Operationalising the Data-to-Knowledge Package Concept: Visualising FAIR Workflows Across Three Environmental Use Cases

Sadra Matmir [1], Carsten Keßler [1], and Mehrad Moradipour [1]

[1]Department of Geodesy, Bochum University of Applied Sciences, Bochum, Germany

Correspondence: Sadra Matmir (sadra.matmir@hs-bochum.de)

Abstract. Reproducibility in computational geoscience has improved substantially through open data policies and shared source code; however, structured reuse of analytical workflows across domains remains challenging. While data and scripts may be available, they are often difficult to re-execute due to missing documentation, unstandardised environments, or insufficient workflow orchestration. The Data-to-Knowledge Package (D2KP) concept addresses this limitation by integrating FAIR data, modular toolboxes, executable workflows, and virtual research environments into reusable research meta-objects. This contribution presents three heterogeneous D2KPs implemented within the AquaINFRA research infrastructure for marine and freshwater science: (1) dasymetric population refinement for the Elbe river basin, (2) ensemble environmental outlier detection using the *specleanr* R package, and (3) reproducible spatiotemporal trend detection for water transparency analysis in the Gulf of Riga. The poster visualises the AquaINFRA infrastructure architecture, the internal workflow structures of each use case, and the resulting analytical outputs. By embedding domain-specific analyses into a shared infrastructure backbone, the D2KP approach demonstrates how reproducible research can be transformed into interoperable and reusable geospatial services.

Submission Type. Infrastructure; Model; Analysis

BoK Concepts. *[WB1]* Web services; *[DM1]* Foundations for Data Modelling Storage and Exploitation; *[GC4]* Open Science; *[AM7]* Spatial statistics

Keywords. AquaINFRA, Data-to-Knowledge Package; reproducible workflows; Galaxy; dasymetric mapping; environmental trend detection

1 Introduction

Reproducibility has become a central topic across scientific disciplines (Goodman et al., 2016). In GIScience, analytical pipelines often involve heterogeneous datasets, spatial transformations, statistical procedures, and software environments. Even when code and data are openly shared, re-executing workflows remains difficult if computational environments and orchestration logic are not clearly defined. The FAIR principles emphasise that research artefacts must be Findable, Accessible, Interoperable, and Reusable (Wilkinson et al., 2016).

The work presented here is developed within the AquaINFRA project, which aims to build an EOSC-based virtual research environment equipped with FAIR multidisciplinary data and services to support marine and freshwater scientists and stakeholders in restoring healthy oceans, seas, coastal and inland waters. AquaINFRA enables cross-domain and cross-country collaboration through discovery mechanisms and spatiotemporal analysis services implemented in Virtual Research Environments.

Within this context, Konkol et al. (2025) introduced the Data-to-Knowledge Package (D2KP) concept. A D2KP links a reproducible basis (data and toolbox functions) with executable workflows and structured deployment environments. This poster demonstrates how the D2KP concept can be operationalised across three heterogeneous environmental use cases while maintaining a unified infrastructural backbone.

2 AquaINFRA Infrastructure Architecture

Figure 1 provides the conceptual anchor of the poster. The architecture consists of three interconnected layers:

(1) Discovery Layer through Data Discovery and Access Service (DDAS) and AquaINFRA Interaction Platform (AIP): datasets, toolboxes, and workflows are indexed and searchable.

(2) Execution Layer – Virtual Research Environment (VRE): workflows are executed within a Galaxy-based infrastructure (Galaxy Community, 2024), ensuring environment consistency and reproducible orchestration.

(3) Archiving Layer – Zenodo: workflow definitions, datasets, and derived artefacts are persistently stored with DOIs.

This layered model aligns with established principles for transparent geoscientific workflows (Gil et al., 2016) and modern workflow systems (Di Tommaso et al., 2017). The infrastructure diagram visually demonstrates how heterogeneous analyses are embedded within a shared reproducibility framework.

3 Three Data-to-Knowledge Packages

3.1 Dasymetric Population Refinement in the Elbe River sub-basins

Hydrological assessments require population data aligned with river catchments, whereas official statistics are reported at administrative levels such as NUTS3. This spatial mismatch limits accurate estimation of anthropogenic pressure.

The Elbe D2KP applies dasymetric redistribution methods using land cover information (Bonnevie et al., 2024). The workflow integrates Eurostat demographic statistics, GISCO geometries, CORINE Land Cover datasets, and Elbe sub-basin boundaries. Processing steps include harmonisation, spatial filtering, area-weighted redistribution, and aggregation to hydrological units. The workflow structure is illustrated in Fig. 2.

3.2 Specleanr: From R Package to FAIR Workflow

Environmental outliers in species occurrence data can bias ecological inference (Basooma et al., 2025). The *specleanr* R package integrates 20 detection methods spanning ecological range-based, univariate, and multivariate approaches.

Within AquaINFRA, the standalone R package is transformed into an executable Galaxy workflow. Occurrence records are harmonised, environmental predictors extracted, and ensemble detection methods classify records into graded outlier categories.

