https://doi.org/10.5194/egusphere-2025-4832
Preprint. Discussion started: 7 October 2025
© Author(s) 2025. CC BY 4.0 License.

A database-driven research data framework for integrating and processing high-dimensional geoscientific data

Dennis Handy[1], W. Marijn van der Meij[1], Mirijam Zickel[1], and Tony Reimann[1]
[1]Institute of Geography, University of Cologne, 50923 Cologne, Germany
Correspondence: Dennis Handy (dhandy1@uni-koeln.de)

Abstract. This paper introduces a modular research data framework designed for geoscientific research across disciplinary boundaries. It is specifically designed to support small research projects, that need to adhere to strict data management requirements from funding bodies, but often lack the financial and human resources to do so. The framework supports the transformation of raw research data into scientific knowledge. It addresses critical challenges, such as the rapid increase in the volume, variety and complexity of geoscientific datasets, data heterogeneity, spatial complexity, and the need to comply to the FAIR (Findable, Accessible, Interoperable, and Reusable) principles. The approach optimises the research management process by enhancing scalability and enabling interdisciplinary integration. It is adaptable to evolving research requirements and it supports various data types and methodological approaches, such as machine learning and deep learning, that have high requirements on the used data and their formats. A case study in Western Romania presents the data framework's application in an interdisciplinary geoarchaeological research project by processing and storing heterogeneous datasets, demonstrating its potential to support geoscientific research data management by reducing data management efforts, improving replicability, findability and reproducibility and streamlining the integration of high-dimensional data.

1 Introduction

Recent technological advancements have driven rapid growth in the volume, variety, and complexity of geoscience research data (Vance et al., 2024) that fuels new data-driven approaches to generate scientific knowledge (Gahegan, 2020) and requires a considerable investment in equipment and expertise, as it accumulates large and heterogeneous volumes of data stored in various formats and databases (Gärtner et al., 2001). This trend coincides with the increasing, fundamental discussion about how scientific data should be published in order to ensure its reusability (e.g. Faniel and Jacobsen, 2010; Murillo, 2019). In 2016, Wilkinson et al. introduced the FAIR principles - Findability, Accessibility, Interoperability, and Reusability - to enhance the reuse of research data and highlighted the importance of effective data management. The authors stress the need for a cultural shift in research data management but also specifically highlight the importance of computational capabilities in data-rich research environments (Wilkinson et al., 2016). The growing complexity of research data and the increasing publication requirements pose new challenges, particularly for smaller projects that lack their own research data infrastructure. This results in a growing demand for straightforward methods of storing and processing data.

While the enormous growth of data volumes and the fulfilment of the Fair Principles are not exclusive to the geosciences,

https://doi.org/10.5194/egusphere-2025-4832
Preprint. Discussion started: 7 October 2025
© Author(s) 2025. CC BY 4.0 License.

geoscientific data and geoscientific research data management have some unique characteristics and challenges that require the development of tailored methods for storage and processing. As Geosciences analyse complex, coupled processes across different spatial and temporal scales, the heterogeneous nature of its research objects often leads to a high-dimensional parameter space; a high number of potentially interdependent variables (Degen et al., 2020). The dimensionality and heterogeneity of geoscientific data requires significant computing resources, and requires multiple processing steps due to its complexity (Li et al., 2015). *Spatial Dependence*, *Spatial Heterogeneity* and *Scale Dependence* further challenge storage and analysis of spatial data. Understanding these spatial properties is essential, as they necessitate adapted methods and workflows throughout the entire data lifecycle: Spatial dependence refers to the autocorrelation of spatial data, requiring specialised methods to analyse spatial relationships, whereas spatial heterogeneity refers to non-stationary spatial patterns influenced by spatial anisotropy or overlapping processes across different scales (Nikparvar and Thill, 2021). Interpreting and sharing spatial data relies on the traceability of the coordinate reference system (CRS) for its location component. Without being able to reproduce the CRS, the data might be plotted in an incorrect location, or the accuracy, precision, and scale are misjudged, leading to misleading results. For this reason, spatial data must have explicit locational information in its metadata (Van Den Brink et al., 2018). Spatial data thus requires reflective methodological approaches to ensure meaningful analysis such as spatial data workflows, where the transformation of raw geospatial inputs into scientifically interpretable outputs depends on addressing the unique properties of spatial data. This calls for systems that account for spatial complexity at every stage of the workflow.

These trends and challenges in geoscientific research data management underscore the urgent need for structured approaches to data storage and analysis, particularly in small and interdisciplinary research projects that must comply with FAIR principles and funding requirements (e.g. Deutsche Forschungsgemeinschaft, 2022). Many projects do not establish sustainable data management structures due to limited funding and expertise. Though specialized databases and methods exist for particular geoscientific subfields, a comprehensive, interdisciplinary approach has been missing, integrating research data from all areas (Nordsiek and Halisch, 2024). While research groups, projects, and laboratories may implement backup procedures or repositories to prevent the physical loss of data, the risk of losing information regarding its existence (Gärtner et al., 2001; Murillo, 2019), its usability and discoverability remain significant as the data's publication not merely implies its accessibility. Metadata might be incomplete or missing completely, links might be broken, or the information needed to make the data usable is missing (Tedersoo et al., 2021). The failure of making research data FAIR is not just an inconvenience but has a measurable financial impact. Non-standardised research data costs billions of euros every year (Klöcking et al., 2023; European Commission et al., 2018).

Databases partially address the aforementioned challenges but focus primarily on static data storage rather than the researchers' entire workflows, the data's provenance (Mitchell et al., 2022) and how the data evolves through its lifecycle, from generation and transformation to storage, analysis and reuse, throughout a research project. Therefore, we propose broadening the perspective of geoscientific research data management by modelling a framework for geoscientific research data encompassing the entire data lifecycle. In contrast to established methods, rather than considering FAIR principles from the perspective of publication, we consider these principles proactively throughout the entire data lifecycle. We aim to:

https://doi.org/10.5194/egusphere-2025-4832
Preprint. Discussion started: 7 October 2025
© Author(s) 2025. CC BY 4.0 License.

1. address challenges arising from the rapid growth in the volume and complexity of data. To this, we provide a standardised approach to store scientific data throughout a research project's lifecycle by implementing data storage systems specifically designed for geoscientific data and providing online interfaces to access the data. By requiring comprehensive metadata, we are aiming to ensure the findability and reusability for future use.

2. address the challenges associated with processing increasingly intricate data workflows, we introduce data pipelines. These pipelines formalise recurring data workflows in code, thereby ensuring reproducibility and scalability while minimising errors.

This requires close consideration of the specific characteristics of geoscientific data, which pose pronounced challenges due to their spatial, temporal, and computational complexity, as well as the understanding of requirements specific to geoscientific research. To this end, we first identify the user requirements (Chapter 2). Then, we will describe how we used these requirements to design and implement the framework (Chapter 3). We test and demonstrate the practical application of our framework using a geoarchaeological project in western Romania (Chapter 4). Finally, we discuss how the framework addresses the outlined challenges and provide an outlook on future developments (Chapter 5).

2 Challenges and Requirements for FAIR Geoscientific Data

A software framework designed to support researchers should not only reflect theoretical concepts of scientific practices, such as the FAIR principles, but should also consider financial, technical, organisational and practical limitations that researchers face in their everyday work. Therefore, to identify the key challenges and requirements of research data management in geosciences, we gathered the following insights on existing challenges in research data management, current best practices, and potential solutions through a literature review. Our review reveals that the identified challenges are not specific to the investigated environment but reflect systemic issues in geoscientific data management:

Volume and complexity Geoscience data is experiencing a significant growth in data volume and complexity (e.g. Klöcking et al., 2023; Vance et al., 2024), including large multi-dimensional and spatio-temporal datasets (Degen et al., 2020; Li et al., 2015). The volume and complexity of data poses major challenges for computer-based methods and data management approaches (Li et al., 2015; Liakos and Panagos, 2022).

