right approach but this is the first stab
    - Start off with bulk density or put everything in?
        - Location, observation time (standards have already been developed so we need to pick with is most robust), depth, sample aggregation, bulk density, organic carbon fraction, KSAT
- Put in the use cases with the above descriptors and a snapshot
- New issue with bulk density, organic carbon fraction, KSAT examples
- Excel was selected because this is what ISCN used
    - Showing eccentricities within a big database
- Goals
    - Would like to pull analogous columns from Kellog and SoDah

        - Get a collection of half a dozen datasets and compare how they use these descriptors
        - Reply to the issue to add them
    - Would like to have the tables generated with the data providers
        - Hesitant to have an ID instead of a label
    - Would like to substitute the variable column for an ontology
    - Dr. TB goal is to point a script to a download location, automate the transfer to rename the columns, change the units, etc
        - Would have to rename the columns if we dropped the tables because columns that are named the same are different across tables
- Mark wants to add a URI column
    - Dr. TB sees this as a second step once we have everything described would be to construct some sort of relational ontology
- Variable column is a local internal standard then we can match it to an external one
- First activity is pulling everything to do with bulk density or organic carbon fraction
- Are you interested in capturing profile descriptions or headons
    - Yes, simon has talked at great lengths or horizon standardization
    - Leave horizon descriptors untouched for now but not only focusing on organic layers
- Is a horizon descriptor part of geo-location for a sample?
    - Depth is considered to be part of it
    - Physiographic location was not included in the ISCN example
        - There could be an argument to include it because not all samples have lat/long
- Need top-depth, bottom-depth?
    - Some have midpoints
        - There's some math you can do to back out of this
- Ontology for soil sampling protocols?
    - Could we do a little of this and a bit of the bulk density, OC percentage, KSAT
    - Setting up a separate table that summarizes all of this?
        - Build a dictionary for that
    - The sampling scheme would be described in the data_meta table
    - Is this a critical first step or can it be left on the table?
- Next call is a presentation

10 March 2021 1700 UTC

Recap:
Hello wordy soils folks! Very brief recap today Emilio Mayorga gave an amazing overview of work he's done with the CZIMEA a project associated with EarthCube and the CZOs. I personally am hoping that he decides to stick around the cluster a bit because it looks like very similar issues to what we are hoping to tackle!

I will add the link to the zoom recording to the meeting notes https://docs.google.com/document/d/1_SIm_3e9xpni_64zdAU1DZrFEd8YclerbsihDJq4qwY/edit?usp=sharing when it becomes available and please go to the meeting notes to see the abstract, notes, and reference links from the talk.

Our next meeting is **24 March 2021 at 2300 UTC**. This will be a working meeting where we will review data models identified on the new git repository https://github.com/ESIPFed/SOI_harmonization/issues

See you then!
-Kathe T-B

“Cross-Site Soil and Microbial Ecology Cyberinfrastructure for the CZIMEA EarthCube Project”
Emilio Mayorga (Applied Physics Laboratory, University of Washington)

Abstract: The CZIMEA (Critical Zone Integrative Microbial Ecology Activity, 2015-2020, https://github.com/BiG-CZ/CZIMEA/) EarthCube Integrative Activities project carried out a cross-site soil sampling and analysis research activity involving many universities affiliated with the 10 Critical Zone Observatories (CZO) across the US. The scientific goal was to gain insights into the differences between soil microbial communities as they vary across ecosystems and with depth along a profile, using a wide range of soil and environmental measurements and both metagenomic and amplicon high-throughput sequencing to analyze nearly 200 unique soil samples.

Cyberinfrastructure and interoperability were important components of this EarthCube project. To facilitate management, integration and dissemination of the cross-site metadata and data generated by CZIMEA, we set a goal to store information using the Observations Data Model 2 (ODM2, http://www.odm2.org) system with the CZ-Admin User Interface (http://odm2.github.io/CZ-Manager/), deployed on the cloud. ODM2 is enriched with cross-linkage to external data systems using universal identifiers that include DOI’s, ORCID’s and IGSN’s, in addition to ODM2-managed controlled vocabularies described online (http://vocabulary.odm2.org). The CZIMEA ODM2 relational database holds extensive, structured sampling and sample metadata, and initial steps were taken to store environmental measurement results as well, including soil moisture, field capacity, texture, carbon and nitrogen content and pH. The plan was to use universal identifiers to link to genomic results stored externally in dedicated systems, particularly the MG-RAST metagenomics analysis server. We also developed the initial elements for storing landscape information derived from continental scale geospatial datasets, including site climatologies, land cover type, soil types, and terrain properties.

I will describe the architecture we developed, progress made and challenges encountered, hopefully illuminating common challenges and opportunities in soil informatics.

Notes

- Recording the presentation to the cloud
- Broad announcements
    - ESIP summer meeting will be virtual July 19th - 23rd
    - Call for sessions will be coming out soon
- Data we're working with/new git repository
    - Collecting new data products and giving example of structure
        - Goal: start to put in existing ontology or going to a place like ENVO(sp?)
        - Interested in bulk density, location data…
        - If you have datasets you would like to contribute, drop a link into the github
            - Datasets should have the doi already
                - Not looking to publish new data
            - Focus first on point data
                - Model data is not focus
            - Lean towards publicly available data
        - What is the structure of your data model?
        - Send an email if github clarification is needed
- Presentation
    - Goal: serve as an interesting case study
    - Emilio Mayorga
        - Research Scientist at UWashington
        - Work primarily around environmental data
            - Ranging from marine to freshwater applications
        - Background in biogeochemistry
    - EarthCube
        - Has been going on for 10 years
        - Is not one system, a community around cyberinfrastructure
        - Aimed to deploy a general cyberinfrastructure
        - Interconnect information across systems
    - Project
        - Objectives
            - Enable Critical Zone to use Earth Cube software
            - Facilitate cross critical zone project comparison
        - 10 Critical Zone sites
            - Focused on microbial ecology drivers and diversity
            - Has produced two integrated papers
        - ODM2
            - ODM1 was focused on time series data and didn't have much sophistication
                - ODM2 added sampling data capabilities
            - Helped build partnerships

        - Appropriate for budget
    - Walkthrough of CZIMEA ODM2 Admin
        - Not public access, requires login information
        - Shows Methods and their types
        - Ability to plot data
        - Not much data to plot, so just showed map interface
    - Older versions of code on the website, not the most recent versions
    - Challenges
        - Group passed limited funding and had turnover, making it difficult to complete
        - Dataset was not hoped for finished project
        - Were at the cusp of exploring linkages across systems
    - Q/A or comments
        - Did you extend any ___ to vocabulary registry
            - No, used in a particular way
        - Given what you know now with the project over, I have a student who wants to publish soil corps data, where would you suggest she do it so that interface can be seen forever?
            - At least in context of US, there is no ideal repository for soil data
            - EDI comes to mind, haven't worked with directly
                - Looking for format presently
            - QUASI is playing a central role in sharing data

Next Meeting is on 24th in working group meetings

24 March 2021 2300 UTC

Recap:

Small but dedicated turnout today! Between the time change, stale links, and conflicting ESIP calendar we only had a few of us call in today. Tanja W brought up the importance of capturing the structure/design of the soil sampling which led to an extensive discussion about landscapes and landforms with David W and Irfad A.

