Geosci. Instrum. Method. Data Syst., 15, 165–181, 2026
https://doi.org/10.5194/gi-15-165-2026

Geoscientific Instrumentation Methods and Data Systems
Open Access
EGU

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

Dennis Handy, W. Marijn van der Meij, Mirijam Zickel, and Tony Reimann

Institute of Geography, University of Cologne, 50923 Cologne, Germany

Correspondence: Dennis Handy (dhandy1@uni-koeln.de)

Received: 30 September 2025 – Discussion started: 7 October 2025
Revised: 22 January 2026 – Accepted: 13 April 2026 – Published: 20 May 2026

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, providing a bottom-up solution that empowers individual teams 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 with the FAIR (Findable, Accessible, Interoperable, and Reusable) principles. The framework uses a dual-component architecture. First, an Online Transaction Processing (OLTP) system features a user interface and a persistent relational database, ensuring accurate and consistent data storage when capturing and managing diverse geoscientific research data. Complementing this, an orchestration layer manages automated data pipelines to process the stored data and generate dynamic in-memory Online Analytical Processing (OLAP) databases that allow flexible, high-performance analysis. It is adaptable to evolving research requirements and supports various data types and methodological approaches, such as machine learning and deep learning, that place high demands on the data and their formats. A case study in Western Romania demonstrates the application of the data framework in an interdisciplinary geoarchaeological research project by processing and storing heterogeneous datasets, thereby reducing data management efforts, improving findability, replicability, and reproducibility, and streamlining the integration of high-dimensional data for small, interdisciplinary teams.

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). The FAIR principles, findability, accessibility, interoperability, and reusability, which emerged from the FORCE11 community, aim 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, 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, in addition to conventional metadata describing the data, spatial data must also include explicit locational information (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 management, particularly in small and interdisciplinary research projects that must comply with FAIR principles and funding requirements (e.g. Deutsche Forschungsgemeinschaft, 2022). However, project-based funding models pose a significant challenge to establishing sustainable data management structures. Many valuable domain-specific data systems originating from research projects struggle to survive due to limited ongoing funding (Klöcking et al., 2023). Though specialised approaches for data management exist for particular geoscientific subfields, an interdisciplinary approach has been missing, integrating research data from various areas (Nordsiek and Halisch, 2024), as required in integrative disciplines such as geomorphological and geochronological research. 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 inability to make research data FAIR is more than just an inconvenience, but has a quantifiable financial cost. Non-standardised research data results in billions of euros in expenses annually (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:

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 the automated orchestration of 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 (Sect. 2). Then, we will describe how we used these requirements to design and implement the framework (Sect. 3). We test and demonstrate the practical application of our framework using a geoarchaeological project in western Romania (Sect. 4). Finally, we discuss how the framework addresses the outlined challenges and provide an outlook on future developments (Sect. 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 the geosciences, we gathered insights on ex-

isting challenges, 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.

2.1 Reusability, interoperability and preservation

With geoscientific data being inherently diverse (Klöcking et al., 2023; Nordsiek and Halisch, 2024), the recent 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), poses significant challenges for research data management and computational methods (Li et al., 2015; Liakos and Panagos, 2022). For example, they are often stored in incompatible or hard-to-access locations, such as personal computers or local databases. This phenomenon, called *data silos*, impedes the discovery and reusability of data (Klöcking et al., 2023). This issue of isolated data applies even to formal database systems. The often isolated subject-specific storage makes cross-domain evaluation difficult, and the complexity of data models causes users to develop error-prone workarounds (Kingdon et al., 2016). Geoscientific data is often stored in different, often isolated formats (GIS, databases, specialised software, proprietary formats), leading 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).

In contrast, the trend towards numerous general-purpose data repositories, while offering availability, can exacerbate discovery and reusability issues because they often don't integrate or harmonise deposited data (Wilkinson et al., 2016). 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 grain sizes from 0.002 to 0.05 mm as silt (Soil Science Division Staff, 2017), the World Reference Base for Soil Resources (WRB) draws the boundaries at 0.002 and 0.063 mm (IUSS Working Group (WRB), 2022).

