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

University Projects - Data Management Challenges and Needs - Session Notes

---

Agenda

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?

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.

Introduction and Challenges

Tianze Song - Georgia INstitute of TEchnology Environmental Microbiology

- Challenge : Uploading diverse data to ESS-DIVE from amplicon/genomics to biogeochemistry

Sherlynette Perez Castro: MBL

- Challenge: Linking to other repositories - NCBI and EMSL.

Nicole Lau - UC Berkeley, Kueppers Lab

- Challenge: Obtaining reliable final data from authors, time spent on unpublished data
- Archive rom published papers, spatial dataset, vast meteorological, datasets in ESS-DIVE
- **Vast amount and range of data**

Eric Ward: USGS Wetland and Aquatic Research Center

- Challenge: Different data types, different spatiotemporal scales, multiple Co-PIs and institution types, each with different policies (USGS, university, national laboratory), data linking and linking between different components

Matthew Ginder-Vogel: University of Wisconsin, Madison

- Challenge: Have collaborators at PNNL and different universities and lots of data types. Model development potentially using existing modeling programs and custom software. Right now, everyone sends their own data to Matthew, no dedicated data manager - how does everyone in the project have a workflow for their data to make sure it is uploaded to a central place? How do we make sure everyone is doing the right and same thing with the data - do we upload by manuscript, by sample + how do we make the data accessible in the long term?

Katherine Duchesneau: Georgia Institute of Technology (Kostka lab)

- Challenge: Interested in making large datasets more easily accessible to a broader audience prior to publication. Also interested in accessing SPRUCE data that other people may have.

Dilys Vela Diaz: UC Irvine

- Communication/coordination among different groups, data location - labs have their own way to standardize and label, how to organize among multiple labs. Working multiple labs, dealing with different standards. Connecting data across different data systems and being able to link these different data.
- Future - relate field and laboratory data collected through different projects but from the same sites and updating/relating data over time.

Caitlin Petro: Georgia Institute of Technology (SPRUCE, Kostka lab)

- Uploading datasets from non-ESS projects
- Want to link smaller datasets (isotopes etc.) to sequencing data
- Data is affiliated with ESS projects (DOE adjacent data) - what is the process of storing data with ESS-DIVE
- **NOTE: Link to project storage request form: https://docs.google.com/forms/d/12LiIt2eAP-8kyKXx9PO0NV8hEs0GVWU0oVBPZPA4vIo/edit?usp=forms_home&ths=true**

Isobel Simpson: UCI Rawland-Blake group

- 9,000 samples. Former CDIAC data contributor
- Transition from CDIAC
- How to update/correct existing data or historical datasets (seasonal etc.)
- DOI identifier for a continuous dataset

Michael Rawlins: University of Massachusetts-Amherst

- Challenge: Determining what format will be best for majority of users (ASCII text, netCDF, and/or ARC-GIS files
- Assistance with uploading for the first time

Indira Paudel

- Sapflux data

Common Themes

Are there some challenges identified in presentations that particularly resonate with your data management or publication?

- Getting whole team on the same page with data management
- A lot of microbial data - sequencing through NCBI
- NMDC - another data system through Berkeley Lab (BSSD within BER) - JGI, NCBI - build hooks to pull sequencing data - try to link data on ESS-DIVE with those on NMDC
- Think about use cases- needs in this case - understand what the linkages are
- Examples of how to link biogeochemical data with sequence data
- Flux rate, etc. about the location - found in ESS-DIVE
- Ameriflux site ID
- Starting to work on reporting format for locations
- Linking sites to other data

What training or data support do you have for managing and publishing data within your project?

- What data management training have you received?
- What data support do you wish that you had?
- Interest in forming working groups around certain data types/projects/topics?

- What developments in recent years have made data publication easier?

Discussion

For groups that do not have a dedicated data manager, what are your challenges and what tools would be helpful?

- Tutorials that are clear enough for grad students and postdocs to follow
  - ESS-DIVE seems to get quickly overwhelming
  - Back and forth communication to work through metadata for samples seem to go on and on
- A lot of focus on field sample data and not bringing things back to the lab and generating data from the samples - haven't seen documentation focused on the incorporation of those data types (not just characterization of samples, but longer term and ongoing experimental systems with samples)
- Do we structure entry into the system under the rubric of an entire project with sub-modules, or publication by type/focus
  - Do we break down by experiment? Assuming that if there are questions about best management practices/advice, will be able to reach out
  - Good news - young people will get used to it :)
- With multi-faceted projects, a way to maintain a file structure and hierarchy - otherwise, getting just the data that a user wants becomes laborious (more of a user issue)
  - Eg. just want core experiments
  - Charu: For large data- may want to break it up, instead of one large dataset, consider instead of large hierarchy
  - Linking data packages
  - Data in paper has to be public at time of publication
- Stephanie - feels like double dipping to cite dataset in paper - not at all, this is the standard process
- Mirror metadata in ESS-DIVE, then linking to where you put the data
  - E.g. raw sequencing data goes in NCBI, NMDC
  - Amplicon data standards

