The Journal of Open Source Software

# Data-to-Science (D2S): An open-source ecosystem for collaborative geospatial data science research

**Minyoung Jung** [1], **Benjamin G. Hancock** [1], **Zhenyu C. Qian** [2], **Na Zhuo** [2], **Ziqian Gong** [2], **Jarrod S. Doucette** [3], **and Jinha Jung** [1¶]

**1** Lyles School of Civil and Construction Engineering, Purdue University **2** Rueff School of Design, Art, and Performance, Purdue University **3** College of Agriculture Research Services, Purdue University ¶ Corresponding author

**DOI:** 10.xxxxxx/draft

**Software**

- Review
- Repository
- Archive

**Editor:** James Gaboardi

**Reviewers:**

- @DeshpandeK
- @morrowcj

**Submitted:** 15 July 2025
**Published:** unpublished



## Summary

Recently, geospatial data has begun to be used across a wide range of research fields; however, its large size and unstructured nature present challenges in fostering cohesive collaboration among diverse disciplines. The **Data-to-Science (D2S)** ecosystem is an open-source package that offers an easy-to-use web application and additional client applications, specifically designed for managing comprehensive geospatial data and thereby supporting a broad range of research applications. The D2S web application serves as the primary interface of the D2S ecosystem, originally intended for archiving and visualizing geospatial data, particularly uncrewed aerial system (UAS) data, which often poses management challenges for individual researchers. To assist those who wish to comprehensively analyze both archived data within the D2S and other external geospatial datasets, the current D2S ecosystem also includes three additional components: the D2S Python module (*d2spy*), the QGIS plugin (*D2S Browser*), and a public STAC catalog accessible via both API and browser interface.

## Statement of Need

With advances in sensor technologies and the growing emphasis on open science, the use of geospatial data is rapidly expanding across various research fields (Breunig et al., 2020). Publicly available geospatial datasets, such as Landsat and Sentinel satellite data, are generally well-managed and distributed by their respective agencies. However, managing UAS-based geospatial data collected by individuals presents significant challenges because these data are often unstructured and large in size. Despite growing evidence for the usefulness of geospatial data in enabling multi-disciplinary research (Duarte et al., 2022; Ecke et al., 2022; Mohd Noor et al., 2018; Molina et al., 2023), the size and complexity of such data often hinder smooth collaborative research. As a result, there is a pressing need for infrastructure that simplifies the management of UAS-based geospatial data collected by individuals or small research groups. To address this need, D2S was developed as a web-based geospatial data management system designed to make handling and sharing such data more efficient and accessible.

## Data-to-Science Features

D2S uses a Python backend with a REST API built on open standards, connecting client applications to a PostgreSQL database with PostGIS for managing core application data and references to user-contributed datasets stored on the local file system. Full documentation on the system architecture is available in the GitHub README. The D2S frontend web application provides researchers with an intuitive interface, developed in collaboration with

a dedicated UI/UX team and informed by user interviews. The design efforts focused on creating an intuitive interface tailored for geospatial researchers, enabling side-by-side view comparisons and supporting analytic workflows through user-centered navigation and clarity of visual feedback. Built around this tailored interface, the current D2S web application (v1.0) offers five core categories of functionality, as outlined in the table below:

| Category | Description | Functionalities |
|---|---|---|
| **Catalog** | Dynamic spatiotemporal cataloging of geospatial data in a cloud-optimized format, enabling users to seamlessly access, browse, and visualize datasets without downloading them | ▪ Sorting geospatial data by time and location: 2D raster data (.tif), 3D point cloud data (.las, .laz), and vector data (.geojson, .shp)<br>▪ Raster and vector data visualization with symbology configuration<br>▪ Swipe comparison of geospatial data products across time or data type<br>▪ Visualizing 3D point cloud data |
| **Collaboration** | Sharing data with others within the D2S web application or by sending a sharable link | ▪ Managing teams and members<br>▪ Creating accessible links and/or QR codes for shared data<br>▪ Granting public access to data with no account or API key required |
| **Preprocessing** | Producing geospatial data products, such as dense point clouds, Digital Surface Models (DSM), and orthorectified images, from raw UAS data | ▪ Connecting to a photogrammetry pipeline based on open-source OpenDroneMap (ODM) via ClusterODM<br>▪ User configurable settings for the ODM pipeline |
| **Postprocessing** | Basic analysis of geospatial data products | ▪ Calculating vegetation indices (NDVI[1], ExG[2], VARI[3]) and hillshade from raster data<br>▪ Generating Digital Terrain Models (DTM) and Normalized Difference Height Models (NDHM) from point cloud data<br>▪ Zonal statistics based on vector data |
| **Publishing** | Publicly publishing data to the D2S STAC catalog | ▪ Generating and pushing STAC catalogs of datasets to be publicly published |