2.2 Workflows complexity

Complex data workflows have become standard, driven by a significant shift in the computational landscape due to the growing prominence of data-intensive sciences (Mork et al., 2015). However, due to the application of recent computational methods, such as machine learning, data workflows are even increasing in volume, velocity, and complexity (Suter et al., 2026). Even for domain experts, the transition from basic science to interpretation is complex. Throughout the workflow, strategic decisions include not only the definition of a model's structure and assumptions, but also the implementation of its code, the selection of parameter values, and the generation of outputs (Mitchell et al., 2022).

2.3 Institutional barriers

Scientists often rely on adapted 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 perceived burdens 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 that focus primarily on publishing papers rather than 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 for small teams that balances standardisation for interoperability with flexibility for disciplinary requirements. Such a framework should support 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 an orchestration framework for data pipelines rather

than focusing solely on database storage. These pipelines consider the ingestion of raw data, its processing, transformation, storage and – eventually – automated retrieval and analysis (Fig. 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 (Sect. 3.1, Fig. 2a), and derived online analytical processing (OLAP) databases implementing a denormalised data model called star schema (Sect. 3.2, Fig. 2b). 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 an 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.

3.1 Online transaction processing (OLTP) module

The Online Transaction Processing (OLTP) module consists of two components: a relational database, serving as the central repository and single source of truth, and a user interface that acts as an intermediary 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. The user interface was implemented with *Django Unfold*. Our OLTP module is tailored to the needs of geomorphological, geoarchaeological and geochronological research, but can be easily adapted to other systems and requirements.

3.1.1 Database

The core of the OLTP module is a normalised relational database that adheres to the fundamental principles of relational theory. This approach was chosen to ensure data integrity, minimise redundancy, and enable set-based queries throughout the entire lifecycle of geoscientific research data (e.g., Codd, 1970; Kingdon et al., 2016). The conceptual starting point is the model of a modular geoscience laboratory database developed by Nordsiek and Halisch (2024), but we go beyond their approach in two fundamental aspects. Firstly, we expand the data model's focus from a pure laboratory database to the entire life cycle of research data, which, in our domain, usually begins with field data collection. Secondly, and this is the core of our contribution, we adopt a stricter relational approach to data storage. While we also archive raw data and refer to it from the database, as recommended by Nordsiek and Halisch, we additionally store the processed and structured data directly in the database in accordance with relational theory. This enables complex queries on the actual measured values, not just on metadata that refers to files.

This model is structured around the sample as a physical specimen. It serves as the central table, from which links are established to related entity types, such as locations, (stratigraphic) layers, various analyses, and the overall project context. This relational structure creates a network of context-related information, ensuring that no entity can exist in isolation. Every piece of data, from field observations to laboratory measurements, can be traced back to its origin and unambiguously linked to all other relevant information. If the sample has already been published, providing a reference to the publication or its *International Generic Sample number* (IGSN) [1] directly links the internal context to the public record.

Central and recurring methods are directly mapped in the model by separate entity types, such as grain sizes. In addition, it implements a flexible design through a generic measurement table, allowing the addition of similar automated processing procedures for other analysis methods in the future without changing the core data model. To this end, the *GenericMeasurement* table references the *Parameter* (measured variable, designation, physical unit, and minimum and maximum limits) and the *Method* tables.

The model distinguishes between internal and external data sources to ensure transparent data provenance and supports secure collaboration by using role-based access controls to manage pre-publication data. A detailed description of the data model, including all entities and analysis methods, is provided in the Appendix.

3.1.2 Initial data processing

Furthermore, the OLTP module anticipates the data pipelines (see Sect. 3.3) by integrating initial data processing steps at data entry, where appropriate. As a proof of concept for this integrated approach, the system can currently automatically parse raw data from our laser diffractometer for particle size analysis. The user interface provides a means for uploading, after which the system extracts the critical data, reclassifies it into grain-size classes, and stores the results in respective database fields. When called up via the web interface, the system automatically generates a particle size distribution plot, providing users with direct visual feedback.