Model data archiving

- Constantly changing, may not make sense to archive some outputs in long run
- Have a session to discuss the guidelines that we came up with - rubric

Address some of the data management linking challenges with Portals

- Way to link data collection - could be project data

- Documentation: https://docs.ess-dive.lbl.gov/portals
- Would help university projects
- Any number of data packages can be added to a portal
    - You can also add information about your project within the portal
    - You can preemptively create a portal with content even if you have not published data packages - set query to filter for your project name, then when you publish data the packages will automatically be added to your portal
- Field campaigns or sites within projects can also have their own portals
    - Create one for project, one for location, etc. -> creating collections
- No limit to the number of portals you can create
- Val: metadata in data packages can help make portals more powerful - keywords, variables, project, related references- can create a query with - working with people outside of project - agree on the kind of metadata to put into packages for searches
    - Charu: More of an issue with field sites - hard to create a query - think about keywords uniquely target the field site, not rely on some site iD that may be common across data packages elsewhere - think of ahead of time
- Coming up: Project spaces will allow project team members to view and edit data packages within the project to collaborate

What kinds of tools and capabilities do you want for data management, what will help now, future?

- A dump where data could be put and then prioritized for publication as they become ready
    - Create levels - raw/messy data, cleaned up data, data + metadata, ready to publish
- Offline metadata editor for collaborative editing, and then someone can
- People wanting tutorials recorded
- Will there be capabilities for team editing of data packages?
    - Yes, this is in the works and is a common pain point
    - Plans within the next year to create features that will enable sharing and permissions
- Co-owning data package: two authors - possible for multiple people to be the contact/owner
    - Charu: part of roadmap- multiple people to get editing privileges, but equal credit is a science citation issue
    - Yes to editing rights in the future

Name of project - goes into data citations, all datasets - is that the name you want living with data packages?

- Projects do work curating, QAQC'd
- As long as author list is accurate - more important
- Descriptive name of dataset is more important
- Can go either way in terms of project versus ESS-DIVE
- DOI also can be used for this
- Association with the grant number? WHat is the project name relation to the grant?

- Nice to have project title in the citation - don't know enough about pros and cons to speak firmly on that
- Benefit of having the whole project - sometimes for the same project, 2-3 different researchers associated with that -
- Challenging because of the complex issues about projects - projects within a project and what do you indicate as the project?
- Now it is flexible - select or write something in
- 2-3 papers for a project: related to main project, each their own - package those as individual publications
- WHen you have datasets in multiple locations - project name from proposal - normally the journal
- From Chat:
    - I think it's useful to have the project name in the title, even though it does make the citation a little bit chunky/repetitive. As Dilys said, I think it makes sense for others to understand that all of these packages came from the same project with the same funding. (Maybe this will be different if public portals were used, and have all data packages be nested under one title?)
    - Is there any way ESS-DIVE can be integrated into the DOI, like in Dryad? Maybe that can be a good compromise.

What Features would you like to see out of ESS-DIVE- Chat

- Getting oriented on how to complete submission of updated data and metadata for a long-term project. Today has been very helpful!
- As a new user I would like to have some Guides for Persistent Identifiers to connect different data types to a number
    - +1 to connect different data types to a number
        - Parent and children identifiers linked to a single code
- Lots of guides and tutorials and having multiple people be able to edit packages
- Templates and tutorials
- A defined strategy for making results of experimental studies "searchable" in a way that would allow someone to look for experimental data on a certain kind of process. This may exist already, e.g. in the way of keywords, but I'm not sure. For example, I just did a search on ESS-DIVE with the phrase "soil organic matter decomposition", and nothing came up...

Q: During the beginning sessions tomorrow would I be able to create data packages and portal as hands-on activity?
A: The more advanced tutorial will focus on the API/creating portals. You can switch back and forth between tutorials, so you can move to the creating portals section if you would like.

Q: Say 2-3 papers are coming out and they're related to the main project but each their own thing - then package their own datasets, wouldn't we want the title to be related to the publication rather than the project name?
A: We recommend that you put in your title, "Data: [Publication]" or "[Publication], dataset" The title field is separate from the project field

Additional: I think it's useful to have the project name in the title, even though it does make the citation a little bit chunky/repetitive. As Dilys said, I think it makes sense for others to understand that all of these packages came from the same project with the same funding. (Maybe this will be different if public portals were used, and have all data packages be nested under one title?)

Q: Is there any way ESS-DIVE can be integrated into the DOI, like in Dryad? Maybe that can be a good compromise.
A: We tried to steer away from that because we took over many data packages from cdicc and it was strange to have cdiac in the DOI when the data were in ess-dive. Bottom line is there is no guarantee the data will stay in one place forever..