ESS-DIVE
Deep Insight for Earth Science Data

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

National Lab Projects session notes

Project Data Slide Summaries

- Note and synthesize data types and challenges during the slide presentations
- ***Ask attendees to enter into notes, chat:***
    - What data types do you most often work with in your ESS project(s)?
    - What tools do you use for data management? (e.g. excel, R, python, SQL database)
    - What data standards have you used in the past?
    - Are you familiar with FAIR data?
    - Approximately what percentage of your project budget/time do you spend on data management?

Discussion questions:
Ways to participate - raise hand to discuss, enter into google doc notes, AND/OR enter into chat

1.) Are there some challenges identified in presentations that particularly resonate with your data management or publication?
2.) What training or data support do you have for managing and publishing data within your project?
    a.) What data management training have you received?
    b.) What data support do you wish that you had?
    c.) Interest in forming working groups around certain data types/projects/topics?
    d.) What developments in recent years have made data publication easier?
3.) What kinds of tools and capabilities do you want for data management, what will help now, future?
4.) What scientific questions would you like to answer by searching for data in an open data repository or database?
    a.) How do you want to use data from ESS-DIVE and other repositories more broadly?

    b.) What tools do you need for search and visualization to be able to more easily use data?

5.) Do you have any concerns with credit if people reuse your published datasets?

    a.) How have you cited data in the past?

    b.) Have you encountered challenges citing data?

6.) General questions about ESS-DIVE and how we can help.

How should we decide to share when we have data from active collaborations
We want to be able to host data in a long term repository the things that your research depended on--but this is up to the user/contributors

Ways to start to standardize data coming from sensors. Some projects that have been doing a nice job: ARM, Ameriflux, lots of agencies. Ranjeet: Their sensor data gets formatted with NetCDF formatting as the are ingested

Questions and Answers

Q: Is there a wider effort to homogenize terminology used across data repositories, not just within ESS-DIVE? I am wanting to perform a meta-analysis of this data. Consistency of metadata across repositories.
A: Ecological Metadata Language (EML) is the standard in ecology sciences, but not throughout the earth sciences. No, it is not pretty right now and with our reporting formats we are at a starting point for consistency.

Q: If we do get to our general conversation today I would be interested in talking more about support and comments that we are far from supporting files like HDF5 etc. I guess more generally, how does the community help to define priorities/needs for ESS such that we can make support and tools sooner more than later?
A: Shawn, great topic. I feel like a fair amount of this is a sociological problem and not a technology problem.

Q: “How do we provide a flexible and achievable framework for data availability that still adheres to a level of standardization and reproducibility?”
A: The AmeriFlux approach might be useful as one model. https://ameriflux.lbl.gov/data/aboutdata/

** Ben Sulman said that registering his non-ESS funded project was painless on his end.

Q: Downloading model data without downloading a huge zip file would be useful, it’d be interesting to hear more about this later
A: …

Q: How do you deal with datasets that are continuously evolving? I anticipate that my project will be dealing with a lot of data like this. People are thinking “well I’ll just upload it to ESS-DIVE when the paper is published and the product has reached a completed state for now”
A: ESS-DIVE takes a snapshot of the data, where we’re headed is a place where a version string is added to the citation so model versions can be tracked. Establishing an update schedule in your metadata can help users understand what time period the code is reliable for. ESS-DIVE is a publisher, not really designed to handle a continuous stream of data coming in. We can however accommodate regular updates

Q: Could we talk more about what should go in the thinking process behind linking parent samples to multiple subsamples/analysis in a project. Do you have a unique label for every time that a sample goes out to be analyzed? Logistically, it prevents a big issue to add all of the underscores.
A: ..

Q: For model data archival - we could use tiered storage. Something like output could be stored in a deep archive (need to basis). And everything else could be on a spinning disk. When you have TB of model output, they need to be accessible for immediate access on a needs-basis.
A: We are working on tiered storage in the future. Globus allows for tiered storage.

Q: How you write instructional files? Right now, I use a pdf to submit an instructional file. Would it be better to submit it in a different format that is usable and searchable?
A: Seems like pdf would not be the way to go, markdown would be preferred. What would be that markup language for instructional files that would go along with that data? This would be a good effort to look into (David Moulton’s group).
It’d be cool to make this into a working group, David would be happy to contribute.

Comments

Copied directly from chat in some cases

If you collect sensor data in hydrologic monitoring scenarios and can't make it to the reporting format tutorial tomorrow, please reach out to me! I'd love to hear about your challenges and get feedback on the draft reporting format! amy.goldman@pnnl.gov

The Flux Processing (FP) Standard that AmeriFlux and the global FLUXNET community use is described here: https://ameriflux.lbl.gov/data/aboutdata/data-variables/

It might be nice for ESS-Dive to take some of those ARM or Ameriflux and put them into the ESS-Dive preferred format so there’s a ready example at GitHub. Since there would be metadata to draw from in those other programs

ARM data has formal ingest process that allows it to be used across multiple agencies. They can be repurposed for new sensors. NET CDF (format will get you closer to the climate models.

From a model perspective, a download API would be really nice for model research.

More than just sample tracking on a really granular level, attaching significant metadata is a more pressing problem. If it isn't included in the original dataset then it is really hard to find that metadata after the fact.

What kinds of tools and capabilities do you want for data management, what will help now, future? And do you have any tools you use that you think might be useful to others?

- A look-up table that compiles terms used in reporting formats could be a useful tool. They can find variable and/or header names that are already in a template and avoid duplicating those terms using a slightly different format.
    - Agree. A searchable or keyed (like a plant key) look up capability for terms, units, or any types of metadata would be useful -- especially if it can evolve and be added to over time.
- Are the data currently searchable by parameter? e.g. If I wanted to see all datasets that have NO3?
    - If the person who put the metadata in used the keywork "NO3" then you will be able to find it.
- Do you have recommended terminology for variables/keywords? This would provide helpful standardization. The downside being people don't like to follow standardization.

- I was thinking look up in the context of best approaches to input terms/formats but I can see where this would be useful for searching within ESS-DIVE

- I was thinking look up in the context of best approaches to input terms/formats but I can see where this would be useful for searching within ESS-DIVE