3.1.3 User interface

The web-based user interface (Fig. 3) ensures data quality and integrity at the point of entry by translating the database's strict relational rules into an intuitive workflow, providing

[1] formerly *International Geo Sample Number* (Klump et al., 2021; De Castro et al., 2023).

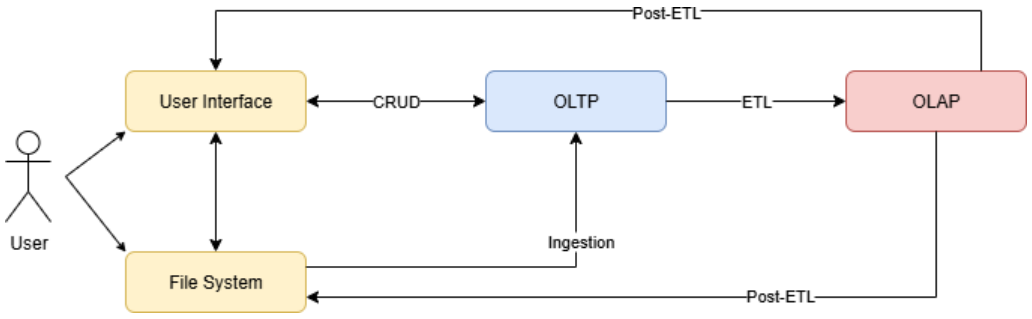

Figure 1. The system architecture illustrates the interaction between user, data storage and data pipelines within the framework. Users initiate the process by uploading data via the user interface or the file system (yellow boxes). The data is processed, validated and linked in the online transaction processing database (OLTP, blue box). An extract, transform, load (ETL) pipeline extract data from the OLTP database, transform it and load it into an online analytical processing database (OLAP, red box). Post-ETL pipelines automatically generate analyses, data visualisations, 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).

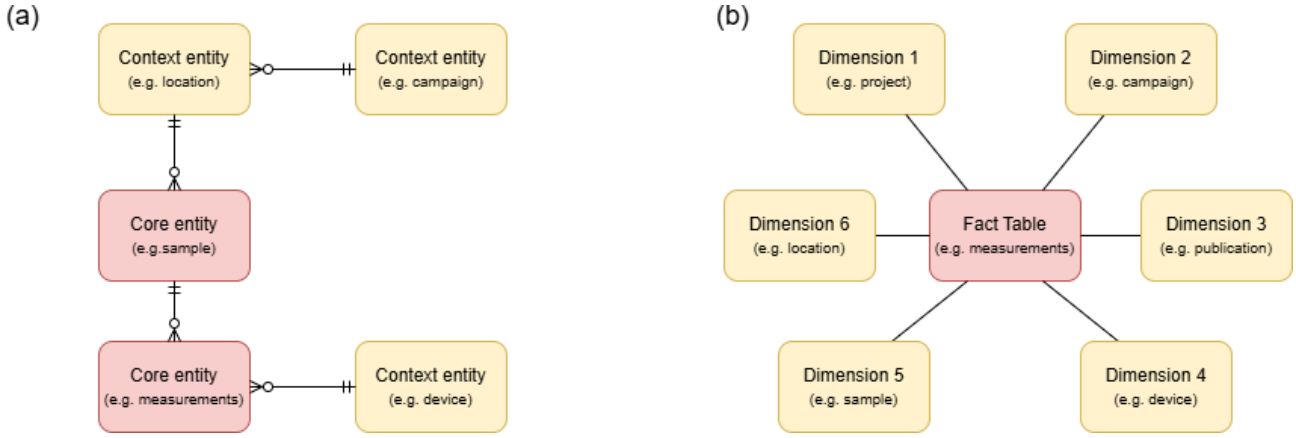

Figure 2. Simplified, conceptual models of **(a)** a relational model and **(b)** a star schema in direct comparison. The normalised relational schema **(a)** organises data across logically linked tables to minimise redundancy and ensure data integrity. To clarify their function within this diagram, we use the illustrative terms *Core entities* for the central objects of research and *Context entities* for the descriptive metadata. 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. This design enables efficient analytical queries by denormalising the data structure, improving performance for aggregations and slicing operations.

convenient support for 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. Moreover, it supports the direct export of filtered datasets in standard file formats, such as CSV, JSON, or Excel tables.

