Abstract: GeoEDF – An Extensible Geospatial Data Framework for FAIR Science

Rajesh Kalyanam, Lan Zhao and X. Carol Song
Rosen Center for Advanced Computing
Purdue University
West Lafayette, U.S.A.
{rkalyana, lanzhao, carolxsong}@purdue.edu

ABSTRACT

The effectiveness and sustainability of a science gateway relies on its ability to support the data acquisition and utilization needs of its users. Workflows that involve a mix of non-reusable code, desktop tools, gateway computing, and intermediate data transfers are a non-trivial barrier to entry for most researchers, and seldom reproducible. With the growing emphasis on FAIR (Findable, Accessible, Interoperable, Reusable) science, workflows that run entirely on a science gateway are key to this effort. Our prior work on Geospatial Data Analysis Building Blocks (GABBs) [1] enabled complex geospatial research workflows in a HUBzero-based science gateway, MyGeoHub [2]. GABBs enhances HUBzero file management with automated metadata extraction, preview, keyword search, and REST API access. GABBs toolkits provide geospatial data processing and visualization capabilities, enabling users to rapidly build and publish interactive geospatial tools. Furthermore, enhancements to HUBzero's middleware enabled direct access to managed data from these tools, allowing them to be run sequentially as a loosely coupled workflow without intervening manual data transfers.

Geosciences and related fields, however, often need to access and utilize large amounts of data hosted in remote repositories, raising some new challenges. Hydrologists utilize precipitation, soil moisture, land cover, and remote sensing data to predict flood damage. Agricultural economists use AgMIP and CMIP model outputs to study climate impact on global food security. Agricultural scientists now use sensor data to study crop health and recommend best agricultural practices. Data sources can range from repositories managed by organizations such as NASA, USGS, etc., to sensor arrays in smart cities, and crowdsourced data from citizen scientists. Due to the massive volume, high dimensionality, and heterogeneous formats involved, and the variability in access protocols, scientists often spend a lot of time manually collecting and processing data using custom, non-reusable code, instead of focusing on core scientific research.

In this lightning talk we will present GeoEDF, our in-progress, extensible geospatial data framework that will enhance GABBs by abstracting away the complexity of acquiring and utilizing data from various sources. A set of extensible GeoEDF *data connectors* will implement common data query and access protocols such as HTTP, OPeNDAP, FTP, Globus, and REST API, supporting both static and streaming data. Data sources can then be configured by simply specifying the data location, authentication, access protocol, etc. Connectors are also parameterizable, allowing reuse for subdataset, time range, and, geospatial region choices from a data source. Extensible GeoEDF *data processors* will implement common and domain-specific geospatial data processing such as resampling, reprojection, or a specific scientific simulation model. A plug-and-play *workflow composer* will allow users to string together data connectors and processors into a workflow that can be executed in various environments including HUBzero tools, HPC resources, or Jupyter Notebooks. GABBs automated metadata extraction and annotation will be integrated into such workflows, supporting FAIR science practices through ease of subsequent data discovery. GeoEDF will enhance interoperability, leveraging data connectors for data transfer between science gateways. By *bringing data to the science*, GeoEDF will accelerate data-driven discovery, while ensuring that data is not siloed.

Keywords—remote data; geospatial data; science gateway; workflow; FAIR science

REFERENCES

[1] Zhao, L., Song, C.X., Kalyanam, K., Biehl, L., Campbell, R., Delgass, L., Kearney, D., Wan, W., Shin, J., Kim, I.L. and Ellis, C. 2017, June. GABBs - Reusable Geospatial Data Analysis Building Blocks for Science Gateways. *9th International Workshop for Science Gateways*.

[2] Kalyanam, R., Zhao, L., Song, C.X., Biehl, L., Kearney, D., Kim, I.L., Shin, J., Villoria, N. and Merwade, V. 2018. MyGeoHub - A Sustainable and Evolving Geospatial Science Gateway. Future Generation Computer Systems, ISSN 0167-739X, https://doi.org/10.1016/j.future.2018.02.005.