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 Sensor Data and QA/QC session notes

---

1. How do you manage sensor data?
2. How often do you (or anticipate) publishing sensor data?
3. How to deal with versioning data from sensors?

Stephanie Pennington - Using dropbox to archive sensor data + posting to github. Considering using R with dropbox. Needs to do QA/QC on these data. I think my *dream* data processing pipeline would be remote raw data to Dropbox -> read into R for QA/QC -> shiny app for data viz and downloading options on Github. Whether it's as seamless as that or not ew shall see ¯_(ツ)_/¯

Les Hook: Spruce has 1000+ sensors - fiberoptic transfer and then archival on ORNL with visualization. System took years to develop. Vista data vision - it is proprietary and has recurring costs. For real-time viz and watching operational issues, it has paid for itself. QA/QC sometimes happens annually. In general data is reliable and seems level 1 even pre QA/QC. Data collected by individual investigators - they process it and review it with Les for consistency. QA/QC is upto investigators. Phenocam data goes off to a separate system. Mini rhizotrons - collect 1000s of images per day per hour and processing that is huge. Only publish final data on ESS-DIVE - those that are not likely to be versioned.

Indira Paudel - Sapflux data archived on dropbox, and used SAS to process the data.Right now uses ICT sensor. Difficult to figure out which sensor is not working when. It would save a lot of time to figure out issues in real-time.

Eric Ward - Duke FACE/ PineMap. Used the campbell software and set up own servers/trailers at the research site. They had cell modems transmit data to offices. They use the Ameriflux framework to publish their data. Most people publish the data when they publish the papers. There are commercial products out there to manage the data.

Stephanie Fulton/ Amy Goldman (PNNL) - Just started working on the project in Nov, and started collecting data this April. Currently collecting water chemistry samples, and sensors - Eureka manta multiprobe - CTD + turbidity. DO with YSI handheld and an s:can spectrolyser - both manually recorded off the screen as well as the signature data. Working with Amy Goldman to set up the ESS-DIVE structure. They have a specific folder data structure on a shared drive. They have 6 sites with weekly data. Turn sensors on and off. They have a continuous dataset of atm. P and T also. They have a QA/QC process that will be automated with R. The naming and hierarchy of folders are consistent and will be used in the scripts. They have metadata sheets which are digitized & is obtained through a google folder. Have photos in the folder as well. Unsure how they are going to package the data and publish this. Planning to do monthly data publications but also annually. Issues with timestamps - daylight etc. Records 1 min-data.. We're going to use R Markdown for visual display using plots

Nancy Merino - Also has CTD data. Works in Savannah river site PondB location as part of LLNL SFA. Taking monthly water column measurements using a CTD depth probe. They also have other probes added on and measuring 8 parameters. Most of the other data is sample-based. Make the data available when the manuscript has been accepted or within 2 years of getting the data.

Nicole Lau: I mostly work with meteorology and soil sensor data (also every 15 minutes), from dataloggers that were installed in the field. We also have some snow data. Field data collection has since ended, so my understanding of how our data were organized was that autologger data went directly into a first computer that ran a first-level automatic QA/QC that then gets sent to a second computer that uploaded directly onto our project website. We're trying to move data off of that website onto ESS-DIVE, but there are definite barriers, just because no real person has really QA/QC'ed all the data besides data that were used in publications. Since data collection has since ended, our plans are really just to store whatever was collected (which is still a lot! ~2008-2016).

Bob Busey: I predominantly have Campbell Scientific hardware and have a couple dozen sites with 30 to 200 sensors per site. Some are telemetered and some aren't. The non telemetered data are downloaded by site visit and mirrored via Dropbox to processing computer. The telemetered end up on debian server/virtual machines. The data there gets broken up into region / then site / then perhaps data tables. I have a bunch of open source python that I've written for automating the data processing. Also have a manual corrections method where I have 1 or more excel spreadsheets for each sensor and manual corrections get applied in a bash script. The intention is that I can fully script the 0 to latest revision by use of scripting and it's repeatable. Then this final product gets reformatted for data archives with other scripts. I have this javascript page which is a bit out of date but I can watch the near real time auto correction data in come in through that and as time allows. http://monitors.iarc.uaf.edu/data/
Examples: https://github.com/frankohanlon/DataPro
Older github: https://github.com/frankohanlon/DataPro

Eric Ward: Les made an interesting aside by mentioning 'Level 1' data. I think every project I've been involved with had multiple levels of QA/QC. I've never been successful completely automating that part.

At some point, humans looking at data and deciding if it makes sense is more efficient, once the obvious errors are removed.

Hannah Blanco - She doesn't pull data from anywhere. Works on NGEE Arctic. On the data management team. Help scientists move data off the sensors off the loggers into ESS-DIVE. Working with metadata mostly at this time.. Not yet working with data yet.

Matthew Ginder-Vogel. - Works with PNNL team. Sensor deployments - pressure transducer, oxygen censor. Big csv files get transferred on SD cards. No automated systems. QA/QC is by manual inspection. QA/QC is done by plotting it in excel. Intermittent field deployments so automation isn't worth it.

How to bundle data for publication
Nicole Lau - Depends on what the author prefer. Sometimes they wan to cutoff the data to just the part for the publication. In general it is what is in the figures. Prioritizing published data. Long-term plans are to archive everything.. Lots to deal with esp with 15-min data for 6-7 years. Data are currently in a hard drive and will be looking at raw data. 3 sites each - two have 20 plots, each with multiple senosrs. The 3rd site they have 40 plots.

Les - One supporting dataset that referred to the larger data series. One had summary data for the plots. Then a table of bacteria archive that was only a table in the publication.

Are there any best practices or formatting guidelines for larger datasets such as datalogger/sensor data?
DA: Ameriflux has a lot of sites with lots of sensors, FRAMES templates, or reporting formats that the community is working on.

Sensor versioning - Eric Ward. Some automated step of QA/QC and then have a human look at it. CAn't automate everything. They have raw data, level 1, level 2 data. Always archived all of them. Not necessarily published all the data. Challenging when you move institutions.

Nicole - Honestly I'm not 100% how things were stored in our server, so I can't answer for sure. We have a "gap-filled" and "not gap-filled" folder, but I haven't talked to our PI about which one to prefer just yet.

DA: Ameriflux - each time they get new data from the site, they consider that a new version and give it a version number. Good idea when creating the version to also add the processing version as well, so it is possible to identify which code was used to process the data when you need to reprocess everything.

Les - in SPRUCE, if they are adding an annual increment to a time-series, take a data release approach. Have a date code indicating the release data as part of the file name. It has a new release code and in the user guide the release history is specified. Data can get messy - columns can get removed or added depending on quality checks.