There hasn't been much contribution to the GitHub page: https://github.com/ESIPFed/SOI_harmonization/issues (Thank you Steven E who has contributed!). I worry that we are starting to run the risk of spinning our wheels a little on this project. Not sure that we have a way through this yet but first step is recognizing the issue. Thoughts and feedback here are welcome (either in this email thread or to me directly). I can continue to add more datasets to the repository but if folks are finding this difficult to use we can continue to try something else.

Our next call is at 1700 UTC on April 7th! We'll be hearing from Kathi Schleidt and Tomáš Řezník - GloSIS, INSPIRE, and ISO 28258 more details to follow next week!

No meeting recording today but as always the notes and call in information is on the meeting notes here: https://docs.google.com/document/d/1_SIm_3e9xpni_64zdAU1DZrFEd8YcIerbsihDJq4qwY/edit?usp=sharing

Agenda

- Introductions and agenda additions
    - Roll call and introductions at the bottom
- Summer ESIP meeting
    - Do we want to submit a proposal? Or something else?
- Review data models on https://github.com/ESIPFed/SOI_harmonization/issues
    - Not much on here, why and what needs to change?
- How are we going to map to semantic resources?

NOTES

- Summer ESIP meeting and USGI meeting occurring at the same time
- Review data sets
    - What does this process/roadmap look like?
- Introductions
    - Irfan Ainuddin- Research in creating app that looks at soil data, working with surveying
    - Kathe Todd-Brown- Background in math, but currently works in soil informatics
    - Tanja Williamson- Works mostly with land use
    - Stevan Earl- Information manager, economists by training; wants to work on describing soil
- Change ESIP calendar to match with meeting time
- Sierra dataset has more complete ontologies and has more complete information
    - Go through how the profile collections are structured
    - Has information about items not in SSURGO
        - Working in SSURGO not ideal for working with profile level data
- Structure of common positions information is just as important as soil profile
    - SSURGO has vocabulary for this, but commonly not used
- Are landscape positions standardized?
    - Don't know, but positions of interests are normally uniform
        - Convex/concave
- APQ
    - Shares aspects with KSL language
    - Bulk density shows up at DB (density bulk)
- Do we break it up or describe all data structures as a whole?
    - Soil profile collections may help with that

    - Start with US profile data instead of trying to piecemeal without context
    - Not found a good way to integrate landscape systems
- ISCN did not have hierarchical terminology, which created confusion
- On GitHub, reply to issues by logging in and treat as conversation thread
- Soil survey description
    - Has a hierarchy with glossary and written documentation, but not policed
    - Has landscape levels, 3 dimensional descriptors, and landforms names
        - Can not use a term twice
    - May not be populated
    - Not standardized
        - EX Austrialians have their own terminology with overlaps
- Create a issue to discuss hierarchy of terminology and documented definitions on GitHub
    - Currently, no standard conventions of GitHub issues
- Best way to move forwards in ontology developments?
    - Two parallel projects
        - Mark's project focused on working with SSURGO data points, but doesn't work with data descriptors
        - This project which focuses of descriptions, terminology, and pedon
    - Two projects come together in SSURGO through connections between pedon and soil mapping data
        - Reach out to Mark who is currently working on C horizon data
    - Bridge NRC data from Mark to GloSIS
    - Talk to speakers about upcoming issues
    - Scale relationship between map units and pedon
        - Map units contain many locations, but pedon you can focus more on specific areas
    - What is pedon?
        - Pit size, generally a meter^2
        - In AQP, there is information about the relationships between descriptors and pedons
        - Slash may indicate a hierarchy
    - Meeting in two weeks to discuss GloSIS at an earlier time

7 April 2021 1700 UTC

Recap:

We had an amazing turn out (22 folks logged on at our peak) with many new faces for a presentation by Kathi Schleidt and Tomáš Řezník on GloSIS soil data model. It was a fantastic deep dive into the details behind the GloSIS mandate, various decisions and compromises made along the way, as well as the current status of the project. So much ground was covered during this presentation that I would encourage you to take a look at their outline on the google doc

https://docs.google.com/document/d/1_SIm_3e9xpni_64zdAU1DZrFEd8YclerbsihDJq4qwY/edit?usp=sharing and watch the video here: https://youtu.be/FsfTiCLEKF0 I have a feeling this will not be the only time we discuss GloSIS!

Key reference links from the chat:

- https://github.com/rapw3k/glosis
- https://www.w3.org/TR/vocab-ssn/
- https://www.ogc.org/standards/om

If anyone wanted to follow up directly with Kathi or Tomas ~~I've cc'ed them on this emai~~l [Kathi Schleidt <kathischleidt@gmail.com>; Tomáš Řezník <tomas.reznik@sci.muni.cz>].

Update on the ESIP Calendar, the Zoom room scheduling is not playing nicely with our two timeslots and it's propagated to issues with the ESIP Calendar. Currently looking at possible solutions to smooth over the issues. As always I'll keep the link on the google doc as our master 'truth'.

Agenda:

- Please add your name to the roll call above and if you are new drop an introduction at the bottom of the document
- Goals of the cluster
    - "The mission of Soil Ontology and Informatics Cluster is to connect the soils research community to informatics tools for better research."
    - "We will do this by 1) promoting research exchange between the informatics and soil science community, 2) creating how-to resources for soil data harmonization, and 3) targeted development of semantic resources such as ontologies and other informatics tools for soil science."
- Presentation (See below)

Presentation from Kathi Schleidt and Tomáš Řezník

Soil properties modelling for Global Soil Information System (GloSIS)
Leveraging existing standardized models

- GloSIS introduction, requirements and use cases
- Status quo of soil properties modelling
    - *List of models analysed*
    - *Focus on ISO 28258 and INSPIRE*
- What is "interoperability" under European INSPIRE SDI perspective vs. exchange harmonized parameters following an agreed domain-specific nomenclature?

    - *General INSPIRE*
    - *Emergence of ISO 28258 from INSPIRE*
- GloSIS data model:
    - *Class diagrams detailing relevant spatial classes including observations thereon*
    - *Additional UML structures to encapsulate the explicit requirements ensuing from GloSIS*
    - *Transformation to semantic structure*
- GloSIS data modelling outputs: UML class diagrams, codelists, XML schema, TTLs
- Governance - Data Provision Options
    - *Codelist registries (include cascade down to regional codelists)*
    - *Data licencing*
    - *Web services/Provision tools (potential for utilizing common tools)*
- Data ingestion - import to central GloSIS repository for further analysis
- Open issues:
    - *Raster formats*
    - *Serialization and provision options*
- Conclusions

Question: Does the model us to ask and answer questions about the properties of soil layers? For example, does the model structure data so we can ask of a particular type of soil, “Show we all the places where Project A found in 2010 a sub-soil layer with a porosity between x and y.”

21 April 2021 2300 UTC

Recap:
Greetings ESIP Soils folks!

