The First GeoEDF Stakeholder Workshop

Workshop Report

Executive Summary

The first GeoEDF Stakeholder Workshop was held at the Hampton Inns & Suites, West Lafayette, Indiana on October 7-8, 2019. The primary goal of the workshop was to engage stakeholders directly and get feedback on the overall direction and key concepts of the project and explore potential opportunities of adoption and collaboration. Approximately 30 people participated in the workshop, including the project team and its advisory board members. The workshop agenda included project presentations, breakout sessions, lightning talks, networking events, and a tour of the Purdue ACRE phenotyping and field imaging facilities. Overall, the stakeholders and project advisory board members agreed that the project is going in the right direction and suggested additional areas to consider and emphasized the importance of engaging the community early during the project.

The workshop agenda, along with all the presentation slides and this report are available at the workshop website: https://mygeohub.org/groups/gabbs/geoedf2019.

Introduction

The NSF CSSI Data Framework project titled “Extensible Geospatial Data Framework Towards FAIR Science” is funded for five years from October 2018 to September 2023. Its vision is to make existing and new valuable, large scientific and social datasets directly usable in scientific models and tools, in addressing the “data wrangling” challenges faced everyday by scientists. Its overarching goal is to create an extensible geospatial data framework to provide seamless connections among platforms, data and tools for the broader research and education communities. By putting easy-to-use tools and platforms into the hands of scientists and students, the project aims to impact the broader research communities in applying the FAIR science principles, making their research Findable, Accessible, Interoperable and Reusable.

At the conclusion of Project Year 1, the GeoEDF team was well-positioned to present its progress to date to a broader set of stakeholders and our advisory board for input, and to explore potential future collaboration ideas and connections with other relevant projects beyond the proposed scientific use cases.

This report is a summary of the workshop activities and the input collected from the workshop to provide guidance on the next steps of developing the GeoEDF data framework.

Workshop Agenda

The workshop featured presentations of a project overview, the four driving scientific use cases, and cyberinfrastructure being developed. Through facilitated small group discussions, the workshop provided a venue for open-ended stakeholder feedback focused on GeoEDF design and its impact, and on collaboration opportunities. The agenda also allowed for several networking opportunities throughout the workshop.

Workshop Agenda

October 7-8, 2019

Meeting Locations:

Day 1: Hampton Inn & Suites-West Lafayette - Wabash / Boiler Room

Day 2: Purdue Agronomy Center for Research & Education

Sunday, October 6, 2019

Travel to West Lafayette, IN

6:30-8:30 PM	Welcome dinner Black Sparrow, 223 Main Street, Lafayette, IN 47901

Monday, October 7, 2019

Hampton Inn Suites Wabash/Boiler Rooms

Breakfast on your own (hotel includes breakfast)

8:30 AM	Welcome and Introductions
9:00 AM	Project Overview (Carol Song, Michael Zentner)
10:00 AM	Coffee Break
10:15 AM	Science Use Cases (Session Chair: Carol Song) ● Near Real-Time Flood Modeling at Multiple Scales from Major Rivers to Urban Streets ● Managing Water Quality Data Collected from Field Sensors ● Real Time Data Collection, Processing, and Decision Making using Crowdsourced Crop Data ● Global to Local Food Security and Sustainability Research Connecting Climate Data and Social Science Data with Crop Modeling
NOON	Group Photo & Lunch
1:00 PM	Design & Architecture Presentation (Rajesh Kalyanam)

2:00 PM	Breakout Session I
3:00 PM	Coffee Break
3:15 PM	Breakout Session II
4:15 PM	Breakout sessions report out & discussion
4:45 PM	Wrap up Day 1
5:00 PM	Adjourn
6:30 - 8:30 PM	Networking Dinner RedSeven Kitchen + Cocktail 200 Main Street, Lafayette, IN 47901

Tuesday, October 8, 2019

Location: *Purdue Agronomy Center for Research & Education, ICSIC Building, Room 1123*

8:30 AM	Day 2 Welcome
9:00 AM	Lightning talks ● Mike Strager, West Virginia University ● Tom Hertel, Purdue University ● Tracy Kugler, University of Minnesota ● Noah Fahlgren, Danforth Center ● Dennis Buckmaster, Purdue University ● Sean Cleveland, University of Hawaii ● Dave Tarboton, Utah State University
10:30 AM	Round Table
11:00 AM	Tour of ACRE Phenotyping facilities ● ICSIC tour ● LeafSpec demo ● Field imaging gantry tour behind the building
NOON	Workshop adjourned