Hetereogeneity and accessibility Geoscientific data are inherently diverse and highly variable (Klöcking et al., 2023; Nordsiek and Halisch, 2024). They are often stored in incompatible and frequently inaccessible individual computers or local databases, which hinders their discovery and reuse without considerable effort (Klöcking et al., 2023). Although relational databases enable secure and structured storage, they often have practical limitations. Data silos, isolated subject-specific storage, make cross-domain evaluation difficult, and the complexity of data models causes users to develop error-prone workarounds (Kingdon et al., 2016). The trend towards numerous general-purpose data repositories, while offering availability, can exacerbate discovery and reusability issues because they often don't integrate or harmonize deposited data (Wilkinson et al., 2016). Geoscientific data is often stored in different, often isolated formats (GIS, databases,

https://doi.org/10.5194/egusphere-2025-4832
Preprint. Discussion started: 7 October 2025
© Author(s) 2025. CC BY 4.0 License.

special software, proprietary formats), which leads to redundancies, integration problems, and data loss. The heterogeneity of standards and tools makes it difficult to perform a holistic analysis, even though combining different data sources could provide valuable insights. A central, networked solution is still lacking. Once projects end or employees change jobs, valuable data is often lost. Thus, data storages become "data cemeteries" (Gärtner et al., 2001).

Interoperability A persistent challenge is the absence of common global standards for data sharing and metadata (Klöcking et al., 2023; Nordsiek and Halisch, 2024). Many datasets lack sufficient metadata, use inconsistent spatial reference fields (Klöcking et al., 2023; Van Den Brink et al., 2018) and semantic differences between datasets create barriers to interoperability. This poses a significant challenge to the fulfilment of the FAIR principles (Lannom et al., 2020). The quality of data varies, and scientists expend effort to assess whether it is relevant, understandable and reliable (Faniel and Jacobsen, 2010; Murillo, 2019). The lack of information about research methods, instrumentation and provenance (i.e. origin and processes) hinders data reuse (Murillo, 2019; Nordsiek and Halisch, 2024). The lack of interdisciplinary standardisation further challenges true interoperability (e.g. Nordsiek and Halisch, 2024). For instance, while the United States Department of Agriculture (USDA) considers graine sizes from 0.002 mm to 0.05 mm as silt (Soil Science Division Staff, 2017), the World Reference Base for Soil Resources (WRB) draws the boundaries by 0.002 mm and 0.063 mm (IUSS Working Group WRB, 2022).

Workflows complexity Intricate data workflows are now the norm, following a dramatic shift in the computational landscape caused by the rise of data-intensive sciences (Mork et al., 2015). Data pipelines increase the overall efficiency of data flow from source to destination by automating the process and reducing the level of human involvement required (Raj et al., 2020). Particularly due to the integration of machine learning methods at multiple levels, data processing pipelines are even increasing in volume, velocity and complexity (Wozniak et al., 2022). Even for specialists, the process of moving from basic science to interpretation, including the selection of parameter values, model structure, assumptions, code implementation, and output generation, is complex (Mitchell et al., 2022)

Institutional barriers Scientists often use specific or customised tools and proprietary data formats, meaning that enforcing a single, uniform system for research data management will not work in a heterogeneous scientific environment (Fitschen et al., 2019). Data sharing is often hindered by a perceived burden of documentation, lack of time, fear of data loss, privacy concerns, legal issues (Faniel and Jacobsen, 2010; Tedersoo et al., 2021) and traditional academic incentives focusing primarily on publishing papers rather than on properly curating and sharing datasets (Vance et al., 2024). Furthermore, scientists sometimes abandon traditional, formally structured databases in favour of more ad hoc solutions, such as spreadsheets (Thomer and Wickett, 2020).

The analysis of actual scientific practice reveals that a large proportion of the identified challenges are already being discussed. However, a crucial challenge is that the implementation of a framework depends on the highly individual nature of science, including the specific combination of methods, available laboratory equipment and the state of the IT infrastructure. These challenges highlight the need for a modular, discipline-specific framework that strikes a balance between standardisation for

https://doi.org/10.5194/egusphere-2025-4832
Preprint. Discussion started: 7 October 2025
© Author(s) 2025. CC BY 4.0 License.

interoperability and flexibility for disciplinary requirements. Such a framework should automate provenance tracking and metadata generation to minimise manual effort, and integrate FAIR principles into daily workflows without disrupting existing practices.

3 Design and implementation

Our system architecture reflects the entire life cycle of research data within a project, from initial acquisition during fieldwork and laboratory analyses to final evaluation. To this end, it prioritises continuous data processing by implementing data pipelines rather than focusing solely on structured storage in databases. These pipelines consider the ingestion of raw data, its processing, transformation, storage and - eventually - automated retrieval and analysis (Figure 1). The framework consists of two separate but linked modules: an online transaction processing (OLTP) module based on a database implementing a normalised, relational data model (Chapter 3.2, Figure 2 a), and an online analytical processing (OLAP) module based on a database implementing a denormalised data model called star schema (Chapter 3.3, Figure 2 b). However, users are shielded from the underlying technical complexity and are provided with a convenient user interface they can use to conveniently create, retrieve, update and delete objects in the relational database. Following the initial data capture, an ETL (extract, transform, load) pipeline is triggered. This critical process involves extracting data from the relational database, transforming it into a suitable format for analysis and loading it into the OLAP database. This, in turn, forms the basis for analytical pipelines that result in either the user interface or the file system. It should be noted that the implementation of the framework should adhere to the specific conditions set by the university's IT infrastructure.

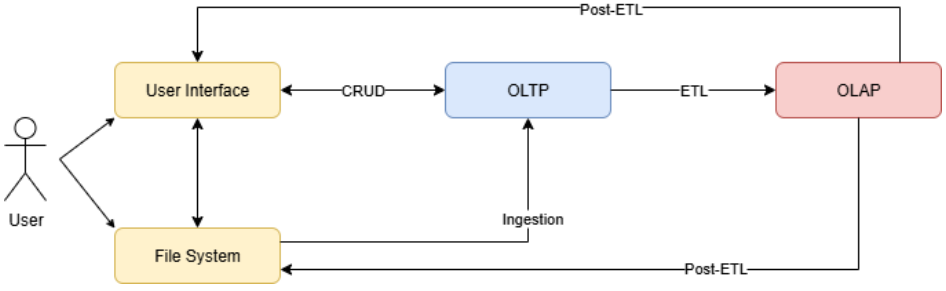

Figure 1. The system architecture illustrates the interaction between user, data storage and data pipelines within the framework. Users initiate the process by uploading data via the user interface or the file system (yellow boxes). The data is processed, validated and linked in the online transaction processing database (OLTP, blue box). An extract, transform, load (ETL) pipeline extract data from the OLTP database, transform it and load it into the online analytical processing database (OLAP, red box). Post-ETL pipelines automatically generate analyses, data visualizations, or specific data products, which are then made available through the user interface (e.g., dashboards) or exported to the file system (e.g., CSV, PDF reports).

https://doi.org/10.5194/egusphere-2025-4832
Preprint. Discussion started: 7 October 2025
© Author(s) 2025. CC BY 4.0 License.

