Contribution ID: **27** Type: **not specified**

Open Science in Action: Hands-On Spatial Data Training with Jupyter Notebooks

Tuesday, September 30, 2025 3:45 PM (20 minutes)

The Jupyter ecosystem for data-driven research

The Jupyter ecosystem offers a powerful and intuitive approach to modern data-driven research. It seamlessly integrates explanatory narratives, executable code, and visualisations within a single, web-based interactive environment, lowering the barrier to engaging with complex data workflows (Kluyver et al., 2016). A centralised infrastructure, such as the Jupyter4NFDI platform (used by initiatives like NFDI4Biodiversity), further enhances accessibility. These platforms provide scalable, cloud-based computing environments. Our training materials are compatible with the Jupyter4NFDI hub and can be accessed and run directly through a web browser. This seamless integration allows users to interact with the training materials in their browsers without needing complex local software installations or managing software dependencies. This makes advanced computational tools more accessible, supports reproducibility across different teams and institutions, and aligns with open science principles by promoting shared, standardised environments.

Furthermore, the training materials offer a valuable blueprint for creating similar open educational resources. The entire process from authoring interactive Jupyter Notebooks to publishing a polished, publicly accessible static website is streamlined using standard Git version control and managed through platforms like GitLab. This automated CI/CD (Continuous Integration/Continuous Deployment) pipeline allows the repository to be easily cloned and adapted, enabling others to efficiently build upon our framework for their teaching needs or develop training in different spatial science domains. This significantly lowers the barrier to creating and sharing high-quality, interactive learning experiences.

Showcasing reproducible spatial data science

This contribution presents Jupyter notebook training materials (see https://training.fdz.ioer.info) developed as a hands-on, open educational resource. They offer interactive tutorials and reproducible workflows for students, researchers, and practitioners conducting fully transparent, reproducible, narrative-embedded data-driven research focusing on spatial data (Dworczyk et al., 2025). The training materials provide users explicit guidance through the entire research output lifecycle (Higgins, 2008). Chapters cover topics ranging from understanding the FAIR (Findable, Accessible, Interoperable, Reusable) principles (Wilkinson et al., 2016) and good scientific practices for handling data to accessing, processing, analysing, and visualising spatial data and, ultimately, publishing findings. Each chapter of the training material features practical examples, including Python code, visualisations, and described processes. For instance, we demonstrate how to make code more broadly applicable and reusable by adhering to conventions like notebook parameterisation. This technique allows code to be easily adapted for different contexts, such as applying the same spatial data processing to a new region, thereby enhancing the reusability and impact of the shared work.

We use illustrative data examples, such as accessing biodiversity observations via the Global Biodiversity Information Facility (GBIF) API, contextualised with data relevant to the Lebendige Atlas der Natur Deutschlands (LAND), and retrieving raster-based land use and environmental data from the Monitor der Siedlungs- und Freiraumentwicklung (IÖR-Monitor). These datasets demonstrate reproducible research practices from programmatic data ingestion (e.g., API querying, handling various file formats) and pre-processing (e.g., data cleaning, geospatial operations like reprojection, clipping, and overlays) to advanced spatial analysis and visualisation (e.g., creating static and interactive maps).

Crucially, the material emphasises the creation of a "Replication Package", a versioned archive containing notebooks, scripts, data subsets, generated outputs, and a detailed README. This ensures the work can be fully understood, reproduced, and cited (e.g., via deposition in repositories like ioerDATA or Zenodo). While accessible via Jupyter4NFDI, we also detail using the versioned Carto-Lab Docker container. A containerised setup can encapsulate the precise software environment, including specific versions of Python and numerous cartographic and geospatial libraries, to guarantee full computational reproducibility and long-term preservation.

The collaborative online framework proved invaluable for efficient teamwork and knowledge sharing. We leveraged Git for robust version control and issue tracking to manage development and ensure incremental improvements. Real-time collaborative editing within Jupyter sessions enabled direct interaction and mentoring between junior and senior team members. This entire ecosystem, supported by a centrally maintained IT

infrastructure, streamlined the research and publication process by offloading technical burdens from individual researchers and simplifying the integration of new collaborators.

Our training materials showcase practical tools and workflows that enhance data stewardship and foster robust reproducibility in spatial science. By providing open-source code, example data, detailed methodological explanations within an executable format, and explicit guidance on creating citable "Replication Packages," these resources actively support the creation and dissemination of FAIR research outputs. The material is designed for open dissemination and re-use, promoting learning, critical engagement with data-driven methods, and the widespread adoption of open science practices across the research community.

Acknowledgement

The development of these training materials was made possible through support from NFDI4Biodiversity (phase II).

Keywords

Jupyter Notebook, Jupyter Book, Open Science, Reproducibility, FAIR Data, Spatial Data, Geodata, Python, NFDI4Biodiversity, Biodiversity Informatics, Environmental Data Science, Training, Open Educational Resources, Carto-Lab Docker, Replication Package.

References

Dworczyk, C., Dunkel, A., Rafiei, F., Syrbe, R.-U., 2025. Exploring Spatial and Biodiversity Data with Python and JupyterLab. https://training.fdz.ioer.info

Higgins, S., 2008. The DCC Curation Lifecycle Model. International Journal of Digital Curation 3, 134–140. https://doi.org/10.2218/ijdc.v3i1.48

Kluyver, T., Ragan-Kelley, B., Pérez, F., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J., Grout, J., Corlay, S., Ivanov, P., Avila, D., Abdalla, S., Willing, C., Jupyter Development Team, 2016. Jupyter Notebooks –a publishing format for reproducible computational workflows, in: Positioning and Power in Academic Publishing: Players, Agents and Agendas. IOS Press, pp. 87–90. https://doi.org/10.3233/978-1-61499-649-1-87

Wilkinson, M.D., et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018. https://doi.org/10.1038/sdata.2016.18

Abstract

Talk

Authors: DELLA CHIESA, Stefano (Leibniz Institute of Ecological Urban and Regional Development); DUNKEL, Alexander (Leibniz Institute of Ecological Urban and Regional Development); DWORCZYK, Claudia (Leibniz Institute of Ecological Urban and Regional Development); SYRBE, Ralf-Uwe (Leibniz Institute of Ecological Urban and Regional Development)

Presenters: DELLA CHIESA, Stefano (Leibniz Institute of Ecological Urban and Regional Development); DUNKEL, Alexander (Leibniz Institute of Ecological Urban and Regional Development); DWORCZYK, Claudia (Leibniz Institute of Ecological Urban and Regional Development); SYRBE, Ralf-Uwe (Leibniz Institute of Ecological Urban and Regional Development)

Session Classification: Talks