Summary of the project team presentations

Project Overview, Carol Song

The project PI, Dr. Song, presented a project overview describing the background, scientific drivers, objectives, and the broader impact of the 'Extensible, Geospatial Data Framework Towards FAIR science' (GeoEDF) project. This project builds on the successful 'Geospatial Data Analysis Building Blocks (GABBs)' project which enables an integrated data management pipeline on the HUBzero science gateway platform, including built-in geospatial data support, visualization, publication, toolkits for rapid application development, and data service interoperability, deployed on the GeoHub (mygeohub.org). While GABBs addresses the day-to-day geospatial computational needs, other challenges still remain, preventing researchers in geosciences to fully utilize the vast amount of existing data resources. These challenges are sometimes summarized as "data wrangling" where researchers could spend as much 80% of their time looking for, accessing, and preparing data for use in their computational codes, and resulting in non-reproducible workflows.

The ultimate vision of GeoEDF is to create an extensible geospatial data framework that will address the challenges by providing seamless connections among platforms, data and tools, hence making valuable, large scientific and social datasets usable directly in scientific models and tools. The primary goal is to put easy-to-use tools and platforms into the hands of researchers and students to conduct scientific investigations following FAIR science principles. The key objectives of GeoEDF are: 1) Develop the plug-and-play data framework; 2) Use the plug-and-play data framework to address domain science needs; 3) Develop interoperability with other CIs and contribute to a national geospatial software community; 4) Disseminate GeoEDF to the broader community.

The key partners and data sources that are being considered for this project are: NASA, USGS, USDA, CUASHI, EarthStat, GEO, CIESIN, EPA, and others. Various deliverables of the project were highlighted such as: plug-and-play geospatial data framework, open-source packages installable on CI platforms, FAIR tool / data linkage, training materials, and CI interoperability.

Mike Zentner from San Diego Supercomputing Center also presented on the current data challenges and future impacts within the HUBzero platform.

Near Real-Time Flood Modeling at Multiple Scales from Major Rivers to Urban Streets

Co-PI Professor Venkatesh Merwade from Civil Engineering Department at Purdue University presented his use case of multi-scale flood modeling using FAIR principles. Flood models (hydrologic and hydraulic) are created at multiple scales to address different research problems from continental scale to a small city to a particular neighborhood. The data and computational needs of these models vary. However, regardless of the scale, many data are common including digital elevation model, land use, soil, hydrography and climate. By developing reusable data connectors and processors to the common data sources, it will greatly reduce the time and effort scientists spend in preprocessing and preparing the data.

Managing Water Quality Data Collected from Field Sensors

Co-PI Jack Smith described the challenges of collecting and sharing water quality sensor data from the field (batch and real-time) to the broader community via the EPA/USGS Water Quality Portal (WQP) using EPA's Water Quality eXchange(WQX) services and APIs. By adopting the GeoEDF framework and using the Aquavit portal (HUBzero-based) as a test bed, he can develop modular reusable data connectors to various sensor devices and databases, data processors to preprocess the data using data standards such as Sensor Observation Service (SOS) and WaterML 2.0, and construct pipeline processes to facilitate the data workflow.

Real Time Data Collection, Processing, and Decision Making using Crowdsourced Crop Data

Co-PI Professor Jian Jin from Agricultural and Biological Engineering Department at Purdue University presented his work of developing a portable handheld hyperspectral imager for accurate crop health measurement. He has collaborated with the GeoEDF team in developing a prototype client server platform for sensor data ingestion, processing, and analysis. In the GeoEDF project, we will map the prototype system into the GeoEDF architecture and develop new data processors and connectors that solve issues related to data privacy, data calibration and reproducibility.

Global to Local Food Security and Sustainability Research Connecting Climate Data and Social Science Data with Crop Modeling

Co-PI Professor Uris Baldos from Agricultural Economics at Purdue University discussed how GeoEDF could help reduce the barriers for global sustainability analysis by developing a set of open source general purpose data connectors and processors for common climate and social science datasets, including gridded data aggregation and crop yield prediction.

Design & Architecture Presentation, Rajesh Kalyanam

Project team member, Rajesh Kalyanam presented the initial design and architecture of the GeoEDF framework that was accomplished in project year 1 and a timeline of future development activities. The design was driven by the requirements from use cases of Co-PIs Merwade and Baldos. The presentation focused on the design and technological choices underlying data connectors and processors and example Python-based connectors and processors that were developed for these use cases. He also presented the proposed YAML syntax for defining GeoEDF workflows and a plan for leveraging a pre-existing scientific workflow library, Pegasus, to implement the plug-and-play GeoEDF workflow framework.