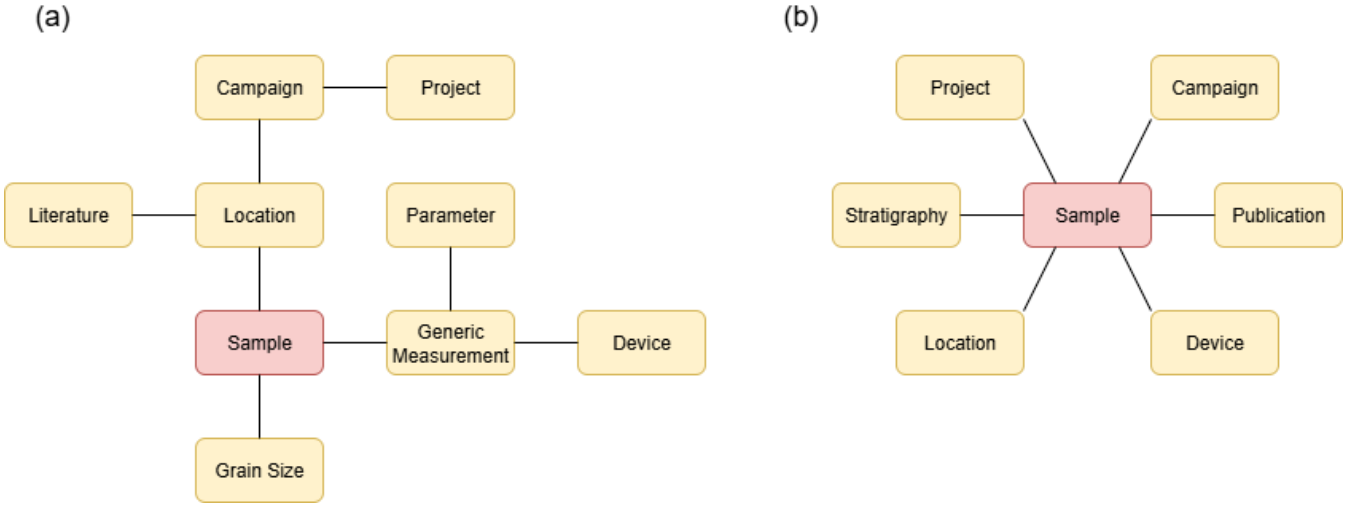

Figure 2. A simplified relational schema (a) as used as in the relational OLTP database and a star schema (b) as used in the denormalised OLAP database. The normalised relational schema (a) organises data across logically linked tables to minimise redundancy and ensure data integrity. Here, the relationships are more granular: a sample is linked to a location, which is then linked to a campaign and, finally, to a project. Although this structure makes analytical queries more complex, it is ideal for transactional operations and ensures consistency in operational workflows. By contrast, the central fact table in the star schema (b) stores quantitative measurements and is directly linked to dimension tables that contain descriptive information on projects, locations, campaigns, and different data types. This design enables efficient analytical queries by denormalising the data structure, which improves query performance for aggregations and slicing operations. In contrast, the normalised relational schema represents a more complex perspective on the relationships between entity types that organises data in logically linked tables to avoid redundancy.

3.1 User interface

Our framework provides a web-based user interface (Figure 3) which supports conveniently managing geoscientific data. The interface enables data capture, validation, and exploration through standardised forms for sample and analysis data (e.g., grain size, luminescence dating) and file upload functionality for raw data (e.g., data tables from laboratory devices). Data entry forms follow international standards (e.g., World Reference Base for Soil Resources) and laboratory-specific templates (e.g., for luminescence dating parameters) and include dropdown menus for controlled vocabularies, mandatory fields and validation rules. Users can assign a sample to a location, link an analysis (e.g., luminescence age) to a sample and its laboratory dataset and document fieldwork contexts (e.g., campaign, stratigraphic layer). The technical connection between the entities is directly checked and established.

https://doi.org/10.5194/egusphere-2025-4832
Preprint. Discussion started: 7 October 2025
© Author(s) 2025. CC BY 4.0 License.

Figure 3. An example of the user interface that show (a) an overview of all available measurements from different projects that are stored in the database, grouped in sedimentological and geochronological measurements, (b) the filtering feature to extract specific data from the data base and (c) the resulting data from the query, in this case geochronological data from the Toboliu project.

3.2 Online transaction processing (OLTP) module

The Online Transaction Processing (OLTP) module consists of a relational database, serving as the long-term data archive and a user interface that intermediaries between the database and users, abstracting the complexity of the data model and the technical implementation. The module was implemented with the Python framework Django and the relational database management system PostgreSQL. Our OLTP module is tailored to the needs of geomorphological, geoarchaeological and geochronological research at the Institute of Geography of the university of Cologne, but can be easily adapted to other systems and requirements. The underlying database implements a normalised relational data model with sediment or soil samples as central entities (Figure 2 a). The relational model is an approach to organising data in unique, logically linked tables, while minimising redundancies through normalisation. It provides users with a set-based query language while concealing the intricacies of physical data storage (Codd, 1970). Normalisation involves organising attributes and tables to minimise data redundancy. This is achieved by dividing relationships into smaller tables that are linked by foreign keys (Kingdon et al., 2016).

Our model builds upon the conceptual model by Nordsiek and Halisch (2024), which was designed as a modular, interdisciplinary geoscientific laboratory datbase (Nordsiek and Halisch, 2024). However, it extends the scope of Nordsiek and Halisch (2024) by including comprehensive fieldwork documentation alongside laboratory data, and by adapting to the particular research setting. The sample entity acts as the core table, linking to related tables that contain information on locations and analyses. Each location is linked to fieldwork details, including field campaigns and associated projects to preserve contextual

https://doi.org/10.5194/egusphere-2025-4832
Preprint. Discussion started: 7 October 2025
© Author(s) 2025. CC BY 4.0 License.

information, such as sampling conditions, for subsequent analysis. The tables used for the analysis are directly linked to the sediment samples. These include specifically modelled analyses, such as grain size analysis and MicroXRF measurements, as well as a generic measurement table. The latter allows the measurement logic to be extended flexibly without requiring modifications to the underlying data model.

Our model distinguishes between internally generated and externally sourced data: If a dataset was generated within the organisation, a location must be linked to a specific project. If the dataset was taken from a publication, the publication must be stored as a data source within the database. Access controls restrict non-published data to project members, while published data is accessible to all users.

Our model currently supports analytical methods, such as sedimentological analyses like grain size distribution, geochemical composition, paleobotanical studies, geochronological dating, and modern imaging techniques such as micro-X-ray fluorescence (MicroXRF) element mapping. Metadata fields document the method, equipment, and procedural steps. The laboratory infrastructure is systematically documented, including the equipment used, calibration procedures, and standard procedures.

3.3 Online Analytical Processing (OLAP) module

Complex queries, such as those requiring joins across multiple tables, can affect performance as the size and structural complexity of the dataset grow. To address this issue, the OLAP module of the framework uses a multidimensional star schema to consolidate normalised tables into simpler structures with controlled redundancy to optimise query performance and data integration (Chaudhuri and Dayal, 1997; Kingdon et al., 2016). In Figure 2 b, Sedimentological samples form the central fact table of the star schema, storing quantitative measurements contextualised by descriptive dimension tables (e.g., campaign, location, project) within a spatial and conceptual framework. Users can navigate from aggregated data to detailed sample records including geographic coordinates, time periods and measurement methods. The module supports filtering by criteria such as location, time or analytical method. For instance, a query could conveniently extract all soil textures from European loess deposits that have been analysed over the past decade. The module is currently primarily based on DuckDB, an open source in-process database designed for analytical workloads (Raasveldt and Mühleisen, 2019). It is closely integrated with Dagster for data orchestration (see chapter 3.4).

