InGeO Webinar #2 Discussion Summary

Reproducibility of Research Results [In Geospace Science]
April 16, 2020

Graduate advisors may not be teaching students new policies regarding sharing code and data and using repositories before publishing papers. There may not be enough instruction on the methodology section in writing. Do advisors discuss the specifics of what it means to apply the scientific method to particular research activities (i.e. purely observational or modeling work that is common in geospacer science)?

GitHub isn't currently a FAIR compliant repository. Do we (the geospace community) want it to be FAIR compliant? If so we need to engage with journals. Maybe GitHub and other more static repositories have different purposes? We can develop community code with GitHub, but also maintain static versioning in other software repositories. Some FAIR-compliant software repositories allow you to link with GitHub projects for versioning.

Is there a way to make non-finalized data available for peer review? Finalizing processing and uploading all data can be extremely time consuming. One option is repositories that allow DOI versioning, such that different versions of the same data set can be published as processing/analysis improves (e.g. Zenodo DOI versioning: https://blog.zenodo.org/2017/05/30/doi-versioning-launched/)

The Geospace community does not have a lot of experience making information (software, data sets, ect.) persist for a very long time (i.e. decades), but maybe other fields do. The traditional research model of a printed journal that can be checked out of a curated collection in a library still works relatively well for retrieving old results, but libraries have also been adapting to provide digital resources and update how information is accessed. Can we look at some of what is done in library science to get ideas of how to make sure information persists across many years? Library of Congress guidelines on preservation: https://www.loc.gov/preservation/digital/formats/

Are there studies about if abiding by FAIR policies actually increases citations or improves the impact of research projects? No studies have been done that we know of in Geospace, but some may have been published in other fields. It is important to continue to look at these guidelines critically and make sure they're actually improving science accessibility and reproducibility as intended. Policies that don't actually accomplish this should be revised.

Publishing data and trying to maintain open data can lead to significantly more collaboration opportunities.

FAIR doesn't necessarily assure all data is publicly and freely available. There has been some discussion that FAIR data policies don't go far enough. Should FAIR be expanded? (https://www.go-fair.org/faq/ask-question-difference-fair-data-open-data/)

One problem modelers face is that they produce VERY large volumes of data, and it is impractical to publish all data with publications. There is not currently a good solution to this in the community. How do climatology/other fields that run big models handle this? One potential strategy is to focus on preserving the details of how exactly the model is set-up and initialized rather than the output, but this only works if there is relatively easy access to community computational resources to run the model and may become problematic if those resources are upgraded. Preserving the computational environment on a large computing cluster is very different than a personal workstation. ML/AI techniques may also rely on a substantial volume (TBs) of input data. Future discussions with modelers/funding agencies/big data experts might be in order?