**Exercise: Students were asked to spend 1 minute jotting down thoughts on a topic and then they switched papers rotating through the following 6 topics related to Data Management: Data, Storage, Standards, Process, Documentation, and Challenges.**

**Transcript of Getting to Know you exercise**

**Data**

* Equipment failure
* Making sure collected data is stored in the right place
* Remembering to copy your fieldbook
* FULCRUM App – but always have paper backup
* Collecting multiple samples from field -> Properly stored and organized
* Troubleshooting
* Remembering to take all the required equipment into the field
* Weather constraints/restriction
* Systematic errors/measuring errors related to measurers
* Standardized versus tailored to individual projects
* Storage
* I just collect it how I think it should be done but there must be a better (and standardized) way? What information should go along the raw data?
* Random pen and paper observation is not enough
* File naming structure
* Traceback to person who collected data and way to contact them
* Clear methods and instrument calibration procedures
* Lidar
* Remote Sensing
* Destroyed samples
* RA’s
* I think about this in a primary data collection sense:
  + Details always matter
  + Where in space and time were the things done/collected/who was there
  + Data collection is the easy part! How do you properly store, share, clean, communicate.
* Hard to sometime partition out what is important from the data you collected
* Lack of STATS training/knowledge
* Hindsight (I wish I collected x instead of y)

**Storage**

* Multiple Copies/Backups
* External Hard drives
* Paper Copies
* Online storage (Google Drive)
* Multiple samples/repeat samples
* Ensure that all data is safe since it can sometimes contain sensitive information
* Scan all the original records from the notebooks
* Who needs access to the data? Affects where/how it can be stored (e.g. shared system)
* 3 copies, 2 types of storage, 1 offline
* Need better storage systems, /consistency/what do we actually need to store, all raw data??? QA’d/QC’d data??? Metadata?? I need to know!!
* Frequent switching of storage systems/methods
* Database is better than excel - > Not just data but metadata storage
  + We often don’t even know that a database exists… or what is in it, or how to add out data to it.
* Back-ups
* Use after person who collected or processed it has left the organization
* Servers, access to servers, access to write to server
* Hard drive
* The cloud
* Open data sharing
* File name, metadata, state of cleanliness, accessibility, backups, file types
* How long do you store for? Who can access? \*Making sure students don’t “Store” things on personal computers only i.e. after graduation

**Standards**

* Why are these important?
* Develop a base for all research so that comparison is easier between similar projects
* Make best practices become more common
* How data were analysed should be standard
* Evaluating quality of data, how they’re treated
* Make sure outside data comes from reputable sources
* Be consistent -> If multiple people are collecting, make sure everyone has the same standards
* Check the literature for common practice
* It feels like none exist in my lab group
* These should be communicated and easy to access
* Standards are useful
* Document or file naming structure – YES!
* Potentially limiting
* 1 efficiency
* Good for reproduces
* Seem to exist for somethings (ie. Water sampling EDA…) but not others, or are not enforced -> What would proper enforcement look like?
* Useful only when used by all and well documented
* Need more consistent procedures/management within labs AND between labs!!!
* Important for longevity of projects/data (comparable between years), building temporal story
* Standards what? It seems like standards are not a thing but IT WOULD BE SO USEFUL TO HAVE SOME GUIDANCE

**Process**

* Determine what needs to be done/questions to be answered
* Decide what methods would be best to do so
* Collect the instrumentation necessary
* The way those data were analysed
* Determine the template of data sheets
* Have a list of what you need so you do not forget any measurements
* Be systematic -> Do it the same everytime
* Frequent communication with any other team members/collaborators
* Organization and proper labelling
* How do we document processes – do we just write equations or is this about the process of data management? I think we often just do thing and don’t really document how we do it?
* Should be reproducible
* Scripts that run on data to produce results
* How data is/was prepared
* Reproducible code
* Documentation
* Trust the process
* So important yet not well documented always (State of data ie. Was it cleaned?)
* Can someone else easily pic up your process (whether it’s fieldwork or code /data analysis)
* Every script/file = read me file (keep up to date!) with what’s been done – Helpful to keep a field/lab/data log of what you did each day.
* What is process?

**Documentation**

* Where the data were collected
* When
* Hourly/daily + YEAR
* Duration
* Who collected the data
* Versions (clean, raw, what’s been done to it)
* Who has worked on files
* TIME! -> Sometimes daily/hourly isn’t enough. Know how precise you need to be
* Properly storing files so others can find/use/reference data in the future
* Detailed file ontes if needed
* Weather during data collection
* Providing proper sources
* Proofreading your own work (and others!) to catch small mistakes that can cause confusion and problems.
* Having multiple copies in multiple places
* I always think things are well documented, but looking at it again I still think ‘huh’?
* Consistency is key – lists/check lists are awesome to ensure clear documentation between multiple people!!!
* Retweet – consistency so that the data is useful outside of the “Life time” of a single student
* Read me files/pages
* Orientation on documentation prior to beginning thesis
* What goes in a readme.txt? How do I make sure it has info useful to other people that might use my data?
* Doucmentation for “post processing chain” would be useful like if I spend time adapting/improving a product
* Readme files
* Markdown in code files
* Details of decisions made in model parameterization and links to code or files where parameter calculations were made
* Has to be clear
* Clear examples
* Add pictures

**Challenges**

* Things that suck: Incomplete datasets, unclean data
* Lost data
* Knowing **where** things were done and what QA/QC was performed
* Not Having **raw** data
* Documentation of metadata
* Photo organizing – still annoying and I have no good system
* Lack of knowledge transfer/training between senior members/new students -> Lack of consistency in documentation/storage etc.
* See above – willingness to adopt new stragegies
* Getting random data with no or incomplete metadata – How do I trust it or use it? My supervisor might say it’s good but how do I justify that?
* Do I have to re-invent the wheel or did someone else already figure it out?
* Sifting through data (often missing)
* Having to ask someone for files you know exist but are not on the server…which then reveal questionable methods
* Repeat data when multiple people dump into an “Archive” folder
* Keeping file naming consistent
* Moving files and not updating working directories in XX
* Metadata that is detailed enough for others to understand….Unitis!
* Storing data in a public place for others to access if needed
* Finding niche data and determining if it is relevant/can be used
* Sharing data when methods/standards aren’t the same across multiple institutions and backgrounds
* Descriptions are not clear enough
* Version control when sharing data