We had about 6 folks join us for a spirited discussion of the different projects that conversations that have been going on in the cluster over the last several months. We can pat ourselves on the back that we’ve gathered some very impressive international collection of how soil data is being worked with and described across a wide range of projects.

Particularly useful for me was to get my head around skos (term-definition pairs) and rdf (richer relationship descriptions). Just to state the obvious here skos are easier to set up and easier to sell as an initial scope of work, where as a richer rdf/owl is a longer commitment and harder to sell without a demonstrated need (ie do the skos and then you demonstrate the need for the rdf/owl). I’m sure there are still things wrong there but it seems like a good start.

Brandon uncovered a really interesting vocabulary here https://www.soils.org/publications/soils-glossary/# and Emilio M dropped a link to a new data paper on ksat here: https://essd.copernicus.org/articles/13/1593/2021/. Tanja and Irfan reminded folks that AQP might be another really good resource to keep in mind http://ncss-tech.github.io/AQP/ for future works. I'm reaching out to some of the folks I know in sssa to figure out how the soils glossary is currently licensed and maybe collaborate in extending it.

I put in an ESIP proposal to talk about some of the work that a NASEM workshop I was involved with and since it's related to the work done in this cluster I figured we could shamelessly merge the two. I'm still working out what exactly that will look like but it will probably be some combination of a summary of the data products covered in the NASEM listening sessions with the community presentations we've heard here in the cluster. Merged with a bit of a brainstorm on next steps for moving towards better semantic tools for soils.

Some logistics, for various reasons we need to set up two zoom rooms for the different meeting times. I've updated the google doc with the two links and these should line up with the ESIP calendar now. Please be sure to update any calendar reminders you might have set up. For those of you who want calendar invites, the google doc now has links to invites you can download!

As always here is the google doc: https://docs.google.com/document/d/1_SIm_3e9xpni_64zdAU1DZrFEd8YclerbsihDJq4qwY/edit?usp=sharing

Our next presentation will be 1700 UTC on the 5th of May. Raul Palma from Poznan Supercomputing and Networking Center will be speaking a bit more about GLOSIS. See you then!

-Kathe

Agenda:

- Cluster logistics, revisit pending topics
    - https://docs.google.com/document/d/1wyplA7K7IHIY_yPXTydfFYNTwiXyrZY2R_9lu3LjRhk/edit?usp=sharing
    - We had three deliverables, two are relevant right now. How are we doing?
        - Community presentations "we can stand up 12 calls in the first year with at least 20 researchers attending and generate quarterly outputs summarizing these presentations." - yes we do this well :D
        - Ontology/vocabulary work: "To have a successful cluster, we will develop an ontology or identify an existing ontology to describe the top 10% of soil metrics reported in at least 10 distinct and varied data sets each covering >500 sample sites by the end of our first year."

- ESIP summer meeting
    - Session proposals: https://docs.google.com/spreadsheets/d/1T9hKbFIzcSIgMpvzfgTnqmUvf-8nTrZmVbAn1aB9G-Q/edit?usp=sharing
    - Potential SWEET integration? Mark S says yes!
- Sierra Navada algorithms, quantitative methods, and pedology - CR Transects
- https://www.soils.org/publications/soils-glossary/# - Brendon might skos-ify this list

5 May 2021 1700 UTC

Recap:

Thank you Raul for a fantastic overview of the GloSIS ontology. If you missed it you can find the recording here: https://youtu.be/q79p_oA5rtU and Anne D took fantastic notes in the logistics document.

Brief housekeeping: We now have ONE zoom room for both meetings and the easy and convenient calendar generating links might not auto update. Sorry that this information keeps changing, please double check your calendar to make sure it matches the information on the google document https://docs.google.com/document/d/1_SIm_3e9xpni_64zdAU1DZrFEd8YcIerbsihDJq4qwY/edit?usp=sharing

We will be meeting next on 19 May at 2300UTC. If no one comes forward with a different suggestion, on the agenda will be pulling together a soil semantic resource list for the ESIP Summer meeting (and ourselves!).

We are looking for other presentations! If you know of anyone working with soil data who would be interested in presenting please send them my way!

-Kathe

Presentation from Raul Palma - Poznan Supercomputing and Networking Center -

GloSIS Ontology Overview

Abstract/outline: The presentation will provide an overview of the process for transforming the Glosis UML data model into an OWL ontology. The transformation was carried out semi-automatically, following best practices in ontology engineering, reusing whenever possible existing standard ontologies and vocabularies. The presentation will also provide an overview on how the ontology is being used as the underlying model to generate Linked Data from soil

databases, which can then be easily connected to other existing and related datasets in the Linked Open Data Cloud.

Recording: https://youtu.be/q79p_oA5rtU

Agenda:

- Housekeeping
    - Roll call and introductions
    - YouTube channel: https://www.youtube.com/playlist?list=PL8X9E6I5_i8iJW_fzrdsUBR6cBs3LT5Tz
    - Summer meeting session outline
        - 5 minutes - Welcome and purpose
        - 10 - **Interoperability dreams:** FAO GLOSIS overview [Fenny van Edmond]
        - 5 - Break out question brainstorming
        - 10 - **Data access and delivery:** USDA-NRCS data tools [?pending confirmation]
        - 5 - Break out question brainstorming
        - 10 - **Data integration:** New Zealand ag soil data [Megan Wong]
        - 5 - Break out question brainstorming
        - 5 - *Break [moderator curate/synthesize questions]*
        - 25 - Moderated panel discussion - fishbowl
        - 10 - Recap, resources report from the SOI Cluster, and next steps
- Presentation from Raul Palma (recorded)
- Q&A

Meeting Notes:

- Presentation notes:
    - GloSIS: spatial data infrastructure that bring together soil information collected by national institutions
        - Interconnecting 340 soil properties in a model structure
        - Implemented as UML data model, indicating spatial object types and various observations gathered from soil sampling
        - GloSIS defines 9 spatial object types
        - Container classes defined to groups relevant for a specific spatial object type
        - SIEUSOIL is implementing an observatory platform that will provide Open Linked Data
            - Linked data improves accessibility and improves information integration
            - Allows for increased use of data through links and inferencing
            - Linked data defines the principles and technologies that will be used to represent data
            - GloSIS assists in data modeling

- Transform GloSIS data model into ontologies through a semi automatic process through processing (alignment with SOSA/SSN) and defining code lists
- Spatial types and “container class” modeled as OWL classes
- Post processing: container classes removed; properties were transformed and aligned to SOSA/SSN, yielding five different property types
- Container classes were restricted using owl properties
    - First restriction was uniform across subclasses: connected sub-class to associated spatial object
    - Second: represented differently based on result type, which are string values, code list values, and numeric values
        - Code: link auxiliary class with code list values, then model codelist with a class and a concept scheme
            - Each codelist value modeled as an individual and a concept scheme
    - Third: linking observation wit observed property (few cases of link to code lists)
- Spatial features were aligned with the existing ISO/TC 211 Standard ontologies
    - Additional tools were created to introduce ISO features
- Ontology check for consistency allowed for emergence of typos and other simple errors
- Final ontology is published with persistent identifiers and includes five main modules with separate code list module, common module, the main ontology, and can add additional lists if desired
- Linked data publication pipelines implemented to create linked data: take data sources and integrate them into a common ontology using GloSIS
    - Possibility to use existing pipelines but room to customize if desired