3.4 Data Pipelines

The increasing complexity of geoscientific data workflows, characterised by heterogeneous data sources, multiple processing steps, and interdisciplinary requirements demands automated solutions for data processing. To address these challenges, we have implemented a pipeline-based architecture that standardises recurring workflows in code to ensure reproducibility and minimise manual labour and errors. Data pipelines are automated, interconnected processes designed to manage dynamic data flows from source to destination. They extract, transform, validate and combine data, with each stage's output feeding the next. Pipelines can handle various types of data (continuous, intermittent or batch) and facilitate real-time monitoring, error detection and correction. Their applications range from data storage to visualisation or machine learning (ML) models (Raj et al., 2020).

https://doi.org/10.5194/egusphere-2025-4832
Preprint. Discussion started: 7 October 2025
© Author(s) 2025. CC BY 4.0 License.

Data workflow automation occurs at two levels within the framework. The OLTP module contains a basic data pipeline for automation of transactional operations, processing and transforming data during storage (e.g. via uploads through the user interface) and retrieval (e.g. for queries or exports). In contrast, more advanced data pipelines are logically and infrastructurally separated from the OLTP module. They run on the same technical infrastructure as the OLAP module and are orchestrated using Dagster. Dagster is an open-source data orchestrator that is designed for constructing, operating and observing efficient, transparent and replicable data pipelines (Picatto et al., 2024). These pipelines manage the flow of data from the relational database through extraction, transformation and loading into the OLAP. This is followed by further transformation into analytical products, such as tables, visualisations or machine learning inputs. Pipelines can be triggered either by user interactions, on a predefined schedule or manually.

In terms of their purpose, pipelines can be divided into ingestion, ETL and post-ETL. Ingestion pipelines feed data into the relational database, processing measurement files from laboratory devices that users upload to a designated file system (Figure 1). The pipelines read these files, extract relevant information, and link it to corresponding database objects. ETL pipelines extract data from the relational database (and other sources, if applicable), transform it through aggregation, validation, filtering, and cleaning and load it into the OLAP database (Figure 4). Post-ETL-Pipelines transform the processed data from the OLAP database into analysis-ready formats, such as feature-engineered tables (e.g. for machine learning), normalised matrices for statistical analysis (e.g., PCA) and curated datasets for visualisation (e.g. depth plots, grainsize plots), or completely automate analysis.

While the technical complexity is abstracted through user-friendly interfaces, the pipelines themselves remain fully transparent as they are documented in code, allowing researchers to review and understand every processing step and modify or extend pipelines to meet project-specific requirements. This modular design ensures that the pipelines are adaptable without compromising the stability of the overall framework.

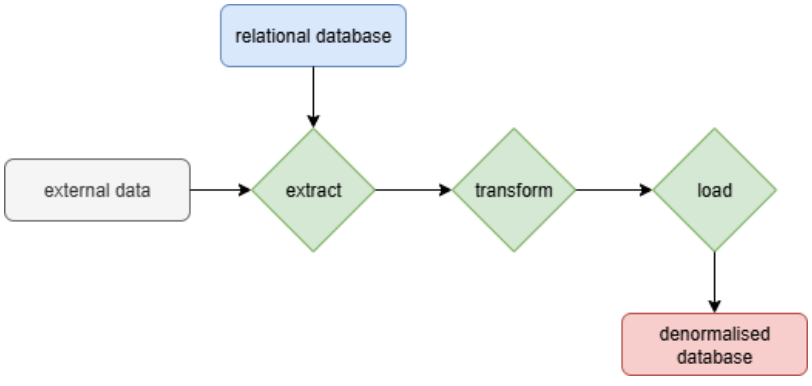

Figure 4. A simplified extract, transform, load (ETL) pipeline that extracts data from the relational database or external data, transforms it through aggregation, validation, filtering, and cleaning and eventually loads it into the denormalised OLAP database for further use.

https://doi.org/10.5194/egusphere-2025-4832
Preprint. Discussion started: 7 October 2025
© Author(s) 2025. CC BY 4.0 License.

4 Case Study: Geoarchaeological Data from Toboliu, Romania

The DFG-funded archaeological research project "Living Together or Apart?" investigates a Bronze Age Tell settlement near Toboliu village in western Romania, focusing on its chronological and spatial development, and social organisation (Glaser et al., 2020). In addition to geoarchaeological analyses of the Bronze Age site, the project focused on landscape evolution to trace back prehistoric human-landscape interaction. The study area, situated in the eastern Carpathian Basin, features a complex stratigraphy dominated by loess derivates and dominantly Phaeozem soils. According to Köppen-Geiger, the study area has a warm humid continental climate with warm summers and cold winters (Dfb). In the past, wetlands and grasslands were widespread, while the current landscape is heavily impacted by intensive agriculture. Remaining wetland areas are grazed but face degradation. The project employs interdisciplinary methods to explore landscape evolution through satellite imagery and geoscientific drilling at 15 sites, utilising data from sedimentological, geochemical, palynological and micromorphological analyses, as well as radiocarbon and luminescence dating (Zickel et al., 2025).

4.1 Research Data Management Challenges in Toboliu

Managing the heterogeneous and high-dimensional research data from the Toboliu project presents various challenges, related to data volume, complexity, heterogeneity, institutional barriers, while adhering to fulfil the research data management guidelines of the DFG.

Volume and complexity: The high volume and heterogeneity of geoscientific data generated during fieldwork and laboratory analysis exceeds the capacity for effective analysis, especially given the limited number of personnel available. In the Toboliu project, as in many small projects, managing the entire data lifecycle, from collection and preprocessing to analysis and interpretation, across multiple subdisciplines, including sediment analysis, geochronology, and palynology, lies with a doctoral student. Given the limited personnel resources and the sheer quantity of data, it is impossible to process and analyse the complete dataset in detail. Instead, researchers prioritise subdatasets based on preliminary trends or representative drilling cores to focus their efforts efficiently. However, a number of challenges complicate the exploratory data analysis and pattern identification required for this.

Heterogeneity and fragmentation: The interdisciplinary nature of the project results in fragmented data spread across numerous files and storage systems, making it prone to discrepancies. Raw data from instruments like a tacheometer for measuring locations might be converted into shapefiles for GIS, while the related documentation of a drilling location is documented in a field diary, a paper form or an Excel-sheet. Potential corrections need to be made in all files aross the different storage media. Discrepancies arise if values are not changed consistently across related files, which makes it more difficult to identify errors. This proved particularly problematic when pre-processing raw laboratory results, as it was necessary to reconcile the data with the logically related metadata.

Complexity of workflows: The large amount of high-dimensional data, combined with a multi-method approach, means that the data must be constantly restructured and transformed to meet the specific requirements of the various methods. The

https://doi.org/10.5194/egusphere-2025-4832
Preprint. Discussion started: 7 October 2025
© Author(s) 2025. CC BY 4.0 License.

introduction of new methods or the correction of values in the existing data set leads to an enormous amount of effort in re-executing entire workflows.

Long-term data storage: The Toboliu dataset faces a critical long-term challenge: while data from selected master cores will be published to address the project's research questions, a substantial portion of the collected field and laboratory data helps to overview patterns in the landscape or stratigraphy but proved incidental to the immediate goals. Although these data are physically stored, they remain only partially processed, documented, and unpublished. This causes indirect data loss: The data exists but is neither findable nor usable because it lacks contextual metadata or is only partially processed.

