Reproducibility in Geospace Science: Best practices for Data Stewardship

Goal

The primary goal of this workshop is to advance discussions on computational reproducibility in CEDAR science, including challenges posed by access to and usage of data and diverse stake-holders' needs.

Notes:

Data distribution and licensing

- Leslie Lamarche: introduced the topic and several of the challenges associated with distribution and licensing
    - Funding agencies/journals require data to be openly available
        - What data is exempt from this? Low level data exists, large volume, complex. Model outputs, etc.
- Kathryn McWilliams discussed SuperDARN:
    - Tomoko Matsuo identified a challenge with properly attributing each PI that provides data into the SuperDARN network
    - SuperDARN is publishing all rawacf and fitacf with DOIs, including metadata that identifies which PI to credit, what versions of software were used to produce the files, etc.
        - Example test dataset, SuperDARN Canada: https://www.frdr-dfdr.ca/repo/collection/superdarn
    - Developing a formal data policy for the network:
        - Working within an initial test system
            - Provides data access reports to help facilitate reporting for PIs
    - Question of who funds these efforts? Where do the resources come from to do this
        - Commercial services don't have the reporting tools that are needed, very fixed and limited and expensive
        - Michael Hirsch: "As was probably mentioned earlier, the data download charges get really extreme with commercial services, like more expensive

than the instrument perhaps, for only downloading the whole data set 10-100 times say”

- Bill Rideout:
    - There is a push and pull between data providers needs and user needs
        - Madrigal compromised by not requiring users to log in with passwords
        - Some data repositories
- Eric Donovan:
    - Looking at long term, funding both software and data infrastructure effort required to build and maintain this infrastructure
        - DOIs are great, but how do you make sure they actually persist
        - DOIs can help providers track who is using the data
    -
- On publishing data:
    - What if someone publishes your data without your permission?
        - Eric: “Regarding republishing data of someone else... I argue that if someone does it without PI permission that should be considered academic misconduct.”
        - Asti: this “is something that needs to be codified or elaborated on. Because it may also create barrier to doing research.”
    - Kathryn: “Publishing data depends on license of the software/data. Some licenses allow ownership of data by data user, but the user MUST acknowledgment the data source as per the license.”
    - Liam Kilcommons: “Simon and Eric, thoughts on secondary data products? What if someone, for instance, made an geomagnetic index using your data and made that available online somewhere?”
        - Eric: “In my experience it really helps me as a PI... demonstrates usefulness. Christine Gabrielse and Toshi Nishimura are deriving conductivities, energy flux, etc. from the THEMIS-ASI data. It's nice that they include me in the first paper (not essential but I think that's a 'best practice').”
    - Tomoko: “I think it is also important to make it “easy” to get information about data sources/provenance so that users can acknowledge them property.”
    - Simon Shepherd: “Acknowledgement of PIs providing data is critical to the continuation of those data being provided. Within SuperDARN we have “agreed” that any study in which an individual radar can be identified should acknowledge the radar’s PI. For studies that use aggregated data, a general acknowledgment is sufficient”
- A common concern amongst speakers is that we often rely on some unpublished rules of the road for usage of data
    - Maybe we need some kind of more rigid structure?

Common data repositories

- Bill Rideout:
    - What are people's thoughts on what users and providers require from data repositories
    - File and data standardization
    - Start off with unstandardized format
    - Standard scientific format (hdf5, netcdf, ect) - well defined readers, but no standards of how they're formatted
        - Use a standard format, but provide enough metadata for a human to understand it
        - Standardize everything enough so a machine can understand it
    - Madrigal has been attempting to achieve both.
    - Madrigal contains very heterogenous data
    - Data is either a scalar or multidimensional
        - All dimensions must be defined
        - Can have nice array layouts
    - Madrigal has export software that lets you export metadata in a particular format
    - Easily discover data - website that lets you search through what kinds of instruments/data there are
    - Try to make data fully self-describing - challenging
    - Accessing data - madrigal has a standard API
        - Alternatively, scripts that general call signal from clicking on website
    - Distribution - have to enter email to get data
    - Database that maps collection of files to single DOI
- Jeff Klenzing
    - Pysat prompts users for name/email address
    - PIs have interface to access usage for their own instruments in madrigal
- Asti
    - Does this tracking work when people access through a third party API?
        - In reality, third party APIs are just wrappers around the Madrigal API, so everything is tracked as normal
- Russell Stoneback
    - Pysat uses Madrigal API
    - Helpful to NOT require users to have account, just username and email
    - Honor system vs password - so far, honor system seems to have worked well for Madrigal
- Asti
    - Very challenging to access data from instruments NOT on madrigal

