An Early Career Investigator Community Vision for the Future NSF Geophysical Facility: Data Services Needs

Authors: H. A. Ford, M. Floyd, D.S. Stamps, M. Mendoza, E. Bozdag, D. Bowden, J. Byrnes, W. Fan, H. Kehoe, E. Chaussard, N. J. Lindsey, S. Wei, G. Barcheck, T. S. de Smet, H. Janiszewski, E. Lindsey, J. K. MacCarthy, K. Materna, S. Naif, D. Portner, D. Trugman, I. Wang

Citation: Ford, H.A., M. Floyd, D.S. Stamps, M. Mendoza, E. Bozdag, D. Bowden, J. Byrnes, W. Fan, H. Kehoe, E. Chaussard, N. J. Lindsey, S. Wei, G. Barcheck, T. S. de Smet, H. Janiszewski, E. Lindsey, J. K. MacCarthy, K. Materna, S. Naif, D. Portner, D. Trugman, I. Wang (2020). An Early Career Investigator Community Vision for the Future NSF Geophysical Facility: Data Services Needs. White Paper, 3 p., 10.6084/m9.figshare.12398321.

1.0 Introduction

This white paper has been developed based on a compilation of input from ~45 Early Career Investigators (ECIs) from various institutions who participated in the "*Early Career Investigator Virtual Workshop on a Community Vision for the Future Geophysical Facility*" held April 23-24, 2020 and 59 respondents to a follow-up survey for ECIs distributed via IRIS and UNAVCO list-servs. Our aim is to identify the critical Data Services that need to be within the scope of the future NSF Geophysical Facility such that scientific objectives determined by today's ECIs can be met.

2.0 Data and Data Products Archiving

ECI's scientific advances rely heavily on the availability of both raw data that require specialized processing (e.g. continuous seismic waveforms, magnetotelluric time series, raw GNSS data sets, SAR, meteorological data, real-time data streams), which are fundamental to the function of any Facility, and data products produced by PIs and facilities (e.g. seismic velocity models, magnetotelluric transfer functions and resistivity models, GNSS velocity solutions and time-series, interferograms and time series, synthetic databases of Green's functions) through IRIS and UNAVCO. Both IRIS and UNAVCO host data and data products generated through NSF- and non-NSF-funded projects, including ingestion of international data. We highly recommend the future Geophysical Facility (FGF) ***continue to archive non-NSF funded geophysical datasets*** as requested by PIs.

We suggest the development of ***a single data archiving portal for geophysical data and data products*** that international and domestic scientists could use for archiving at the new FGF. A new single portal for uploads should accommodate standardized file formats, historical data, and updatable metadata that describes the data or data products, as well as any associated information from the originator. A corroborating necessity is the creation of Digital Object Identifiers (DOIs), by the FGF or by the originator, so that the data, data products and related resources are discoverable and citable (following the example of UNAVCO's WInSAR). The service of providing DOIs for data and data products assists in abiding by FAIR data principles and the FORCE11 data citation principles, which are being widely adopted by publishers and is therefore requested by the ECI community. We note that training would be necessary for users of the portal.

ECIs are highly ***supportive of community standards*** as they evolve, such as alignment with the International Federation of Digital Seismograph Networks, SINEX formatted combined velocity solutions, seismic velocity models and magnetotelluric data in HDF5, NetCDF, and ascii formats, and the use of other upcoming standards like GeoCSV and GeodesyML/TimeSeriesML. We suggest that the FGF play a leading role, in cooperation with researchers, in determining, developing, and promoting the most useful standardized data formats that conform to Open Geospatial Consortium and modern High Performance Computing standards.

A centralized repository for diverse derived data products is also of great importance to ECIs. As publishers increasingly require access to these products in FAIR-aligned repositories, a common location provided by the FGF would ensure that the products remain permanently accessible and discoverable by other researchers. Storing published or otherwise unavailable processed solutions are highly recommended for rapid responses to geophysical events.

We envision that the scientific questions addressed by ECIs will increasingly require ***storage of large-volume datasets*** as emerging instrumentation technologies mature (e.g., three component nodal arrays, distributed acoustic sensing, real-time GNSS data, and InSAR time series). Therefore, the FGF will need the capacity to store large datasets and the single archiving portal to upload them. If the costs of storing the large datasets is an issue, additional funding and technical support for storage should be requested by PIs, with consultation with the FGF, when submitting grants to allow the data to be stored in perpetuity.