- Discussion:
    - There is a way to evaluate you are complying with the model
    - Kathe Todd Brown asked how do you break out method procedure from lab procedure?
        - GloSIS asks what is the associated procedure
        - Different codelists from type of values to type of procedures used to generate values
    - Is there a way to represent the hierarchy of how a sample is drawn and processed or lumped into a general procedure?
        - According to model, way to model methodology and samples in similar ways, but don’t have this yet in UML GloSIS ontology
        - Sample treatment up until now hasn’t been considered yet, what has been considered so far is procedure (ex. pH) - Luis de Sousa
    - GloSIS is a continuous and collaborative project
    - Mark Schildhauer: discussed use cases previously, are the observations integrated yet into the ontology, and how are the subclasses defined as of now? (without definition possible to have confusion)

        - Going in the direction of the principles of linked open data, using W3ID GloSIS ontology, working towards referencing ontology as well as referencing terms
    - Is linework envisioned to be able to be sorted using case specific results from multiple datasets?
        - Once able to represent multiple datasets using GloSIS, will be able to respond to these kinds of queries; need to define most common use cases first
        - De Sousa- now gathering literature sources to backup various procedures referenced; generating eventually a hub that will compile federation (will take time to generate answers)
    - One one hand can take guidelines as basis, but the guidelines are not entirely consistent and include inconsistencies - Tomas Reznik
    - The inventory of inconsistencies can be worked through collaboratively, including through the scientists who used procedures. Europe has more conflicts regarding individual codelists based on country classification that will need to be merged - Fenny van Egmond
    - Referencing to crops but no linking to agricultural ontology systems; may be useful to link to existing FAO vocabulary- Schildhauer
    - Using domain and range quantifiers and other quantifiers? Yes

19 May 2021 2300 UTC

Recap: My apologies for a very late recap of the 19 May meeting. We had a great conversation last meeting with folks talking about the difference between 'A'-assertion and 'T'-terminology ontologies (notes on the google doc). And touched briefly on the session proposed for the summer meeting (remember to registar!).

If you are interested. I'm trying to go through our old notes documents and compile a summary of soil resources that we've talked about. https://docs.google.com/document/d/14hF5Xubd1dny7g2aK2CSReXBzaM1P9pkiP0b_TUZS8g/edit?usp=sharing Still in progress but I got through some of the early notes and started generating a resource list. If anything is glaringly missing please feel free to add to it!

We have a presentation by Kai Blumberg (UArizona) on ENVO and OBO foundry workflow and tools tomorrow 2 June at 1700 UTC. Hope that you can make it!

Agenda:

- Roll call
- Summer meeting session outline
    - 5 minutes - Welcome and purpose

    - 10 - **Interoperability dreams:** FAO GLOSIS overview [Fenny van Edmond]
    - 5 - Break out question brainstorming
    - 10 - **Data access and delivery:** USDA-NRCS data tools [Dylan Beaudette and Andrew Brown]
    - 5 - Break out question brainstorming
    - 10 - **Data integration:** New Zealand ag soil data [Megan Wong]
    - 5 - Break out question brainstorming
    - 5 - *Break [moderator curate/synthesize questions]*
    - 25 - Moderated panel discussion - fishbowl
    - 10 - Recap, resources report from the SOI Cluster, and next steps
- Review previous meeting notes and aggregation of soil semantic resources for ESIP summer meeting
    - List of soil vocabulary lists
    - List of soil ontologies
    - Recap of status of soil representation in OBO/ENVO(?)
    - Status/update
        - GLOSIS
        - NRCS
        - CISRO??
    -

Notes:

- Introductions
- Hackathon
    - Registration mandatory, 1-4 hour commitment per week for four months
    - Challenge/Idea: hackathon agrees on definitions and registers preliminary definitions
    - Aiming for soil researchers looking for datasets, will shift hackathon to end use thinking
    - Restor Effort - shared by Rodrigo Vargas
        - http://www.restor.eco/
    - Common question: there is so much research out there, how do we harmonize all of this?
    - It seems some groups are discussing the same points/discussing rather than searching for solutions
        - “Everybody is buzzing about FAIR, but are people actually reading it front to back?”
            - This paper endorses best practices, a significant nudge in the right direction
                - Idea: start with FAIR, then move on
- Soil changing in the future discussion, started by Bhavna Arora
    - Sounds more like an A box idea, rather than T box
    - Ontology should be built before the knowledge graph; T box is important
        - Science should be able to define, or else ontology won’t be able to

        - T box
            - Terminology
                - Example: Defining amendments
                - Should have a spreadsheet built that has ultimate definitions
                    - Validates agreed upon definitions
                    - Should be linked so it is not lost
                        - Struggles: what is part of methodology vs note/annotation or treatment
                            - When do we spin off a new definition vs when we split off into a new definition?
        - A box
            - Assertions
                - Empirical knowledge, as it builds
                    - No longer considered an ontology, considered a knowledge graph
                    - Hesitancy: there will be multiple hypotheses for understanding soils in systems in an A box format
            - What should the A box be able to answer?
                - Find samples and their locations
                    - Helps if ontology is well defined
- Compilation
    - List of players
        - Current ontology tools
        - What are we not seeing in the field
        - What do we see that we like
    - Next steps
        - Request for best practice to build ontology taking in all of these considerations

2 June 2021 1700 UTC

Presentation from Kai Blumberg - University of Arizona, Woodwell Climate Center - ENVO and OBO foundry workflow and tools https://youtu.be/tyJphUclr1E with slides 20210602_ESIP_SOI_Blumberg.pptx

Recap:
Greetings wordy soil folks!

We had a great presentation from Kai Blumberg Wednesday 2 June on how to fit further development of a soil ontology resource into the current landscape with special attention to how ENVO has incorporated topic specific ontologies. You can see the full presentation here: https://youtu.be/tyJphUclr1E and notes are in the google doc here https://docs.google.com/document/d/1_SIm_3e9xpnj_64zdAU1DZrFEd8YclerbsihDJq4qwY/edit?usp=sharing

As a reminder we are pulling together a list of resources here (https://docs.google.com/document/d/14hF5Xubd1dny7g2aK2CSReXBzaM1P9pkiP0b_TUZS8g/edit?usp=sharing) from earlier cluster presentations for a session at the ESIP summer meeting (https://2021esipsummermeeting.sched.com/info)

Our next meeting will be 16 June at 2300 UTC and we'll be hearing from Cheryl H. Porter (University of Florida) about work on the Agricultural Research Data Network: harmonization of agronomic data

Hope to see you there!
-Kathe

Agenda:

1. Resource development for ESIP summer meeting: https://docs.google.com/document/d/14hF5Xubd1dny7g2aK2CSReXBzaM1P9pkiP0b_TUZS8g/edit?usp=sharing
2. Presentation

Notes:

- Housekeeping - want to compile a 1-2 page of the resources we've aggregated
- We have a presentation by Kai Blumberg (UArizona) on ENVO and OBO foundry workflow and tools
    - FAIR
        - Data and vocabulary should be accessible to people and machines
            - Findable
                - Terms are found and discovered, with a global unique identifier
            - Accessible
                - Behind a paywall vs. open access
                - Availability beyond hard copy
            - Interoperable
                - Metadata using formal but accessible language for knowledge representation
            - Reusability
                - Including more than just labels
        - OBO
            - Open biological and biomedical ontology foundry and library
            - Do not reinvent the wheel
            - Every domain should have people adding relevant inputs
            - Environment Ontology (ENVO)
                - Good example of FAIR
            - Ontology Design Patterns

        - Generate ontology classes from csv files
            - Robot templates
                - Guides through the process of creating new terms
                - Intended for non-ontologists
            - Dead simple ontology design pattern
    - Possibilities for soil-info cluster
        - Build a new ontology from scratch (not recommended)
            - Need a URL server
            - Hard to reuse existing tools
            - Extra work to make interoperable with existing systems
        - Working within/into an existing ecosystem (recommended)
            - Fosters FAIR principle
    - Ontologies that spun off from ENVO
        - Foodon
            - CDN spun off from Foodon
        - AgrO
    - Q&A
        - Do you also have a SOSA/SSN compliant template?
            - Some of these tools could be used to create SOSA like terminology (robot template)
        - How to represent things like sample collection/methodological differences?
            - This framework (the example provided) is not doing that, but OBI does have capability to handle such inputs
            - Would have to register the methodologies as a device
                - Though only one way to handle this
        - How do you measure the uptake of the ontologies that get built?
            - People do not always report
            - Hard to know, actual statistics are hard to know
            - Google Analytics could be a solution

Discussion afterwards about EnvO model and ROBOT tool

Link to Agriportal that may have vocabs/ontos with useful terms relative to Land Use: http://agroportal.lirmm.fr/ontologies

Agmip-- https://agmip.org/
 for reconciling Ag models, but doesn't appear to be any "semantic" approach.

16 June 2021 2300 UTC

Recording: https://youtu.be/tEmmJzmEQ_c
Presentation: 20210616_ESIPSoilOntology_ARDN_intro.pdf

Presentation from Cheryl H. Porter - University of Florida
Agricultural Research Data Network: Harmonization of Agronomic Data

Abstract: Agricultural research data are valuable resources that are often lost after the original research objectives are complete. Numerous datasets have become findable and accessible through data portals such as CGIAR's GARDIAN and the USDA's Ag Data Commons but many of these legacy research datasets are difficult for researchers to reuse due to the diversity of content, vocabularies, units, domains, and schemas. ARDN provides dataset annotation protocols that allow quantitative agronomic data to be machine-interpreted and translated to a common format with a standardized vocabulary. The original raw dataset, often associated with a DOI, are maintained without modification but are linked to the new interoperable data products. Using these data annotation protocols, a core set of interoperable agronomic data from multiple datasets can be discovered, harmonized, and aggregated into new data products for quantitative research including modeling, data analytics, and meta-analyses.

These ARDN protocols expand upon data interoperability standards developed as part of the Agricultural Model Intercomparison and Improvement Project (AgMIP) and are being implemented in Ag Data Commons; GARDIAN Labs as part of the CGIAR's Platform for Big Data; and at the International Fertilizer Development Center (IFDC) as part of their open data efforts.

Notes:

- Pre-presentation discussion:
    - ENVO is the most comprehensive ontology, but has issues on how they treat soil
    - Minimum data principle: You may need to collect some specific data, but while you're out there collect some minimum soil data
        - Soil density and LLI
    - It's important to get as many different disciplines in the room and getting them connected and working together
- Announcements:
    - Join Slack channel for reminders!
- Motivations for ARDN
    - Major gap between potential value of data collected through agricultural experiments and the current value obtained with the data
    - Data isn't typically used beyond its original purpose
    - Many legacy data sets have a lot of value and could have even more value if they were combined across locations, time, genetics, etc.
    - The goal is to bridge the gap between data supply (multiple datasets, research experiments, farm surveys, government and NGO data, etc.) and data demand (modeling , data analytics, meta-analysis, cross-domain analyses, etc.)
        - Challenges: Formats, database schemas, locations, vocabularies, etc.

        - Solutions: Annotate data with standard vocabularies, new translator for conversions, and additional translators for end users
- ARDN History
    - Started with AgMIP to address global problems (ex: food insecurity)
        - Started using multiple agricultural models to view climate change and found that the mean of all of the models is better than any individuals model
    - Worked with CGIARS Big Data Platform to annotate legacy data for early annotation protocols
    - IFCD has now adopted these tools
- ARDN data characteristics:
    - Point based data
        - Can be collected at any scale
    - Agronomic research data
        - Field crop experiments
        - Farm survey data
    - Methods, protocol, software also have impenetaions in other disciplines
    - Specifically interested in plots of data associated with latitude and longitude
        - Can be used with precision data (ideal) or generalizing an entire field
- ARDN annotations
    - Contains core metadata which meets a quality standard, allowing datasets to be discovered and access
    - Sidecar file #1 allows data comets to be discovered in an automated way through the semantic web
        - Dataset variables are mapped to ontologies for meta data
    - Sidecar file #2 presents annotated data using AgMIP translators and is produced using VMapper
        - Not meant to be human readable, instead meant for computer extraction
    - Sidecar file #3 creates a road made for searchable data that can be used in many different way
        - Contains an indexing section, extracted data section in a tab delimited format, and summary statistics
    - Data is yet to be handled sub daily, but its is something that is currently being worked on
- ARDN data sharing workflow
    - Starts with someone wanting to map data
    - They use VMapper to annotate data
        - An app developed to take data from CSV and map it to ICASA variables
        - Unit conversion is taken care by VMapper
        - User can define variables and conversions
        - Allows multi table relational linkages
    - ACEB translators turn the data into a ACEB format file, which is then translated by AgMIP translators into DSSAT input and more

    - The interpreted annotation can be used an ontology mappers to form semantic linkages
    - The Sidecar #3 extraction brings out data subset summary and statistics
- AgMIP data interoperability tools
    - Keeps assumptions differents from the data
    - Designed to work with inconsistent data sources
- AgMIP harmonized data format
    - Uses key-value pairs stored in JSON objects
        - JSON is split into soil layer data, initial conditions, weather, and other information
        - Key is ICASA variable
            - ICASA
                - Data dictionary held in a Google Doc (www.tinyurl.com/icasa-mvl)
                - ICASA terms are developed using a modeling perspective
                - Standards include documentation of environment
                - Not yet able to ontologize and not yet focused for working on microbial data
                - Working on best way to curate
    - Non-relational schema works with irregular data from a variety of sources
    - CGIARS works more with ontology and has developed the agronomy ontology
        - Works closely with CGIARS to ensure compatibility
    - Need a lot of variables for the models that are still not well categorized in the ontologies
    - Will soon work to organize a schema
- Hackathons
    - AgMIP hackathons were held to develop model specific translators and were super effective
    - Crop modelers were paired with programers in order to maximize productivity and create a usable translators within a week
    - 5 multi model development sprints have produced 13 different translators for 13 different models
    - ARDN hackathons are now monthly data sprints (4-5 hours) in an online Zoom format
        - 4-5 hour format works much better than a week long Zoom conference
        - Post COVID, weekly Zoom meetings may continue
    - Everything is open sourced and found on GitHub
        - https://agmip.github.io/ARDN/
        - AgMIP is found on GitHub as well
        - ICASA is being converted into a GitHub repository

30 June 2021 1700 UTC

Coffee chat! No formal agenda. What are you working on?