    - What do we need to do to either get all data on madrigal, or find some other solution?
    - Optical data takes lots of space - not practical to store it in multiple places
    - Madrigal has metadata links
        - Not a perfect solution - user has to take extra steps to go get it
- Susan Nossal
    - One challenge is to be able to work with users of data to help them understand uncertainties and caveats. At the same time, it is helpful to archive data at early stages of analysis so that the data can be reanalyzed in the future using updated analysis approaches. Do you recommend posting both early stage and calibrated data?
    - To clarify, I think that it could be important for users of the data to contact people with knowledge of the uncertainties so as to be able to accurately use and interpret the observations.
    - Bill: often users need to talk to the PI to understand and work with the data
    - Michael Hirsch: Also, uncertainties are themselves complex. My cameras control temperature to 0.01 degree C, but the temperature affects multiple parameters
    - Russell: Instrument uncertainties are complex, even when there is local access to the instrument, and definitely when the instrument is in space.
    - Susan: In my experience, assessing uncertainty can be quite challenging and time-consuming. Potentially, we could share practices that we found to be helpful.
    - Ashton: This is an excellent point. Usability of the data also necessitates sufficient documentation about that data, whether that includes publications about how the data was processed, and/or contact information for experts you can talk to. Probably we need a more holistic approach. And we should probably demand funding for data providers to put in the extra effort needed
- Tomoko Matsuo
    - Uncertainty has to be published with data sets
    - Fusing data sets - also need to understand integration time
    - Track down uncertainty by working closely with instrument PI
        - Not funding for this
        - Requirements (uncertainty, flags, ect) pile up
- Eric Donovan
    - Virtual observatories
    - Organized around data type, rather than science
    - Horizontal vs vertical observatories
    - API links between virtual observatories
    - Community challenge - get all PIs to do everything the same way

    - Re "community challenge"... wouldn't it be great, e.g., if all magnetometer data was presented in the same coordinate system... we're trying here in Calgary to have all of our optical data (TREx, THEMIS-ASI, REGO, NORSTAR, Rainbow, CANOPUS ASI, POCA, etc) readable by one readfile…
- Michael Hirsch
    - As was probably mentioned earlier, the data download charges get really extreme with commercial services, like more expensive than the instrument perhaps, for only downloading the whole data set 10-100 times say
    - And that is not for imaging instruments, where downloading even a fraction of the data once could be too expensive from commercial archiving. That's for 1-D data from multiple 10..100Hz instruments where downloading the whole data set 10 or 100 times gets too expensive
    - Downloading whole data sets (or substantial fractions of whole data set) is a more important use cases these days with machine learning / data science approaches
    - Asti: Those are great points Michael. Those costs can be prohibitive and therefore can privilege researchers/institutions with more resources. Do we then think it would be useful to have NSF-supported repositories for raw data storage?
        - Asti, I think an American equivalent of Zenodo (even if much less general) would be helpful. Zenodo's main problem is it's 100x slower than Dropbox and similar services, and the dataset size limitation
    - Leslie: Thanks Michael. One thing I saw promoted a lot at AGU last year was this idea of cloud based computing, so all data and some kind of computing environment exists in the cloud, and users access it through accounts. This removes the need to download data, but adds additional costs that I assume are equivalently prohibitive. It’s also not entirely clear to me how helpful those platforms are for reproducibility.
        - Leslie, right if one has like a next generation CCMC in AWS where you have small input config files, and modest output files, with the processing done in the cloud, that could work.
        - However, for users that want to work with the data on their own HPC or in novel ways, they need to exfiltrate the data from the cloud generally, and that's where the huge expense is (to the PI)

Data citation and attribution

- Ashton Reimer
    - Introducing the topic
    - Understanding where the data come from and how it is produced is important
    - Provide attribution to the people that put together the data

