Community Data Workshop 2021
May 24th-25th
Day 2, 12:00 - 1:00 PST

For announcements on ESS-DIVE activities (i.e. webinars, publications, new feature announcements)...

Follow ESS-DIVE on Twitter! @ESSDIVE

Join ESS-DIVE's Community Mailing List!
https://groups.google.com/a/lbl.gov/g/ESS-DIVE-Community

ESS Sample Data session notes

What is your process for tracking samples sent to different labs, collaborators, etc. for analyses and publication? What do your sample names look like, and how do you use them?

What (if any) challenges have you encountered with sample tracking?

- Long term data- people coming and going, changing sample names over time
- Large projects - the practices you use in small projects depend on person-to-person knowledge that don't scale (Ben BL)

Sample names

- Ben: flexible, as long as it's machine-parseable and maps to single authoritative info (plots, treatments, etc)
    - Generally ad hoc system
- Isobel: can numbers - no consistent way of labeling samples
    - Global monitoring - organize into lat long bins
    - Struggle with ways of rounding up data
- Ricardo: indicative of field site, date, shorthand for plots, depths

Where is your sample data stored or published?

- Ben: most often data and code are on github and archived in zenodo; sometimes use other repos like Figshare. Links between data and papers just in descriptive text but not machine readable.
- Kristin: SFA in DOE. We are still tring to figure this out. Have some code savvy people using github and other place. Some of us are either working with google drive or storing data in our

own personal computers. Everyone uses their own sample data formats. I'm involved in reporting formats. Right now it is not organized.

- Ben: How do we track physical locations for samples?
- Joan: How do you compile the data and know where the data goes
- Ben: have sample numbers and able to meticulous, time consuming, have to be careful with lists and mapping
- Roser: legacy of sample archived - unique identifier, know metadata - go back and create a unique persistent identifier
    - Tools to standardize: crosswalk between current fields and
- Julie: individual samples used in different analyses, different papers, group papers, BUT often take multiple samples from the same soil horizon at the same coordinates
    - Some samples analyzed for different things, average value for a measurement - if you have average of 10 samples
    - AGGREGATE samples

What tools do you think would be useful for your sample management and tracking?

- Julie: Go to a paper - see 100 IGSNs, see what other papers might have data for those samples
    - Ben: this would be COOL and a great demonstration
    - Start as simple as possible?
- Joan: not done automatically. It is where we would like to go. In order to do this we need to build a foundation for use with better tools
- Ben: perfect is not necessary to provide value. Perfect world each sample links to a treatment mean to a result (figure). But that's hard
    - What are the low hanging fruit that would build community support?
- Kristin: Have not interacted with ISGN very much. Wondering if there is a way to make it searchable. Like have a different set that you associate with the sample names so that a human could understand.
- Joan: IGSN has a relationship with publishers for these links
- Kristin: This seems sample by sample and very tedious to navigate parent child relationships. It would be good if there was a way to link or compose a google sheet or database where you could see that relationship.
- Joan: There are other ways to link samples. 1) One of those is to have a genarlized collection identifier that you define. 2) You can link samples by location ids. 3) Create a data set for the sample metadata. Another idea is to see sample collection pages.

Tuesday, May 25th, 2021

Sample Tracking and Analysis Workflows

Sample Tracking Challenge

Challenge - lack of a practical, standardized sample tracking system. Difficulty tracking samples as they are sent to collaborators, labs, etc. Integrating data as they are added. Difficult in findability.
Solution: International Geo/General Sample Numbers (IGSNs); a persistent identifier

QA:

1. BB: That's not many characters in the IGSN code! I'm surprised they can guarantee global uniqueness.
    a. Unique user code issued, standard 9 chars. Central registry that assigns these identifiers.
2. Roser Matamal: Do you get one if you don't have one?
    a. Register through central registry to get one

Discussion:

1. Current sample tracking approach - how do you share your sample names with collaborators and how do you use them?
    a. Ben Bond-Lamberty (BB) - spreadsheets, but doesn't scale well. For small teams, people know what the names or terminology is, but as teams grow it makes it harder to track
    b. Ricardo Eloy Alves (REA) - long running projects and many people have sampled diff stuff at diff times, and then creates spreadsheets with whatever label; people help with field work. Identifiers and metadata, no standardized or followed across all people collecting the data
    c.
2. How do you use your sample names? Connecting related samples? System of location identifiers?
    a. BB: don't care, as long as it maps to a single reference it is fine. "Ad-hoc"
    b. Isobel Simpson (IS): no systematic way. Global monitoring, break world into smaller bins. Struggle with rounding up data
    c. REA: abbreviation of field site, always have date of sampling, shorthand of plots, depths, etc. Keeps this structure even if physical samples are collected somewhere/someone else, they will always know what it is without having to refer to a key or list
3. Where is your sample data stored or published? Is data clearly linked in some way or currently disconnected?
    a. BB: often data and code for publication are on github and archived in Zenodo as a tag. No good links between papers, there is inscriptive text.
        i. *Clarifying question to Joan*: Sending samples to other labs, how do we track samples going someplace else?
            1. As long as you are meticulous with list and mapping files, be careful because of the lack of global uniqueness.
    b. Kristin Boye: still trying to figure out, some code savvy people putting code on github and other places. Others, either working with Google sheets or Drives to

share things or to store on personal computer. Everybody uses their own sample naming format, currently disconnected.

c. RM: we have legacy of samples that have been archived, has unique id so we know where the come from, metadata, etc. How do we go back and create unique persistent identifier? Seems like a lot of work. Is there a way to quickly do this?

i. Joan: Likely map to what we have, standardize the file with sample name and metadata. Taking legacy data and standardizing it. Useful to have a persistent identifier in publications and future sample tracking that the data is linked. Moving forward, fusion db and working on reporting formats - it would benefit from more advanced search query.

ii. Val: Create a “Crosswalk” - spreadsheet that lists and maps all your metadata. Mapping fields to standard fields

d. Julie Jastrow (JJ) : individual samples used for multiple types of analyses. Likes persistent identifier idea. We take multiple (ie: 10) samples from soil horizon from same coordinates - some of those samples will be analyzed for other things. But for papers, we come up with an average value (aggregate val for paper).

4. What tools do you think would be useful for your sample management and tracking?

a. JJ: If someone went to a paper and identified data is associated with 100 examples, is there a way to see which papers are also associated with having data for those samples? Is the person responsible for those samples responsible for the IGSNs.

i. Manual process currently, adding links . Would like to get better tools, first need people to use persistent identifiers.

b. BB: What are the things we can do to convince people this is worth the effort and community support? Needs a bunch of publications to demonstrate the power.

c. KB: IGSNs are difficult to remember and associate with anything, is there a way to make it searchable? Or associate and understand what these codes mean and see the relationship.

Unique Identifier: provides meaningful project specific unique ID to organize data
Metadata: descriptive info about data
Persistent Identifiers: globally unique IDs with permanent link/landing page, associated metadata (ie: ORCID: People, DOI: Data, publications)

Persistent IDs: Landing Pages

Benefits of Persistent Identifiers:

1. Link and expand access pathways
2. Avoid dupes of info across platforms
3. Interpretation and reuse

- Looking for solutions to link and exchange related data across multiple systems
- Enabling people to better compile data and link across systems

- Add links to related data

-