Summary of the lightning talks

Mike Strager, West Virginia University

Mike Strager presented work on the development of a high-resolution land cover classification database that is based on NAIP imagery collected every two years. This 1 metre resolution data has been used to study water quality issues in watersheds and is of interest to hydrologists. The database allows spatial data filtering based on a user-provided shapefile. While the data is intended for local analysis and usage, future work may involve data hosting in EPA's Streamcat database or HydroShare and the development of data connectors to access this data from the GeoEDF framework.

Tom Hertel, Purdue University

Tom Hertel presented work on the modeling of the effect of global forces (climate, population, income) driving local stresses on crops and other resources. As a specific example, he presented the issue of nitrate leaching from field drains to the dead zone in the Gulf of Mexico. Solutions to this issue involve various local policy changes including the use of controlled drainage and establishment of managed wetlands. Tom also presented the challenges of establishing communication between four models that are involved in modeling the effect of global stresses on various resources such as water, energy, and food. In particular, there is a need for consistency between the datasets utilized in each of these models as well as seamless data transfer from one to the other. These are areas where the GeoEDF framework can help.

Tracy Kugler, University of Minnesota

Tracy Kugler presented work by IPUMS on developing the largest repository of demographic data based on census and survey data. This data includes complete census responses between 1850 - 1940 and sample responses and summary data from 1950 onwards. An API is now in development for the NHGIS data which includes census tables and GIS boundaries down to the administrative levels. Future development will include a Python and R API for the IPUMS US data and a spatial database. IPUMS also supports various cross tabulation operations and aggregation across grids. Collaborations with the GeoEDF project include the development of connectors to interact with the IPUMS APIs and GeoEDF processors to implement the aggregation of gridded data.

Noah Fahlgren, Danforth Center

Noah Fahlgren presented work on plant phenotyping in the TERRA-REF project. He described the use of leaf imaging data from multiple scales including X-ray and CT scans. He also described the various CI components of the TERRA-REF project including the PlantCV modular image analysis library, Jupyter notebooks for composing image analysis workflows, and the use of the Nextflow and Parsl workflow engines. Data from the project is available via Globus or through direct access of the BETYdb database.

Dennis Buckmaster, Purdue University

Dennis Buckmaster presented a logistics approach to agriculture that uses a graph based model to support decision making. This model is used to improve interoperability between the various resources used in farming. He presented various mobile applications that can be used in the field to delineate watersheds and share topography data.

Sean Cleveland, University of Hawaii

Sean Cleveland presented work on models that study water sustainability using data from rainfall stations, water quality and well monitoring sites. He presented their CI platform that is built on the Tapis framework. His most recent work involves the extension of Tapis to streaming data to support both the collection, and processing of streaming data in response to various triggers. Support for streaming data in PI Smith and Jin's use cases will utilize a similar design, borrowing from Sean's work in integrating with EarthCube's CHORDS framework.

Dave Tarboton, Utah State University

Dave Tarboton presented Hydroshare, a CI for hosting hydrologic models and resources, as well as applications. Hydroshare is focused on supporting both reproducible workflows, as well as FAIR science through its requirements on sufficient resource annotation. Hydroshare utilizes both Schema.org JSON-LD and the Dublin Core metadata vocabulary. Hydroshare is a collaboration partner on GeoEDF and future work will include the development of connectors for Hydroshare resources, as well as the development of processors as Hydroshare web applications.

Summary of the breakout sessions

There were two breakout sessions in the workshop where the attendees were organized into groups based on their science domains. Small-group discussions centered around two themes: the impacts of GeoEDF and community engagement. We used a list of topics and associated questions (below) to help start and focus the discussions:

Topic 1: What features presented could impact your work?

- What are your key priorities?
- Collect feedback on GeoEDF features, capabilities, and our approach
- We aren't discussing collaboration ideas yet, that will come in session 2
- Do you have ideas on better ways to do something?
- Science PIs can ask about how their proposed work will impact the domain in general

Topic 2: What data / CI issues are you facing that GeoEDF could incorporate?

- What have we missed?
- Can we implement something (from scratch or by integration) that would make you more likely to use GeoEDF? (either directly or linking in some manner?)