4.2 Application of the framework

The aformentioned challenges were identified in an early stage of the Toboliu project and were actively addressed throughout fieldwork, laboratory measurements and data analysis using our geoscientific data framework.

4.2.1 Applying the OLTP module

The web-based user interface had three critical functions in the Toboliu project: i) structured digitisation of field data to replace paper records with digital formats, ii) automate validation and standardisation during data entry, and iii) enforce consistency of data relationships: Field notes typically recorded relationships between samples and their contextual metadata (e.g. locations, stratigraphic layers) as free-text annotations (e.g. Location Y, Layer Z). However, inconsistent notation among project partners (e.g. different abbreviations for locations or layers) introduced the risk of ambiguous or conflicting identifiers. Upon database entry, these informal references were systematically converted into unambiguous, machine-readable relationships through the use of unique sample identifiers and automated validation, which ensured the existence of the referenced entities and consistency. This process eliminated ambiguities while preserving the original contextual information in a queryable, FAIR-compliant format.

Referential integrity and machine-readable relationships are required to flexibly filter entities based on their direct and indirect relationships. For instance, all grain size measurements are uniquely assigned to a sample. As the sample is also assigned to a location, it is possible to filter the measurements by location, campaign or project (Figure 5). Since the interfac ereturns key parameters such as the 'sample concentration' of a grain size measurement, i.e. if the sample's concentration was within the target range when measured, it helps users navigate the constantly evolving data directly, assessing the progress of the analysis and monitoring the data quality.

4.3 Applying the OLAP module

The OLAP module, specifically its denormalised star schema, played a key role in further simplifying access to the vast and complex dataset. The schema enabled us to derive sub-datasets quickly and efficiently within a structure tailored to the requirements for a specific analysis. This allowed for significantly more flexibility in the project's exploratory phase, as manual data integration and transformation were almost entirely eliminated. While creating a view of the grain size measurements in

https://doi.org/10.5194/egusphere-2025-4832
Preprint. Discussion started: 7 October 2025
© Author(s) 2025. CC BY 4.0 License.

Figure 5. An excerpt from the database showing grain size measurements in the user interface. In addition to the location and project assignment, the color of the measured quantity (sample concentration) indicates wether the value was within the accepted range for this property. Referential integrity and machine-readable relationships allow users to filter entities flexibly based on their direct and indirect relationships. For example, an analysis can be filtered by the location, campaign or project of its sample.

the relational database still required several joins, the star schema in OLAP allowed the query to be reduced to a simple select statement. While not all workflows were fully automated (Chapter 4.4), this approach still accelerated critical analyses, enabling faster iteration and deeper insights.

A key advantage was that derived data sets for further analysis were not generated ad hoc, but rather had a clearly defined and reproducible data structure. This meant that sub-records could be regenerated as required, even when the underlying data evolved due to new measurements or corrections. This decoupling of processes meant that the researcher could begin analyses, such as writing and testing code, while additional laboratory data was still being generated. Consequently, the project timeline was significantly shortened and iterative improvements became possible without delay.

4.4 Applying data pipelines

Data pipelines were specifically designed and implemented to automate complex, recurring data workflows in code. This ensured reproducibility, minimised manual errors and was crucial given the project's specific challenges.

Ingestion pipelines The project built in pipelines to automate the processing of standardised data sources (e.g. laser diffraction raw data), while custom pipelines extended the system's capabilities by directly ingesting raw measurement files from various laboratory devices, such as XRF spectrometers used for geochemical analyses, into the relational database. The focus of these pipelines was on capturing, parsing and storing raw data with minimal transformation, thereby ensuring

https://doi.org/10.5194/egusphere-2025-4832
Preprint. Discussion started: 7 October 2025
© Author(s) 2025. CC BY 4.0 License.

full traceability and data integrity from the source. Supporting a variety of input formats (e.g. CSV and proprietary binary files) enabled the seamless integration of heterogeneous laboratory datasets.

ETL pipelines ETL pipelines extracted data from the relational database and applied transformations for data handling and validation, such as unit normalisation, outlier detection and referential integrity checks, before loading the processed data into the OLAP database. This step was crucial in preparing the datasets for complex analytical queries and ad hoc exploration by structuring the data into dimensions and facts optimised for OLAP operations.

Post-ETL pipelines Post-ETL pipelines processed the data from the OLAP database further to generate analysis-ready datasets, including feature-engineered tables and normalised matrices. These pipelines not only enabled the use of advanced analytical techniques, but also automated full analytical workflows. Analyses such as texture classifications, principal component analysis (PCA), cluster analysis and interactive visualisations were automated directly as data pipelines within the data orchestration to decouple data management and data analysis. For instance, K-means clustering was applied to geochemical fingerprints to identify distinct stratigraphic layers or flag anomalies in sample sequences. Through continuous data integration and iterative model retraining, these pipelines gradually improved the accuracy of their analyses, revealing emerging patterns such as spatial correlations of sediment layer sequences or chronostratigraphic trends of geochemical signatures.

This dynamic process empowered data-driven research decisions from the project's earliest phases, such as targeted prioritisation of laboratory analyses, ensuring that insights became more precise and reliable as the dataset evolved.

5 Discussion

5.1 Comparison to other database approaches

Database technologies have been used in the geosciences for decades to manage scientific data. In the past thirty years alone, the disciplines of geochemistry and cosmochemistry have spawned several repositories, which exist alongside numerous more recent national initiatives. In contrast to geoscience laboratory information systems such as AusGeochem, StraboSpot and Sparrow, these repositories and databases are used solely for the publication of data or the compilation of published data (Klöcking et al., 2023). Thus, existing technologies are not always capable to address contemporary challenges in research data management holistically, as as they focus on isolated stages of the data lifecycle. Moreover, in many scientific data management concepts, metadata is managed in a database, while the actual data is stored in file systems (Li et al., 2015). Large geoscientific projects often have designated database systems (e.g. Willmes et al. (2014)). These databases primarily aim at long-term archiving and publication of data from these large interdisciplinary projects, but also provide an integrated database and infrastructure to facilitate and support research within the projects (Curdt, 2019). Recent approaches, such as LinkAhead, establish a more agile and holistic perspective on research data management, that try to encompass the whole research data lifecycle (Hornung et al., 2024). In contrast, our framework differs significantly by regarding research data management holistically:

https://doi.org/10.5194/egusphere-2025-4832
Preprint. Discussion started: 7 October 2025
© Author(s) 2025. CC BY 4.0 License.

1. **Comprehensive scope:** By standardising the approach to managing scientific data throughout the whole research data lifecycle, from fieldwork to data analysis and supply, we address research data management as a multifaceted challenge that extends beyond mere storage.

2. **Integration:** Unlike established systems, which often store data in inconsistent formats of varying quality, our framework ensures the seamless integration of diverse data through the strict application of a modular and extendable relational data model.

3. **Data-centric design:** We consider data to be dynamic rather than static. Therefore, our approach not only considers its structure and inherent logic, but also how researchers work with it and their increasingly complex workflows.