3.2 Online analytical processing (OLAP)

Complex queries, such as those involving joins across multiple tables, can degrade performance as datasets grow in size and structural complexity. To address this issue, the framework generates on-demand analytical databases that use a multidimensional star schema. This consolidates normalised tables into simpler structures with controlled redundancy to optimise query performance and data integration (Chaudhuri and Dayal, 1997; Kingdon et al., 2016). As shown in Fig. 2b, the central fact table of the star schema stores the core quantitative measurements of the research, which are contextu-

Figure 3. An example of the user interface implemented with Django Unfold 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.

alised by descriptive dimension tables (e.g., campaign, location, project), providing a spatial and conceptual framework. Users can navigate from aggregated data to detailed sample records, including geographic coordinates, time periods and measurement methods. This analytical layer supports filtering by criteria such as location, time or analytical method. For instance, a query could efficiently filter across multiple dimensions, such as location, time, or analytical method, to aggregate or extract specific measurements, such as all geochronological ages measured in the luminescence laboratory over the past decade. These dynamic databases are implemented using DuckDB, an open-source in-process database designed for analytical workloads (Raasveldt and Mühleisen, 2019). It is closely integrated with Dagster for data orchestration (Picatto et al., 2024, see Sect. 3.3).

3.3 Data orchestration

The increasing complexity of geoscientific data workflows, characterised by heterogeneous data sources, multiple processing steps, and interdisciplinary requirements (see Sect. 2), demands automated solutions for data orchestration. This concept encompasses the deployment, scheduling, and execution of a workflow's computational steps (Suter et al., 2026). Or in technical terms, a data orchestrator enables the "construction, operation, and observation of [replicable] data pipelines" (Picatto et al., 2024). 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 continuous, intermittent or batch data 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). This enables the replicable automation of digital processing steps within complex scientific workflows.

The framework implements a comprehensive and generally applicable architectural pattern by dividing data workflow automation into two levels. The OLTP module already contains basic transactional data pipelines for automating transactional operations, processing and transforming data during storage (e.g., via uploads through the user interface) and retrieval (e.g., for queries or exports). It is a self-contained component that, by design, is independent of the specific institutional IT infrastructure (see Sect. 3.1.2). However, effective data orchestration must be understood in the context of the specific IT environment it operates within. Automating tasks relies on local infrastructure, including connecting to specific laboratory instruments, accessing locally stored files, or saving results in the university's backup systems. Therefore, the framework introduces a second layer specifically to manage more complex and computationally intensive tasks that are inherently dependent on that infrastructure. This layer is implemented using the open-source orchestrator Dagster (Picatto et al., 2024).

Although the primary workflow involves managing data flow from OLTP to OLAP, the orchestrator's capabilities extend beyond this one-way process. For instance, ingestion pipelines can monitor a file system, automatically process

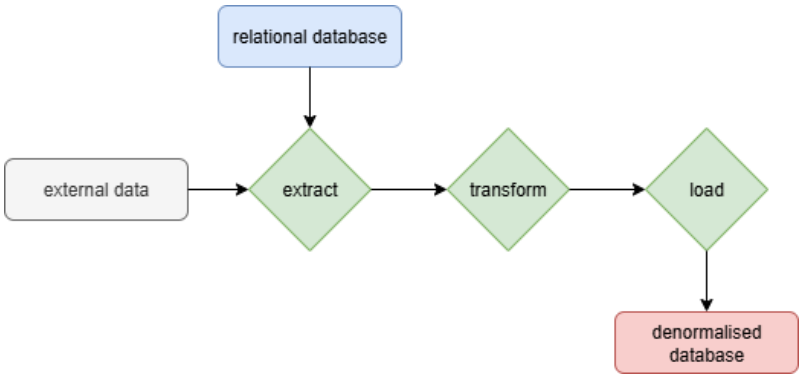

Figure 4. A simplified extract, transform, load (ETL) pipeline that extracts data from the relational database or external data, transforms it through aggregation, validation, filtering, and cleaning and eventually loads it into the denormalised OLAP database for further use.

files and feed the data into the relational database (Fig. 1). ETL (extract, transform, load) pipelines then extract data from the relational database (and other data sources, if applicable), transform it (e.g. aggregation and validation) and load it into an OLAP database (Fig. 4). Analytical pipelines transform processed data from these 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, grain size plots), or completely automate analysis.