    - Challenges: Not enough funding to do this properly. There are limits put on the data management part of projects.
    - Incentives: Current ones are maximize the number of publications …
    - Rules of the road:..
        - Authorship (when, why?)
- Topic expert: Lan Jian, NASA SPDF
    - NASA Goddard Space Flight Center
    - Improve reproducibility
    - Heliophysics Digital Resource Library (HDRL)
        -
    - Purpose of HPDE
        - HPDE Uses DataCite (https://datacite.org) to Mint DOIs for Heliophysics Data
        - Non-for-profit global initiative
        - Open, collaborative, community driven
        - Provides services for over 1900 repositories
    - NASA's Space Physics Data Facility (SPDF)
        -
        - Coordinated Data Analysis Web(
    - About Data Generated from NASA ROSES award
        -
    - Heliophysics Data Portal (HDP)
        - Data authorities are defined in Space
    - Service costs a few thousand $ per year
    - Alternative Way to Reach
    - Open questions
        - Different ways of registering DOIs (r.g. NASA vs. ESA)
        - Some data providers are not keen to making DOIs
        - Frequent changes of dataset (e.g. Change of dimensions)
        - Different version of data
        - Non-typical data (e.g. catalogs)
        - Different requirements of journals for the citations of data
- Michael Hirsh (chat)
    - Asti, I think an American equivalent of Zenodo (even if much less general) would be helpful. Zenodo's main problem is it's 100x slower than Dropbox and similar services, and the dataset size limitation
- Asti:
    - Challenges for ground-based data. There is no clear mandate to put the data in a single place.
- Tomoko:
    - GDC has strong requirement on open source code
    - What role with SPDF play in NASA mandating open source data?
    - Additional directories to archive software
    - Good user guide/documentation

    - Trust that the missions will do this correctly - large amounts of oversight is impractical
- Asti:
  NCAR has a large community data repository for atmospheric data, and tools to search them as well. Does anyone have a perspective on expanding use of NCAR's repository for geospace data?
    - Tomoko: NOAA has ones too. NCAR's interests are for their community model output. It makes sense that NASA hosts data produced by NASA missions.

FAIR data in geospace

- Asti - Introduction
    - Findable
    - Accessible
    - Interoperable
    - Reusable

- Liam Kilcommons - AMGeO, experience dealing with FAIR principals
    - How is it implemented in practice
    - Illustrating with a student accessing your data
        - Can she write code to retrieve any interval? (findable, accessible)
        - Able to write code to read/analyze in preferred language? (interoperable)
        - Metadata help understanding? (reusable)
    - Data is essential to the AMGeO project
- Open discussion
    - Asti - how challenging for AMGeO to access and acquire data?
    - Liam - Findable focused on machines, but need other information
        - Dealing with uptime issues with providers
        - Communicating these issues to users
    - Bill Rideout - How to find attributions of constituent data products?
        - Liam - links to attribution statements for products (i.e. Superdarn)
        - Combined DOI issues
    - Tomoko Matsuo -
        - Automatic generation of acknowledgment with data products, need that information to be passed through programmatically
        - Risk of losing accessibility to data sets if not acknowledged
        - Data providers not funded for these types of efforts
    - Asti - Describing the InGeO citation helper
      https://github.com/EarthCubeInGeo/citationhelper

        - Ashton - We've been borrowing a lot of ideas from software, but have we seen similar licensing for data? Rules of the road to help reviewers (i.e. is data properly attributed according to the rules)

Learning from other disciplines - GEOCODES

- Kenton McHenry
    - EarthCube GeoCODES
    - NSF EarthCube program - advance cyberinfrastructure to support geoscience research
    - Challenges:
        - outreach/exposure not far reaching
        - Preserving outcomes after end of projects
        - Reducing duplication of effort
    - Software side:
        - Tools inventory
        - Research registry
        - Challenging to keep updated, easy to navigate
    - Data side: GeoCODES
        - Making data repositories discoverable
        - Crawl data repositories
        - Stand up portal in front of repository
        - Adoption of schema.org
        - Body of repositories implementing this, plus crawler and portal
    - Incorporate tools aspect with data aspect
    - Guidance document - "Recommended Standards and Specifications for EarthCube Projects"
    - Peer reviewed calls for notebooks
        - Submit tools as a jupyter notebook
    - How do you add your repository after setting up schema
        - Let GeoCODES know
        - Repo needs to provide site map
    - Jason Der: "This is outstanding. I'd LOVE to see this for the space/magnetospheric/ionospheric physics community."
- Leslie Lamarche - are there two different philosophies - putting all data into one repository or crawling through multiple repositories?
    - This is more for sustainability of efforts so you don't need centralized way of doing things
- Supplements do support coding to implement schema.org
- Kathryn - how does this help funding agencies track the usage?
    - There is acknowledgement that can be provided as part of crawling.
- Tomoko - AMGeO tried to implement acknowledgement esp. For SuperDARN data.