[1] NDVI = Normalized Difference Vegetation Index
[2] ExG = Excess Green Vegetation Index
[3] VARI = Visual Atmospherically Resistant Index

Furthermore, the Python module, *d2spy*, is available through PyPI (https://py.d2s.org/), and the QGIS plugin, *D2S Browser*, is also available at https://plugins.qgis.org/plugins/d2s_browser/. Notably, with *d2spy*, researchers can comprehensively analyze the geospatial

data stored within the D2S ecosystem as well as external public datasets, such as Landsat, by seamlessly integrating with other Python packages like *geemap* (Wu, 2020) and *leafmap* (Wu, 2021). Additionally, datasets, such as 3DEP and NAIP datasets, can also be incorporated into collective analyses as they are provided via the D2S STAC catalog (https://stac.d2s.org/) as part of the D2S ecosystem.

## Data-to-Science Tutorials

The D2S web application is containerized using Docker, enabling consistent deployment across both Linux servers using Docker Compose and cloud environments orchestrated with Kubernetes. A single Docker Compose file enables local deployment, while public Docker images and minimal configuration make D2S easy to integrate into cloud infrastructure. Step-by-step instructions are available in the GitHub README. The basic user manual for the D2S functionalities (as described in the table above) is available at https://docs.gdsl.org/data-to-science-user-manual with the publicly available sample data. A collection of example guides for using the D2S Python module, d2spy, is also available at https://py.d2s.org/guides/. In addition, a range of real-world application cases using the D2S ecosystem is provided as video tutorials at https://d2s.org/workshop.

## Acknowledgements

This work was partially supported by the Purdue Plant Science 2.0 Initiative, the Institute for Digital Forestry at Purdue, the PERSEUS grant (#2023-68012-38992) under USDA NIFA, the EFFICACI grant (#NR233A750004G044) under NCRS, and the National Agricultural Producers Data Cooperative (Award 2023-77039-41033; Sub-award 25-6231-0428-008) under USDA.

## References

Breunig, M., Bradley, P. E., Jahn, M., Kuper, P., Mazroob, N., Rösch, N., Al-Doori, M., Stefanakis, E., & Jadidi, M. (2020). Geospatial data management research: Progress and future directions. *ISPRS International Journal of Geo-Information*, *9*(2), 95. https://doi.org/10.3390/ijgi9020095

Duarte, A., Borralho, N., Cabral, P., & Caetano, M. (2022). Recent advances in forest insect pests and diseases monitoring using UAV-based data: A systematic review. *Forests*, *13*(6), 911. https://doi.org/10.3390/f13060911

Ecke, S., Dempewolf, J., Frey, J., Schwaller, A., Endres, E., Klemmt, H.-J., Tiede, D., & Seifert, T. (2022). UAV-based forest health monitoring: A systematic review. *Remote Sensing*, *14*(13), 3205. https://doi.org/10.3390/rs14133205

Mohd Noor, N., Abdullah, A., & Hashim, M. (2018). Remote sensing UAV/drones and its applications for urban areas: A review. *IOP Conference Series: Earth and Environmental Science*, *169*, 012003. https://doi.org/10.1088/1755-1315/169/1/012003

Molina, A. A., Huang, Y., & Jiang, Y. (2023). A review of unmanned aerial vehicle applications in construction management: 2016–2021. *Standards*, *3*(2), 95–109. https://doi.org/10.3390/standards3020009

Wu, Q. (2020). Geemap: A python package for interactive mapping with google earth engine. *Journal of Open Source Software*, *5*(51), 2305. https://doi.org/10.21105/joss.02305

Wu, Q. (2021). Leafmap: A python package for interactive mapping and geospatial analysis with minimal coding in a jupyter environment. *Journal of Open Source Software*, *6*(63), 3414. https://doi.org/10.21105/joss.03414