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 in the study area. Situated in the eastern Carpathian Basin, the investigated area has a fully humid temperate climate (Cfb) but already in close spatial proximity to a fully humid snow climate with warm summers (Dfb) (Kottek et al., 2006). The project investigated the complex stratigraphy of loess derivates and historical environmental conditions, including the presumed widespread presence of wetlands, using interdisciplinary methods. These methods comprised satellite imagery analysis and geoscientific drilling at 15 sites, which produced data from sedimentological, geochemical, palynological, and micromorphological analyses, as well as radiocarbon and luminescence dating.

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 exceed 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 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 aforementioned 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 (Fig. 5). Since the interface returns 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 OLAP

An OLAP database, 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 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 (Sect. 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 subrecords 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 While the OLTP already automated the processing of laser diffraction raw data, 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 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.

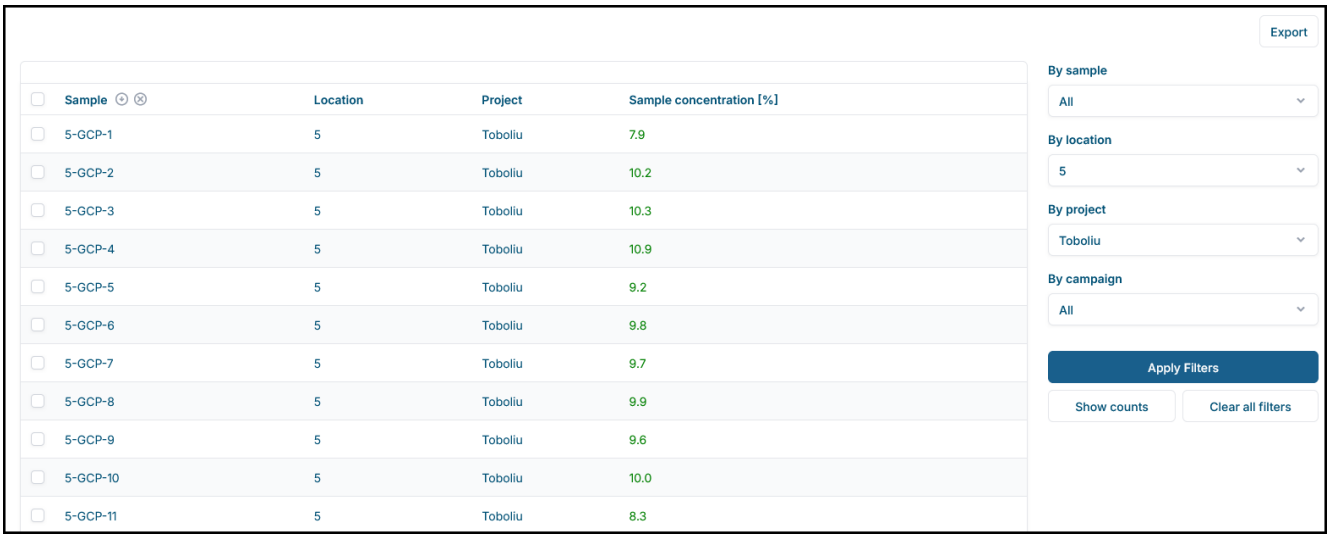

Figure 5. An excerpt from the database showing grain size measurements in the user interface. In addition to the location and project assignment, the color of the measured quantity (sample concentration) indicates wether the value was within the accepted range for this property. Referential integrity 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.