4. **Project-independent design:** Our framework is designed to be inclusive and adaptable. It caters not only to large-scale projects, but especially smaller research projects, as it is explicitly not project-related. This addresses a critical gap, as many smaller projects, such as the Toboliu case study, must adhere to strict data management requirements, but often lack access and funds to dedicated data management infrastructure

5.2 Addressing challenges in data management

Through its holistic approach, our framework addresses a wide range of contemporary key challenges in the management of research data, regarding data heterogeneity, data interoperability, workflow complexity, data loss and reusability. Here, we demonstrate how we address these with our framework, illustrated by the case study in Toboliu:

5.2.1 Data heterogeneity

Challenges of data heterogeneity, as identfied e.g. by Nordsiek and Halisch (2024) were also evident in the Toboliu project. For instance, Toboliu's heterogeneous, high-dimensional data from sedimentological, geochemical, and dating analyses were prone to inconsistencies due to their distribution across numerous files and irregular updates. Plotting grain size composition along a drill core depth required manual integration of tachymeter geodata, sample and stratigraphic unit information (field documentation), and derived measurement results. Our framework's unique focus on automating data linking and contextualisation significantly reduces the manual effort required by researchers, making it more practical for diverse datasets. Before applying our framework, corrections or remeasurements necessitated repeating the entire data integration and transformation process, increasing the risk of manual errors. By viewing databases as a system that accompanies the entire research process rather than as the final destination for finalised data, we have been able to maintain data integrity from the outset and identify errors immediately. In contrast to Nordsiek and Halisch (2024), our model explicitly considers fieldwork as the beginning of the research data lifecycle and a crucial link between all entities.

Our framework addresses the limits of relational databases, as described by (Kingdon et al., 2016), by logically separating data management and data analysis through the introduction of a denormalised OLAP database. However, unlike their proposed solution of a data warehouse, we do not consider our OLAP module to be independent; rather, we consider it to be integrated with the OLTP through an ETL pipeline. Integrating the database throughout the research process enabled us to create

https://doi.org/10.5194/egusphere-2025-4832
Preprint. Discussion started: 7 October 2025
© Author(s) 2025. CC BY 4.0 License.

method-specific data views (e.g. grain size analysis, clustering and PCA), independent of the underlying data state. Rather than generating static datasets for each analysis, we defined the necessary data structures that drew directly from the most recent dataset with every execution. In the Toboliu project, this enabled parallel work in the laboratory and on data analysis.

5.2.2 Interoperability

Many scientific data storage approaches store metadata in a database, while they store the actual data in file systems (Li et al., 2015). This limits the accessibility, machine-readability and interoperability. Our framework overcomes this issue by integrating metadata and measurements into a unified data model, ensuring that all information is stored consistently and is easily accessible.

Building on the approach by Nordsiek and Halisch (2024), our framework includes tools to manage diverse parameters and units. This ensures the system is applicable across disciplines. The framework's modular, discipline-specific design adds further flexibility, allowing different scientific fields to adapt it to their needs. By standardising discipline-specific terms, we improve semantic interoperability; the ability to exchange data across different systems. This addresses a key challenge highlighted in the FAIR principles (Wilkinson et al., 2016). However, full interoperability requires formal ontologies (structured definitions of terms and their relationships; Guizzardi (2020)). Although our framework standardises the storage of internal data, integrating external or legacy data remains challenging. These datasets often contain semantic differences (for example, the same term may have different meanings of definitions in different fields). Future work will focus on connecting external data using discipline-specific thesauri, as suggested by Nordsiek and Halisch (2024). Yet, resolving deep semantic differences in legacy data will require more than technical solutions such as community-driven efforts to map and reconcile semantic schemas (Lannom et al., 2020; Klöcking et al., 2023).

5.2.3 Workflow complexity

Our framework addresses the complexity of geoscientific workflows by modelling the entire research data lifecycle, including data storage. It thereby streamlines the transformation of raw data into scientific knowledge. In the Toboliu project, the automated processing of raw data, such as particle size measurements, accelerated re-measurement of samples if the direct feedback on measurement quality indicated any issues. Moreover, the automated processing of the measurements enabled early insights in the data, which helped to make strategic decisions on measuring additional samples in an early stage. The ETL pipelines virtually eliminated the need for manual data transformation. The integration of all code-based analyses, such as grain size analysis, depth plots, principal components analysis and cluster analysis, into the post-ETL pipelines enabled code versioning and significantly simplified and accelerated the process. The researcher could start with data analysis, while laboratory work was still in progress, test it on incomplete datasets, and finally, without any additional effort, execute it on the final data. This allows to focus on analysis rather than manual data integration, which accelerates the research process. Our framework's approach to automatically generating detailed provenance aligns with best practices, such as the FAIR Data Pipeline as it tracks data versions, software commits, and research outputs, and annotates runs with detailed metadata (Mitchell et al., 2022). Fitschen et al. (2019) emphasize the need to built Research Data Frameworks around scientists' workflows. They need to be able to sup-

https://doi.org/10.5194/egusphere-2025-4832
Preprint. Discussion started: 7 October 2025
© Author(s) 2025. CC BY 4.0 License.

port changes and to adapt to evolving standards (Fitschen et al., 2019). We developed our framework from a domain-specific perspective and aimed to strike a balance between standardisation for interoperability and the flexibility required to meet disciplinary and institutional requirements such as the unique conditions of our laboratories and specific requirements from research projects. For instance, we prioritised optimising workflows for Toboliu's specific context, such as automating data processing for the project's laboratory equipment, over generic applicability. In practice, this involved developing customised solutions for specific laboratory devices, which streamlined local operations but limited direct transferability to other settings. While this trade-off was justified by the efficiency gains achieved in the Toboliu project, we recognise the need for additional adaptation when applying the framework elsewhere. There will need to be continuous effort to provide user-friendly interfaces and robust documentation practices that overcome the tendency for custom scripts to be poorly documented, as observed in cloud environments (Mork et al., 2015). Nevertheless, the framework's capacity to automatically capture provenance and generate metadata during the research process, rather than necessitating post-hoc documentation, signifies a substantial advancement in making reproducibility the norm rather than a difficult task.

5.2.4 Data loss

In Toboliu, significantly more data was collected and generated than was ultimately necessary to answer the research questions. Ultimately, only two of the 15 drill cores were relevant for the analysis and description of landscape development in high detail, while two others were used for contextual information. Data from four other drill cores was included in the analysis to some extent (Zickel et al., 2025). The remaining seven cores were found to be redundant for addressing the objectives of the research project at this stage. Much of the data, which was collected at great expense and processed in the laboratory, has therefore not yet been published, despite potentially being scientifically relevant in the future. Thanks to the structured storage and detailed documentation of the processing, however, the data is not only preserved and retrievable, but also findable, accessible, interoperable and reusable within the organisation. Thus, our framework prevents "data cemeteries", as highlighted by Gärtner et al. (2001), by standardising data management and storage throughout the entire research data lifecycle.

5.2.5 Reusability

Our framework enhances reusability by optimising the entire research data management process to ensure interoperability, thereby facilitating the combination and analysis of data. It provides comprehensive context for each data point by capturing the entire research data lifecycle, from generation and transformation to storage, analysis, and reuse. A core strength lies in the metadata validation and provenance tracking, which minimises manual effort and is crucial for ensuring the long-term reusability of datasets. The FAIR principles advocate for machine-actionability and reusability for both humans and machines, relying on persistent identifiers and standardised metadata (Wilkinson et al., 2016). Our framework's use of machine-readable standards and detailed, automated metadata generation aligns well with these foundational tenets. Whereas curated domain repositories employ human curators for quality assurance and discoverability (Klöcking et al., 2023), our framework's automated metadata generation and provenance tracking provides a scalable, built-in mechanism to achieve similar levels of data quality and discoverability, reducing reliance on manual curation post-hoc. Furthermore, our framework offers a robust,

https://doi.org/10.5194/egusphere-2025-4832
Preprint. Discussion started: 7 October 2025
© Author(s) 2025. CC BY 4.0 License.

integrated system that facilitates the creation of well-documented datasets, even for unfinished projects, that are suitable for publication.