This transformation illustrates how research-oriented software becomes an interoperable infrastructure component embedded in a reproducible workflow environment.

3.3 Spatiotemporal Trend Detection in Water Transparency in the Gulf of Riga

The Gulf of Riga use case investigates whether long-term changes in water transparency can be detected in this semi-enclosed Baltic Sea basin. The D2KP integrates in situ Secchi depth measurements from Latvian monitoring programmes with Baltic Sea Assessment Unit polygons.

The workflow performs spatial aggregation of point measurements to assessment units, seasonal grouping, and calculation of mean transparency per unit and season. Time series are filtered based on temporal coverage, and missing values are interpolated where appropriate. The final analytical step applies the non-parametric Mann–Kendall test to detect significant monotonic trends in transparency.

Executed within the Galaxy-based VRE and archived via Zenodo, the workflow ensures parameter transparency and reproducible execution. The poster visualises the workflow graph alongside resulting trend maps and Kendall's tau charts, illustrating the transformation from monitoring data to statistically interpretable environmental indicators.

3.4 Data and Software Availability

All three use cases are implemented as AquaINFRA Data-to-Knowledge Packages (D2KPs) and are archived on Zenodo with persistent identifiers:

- Gulf of Riga transparency trend detection: zenodo:17175368
- Ensemble outlier detection (*specleanr*): zenodo:17175591
- Elbe dasymetric population refinement: zenodo:18607604

Each record contains the workflow definition, input data references, and execution environment specification required for reproducible re-execution. Workflows are deployable within the AquaINFRA Galaxy-based Virtual Research Environment.

Declaration of Generative AI in Writing

Generative AI tools were used exclusively for language editing and structural refinement. All scientific content and interpretations remain the responsibility of the author.

Acknowledgements

The authors are affiliated with the AquaINFRA project, a Horizon Europe project funded by the European Union under Grant Agreement No. 101094434 (DOI: 10.3030/101094434).

References

Basooma, A., Schmidt-Kloiber, A., Domisch, S., Torres-Cambas, Y., Smederevac-Lalić, M., Bremerich, V., Meulenbroek, P., Tschikof, M., Funk, A., Hein, T., and Borgwardt, F.: specleanr: An R package for automated flagging of environmental outliers in ecological data for modeling workflows, Ecography, https://doi.org/10.1002/ecog.08221, 2025.

Bonnevie, I. M., Hansen, H. S., and Schrøder, L.: Dasymetric algorithms using land cover to estimate human population at smaller spatial scales, ISPRS International Journal of Geo-Information, 13, 427, https://doi.org/10.3390/ijgi13120427, 2024.

Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., and Notredame, C.: Nextflow enables reproducible computational workflows, Nature Biotechnology, 35, 316–319, https://doi.org/10.1038/nbt.3820, 2017.

Galaxy Community: The Galaxy platform for accessible, reproducible and collaborative data analyses: 2024 update, Nucleic Acids Research, 52, W83–W94, https://doi.org/10.1093/nar/gkae410, 2024.

Gil, Y., David, C. H., Demir, I., Essawy, B. T., Fulweiler, R. W., Goodall, J. L., Karlstrom, L., Lee, H., Mills, H. J., Oh, J.-H., Pierce, S. A., Pope, A., Tzeng, M.-H., Villamizar, S. R., Yu, X., and Yu, Y.: Toward the geoscience paper of the future: Best practices for documenting and sharing research from data to software to provenance, Earth and Space Science, 3, 388–415, https://doi.org/10.1002/2015EA000136, 2016.

Goodman, S. N., Fanelli, D., and Ioannidis, J. P. A.: What does research reproducibility mean?, Science Translational Medicine, 8, 341ps12, https://doi.org/10.1126/scitranslmed.aaf5027, 2016.

Konkol, M., Labuce, A., Domisch, S., Buurman, M., and Bremerich, V.: Encouraging reusability of computational research through Data-to-Knowledge Packages: A hydrological use case, Open Research Europe, 5, 123, https://doi.org/10.12688/openreseurope.20221.3, 2025.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., et al.: The FAIR Guiding Principles for scientific data management and stewardship, Scientific Data, 3, 160 018, https://doi.org/10.1038/sdata.2016.18, 2016.