3.0 Data Distribution

Stable, reliable, and free data retrieval is crucial for geophysical research. ECIs envision the FGF with an intuitive, single-access data portal that will encourage the use of diverse datasets and data products. We suggest that the FGF maintain (or develop) the ability to quickly visualize data and data products with online tools, as well as the capability to access these resources remotely via existing standards (e.g., DaaS, OPeNDAP, HTTP) to minimize disruption in current workflows. We request the FGF to ***offer data retrieval services that cater to various demands***. For example, the FGF would need to be capable of efficiently and securely distributing or providing in-situ access to large volume (multi-TB) continuous data, like the entire TA legacy records. Also, the ability to download portions of larger datasets (e.g., a spatial component of time series data, a subset of stations of a larger network, stations within geographic coordinates), rather than full data archives, is crucial to eliminating unnecessary data requests. We specifically encourage efforts towards efficiency of both storage and tailored *ad hoc* data requests by developing a back-end system alongside an optimized internal database to derive the requested data or product output at a resolution and in a format specified by the user through the front-end. Data storage and requests could then be made more efficient by providing such capability to translate, decimate, or other basic on-the-fly data manipulations as part of distribution, including provision of real-time streams if available.

Due to the inherent diversity of geophysical data sources, the necessity of both raw data and derived data products (see Section 2), and the importance of data quality, we suggest that provenance and attribution be explicitly documented and accessible. For example, when a dataset or data product is accessed the citation and any prior data manipulation should be clear.

We envision that the FGF will continue to develop a variety of tools to assist with the download of data via non-interactive command line calls in several languages used in this community (such as MATLAB, Python, and Julia) as well as interactive GUIs. This may also serve as a single-access front end to facilitate downloads no matter where the data are physically archived (in some instances, this could be on a PI's server that is set-up with external access through cyberinfrastructure such as Hyrax), which in turn may ease the transition to handling large data sets.

4.0 Software Resources and Support

We recommend that the FGF serve as a public face for the data stored and distributed through it. Ideally, there will be simple browsing interfaces (e.g. IRIS Wilber 3, the UNAVCO Data Archive Interface) with interactive tools such as time-series (and, for GNSS and InSAR, velocity) plotters and visualizers. ECIs consider centrally hosted software and some facility-supported software (i.e. translators, portals) to be an asset of the FGF. Examples of desirable software resources include: (1) tools that aid the user in accessing (meta)data archives; (2) tools for appraising data quality; (3) tools that allow flexibility for the user to interact with data, e.g. email,

query form, API's, near/real-time monitoring; and (4) a dedicated webpage containing links to externally- and facility-supported software that are commonly used in the geophysical community for education and research. We note that software built to handle data (e.g. Obspy, Antelope, teqc) and collect it (e.g. data loggers, GNSS receivers, clocks) require continuous updates and a degree of understanding in order to utilize properly, therefore ***software support services will be essential at the FGF.*** We also request the FGF continue making standard configuration template files for common equipment in pools and stations that are part of regional/global networks open to the community. We find that it will be useful for the FGF to maintain support for ECIs by way of online documentation/references and technicians/engineers who are contactable via email, phone, or in-person.

5.0 Community Governance:

It is essential that the FGF be responsive to the changing data services needs of its users. We support a ***community governance model*** that pairs facility guidance with community input via an oversight-empowered standing committee made of community member stake-holders, including ECIs. This system ensures detailed, two-way feedback between the FGF and the community, assists the FGF in responding more nimbly to changes or expansions in community science emphases, and enhances community investment in (and usage of) FGF services.

6.0 Preparing for Future Science:

As collaboration amongst geophysics researchers becomes increasingly more common through support from the FGF, ECIs would like to see centralized and standardized methods of hosting, requesting, and downloading data. ECIs are in agreement that, as the volume of geophysical datasets being stored and downloaded inevitably increases, from current to emerging technologies (e.g., distributed acoustic sensing, large-*N* nodal arrays), continued support of existing online services and expansion into efficient data storage and cloud computing tools will both be essential. Transfer and processing of ever-increasing volumes of data is rapidly becoming impractical without community ***access to computing power adjacent to data storage*** and a central repository of benchmarked, open-source code for data-intensive processes, including the potential to provide access to community-developed software on such systems. ECIs agree that a transition to cloud-based storage is a viable option to accommodate fast access to large data volumes and effectively addresses (near-)real time processes and big data projects. While ECIs are excited by the additional possibilities associated with cloud-based storage and computing, there is concern regarding the pricing model of commercially available storage solutions and the long-term autonomy provided by any given operator. ECIs prefer to see cloud-based storage allocated via NSF-supported resources, such as XSEDE, with contingent means to interoperate with other cloud storage and HPC providers.

Managing a diverse range of data types and products will require careful planning in regards to data formats and request tools. We suggest the FGF stay at the ***forefront of data formatting standards***, such as those needed to conform to HPC standards or the cloud. For example, the ASDF standard for seismic data is built on HDF5 containers and has gained traction in recent years, in large part due to its flexibility and scalability on HPC systems. Adapting to new data formats implies that software support to convert back to other formats will be important for many existing software packages.