5.3 Applicability of the framework: Implications for other projects and academic institutions

Beyond its original scope, the framework has already demonstrated its modularity and scalability within our organisation. It is now used to manage and integrate datasets across multiple projects. It has also enabled the standardisation, long-term preservation and accessibility of legacy projects, ensuring compliance with FAIR principles and safeguarding valuable research assets. These capabilities facilitate interdisciplinary collaboration and support future usability of valuable research data.
While designed and implemented in compliance with the FAIR principles, we tailored it to the needs from the Toboliu project, the technical requirements and limitations of our IT infrastructure and research environment, and the anticipated future uses of the framework, such as serving as a foundation for more advanced data analysis and data mining using machine learning. Although initially developed to meet local requirements, its successful application in Toboliu shows that modest investments in data infrastructure can significantly improve research efficiency and outcomes. While the framework provides a robust, adaptable approach for managing complex datasets, institutions need to develop their adapted implementation based on their own research environments, including their research methodologies, data types, laboratory equipment and IT infrastructure. Exclusive investment in the development of general-purpose software is insufficient and does not reflect the complexity and constant changeability of research environments. Strategic investments in dedicated data engineers to orchestrate increasingly complex data storage and flows across projects are essential to unlock the full value of research data, pave the way for its broader adoption and ensure sustained success in interdisciplinary research. We want to emphasize that our framework is intended to promote a shift in the technical and organisational implementation of research data management in the geosciences, rather than providing a fully functioning end-user application or data models.

6 Conclusions

We present a database and data pipeline-driven framework for geoscientific research that addresses challenges related to the increasing volume and complexity of geoscientific datasets. It focuses on data heterogeneity, spatial complexities, and adherence to FAIR principles (Findable, Accessible, Interoperable, Reusable), transforming raw data into scientific knowledge. Our framework demonstrates, by its successful application in the Toboliu project, that it

- enhances integration of high-dimensional datasets,
- streamlines data management,
- improves replicability and reproducibility,
- promotes FAIR principles, scalability, and transparency and
- benefits small research projects with limited resources by simplifying adherence to data management requirements.

https://doi.org/10.5194/egusphere-2025-4832
Preprint. Discussion started: 7 October 2025
© Author(s) 2025. CC BY 4.0 License.

Although sustained institutional investment in IT infrastructure and expertise is necessary for long-term scalability and sustainability, the framework's proven efficiency and adaptability provide academic institutions with a clear pathway to increasing scientific impact and expanding interdisciplinary collaboration.

Code availability. The specific implementation of the framework is tailored to the research environment, IT infrastructure and security guidelines of the university. Consequently, its disclosure for external utilisation is not feasible.

Author contributions. DH, MM, and TR designed the research objectives and outline of the project. TR and MM provided resources and supervised the project. DH and MM conceptualized and implemented the framework. DH and MZ gathered present challenges in geoscientific research data management and provided data for the Toboliu case study. DH prepared the paper draft and all authors contributed to writing, reviewing, and editing the manuscript.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. This project was funded by the Key Profile Area 'Intelligent Methods for Earth System Sciences' at the University of Cologne (grant number 006). The case study in Toboliu was funded by the Deutsche Forschungsgemeinschaft under project number 436834905. We thank Dr. Katja Sperveslage for organizational support, and Dr. Stephan Opitz, Marie Gröbner, Dr. Dominik Brill, Dr. Anja Zander, and Florian Steininger for technical assistance. We are grateful to Professor Christina Bogner, Nicodemus Nyamari, and Felix Reize for test datasets, Christina Stollenwerk for system testing, and the IT staff for technical support. Special thanks to external IT experts Mark Handy and Thomas Schmidt for their critical reviews.
AI tools were used to improve the manuscript's grammar and readability.

https://doi.org/10.5194/egusphere-2025-4832
Preprint. Discussion started: 7 October 2025
© Author(s) 2025. CC BY 4.0 License.

References

Chaudhuri, S. and Dayal, U.: An Overview of Data Warehousing and OLAP Technology, ACM SIGMOD Record, 26, 65–74, https://doi.org/10.1145/248603.248616, 1997.

Codd, E. F.: A Relational Model of Data for Large Shared Data Banks, Communications of the ACM, 13, 377–387, https://doi.org/10.1145/362384.362685, 1970.

Curdt, C.: Supporting the Interdisciplinary, Long-Term Research Project 'Patterns in Soil-Vegetation-Atmosphere-Systems' by Data Management Services, Data Science Journal, 18, 5, https://doi.org/10.5334/dsj-2019-005, 2019.

Degen, D., Veroy, K., and Wellmann, F.: Certified Reduced Basis Method in Geosciences: Addressing the Challenge of High-Dimensional Problems, Computational Geosciences, 24, 241–259, https://doi.org/10.1007/s10596-019-09916-6, 2020.

Deutsche Forschungsgemeinschaft: Guidelines for Safeguarding Good Research Practice. Code of Conduct, https://doi.org/10.5281/zenodo.6472827, 2022.

European Commission, Directorate General for Research and Innovation., and PwC EU Services.: Cost-Benefit Analysis for FAIR Research Data: Cost of Not Having FAIR Research Data., Publications Office, LU, 2018.

Faniel, I. M. and Jacobsen, T. E.: Reusing Scientific Data: How Earthquake Engineering Researchers Assess the Reusability of Colleagues' Data, Computer Supported Cooperative Work (CSCW), 19, 355–375, https://doi.org/10.1007/s10606-010-9117-8, 2010.

Fitschen, T., Schlemmer, A., Hornung, D., Tom Wörden, H., Parlitz, U., and Luther, S.: CaosDB—Research Data Management for Complex, Changing, and Automated Research Workflows, Data, 4, 83, https://doi.org/10.3390/data4020083, 2019.

Gahegan, M.: Fourth Paradigm GIScience? Prospects for Automated Discovery and Explanation from Data, International Journal of Geographical Information Science, 34, 1–21, https://doi.org/10.1080/13658816.2019.1652304, 2020.

Gärtner, H., Bergmann, A., and Schmidt, J.: Object-Oriented Modeling of Data Sources as a Tool for the Integration of Heterogeneous Geoscientific Information, Computers & Geosciences, 27, 975–985, https://doi.org/10.1016/S0098-3004(00)00135-7, 2001.

Glaser, B., Kienlin, T., Röpke, A., and Deutsche Forschungsgemeinschaft: Separiert Oder Integriert? Studien Zur Entwicklung, Organisation Und Sozialen Struktur Der Komplexen Bronzezeitlichen Tellsiedlung von Toboliu, Westrumänien. Teil 2: Naturwissenschaftliche Untersuchungen, https://gepris.dfg.de/gepris/projekt/436834905, 2020.

Guizzardi, G.: Ontology, Ontologies and the "I" of FAIR, Data Intelligence, 2, 181–191, https://doi.org/10.1162/dint_a_00040, 2020.

Hornung, D., Spreckelsen, F., and Weiß, T.: Agile Research Data Management with Open Source: LinkAhead, ing.grid, 1, https://doi.org/10.48694/INGGRID.3866, 2024.

IUSS Working Group WRB: World Reference Base for Soil Resources 2022: International Soil Classification System for Naming Soils and Creating Legends for Soil Maps, International Union of Soil Sciences, Vienna, Austria, 4.edition edn., ISBN 979-8-9862451-1-9, 2022.

Kingdon, A., Nayembil, M. L., Richardson, A. E., and Smith, A. G.: A Geodata Warehouse: Using Denormalisation Techniques as a Tool for Delivering Spatially Enabled Integrated Geological Information to Geologists, Computers & Geosciences, 96, 87–97, https://doi.org/10.1016/j.cageo.2016.07.016, 2016.