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 an 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 classification, principal component analysis (PCA), cluster analysis, and interactive visualisations were automated directly as data pipelines within the data orchestration, decoupling data management from 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 in sediment layer sequences and chronostratigraphic trends in 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 Positioning the framework within the research data ecosystem

Databases have been well-established technologies and applications in the geosciences for decades. They are used as laboratory information systems, repositories, and data catalogues. In contrast to geoscience laboratory information systems such as AusGeochem, StraboSpot, and Sparrow, repositories and data catalogues are mainly used for publishing or compiling published data (Klöcking et al., 2023). However, scientific practice shows that, although scientists apply relational principles, for pragmatic reasons, they resort to ad hoc solutions that are not technically designed as databases. These primarily include spreadsheets or collections of text files (Thomer and Wickett, 2020). Thus, existing technologies and actual scientific practice are not always capable of addressing contemporary challenges in research data management holistically, as they focus on isolated stages of the data lifecycle. This reality reveals a critical blind spot between large, formal infrastructures and the dynamic, iterative nature of local research data management.

While large geoscientific projects often have designated database systems (e.g., Willmes et al., 2014), they primarily aim to archive and publish project-related data for long-term preservation. Though they might also provide an integrated database and infrastructure to facilitate and support research within these projects (Curdt et al., 2019), there remains a blind spot between large inter-organisational infrastructures and local research data management. Recent approaches, such as LinkAhead, adopt a more agile, holistic perspective

on research data management that encompasses the entire research data lifecycle. It achieves this by employing a flexible data model that enables dynamic linking of all research entities (e.g., files, samples, processes) to track their provenance and relationships (Hornung et al., 2024). While this strategy provides flexibility for mapping the complex web of the research process, our framework deliberately adopts a different philosophy. It prioritises data integrity and analytical performance. Instead of a flexible, linkage-focused model, our framework is built on a strict relational OLTP database that enforces consistency at the point of entry via a predefined, yet extensible, schema. This initial investment in structure is a prerequisite for creating performant, on-demand analytical databases and ensures a consistent foundation for quantitative analysis with recent demanding computational methods, such as Machine Learning and Deep Learning. Our framework thus fits seamlessly into the existing research data management ecosystem and closes the gap between structured local archiving, high-performance and transparent analysis, and preparation for ingestion into large, established repositories or scientific workflow systems.

5.2 Addressing challenges in data management

Through its holistic approach, which builds on the positioning outlined in the previous section, our framework addresses a wide range of contemporary key challenges in the management of research data, including data interoperability, reusability and preservation, data heterogeneity, and workflow complexity, as illustrated in the Toboliu project.

5.2.1 Interoperability and reusability

Interoperability is "the ability of data or tools from non-cooperating resources to integrate or work together with minimal effort" (Wilkinson et al., 2016). The challenge involves integrating independently designed information spaces to enable consistent analysis, considering various levels of interoperability: Syntactic interoperability guarantees that systems can agree on shared syntactic formats for processing symbols (Guizzardi, 2020), while semantic interoperability refers to the consistent understanding of the information's context and meaning (Stocker et al., 2018). However, as described above, researchers often resort to local ad hoc solutions, such as individual spreadsheets and text files (Thomer and Wickett, 2020). Furthermore, a significant limitation of established data management systems is that scientific data is stored in file systems that are separate from the metadata, which is stored in a database. This physical separation limits the ability to automatically access, archive, analyse and mine scientific data (Li et al., 2015). Accordingly, interoperability often fails due to challenges that precede semantic interoperability. Our proposed framework establishes the fundamental prerequisites for semantic interoperability by relying on mature, architecturally coherent approaches. By overcoming the structural separation of data and metadata from the ground up, a sustainable, future-proof foundation is created on which true semantic interoperability can be built.