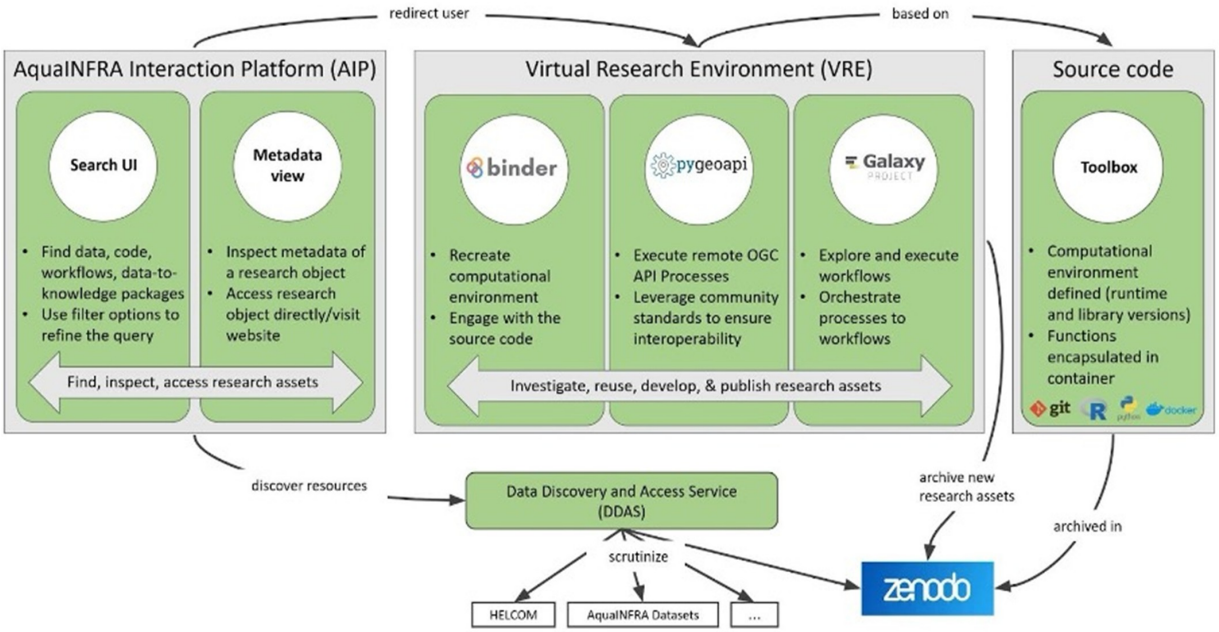

Figure 1. The schematic overview of the AquaINFRA infrastructure architecture.

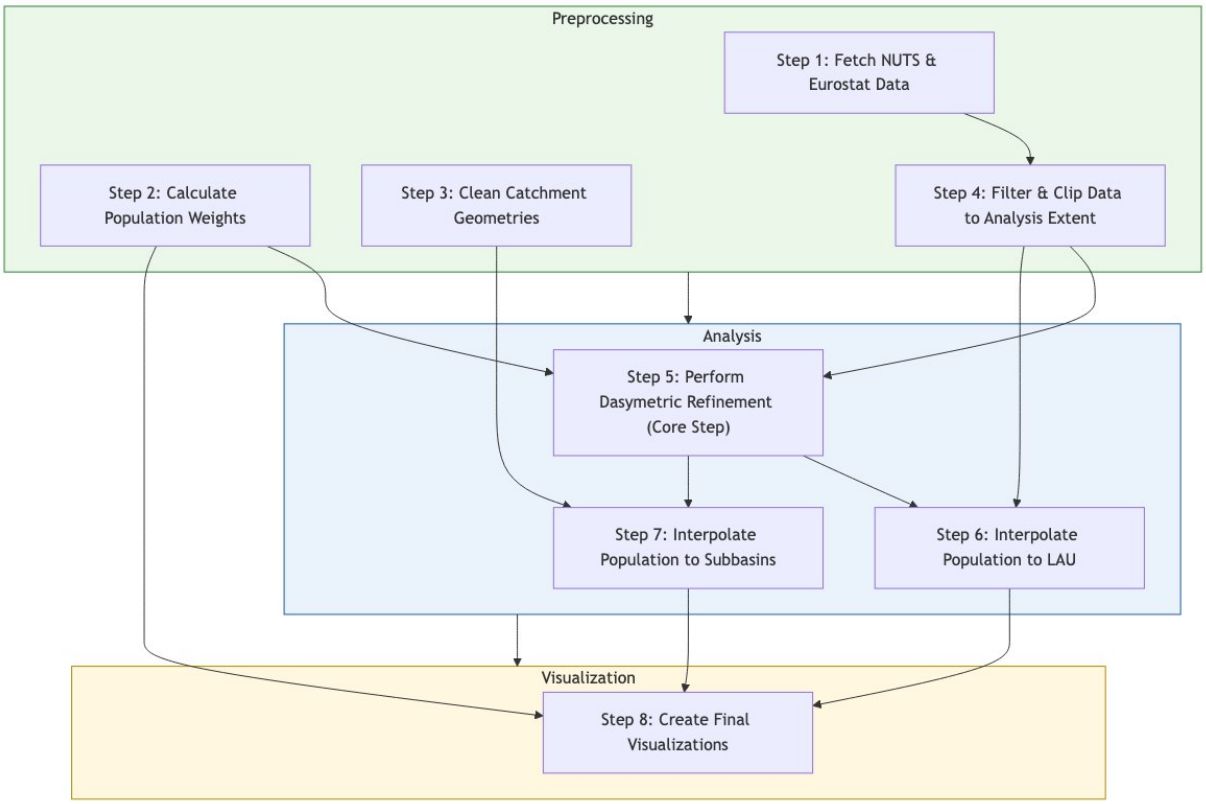

Figure 2. The workflow structure in Elbe D2KP.