# CSSI Framework: Extensible Geospatial Data Framework towards FAIR (Findable, Accessible, Interoperable, Reusable) Science

PI: Carol Song, Co-PIs: Uris Baldos, Venkatesh Merwade, Jian Jin, Jack Smith*; Development team: Rajesh Kalyanam, Lan Zhao, Rob Campbell, Jungha Woo, Nicole Brewer
Institutions: Purdue University, Marshall University*

## Goals



✗ Ad-hoc, monolithic code for wrangling and processing data
✗ Hard to maintain with research group turnover
✗ Mix of desktop tools, HPC, and gateways, requiring cumbersome data transfer back and forth
✗ Rarely reproducible !!!

GeoEDF

✓ Reusable, portable building blocks for data wrangling and processing
✓ Reproducible, extensible workflow clearly delineating data acquisition and processing operations
✓ Operates entirely in a web gateway environment
✓ Leverages gateway tools for data publication and sharing

OUR DATA WORKFLOW - Final
1. Go to the science gateway
2. Define "my-workflow.yml"
3. Ask GeoEDF to execute
4. Manage workflow file and outputs in gateway
5. Share workflow file with collaborators

## Approach

### Design

**Reusable Data Connectors**
Implement various data access protocols, enable data acquisition from popular repositories

**+**

**Reusable Data Processors**
Implement domain agnostic & domain specific geospatial processing operations

**+**

**Plug-and-play Workflow Composer**
Enable the composition of individual connectors & processors into complex workflows

**GeoEDF**
Enable researchers to conceive of geospatial data driven workflows as a sequence of data acquisition and processing steps that can be carried out using pre-existing or user contributed connectors and processors

### Deployment



MyGeoHub

Connector/Processor Registry | Pegasus Workflow Management | HPC

GeoEDF Jupyter Library
Gateway
Leveraged Services

### Results

✓ 30 connectors and processors developed
✓ 25 pull requests for community contributed connectors/processors via GitHub
✓ Workflows developed for diverse domains: hydrology, water quality, resource sustainability, ecology
✓ StreamCI used to support streaming data in ecological and manufacturing system research & training

## Looking Ahead

**Objectives**
○ Ease and scalability of deployment via Kubernetes and Helm charts
○ Ease of integration with other JupyterHub-based computational gateways
○ Extensible data portal for FAIR science capabilities
○ Use of Schema.org for web searchable datasets

➢ Research data portal for FAIR data
➢ Reproducible workflow publication and launch
➢ Embedded schema.org for dataset search indexing

**Canonical GeoEDF Deployment**
To support FAIR science, ease of integration and scalability

➢ GeoEDF deployed to Jupyter notebook server image
➢ Workflows run in local htcondor pool
➢ CILogon integration for user authentication

JupyterHub

➢ Message broker for processing resource publication requests

➢ Metadata extraction and indexing
➢ Resource storage

Google
➢ Google Dataset Search for resource discovery

zenodo
➢ Publication to Zenodo for DOI generation