This leads to the framework's central objective: It enhances reusability by optimising the entire research data management process to ensure internal interoperability, thereby facilitating the combination of data and its analysis. It offers an exhaustive h 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 metadata validation, 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 with these foundational tenets. Generating metadata during the research process, rather than relying on post hoc documentation, marks a substantial step towards making reproducibility the norm rather than a difficult task. By automatically documenting detailed provenance through formalising data storage and workflows in code, our framework aligns with best practices, such as the FAIR Data Pipeline (Mitchell et al., 2022). In our case study, this included creating pipelines adapted for specific laboratory devices and computational methods. While this investment improved all aspects of data handling, it also limited its ease of application in other environments. While this trade-off was justified by these enhancements, we recognise the need for a technical and semantic connection to external systems to further improve the reusability of the data we process and the replicability of the pipelines we apply. For publishing, the research context, including data and metadata, could be packaged into Research Object Crates (RO-Crates) to enable machine-readability (Soiland-Reyes et al., 2022). Specific extensions such as *Workflow Run RO-Crate* record the provenance of computing workflows in detail (Leo et al., 2024). This approach enables the creation and use of FAIR research archives (Soiland-Reyes et al., 2022). While these robust technical frameworks handle the methods and structures of data exchange effectively, they can not address the underlying meaning. Consequently, significant semantic differences in research require community-led initiatives to develop and align semantic schemas (Lannom et al., 2020; Klöcking et al., 2023).

5.2.2 Workflow complexity

Our framework addresses the complexity of geoscientific workflows by modelling the entire research data lifecycle, including data storage and data flow. At a technical level, data orchestration is implemented using the Dagster framework to automate Python pipelines that integrate data storage, processing and analysis. This approach was chosen to

provide researchers with the performance and flexibility of a full programming language, a prerequisite for many machine learning algorithms and data-intensive tasks. To combine this flexibility and performance with standardised portability, a future extension could enable the export of these pipeline definitions to the Common Workflow Language (CWL). This would ensure that pipelines designed on the platform are portable and can also be executed in external, CWL-compatible environments (see Crusoe et al., 2022).

For instance, 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. Integrating the database throughout the research process enabled us to create method-specific data views (e.g. grain size analysis, clustering and PCA), that are always based on the most up-to-date version of the data. 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 denormalised OLAP databases. However, unlike their proposed solution of a data warehouse, we do not consider OLAP to be independent; rather, we consider it to be integrated with the OLTP through ETL pipelines. Rather than generating static datasets for each analysis, we define the necessary data structures that draw directly from the most recent dataset with every execution. This approach not only allows researchers to focus on analysis rather than manual data integration but, as demonstrated in the Toboliu project, it enables true parallel work between the laboratory and data analysis, significantly accelerating the research process. 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 data analysis while laboratory work was still in progress and test it on incomplete datasets.

5.2.3 Data heterogeneity

Challenges of data heterogeneity, as identified 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.

5.2.4 Data preservation

Beyond interoperability for publishing, the framework's structured approach provides a critical defence against the permanent loss of valuable research data. Ultimately, consistently structured recording and detailed documentation of the entire process are the most effective protection. For example, in Toboliu, significantly more data was collected and generated than was ultimately necessary to answer the research questions. Ultimately, only a subset of the 15 drill cores was relevant for the detailed analysis and description of landscape development. The remaining cores were found to be redundant for achieving the research project's objectives at this stage. Much of the data, which was collected at great expense and processed in the laboratory, has therefore not yet been published, even though it may become scientifically relevant in the future. Thanks to the structured storage and detailed documentation of the processing, 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.

5.3 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 as a generic approach in compliance with the FAIR principles, we adapted the implementation to the needs of our researchers, the technical requirements and limitations of our IT infrastructure and research environment, and the anticipated future uses of the framework, such as machine learning. Its successful applica-

tion in Toboliu shows that modest investments in data infrastructure can significantly improve research efficiency. This highlights that exclusive investment only in software products is insufficient for complex modern research environments. Therefore, strategic investments in dedicated data engineers to orchestrate complex data flows are essential to unlock the full value of research data. Ultimately, we want to emphasise 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

This study presented a geoscientific research data framework that integrates data storage and orchestration to address 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.

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.

Appendix A: Data Model

The following tables depict the concrete implementation of the entity types in the OLTP module at the time of this paper's publication, along with their attributes, which take up and expand on the conceptual framework proposed by Nordsiek and Halisch (2024) as described in Sect. 3.1. Attributes that refer to other objects, such as foreign keys or one-to-one and many-to-many relationships, are marked in italics. Django's generic models and m:n-tables are not marked separately. For clarity, detailed descriptions of each attribute have been omitted. This also applies to inherited attributes. Both can be viewed in the repository's source code, which has been published separately and linked.