Current soil data landscape

Notes:

- ESIP summer meeting deadline: July 22nd
- With soil ontologies right now, new model is taking more shape and there is more discussion occurring with how to model procedures (categorical properties and list of chemical properties) and entities more specifically
    - Units and other factors are being confirmed
    - Human readable descriptions are being worked on next
    - Report submitted March 31 (Fenny and Christine)
        - Update on sampling (mainly statistically sampling design) and modeling and overview of data available in European countries (not complete)
- (Christine) working on mapping of soil information, waiting on soil ontologies
    - Uses ISO system (fenny)
- (Fenny) many different countries and languages and issues with translations is an issue that will need to be addressed, default to international standard?
- (Kathe) would be helpful to map international standards of soil sample preparations
    - Harmonized on http://www.fao.org/global-soil-partnership/glosolan/soil-analysis/standard-operating-procedures/en/ (fenny)
        - More information will be made available with new editions (next few years)
- (Stevan) works on a soil project focusing on soil carbon data
    - Post doc Derek working on building databases for SODA CZO's
- (fenny) there is a world soil congress upcoming next year, working group will present on soil data information standards (not distinct to GloSIS)
- (kathe) soil data licensing issues in US soil society, book licensing versus ontologies; how to augment ESIP and hackathons with more soil scientists?
    - (brian) used to be a group focused on biodiversity data that had legal restrictions, might be helpful with insight
    - (sumeet) group is observing start up activity to guess soil health from air (remote sensing), considering stitching together with soil data to help satellite data map it
- (fenny) a project working improving soil organic carbon in topsoil from space, results from workshop will come online soon
- (sumeet) sensor can determine healthy versus unhealthy soil structure
    - (kathe) challenge is soil is deep, and scanning only can measure top easily
- One more meeting before esip

~~14 July 2021 2300 UTC~~

~~Casual chat~~

~~28 July 2021 1700 UTC~~

ESIP summer meeting, no cluster call

11 August 2021 2300 UTC

Purpose: *Recap first year and plan for next year*

1. ESIP Soil Ontology and Informatics Clustering proposal
2. What was done by the numbers:
    a. Number of cluster calls with average number of attendees [95/12 = 2021 and 134/10 = 2020 ~ 10.5 folks per call]
    b. YouTube count: between 15-50 views on 6 recordings
    c. Harmonized datasets: 0
3. What do we want to do?
    a. Calls/presentations
        i. who coordinates
        ii. Who wants to present?
    b. Hackathons
        i. Temporal-geolocation; sampling design; organic carbon fraction; texture; elemental analysis; available nutrients; soil respiration; experimental design; microbial biomass; enzyme activity assays;
    c. Possible future ENVO soil ontology?

Recap:

We revisited some of the metrics for the cluster proposal (ESIP Soil Ontology and Informatics Clustering proposal) and in general it was a very successful first year! We hosted 22 cluster meetings over the past year and had an average of 10.5 folks per meeting, which makes us one of the more active clusters from what I've heard. We have a new YouTube channel with 6 presentations on it and between 15-50 views each, not going to break any records here but I'm quite comfortable with this presence (also points for amusing faces on opening shots). We completely flubbed our harmonization goal this year, which in retrospect was extremely ambitious. But we will fix this!

There was a lot of interest in this call in taking some data (soil pH in particular) and annotating it with several existing ontologies in an effort to 1) teach us poor domain researchers how to do such a thing, and 2) identify the gaps and possible opportunities across these ontologies. Some ontologies that were listed were GLOSIS, ODM, SOSA + SSN, ENVO, and AGRO. I will work up a draft of an announcement and design for a series of working meetings in October with a goal of having this live in September.

I gave a short report-out from the first 'hackathon' that I held this summer to kick around a community-produced soil ontology seed. We came up with

https://github.com/ktoddbrown/CoSoDat which focuses on a heiractical method descriptions within some value-unit-method and input-parameter-function tuples. Bulk density is fairly complete according to the group and dozen or so datasets we tested it against.

Both the ontology mapping and CoSoDat are seed work for a possible ENVO soil slice at some point in the future. Both activities are geared towards developing a core group of collaborative researchers in both soils and informatics who could provide substantive contributions to these semantic resources.

Next steps: Expect a draft in the next week on the soil pH hackathon that will be open for comments. If you have someone you think would be interesting to hear an update on their soil data project, feel free to snag one of the open spots this fall on the meetings document: https://docs.google.com/document/d/1_SIm_3e9xpni_64zdAU1DZrFEd8YclerbsihDJq4qwY/edit?usp=sharing

Best,

-Kathe

25 August 2021 1700 UTC

Flow:

- Cluster announcements:
  - Planning October hackathon pH! BoisterousBeetles_planning
- Presentation from Melisa Cragin

Title: “Preparing for new Federal guidelines on data publishing: Evolving Federal initiatives and the NSF Public Access Repository.”

Abstract: Research data and code are fundamental to scientific review processes, claim verification, and the production of new knowledge. Increasingly, these products are viewed as first class research objects that are integral to community expectations about publishing and public access. Given this, we can anticipate that improving public access to the results of federally-funded research will mean the expansion of reporting by grant recipients to include records about research datasets that underly publications. This will necessitate the selection of a repository and acquisition of Digital Object Identifiers (DOIs) for these data, and submission of a metadata record for into a public records system. Federal requirements on Public Access include, “making copies of articles in peer-reviewed scholarly journals and juried conference papers” within 12 months of publication. The NSF Public Access Repository was designed to meet this aim for NSF awardees, and it is anticipated that the scope will be extended to data resulting from NSF funding. I will give a short presentation on the NSF Public Access Plan and PAR, to be followed by an extended discussion on community readiness for these likely new guidelines. This work is part of an NSF EAGER project, “Amplifying Community Readiness to Increase Public Access to Data.”

Bio: Melissa Cragin is Chief Strategist for Data Initiatives in the Research Data Services division at the San Diego Supercomputer Center (SDSC), at the University of California San Diego (UCSD). Melissa is part of the EarthCube Coordination Office, and the open Storage Network; she is PI for the NSF-funded EAGER project, “EAGER: Amplifying Community Readiness to Increase Public Access to Data.” Prior to joining SDSC, Melissa was the inaugural Executive Director of the Midwest Big Data Hub, based at the National Center for Supercomputing Applications (NCSA) at the University of Illinois at Urbana-Champaign (UIUC). Previously, she was Staff Associate in the Office of the Assistant Director, Directorate of Biological Sciences at the National Science Foundation (NSF), where she also served as an AAAS Science & Technology Policy Fellow.