Klöcking, M., Wyborn, L., Lehnert, K. A., Ware, B., Prent, A. M., Profeta, L., Kohlmann, F., Noble, W., Bruno, I., Lambart, S., Ananuer, H., Barber, N. D., Becker, H., Brodbeck, M., Deng, H., Deng, K., Elger, K., De Souza Franco, G., Gao, Y., Ghasera, K. M., Hezel, D. C., Huang, J., Kerswell, B., Koch, H., Lanati, A. W., Ter Maat, G., Martínez-Villegas, N., Nana Yobo, L., Redaa, A., Schäfer, W., Swing, M. R., Taylor, R. J., Traun, M. K., Whelan, J., and Zhou, T.: Community Recommendations for Geochemical Data, Services and Analytical Capabilities in the 21st Century, Geochimica et Cosmochimica Acta, 351, 192–205, https://doi.org/10.1016/j.gca.2023.04.024, 2023.

https://doi.org/10.5194/egusphere-2025-4832
Preprint. Discussion started: 7 October 2025
© Author(s) 2025. CC BY 4.0 License.

Lannom, L., Koureas, D., and Hardisty, A. R.: FAIR Data and Services in Biodiversity Science and Geoscience, Data Intelligence, 2, 122–130, https://doi.org/10.1162/dint_a_00034, 2020.

Li, Z., Yang, C., Jin, B., Yu, M., Liu, K., Sun, M., and Zhan, M.: Enabling Big Geoscience Data Analytics with a Cloud-Based, MapReduce-Enabled and Service-Oriented Workflow Framework, PLOS ONE, 10, e0116 781, https://doi.org/10.1371/journal.pone.0116781, 2015.

Liakos, L. and Panagos, P.: Challenges in the Geo-Processing of Big Soil Spatial Data, Land, 11, 2287, https://doi.org/10.3390/land11122287, 2022.

Mitchell, S. N., Lahiff, A., Cummings, N., Hollocombe, J., Boskamp, B., Field, R., Reddyhoff, D., Zarebski, K., Wilson, A., Viola, B., Burke, M., Archibald, B., Bessell, P., Blackwell, R., Boden, L. A., Brett, A., Brett, S., Dundas, R., Enright, J., Gonzalez-Beltran, A. N., Harris, C., Hinder, I., David Hughes, C., Knight, M., Mano, V., McMonagle, C., Mellor, D., Mohr, S., Marion, G., Matthews, L., McKendrick, I. J., Mark Pooley, C., Porphyre, T., Reeves, A., Townsend, E., Turner, R., Walton, J., and Reeve, R.: FAIR Data Pipeline: Provenance-Driven Data Management for Traceable Scientific Workflows, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 380, 20210 300, https://doi.org/10.1098/rsta.2021.0300, 2022.

Mork, R., Martin, P., and Zhao, Z.: Contemporary Challenges for Data-Intensive Scientific Workflow Management Systems, in: Proceedings of the 10th Workshop on Workflows in Support of Large-Scale Science, pp. 1–11, ACM, Austin Texas, ISBN 978-1-4503-3989-6, https://doi.org/10.1145/2822332.2822336, 2015.

Murillo, A. P.: Data Matters: How Earth and Environmental Scientists Determine Data Relevance and Reusability, Collection and Curation, 41, 77–86, 2019.

Nikparvar, B. and Thill, J.-C.: Machine Learning of Spatial Data, ISPRS International Journal of Geo-Information, 10, 1–32, https://doi.org/10.3390/ijgi10090600, 2021.

Nordsiek, S. and Halisch, M.: Making Geoscientific Lab Data FAIR: A Conceptual Model for a Geophysical Laboratory Database, Geoscientific Instrumentation, Methods and Data Systems, 13, 63–73, https://doi.org/10.5194/gi-13-63-2024, 2024.

Picatto, H., Heiler, G., and Klimek, P.: Cost-Effective Big Data Orchestration Using Dagster: A Multi-Platform Approach, https://doi.org/10.48550/ARXIV.2408.11635, 2024.

Raasveldt, M. and Mühleisen, H.: DuckDB: An Embeddable Analytical Database, in: Proceedings of the 2019 International Conference on Management of Data, pp. 1981–1984, ACM, Amsterdam Netherlands, ISBN 978-1-4503-5643-5, https://doi.org/10.1145/3299869.3320212, 2019.

Raj, A., Bosch, J., Olsson, H. H., and Wang, T. J.: Modelling Data Pipelines, in: 2020 46th Euromicro Conference on Software Engineering and Advanced Applications (SEAA), pp. 13–20, IEEE, Portoroz, Slovenia, ISBN 978-1-7281-9532-2, https://doi.org/10.1109/SEAA51224.2020.00014, 2020.

Soil Science Division Staff: Soil Survey Manual, no. 18 in USDA Handbook, Government Printing Office, Washington, D.C., 2017.

Tedersoo, L., Küngas, R., Oras, E., Köster, K., Eenmaa, H., Leijen, Ä., Pedaste, M., Raju, M., Astapova, A., Lukner, H., et al.: Data Sharing Practices and Data Availability upon Request Differ across Scientific Disciplines, Scientific data, 8, 192, 2021.

Thomer, A. K. and Wickett, K. M.: Relational Data Paradigms: What Do We Learn by Taking the Materiality of Databases Seriously?, Big Data & Society, 7, 205395172093 483, https://doi.org/10.1177/2053951720934838, 2020.

Van Den Brink, L., Barnaghi, P., Tandy, J., Atemezing, G., Atkinson, R., Cochrane, B., Fathy, Y., García Castro, R., Haller, A., Harth, A., Janowicz, K., Kolozali, Ş., Van Leeuwen, B., Lefrançois, M., Lieberman, J., Perego, A., Le-Phuoc, D., Roberts, B., Taylor, K., and Troncy, R.: Best Practices for Publishing, Retrieving, and Using Spatial Data on the Web, Semantic Web, 10, 95–114, https://doi.org/10.3233/SW-180305, 2018.

https://doi.org/10.5194/egusphere-2025-4832
Preprint. Discussion started: 7 October 2025
© Author(s) 2025. CC BY 4.0 License.

Vance, T. C., Huang, T., and Butler, K. A.: Big Data in Earth Science: Emerging Practice and Promise, Science, 383, eadh9607, https://doi.org/10.1126/science.adh9607, 2024.

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., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J., Groth, P., Goble, C., Grethe, J. S., Heringa, J., 'T Hoen, P. A., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., Van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., Van Der Lei, J., Van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., and Mons, B.: The FAIR Guiding Principles for Scientific Data Management and Stewardship, Scientific Data, 3, 160 018, https://doi.org/10.1038/sdata.2016.18, 2016.

Willmes, C., Kürner, D., and Bareth, G.: Building Research Data Management Infrastructure Using Open Source Software, Transactions in GIS, 18, 496–509, https://doi.org/10.1111/tgis.12060, 2014.

Wozniak, J. M., Chard, R., Chard, K., Nicolae, B., and Foster, I.: Xd/Ml Pipelines: Challenges in Automated Experimental Science Data Processing, in: ASCR Workshop on the Management and Storage of Scientific, 2022.

Zickel, M., Nett, J. J., Röpke, A., Opitz, S., Handy, D., Rische, S., Steininger, F., Mantke, N., and Reimann, T.: Unravelling Pleistocene Landscape Evolution and Bronze Age Land-Use at the Eastern Margin of the Carpathian Basin in Romania, Journal of Quaternary Science, Submitted, 2025.