Geospatial Software Institute and Cyberinfrastructure for the Hydrology Community

David Tarboton

Hydrologic Science seeks to understand the dominant role of water in the earth system, spanning natural and anthropogenically altered environments across the full range of climate and geomorphologic regimes. The grand challenges that hydrology faces are by nature integrative, and require the integration of large datasets and models. Much of this data is geospatial. Advances in hydrology, and other sciences, increasingly require integration of information from multiple sources. The complex, multi-faceted problems faced in hydrology such as predicting floods and droughts in the face of climate and watershed changes demand team science and collaboration, with data and models open, accessible and transparent to support reproducibility and enhance trust in findings and results. Cyberinfrastructure (CI) is needed to help scientists move into this new paradigm of collaborative research. This paper considers some of the opportunities and challenges associated with this CI drawing upon my experience and work with CI for the hydrology community, to inform some of the functionality that should be considered for a Geospatial Software Institute.

From a cyberinfrastructure perspective, a motivating question is "Can we deliver hydrologic research and analysis functionality as a service over the web". One way to think of this is as "clearing your desk", or moving your research work completely into a cyberinfrastructure ecosystem of many interfaces to shared services (Chaudhary and Ramnath, 2017). Collaboration on the editing of documents using Google Docs is one widespread manifestation of this paradigm. CyberGIS has also pursued this vision (Wright and Wang, 2011; Wang, 2010), as has ESRI with ArcGIS online. Research collaborative environments such as Cyverse (http://www.cyverse.org/), SEAD (http://sead-data.net/), and HydroShare (http://www.hydroshare.org) have made some steps along this path, but I think it still remains a challenge, and an opportunity to build out a fully interoperable web of components that support completely web based collaborative workspaces. The size of geospatial data makes it a strong candidate for application of this paradigm. I will comment now on some specific successes and challenges.

HydroShare is a hydrologic information system operated by the Consortium of Universities for the Advancement of Hydrologic Science Inc. (CUAHSI) that enables users to share and publish data and models in a variety of flexible formats, and to make this information available in a citable, shareable and discoverable manner. HydroShare includes a repository for data and models, and tools (web apps) that can act on content in HydroShare providing users with a gateway to high performance computing and computing in the cloud. HydroShare has been developed with NSF support from the Software Infrastructure for Sustained Innovation (SI^2) program as a Scientific Software Integration (SSI) project and faces many of the challenges related to software infrastructure and sustainability that Scientific Software Innovation Institutes (S^2I^2) such as the proposed Geospatial Software Institute are designed to address. HydroShare's functionality and architecture are organized into three categories (Fig. 1): (1) resource storage, (2) resource exploration, and (3) actions on resources. These are implemented using system components that are loosely coupled and interact through application program interfaces (APIs). The loose coupling among HydroShare's components is a variant on Services-Oriented-Architecture (SOA) that enhances robustness, as components can be upgraded and advanced

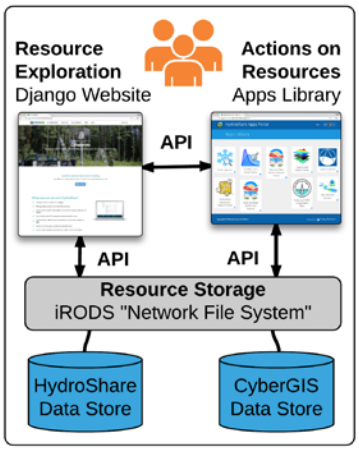

Figure 1. Broad HydroShare Component Architecture.

relatively independently. This architecture lends itself to providing both software (e.g., web apps) and infrastructure (e.g., storage and computational resources) as services to users. The full power of this paradigm is the extensibility it supports, in that anybody can develop a web app that interacts with resources stored in HydroShare. This architecture has the potential to serve the goal of providing fully web based research and analysis functionality.

From working with HydroShare and associated apps, such as the CyberGIS TauDEM app established at the CyberGIS center at the University of Illinios, Jupyter Notebook and HydroShare GIS apps, hosted at RENCI the following considerations and needs have become apparent.

- Web GIS apps need richer basic GIS functionality, for tasks such as subsetting, reprojecting, and control over the presentation of web maps (symbolization). The usability and broader application of advanced computing, without this functionality is diminished.
- Interfaces are needed to accommodate the delays associated with batch processing.
- Simple models are needed for computational unit accounting to enable users to fully take advantage of the HPC capability of the systems that host apps.
- Simpler models are needed to manage where content is stored so that it is transparent to users what of their work is held in app storage, versus collaborative repository (HydroShare).
- Stability, sustained reliability and clear messaging are critical to establishing and growing the trust users need to rely on web based CI.

The CyberGIS TauDEM app lets users open resources (the containers for data) from HydroShare, import and execute a sequence of TauDEM geo processing functions and return the results to HydroShare. All of the above considerations limit, to some degree, the usability of the TauDEM CyberGIS app, where greater capability to use subsetting, projection and visualization functions common in desktop GIS systems pre- and post TauDEM function execution would be of value. The HydroShare Jupyter Notebook app provides a mechanism for content to be read from the HydroShare repository, acted on through the general purpose python or R scripted analysis environment of Jupyter, then results put back in HydroShare. This is a powerful paradigm for collaboration and reproducibility, but requires a file workspace management model that helps users track where their products are in a simple way. It also requires stability of platform software and management of library dependencies to ensure a consistent user experience. GIS Some of these points may not appear transformative, from a research or CI development perspective, but in the experience of HydroShare these have arisen as barriers and pain points that have limited some of the potential for the web based research platform outlined to be realized, and these should be factors considered in the development of plans for software institute to support Geospatial CI.

Chaudhary, V. and R. Ramnath, (2017), "Software Infrastructure for Sustained Innovation (SI2)," SI2 PI Meeting, Arlington, VA, 21-22 February, Figshare, https://doi.org/10.6084/m9.figshare.4676173.

Wang, S., (2010), "A CyberGIS Framework for the Synthesis of Cyberinfrastructure, GIS, and Spatial Analysis," Annals of the Association of American Geographers, 100(3): 535-557, http://dx.doi.org/10.1080/00045601003791243.

Wright, D. J. and S. Wang, (2011), "The emergence of spatial cyberinfrastructure," Proceedings of the National Academy of Sciences, 108(14): 5488-5491, http://www.pnas.org/content/108/14/5488.abstract.