Notes:

- Community level to data publications & metadata records to Public Access Repository
- How can we better utilize data management plans
- 2013 memo from President that requires data access for peer reviewed publications => implementation plans to NSF (2015) and USDA (2014)
    - 2.0 plans are starting to roll out
    - Research data mandated to be open by agencies w/ focus on long lived access
- NASEM - workshops around open science and data - ethics, conduct, reusable, and longevity, fit for purpose
- FAIR data => continuum and aspirational
    - Make it machine ‘actionable’ (metadata, data, software, etc)
    - Nothing about data quality
    - Unrelated to open science - “as open as possible and as closed as necessary”
- NSF Public Access Repository
    - 2016 special report history
    - Results of publicly funded papers, data needs to be access to version of record within 12 months OR access to publication
    - Metadata must be machine readable
    - Outcome reports from NSF now include publications and other research products (ie data)
- Where is it going
    - Researchers will publish in preferred (FAIR and TRUSTed)
    - Repos maintain sustained long term access w/ persistent identifier and descriptive metadata
    - Metadata schema “DataSite”
    - PIs will be responsible for publication
    - Expected to be live for dataset metadata within a few months
- Purpose of archives are not always unified
    - Linking funding to output/products
    - Integration and harmonization
        - Researcher deep dive for fit for reuse
        - AI-ready and machine actionable

- Conversations around thresholds -> how likely is it for data reuse? Do we really need to save all data sets or is there some threshold?
    - Learning exercises are important but…
- Funding agencies (NIH… possibly USDA) are thinking about how do you build on past data - standardization of data access/deposit
- Soils are special because they are at the intersection
- Updates to SOI cluster and LTAR on this project are forthcoming!

22 September 2021 - 1700 UTC

Purpose: Open Discussion

- What are you working on right now related to soil informatics?
- What are you curious about?

Recap:

About half dozen of us called in for a great chat on our current projects. @Stevan Earl pointed out an upcoming LTER seminar series that looks relevant for soils: https://lternet.edu/stories/2021-2022-lter-webinar-series/ . There is an upcoming soil hackathon that I'm organizing on soil pH measurements that is slowly coming together and it was great to check in with folks on their own specific projects. Our next meeting is in October and we'll be hearing about the Early Career CZ network-of-networks from Bhavna Arora.

20 October 2021 - 2300 UTC

Recap: We had a lively introduction with some excellent leads on current workshops and opportunities in soil-related informatics (see notes document https://docs.google.com/document/d/1_SIm_3e9xpni_64zdAU1DZrFEd8YclerbsihDJq4qwY/edit?usp=sharing).

Bhavna updated us on the Critical Zone Network-of-Networks (see https://youtu.be/BAk50R7K8go) and how they are developing an international community through workshops and seminars centered around critical zone observatory sites. One of their sticky points is data and how to make data sets from one site/center comparable to a second.

Finally we ran away over time talking about how to move forward with a paper, seminar series or hackathon to evaluate the current semantic tools for soil annotations. Watch the slack channel for follow ups on this!

We don't currently have a speaker lined up for next month so if you know of anyone you would like to hear talk on efforts related to soil informatics we have room!

See you next month!

-Kathe T-B

Agenda:

1. Round checkin: What is something interesting you've heard recently about soil informatics? [15 minutes]
2. Presentation by Dr Bhavna Arora - Intro to Early career CZ network-of-networks [45 minutes]
3. Planning session for future soil data hackathons [30 minutes]
    a. Identify purpose
    b. Confirm process
    c. Next step action items

NOTES
Round Checkin:

- 2022 ESIP January Meeting: Call for Sessions DUE November 8th. The session proposal portal for the 2022 Earth Science Information Partners (ESIP) January Meeting (January 18th-21st, 2022 ONLINE) is open now through 11/8. For over 20 years, ESIP meetings have brought together the most innovative thinkers and leaders around Earth observation data, thus forming a community dedicated to making Earth data more discoverable, accessible and useful to researchers, practitioners, policy makers, and the public. See, for example, topics and speakers from the 2021 ESIP July Meeting.
- Sci data con links-
    - Scientific vocabulary link: https://www.scidatacon.org/virtual-2021/sessions/351/
    - Keynote for Sci Data Con (FAIR digital objects to establish a global and interoperable data space) link: https://www.scidatacon.org/virtual-2021/sessions/330/
    - I-ADOPT session at Sci Data Con link: https://www.scidatacon.org/virtual-2021/sessions/368/
- Oceanography observation model github link: https://github.com/CSIRO-enviro-informatics/PUV-ont
- STEPS- Science and Technology for Phosphorus Sustainability project link: https://steps-center.org
- CUAHSI Cyberseminar series link: https://www.cuahsi.org/education/cyberseminars/cuahsi-winter-2021-cyberseminar-series-tools-for-integrating-and-synthesizing-data-from-czos-and-watershed-sites/
- Abstracts for World congress on soil science link: https://22wcss.org/conference/abstracts/

Bhavna's presentation:

- Critical Zone observatories are found throughout the world, in North and South America, Europe, Africa, Asia, and Australia
- Separate methodologies through observatories although area zone shares goals and struggles with the same questions

- Purpose of the group is to bring together early career scientists to tackle and think about questions while utilizing similar methods
- Community motivation in the critical zone observatory network
- How can this cross-network and cross-site knowledge be facilitated?
- Came together and created unresolved questions in CZ (critical zone) science to prepare for AGU workshop
- Breakouts sessions were organised at the workshop and hypotheses/ideas/insights regarding CZ science were noted in the AGU workshop breakouts
    - What kind of data is needed to bridge the gaps
- Launched library website for CZ research
- Funding, availability of data/networks, and data harmonization are the main limitations to CZ research collaborations
- For the 2020 workshop, new volunteers joined cohort and community engagement grew
- Goal of linking up with NSF CUAHSI
- 2020 workshop had 5 groups that separately presented their shared interests and moved towards using shared tools and approaches
- Workshop was found to be very valuable (62% on survey)
- Beyond workshops, some of the goals have been met including launching a special issue released in Earth's Future and WRR
- In the works: forward leaning commentary article commenting on survey responses, new CUAHSI Cyberseminar series (link above), new AGU 2021 workshop and town hall
- Must be data cleaning and harmonization between the different networks and groups
- Center for scientific collaboration and community engagement link: https://www.cscce.org
- TERN announced as first Australian DataONE node (harmonizable data) link: https://www.tern.org.au/tern-announced-as-first-australian-dataone-node/

Planning session:

- Hackathons
    - Previous soil data hackathon group focused on bulk density and may pull together another hackathon on soil pH
        - How to annotate data with ontology
        - How to evaluate ontology as fit for a purpose
        - pH data
- Account for what had been done and utilize this
- Is domain knowledge captured; search and recall and knowledge representation perspective, how do soil scientists view the world; share data that you can put in models or capture knowledge to focus on inferencing at scale; can be both but must be stated if both
- Differing methodologies represents an ontology on its own
- Can review or look at pH methodologies in the workshops; craft this into a report or a paper

17 November 2021 - 1700 UTC

Recap:
Greetings wordy soil people,

Sorry for the delay in this recap. Last week we talked through next year activities for the cluster. We decided to move forward with a seminar-like series paired with one page recaps of each presentation. For the presentations we have a list of 14(!) resources we would like to hear from and I'll start reaching out to folks today to get dates nailed down. The format for the presentations is being discussed here: https://docs.google.com/document/d/12OtoXo6YV4zK9a79VeJK8XE7iEf02hF9Uf-ggQ8YIG4/edit?usp=sharing but will include a new position of facilitator.