Topic 3: How can we engage with you and other partners? (examples below)

- If you do not yet use a CI, will you be willing to use GeoEDF/MyGeoHub?
- If you do use a CI currently, can we think of ways to interoperate?
- Can we host any of your modeling tools/code in GeoEDF?
- Do you know of other datasets/tools that may be useful for our science goals?
- Do you know of other groups/researchers we ought to reach out to?
- Are there conferences or venues that could help locate the partners?
- How do you want to be engaged with? (google group, email, via MyGeoHub)
- Are there other people or networks that we can engage with or reach out to?
- Would you like to explore partnering up on proposals in the future?

Summary of the feedback

The discussion focused on three areas as summarized below:

1. The attendees in general agreed that the proposed work will positively impact their work. More specifically, several people wanted tools to provide a generic way of gridded data processing, such as reprojection, interpolation, aggregation, and QA/QC processing. These reusable tools are also desired by data providers such as IPUMS for making their social science data more accessible and usable. Some researchers face challenges with handing large high-resolution data and field data in different formats which may benefit from GeoEDF's data connector and processor libraries and execution platform. Most researchers recognize the importance of making data FAIR in their research, and having easy-to-use tools that meet their needs is the key.

   Other useful GeoEDF features attendees appreciate include

   - exposing GeoEDF as services so that users can invoke (part of a) GeoEDF workflow remotely and enabling ingestion of continuous streaming data for real time applications
   - helping with designing non-trivial workflows that support some amount of customization and choice
   - helping communities where several high-resolution datasets are available, but not in a single location.

2. Other data/CI issues that GeoEDF could incorporate include support of FAIR science through versioning of workflows and resource tagging, identification of popular data repository standards and leveraging them to discover useful datasets, making GeoEDF connectors and processors discoverable from relevant CIs such as HydroShare, as well as suggestion of better/alternative data sources when user searches for a specific dataset. GeoEDF may need to highlight more explicitly how its uses will be FAIR.
3. Collaboration with data providers to help make their data accessible may be explored, including the suggestion of making it easier for researchers to "discover" useful data via spatial queries and cross linking against existing data repositories supporting schema.org or DCMI schemas.
4. In terms of user engagement, the project may explore multiple channels in parallel: conferences such as AGU, Phenome, and ESRI events, communities such as CZO, EarthCube, CUAHSI, and BigDataHub, workshop style meetings, webinars, and more focused session with collaborators, university extension services, courses, as well as hackathons. Partnering with other organizations/projects in hackathons may be considered. It was frequently mentioned that the project should engage users early instead of waiting until year 3 or year 4.

Touring of Purdue’s ACRE Phenotyping Facility north of the campus

Group picture taken by a drone used by Prof. Jian Jin’s group to study crop health from remote sensing data

List of Workshop Participants

Project Lead Investigators	
Carol Song	Research Computing, Purdue University
Uris Baldos	Agricultural Economics, Purdue University
Jian Jin	Agricultural & Biological Engineering, Purdue University
Venkatesh Merwade	Civil Engineering, Purdue University
Jack Smith	Center for Environment, Geotechnical & Applied Sciences, Marshall University
Stakeholders & Advisors	
Aaron Walz	Agricultural Data Services, Purdue University
Addie Thompson	Department of Plant, Soil and Microbial Sciences, Michigan State University
Alfred J Kalyanapu	Civil and Environmental Engineering, Tennessee Technological University
David Tarboton	Civil & Environmental Engineering, Utah State University
Dennis Buckmaster	Digital Agriculture, Purdue University
Dominique van der Mensbrugghe	GTAP, Purdue University
Fran Fabrizio	University of Minnesota, IPUMS
Marian Muste	IIHR-Hydroscience & Engineering, U of Iowa
Michael P Strager	Resource Economics & Management, West Virginia University
Michael Witt	Libraries & Information Sciences, Purdue University
Noah Fahlgren	Donald Danforth Plant Science Center
Peter J. Singhofen	Streamline Technologies, Inc.
Sean Cleveland	Information Technology Services, University of Hawaii
Siddharth Saksena	Civil & Environment Engineering, Virginia Tech
Tracy Kugler	University of Minnesota, IPUMS
Tom Hertel	Agricultural Economics, Purdue University
Technical Team	
Carolyn Ellis	Research Computing, Purdue University
Jungha Woo	Research Computing, Purdue University
Lan Zhao	Research Computing, Purdue University
Marisa Brazil	Research Computing, Purdue University
Michael Zentner	San Diego Supercomputing Center
Nicole Brewer	Research Computing, Purdue University
Rajesh Kalyanam	Research Computing, Purdue University
Rob Campbell	Research Computing, Purdue University
Shirley Skeel	Research Computing, Purdue University