Title: *Building Spatial Data Infrastructure (SDI) for Future GIS Research and Education: A Case Study from Harvard University*

Contributing Organization: Center for Geographic Analysis (CGA), Harvard University

Contacts: Devika Jain, Data Science Project Manager (kakkar@fas.harvard.edu), Dr. Rachel Franklin, Executive Director (rachel_franklin@cga.harvard.edu)

Summary

In response to the growing demand for high-performance, cost-effective spatial data infrastructure, this Office of Vice Provost of Research (OVPR) -funded initiative enhances Harvard's Spatial Data Infrastructure (SDI) to support Spatial AI, Big Data, and Data Science research across disciplines. Leveraging the New England Research Cloud (NERC) and FAS Research Computing (FASRC), we've built a modern, cloud-native SDI that accelerates and simplifies complex spatial analysis. Featuring containerized systems, GPU-accelerated analytics, and open-source technologies, the SDI exemplifies next-generation capabilities and offers replicable strategies aligned with OGC modernization goals by emphasizing automation, interoperability, and cross-disciplinary use. The key features include:

- Scalable big data analysis & visualization on High Performance (HPC) and Cloud computing
- Advanced spatial analytics with AI and Data Science integration
- Support for FAIR (Findable, Accessible, Interoperable, Reusable) principles using GitHub.
- Reproducible, workflow-driven computing using KNIME, PostGIS, Jupyter and more
- Community GIS education using open source workshops and training

OGC Theme Alignment

1. Data & Technology

Harvard's SDI combines FASRC (short-term HPC) and NERC (long-term cloud) resources to support robust geospatial big data processing. Core components include the following (Please follow the URL for application details and use-cases):

- **Standardized Datasets developed**:
  - Twitter Sentiment Geographic Index (TSGI) – A global well-being monitor published in Nature and used for UN's Sustainable Development Goals (SDGs) Today.
  - Harvard CGA's Geotweet Archive v2.0– Collection of 10 billion global geo-tagged tweets collected from 2010-2023.
- **Software Infrastructure deployed on HPC/Cloud**:
  - KNIME Business Hub: On-demand tool for workflow-based, reproducible analytics
  - ArcGIS Enterprise: Secure institutional GIS platform for advanced analytics
  - PostGIS: Open source CPU-based spatial database for data management
  - *Heavy.ai* & *HeavyIQ*: GPU-based real-time analytics and GenAI querying
  - Jupyter Notebook: Scalable Python-based spatial analysis

- **Open-Source Applications Developed**:
  - RINX: Containerized tool for information extraction from raster big datasets
  - RapidRoute: High-performance tool for finding shortest travel times on spatial big data
  - K-NN: High-performance tool for K-NN computation on spatial big data. Nature paper
  - Optipath : An efficient tool for optimal path and network analysis on raster big data

    - Billion Object Platform (BOP): API-accessible tool for real-time spatio-temporal analytics

2. Governance

- **Open Standards**: Uses open and community-compliant data and code formats
- **Responsible Use**: Reproducible, replicable and reusable workflows with containerized systems
- **Institutional Governance**: Integrated with FASRC/NERC for access, security, and compliance
- **Community Access**: Tools, code, and workshops are open-source on GitHub and YouTube
- **Community Participation**: Contributions to University Domain Working Group (DWG) and Cloud-Native Secure Web Gateway (SWG); to align with trusted, transparent governance goals

3. People

The SDI follows a user-centric design, prioritizing usability, scalability, and replicability. We offer open-access workshops on key topics such as Python for GIS/Data Science, ArcGIS Enterprise on NERC, and PostGIS-based geospatial data management. All software scripts are open-source on GitHub to encourage community use and collaboration. To support learning and showcase impact, we provide publications, presentations, and user testimonials. Our work has been featured by NSF/IUCRC, Harvard, MIT News, the Massachusetts Open Cloud Alliance, and the NERC website. Key user resources are available at the links below:

- Workshops
- Self-led Tutorials
- Publications
- Presentations
- Media Coverage
- Testimonials
- GitHub Repositories

Key Operational & Research Impact (Follow links for related publications)

- **UN SDGs Today**: TSGI used globally for well-being tracking and published in Nature
- **Public Health**: RapidRoute used in multiple health care access analysis projects ISPRS
- **Earth Observation Data**: RINX has been widely adopted in processing large scale EO data ISPRS
- **Politics**: K-NN has been used to study first ever individual level partisan segregation in the US Nature
- **Route Planning**: Optipath has been for optimal route planning for pipelines in the US ISPRS
- **Social Media**: BOP has been used for real-time analysis of a billion geotagged tweets OSGeo

Proposed Contributions to OGC

- Open source datasets such as *TSGI* for use by the larger community
- Open-source GIS tools developed by Harvard CGA: *RapidRoute, OptiPath, RINX, and BOP*
- Architecture designs, deployment strategies, and documentation from GIS system deployed
- Reproducible, replicable, and reusable templates and operational workflows
- Publicly accessible workshops and training materials on YouTube
- Peer-reviewed publications and conference presentations
- Open-source software scripts available on GitHub for community access and collaboration
- Institutional insights into SDI governance, sustainability, and education models