Facilitators will host the presentation for that week and work up the initial draft of the 1-page resource summary. If you are willing to do this please let me know or go to the page and sign up for a week that works for your schedule.

Still in the ideation phase is a hackathon-like activity and best-practices manuscript to guide the use/development of these resources by the research community and stakeholders. If you have thoughts on this please move the discussion to the Slack channel https://join.slack.com/t/esip-all/shared_invite/zt-y7pwdwno-IFCN~MELSH0FY_pvGpZK6Q

We will be having a cluster call on December 15th at 1700 UTC. Dr Steven Quiring (Ohio State University) will be delivering a talk titled Building a National Soil Moisture Dataset

Best,
-Kathe TB

Purpose: Plan out the Soil Semantic Exploration Campaign (AKA Boisterous Bettles Hackathon)

- Identify semantic resources related to soils and potential speakers
- Proposed format:
    - Every 2wks invited talks on soil relevant semantic resources: Why was this resource developed? How can it be used?
        - Soil data models
            - GLOSIS
            - OGCs O&M https://www.ogc.org/standards/om
        - Ontologies
            - ENVO
            - Chemical Methods Ontology
            - AGROVOC
            - SOSA and SSN https://www.w3.org/TR/vocab-ssn/
            - ODM2 https://www.sciencedirect.com/science/article/pii/S1364815216300093

        - Infrastructure
            - OBOE (ENVO?) https://www.anaee-france.fr/en/infrastructure-services/modelling-and-data/semantic-reference-framework - Chistine Le Bas
            - Wikidata
            - Science on Schema.org
        - Thesaurus/glossary
            - SSSA - Soil Glossary (US)
            - INRAE Thesaurus (French) https://consultation.vocabulaires-ouverts.inrae.fr/thesaurus-inrae/fr/
            - Aussies ??
            - GEMET (General Multilingual Environmental Thesaurus)?
    - Follow up 1 pg recap on off weeks; intended as SI for publication
    - At end of talks, manuscript submission within 6 months recapping findings
- Timeline:
    - January advertise and finalize speaker schedule
    - February - May: Talks and one-pagers
    - Summer 2022: Draft manuscript “Review of semantic tools for soil data: current status and future needs”
    - Fall 2022: revisions and submission of manuscript

15 December 2021 - 2300 UTC

Presentation: Building a National Soil Moisture Dataset
Who: Steven Quiring (Ohio State University)
Youtube: https://youtu.be/NqDlg20yUcs

Agenda:

- Roll call (please log into RollCalls)
- Update on plan for next year
    - 2022 Meeting Notes - Soil ontology and informatics - ESIP Cluster
- Presentation & Discussion

Steven will lead off-presenter discussions
Metadata by machines (associated with bioportal) - Melissa Cragin
One-page short take away include how talk about soil characterizations? QAQC with measurements, slope positions, depth

Ideas for future topics

Pending scheduling

- ENVO - Pier Luigi Buttigieg
- Andrew Brown - NRCS - Soil data exercise
- Kai Blumberg <kblumberg@email.arizona.edu>
    - https://github.com/EnvironmentOntology/envo/wiki/ENVO-Robot-template-and-merge-workflow workflow in order to facilitate ENVO term requests - 2 June presentation
- Presentation from Nic Jelinski - University of Minnesota - linking public & private soil datasets (example of soil organic carbon / permafrost affected soils in Alaska, USA)
-

Meeting org logistics

Emails addresses for leadership

Brandon Whitehead <WhiteheadB@landcareresearch.co.nz>; Williamson, Tanja N <tnwillia@usgs.gov>; Beaudette, Dylan - NRCS, Sonora, CA <dylan.beaudette@usda.gov>; Kathe Todd-Brown <ktoddbrown@ufl.edu>

How to `SOIl cluster`

1. Try to give 1-3 months lead time to any presenter
2. Update ‘news’ and ‘upcoming activities on the wiki page
3. Email mailing list 1-4 days in advance, if you can make it by the Friday before then we will get a line in the ESIP all-hands weekly mailer
4. Start call, for gotomeeting password contact Kathe or Megan Carter
5. During the call, have someone not presenting/running the call take notes (UF students are currently covering this otherwise ask for volunteers).
6. After a presentation, get permission from the speaker to share the recording. Contact Kathe or Megan Carter to post to YouTube.
7. After each meeting push a 1-2 paragraph recap to the mailing list and slack channel, include recording link if appropriate.
8. Update ‘news’ and ‘upcoming activities on the wiki page.

Introductions

- **Kathe Todd-Brown (UFlorida):** I'm a computational biogeochemist who is interested in large datasets to support soil carbon model development. I'm a professor at the University of Florida and the data coordinator for the International Soil Carbon Network. My interest in this group is to develop formal registered language that I can use to help harmonize data from different sources.
- **Tanja Williamson (USGS)**: Hydrologist-Pedologist. I use soils in most aspects of hydrologic modeling, research on land-use effects, and potential effects of climate change.
- **Gary Berg-Cross, (Ontolog)** Cognitive Psychologist, original cognitive linguistics, long time data and knowledge engineer and conceptual models, work on methodologies to help develop ontology design patterns. Part of the original cryo vocabulary/glacier domain work to harmonize term definitions and EnvO.
- **Margaret O'Brien (UCSB)** Data manager - with the repository of the environmental data initiative (EDI link), LTER Network. organizes data harmonization efforts for EDI. Some experience with ontology (EnvO user). I tend to work mostly with individual scientists rather than observatories.
- **Fenny van Egmond** Soil sensing at ISRIC - World Soil Information and Wageningen Environmental Research. Involved in Global Soil Partnership (GSP) pillar 4 and 5 for building a Global soil information system (GLOSIS), harmonising data exchange (needed for GLOSIS), IUSS Soil Information Standards WG lead, co-lead, GLOSOLAN soil spectroscopy co-lead. Participate in EJP Soil project where we are compiling codelists for soil to aid soil data exchange, connecting systems for INSPIRE and want to use the same ones for GSP.
- **Irfan Ainuddin (Chico)** Soil Scientist. Graduating from Chico State Spring 2021. Looking for a job or PhD program. Interested in utilizing soil information systems for public education in creative ways.
- **Andrew Brown**: Soil Scientist; MLRA Soil Survey office Sonora, CA. Interest in providing technical support with implementation and ingestion of USDA-NRCS / National Cooperative Soil Survey standards for use in new semantic technologies.
- **Jorge S. Mendes de Jesus:** Dev/Op at ISRIC - World Soil Information. Supporting server work for Global soil information system (GLOSIS).
- **Bryan Carlson** (LTAR): USDA-ARS Ecoinformaticst for the R.J. Cook Agronomy Farm in Pullman, WA.

Past years

2020 Soil ontology - Notes and Agenda