The OLTP's Django project, and thus the data model, is divided into several modular applications: analysis, bibliography, field_data, and laboratory. Additional Django models are defined at the project level (Base).

Table A1. Base.

Table	Attributes
BaseModel	created_at, *created_by*, *updated_by*
ResearchGroup	label, *head_of_group*, *auth_group*
Researcher	*user*, academic_rank, position, orcid
Project	title, subtitle, label, *principal_investigator*, *associated_investigator*, *research_group*, *parent*, start_date, deadline, description, status, public
ProjectUserObjectPermission	content_object

At the project level, there are five models that are not assigned to any of the four apps. Defining research groups, researchers and projects enables the basic structuring of the stored data and access control. The *ProjectUserObjectPermission* model enables access control to project-related data. The BaseModel defines the basic properties inherited by all other models: created_at, created_by and updated_by. This allows tracking the creation and modification of each object in the database.

Table A2. Analysis.

Table	Attributes
Algorithm	name, version, description, link, file, programming_language
RawMeasurement	*project*, *sample*, *device*, *accessories*, *researcher*, file, description
RawProcessing	*raw_measurement*, processed_file, processing_description, *processed_by*, processing_date, *preparation_algorithm*, *evaluation_algorithm*, *publication*
Counting	*sample*, *raw_data*, type
Pollen	name, token, name_en, name_ger, name_nor
PollenCount	*counting*, *pollen*, number
LuminescenceDating	laboratory_id, *sample*, *raw_data*, sample_id_cll, mineral, dating_approach, luminescence_age, age_error, signal, protocol, palaeodose_value, palaeodose_error, dose_rate, dose_rate_error, published, year_of_publication, thesis, comments, grain_size_min, grain_size_max, aliquot_size, aliquot_number_used_for_palaeodose, od_percent, od_percent_error, od_gy, od_gy_error, age_model, beta_source_calibration, instrumental_beta_source_error, uncertainty_beta_source_calibration, fading_correction, g_value, g_value_error, lnat_lsat_ratio, dose_rate_measurement_technique, dose_rate_calculation_software, u_ppm, u_ppm_error, th_ppm, th_ppm_error, k_percent, k_percent_error, water_content_for_dating, water_content_for_dating_error, a_value, a_value_error, alpha_dose_rate, alpha_dose_rate_error, beta_dose_rate, beta_dose_rate_error, gamma_dose_rate, gamma_dose_rate_error, cosmic_dose_rate, cosmic_dose_rate_error
RadiocarbonDating	*sample*, *raw_data*, lab, lab_id, age, error
Parameter	name, token, unit, minimal_limit, maximal_limit, classes
MeasurementSeries	datetime
GenericMeasurement	*sample*, *raw_data*, *measurement_series*, sample_weight, *method*, *parameter*, value, error
GrainSize	*sample*, *raw_data*, sample_weight, sample_concentration, method, classes, measured_data, clay, fine_silt, medium_silt, coarse_silt, fine_sand, medium_sand, coarse_sand, mean, mode, median, std, skew, kurtosis, fwmean, fwmedian, fwsd, fwskew, fwkurt
MicroXRFMeasurement	*sample*, measurement_date, *method*, notes
MicroXRFElementMap	*measurement*, element, raster_file, unit

Table A3. Bibliography.

Table	Attributes
Author	last_name, first_name, *user*
ReferenceKeyword	keyword, keyword_ger
Reference	title, year, published, *parent_publication*, *lead_author*, *second_author*, *supervisor*, abstract, journal, volume, number, pages, publisher, location_of_publication, type, *project*, doi, issn, isbn_print, isbn_online, how_to_cite, *keywords*

Table A4. Field Data.

Table	Attributes
Country	name, iso_code, geometry
Province	name, *country*, geometry
Tag	*content_type*, word, slug, *project*
SampleType	word, label
StudyArea	label, *project*, *province*, geometry, climate_koeppen, ecozone_schultz
Site	label, *study_area*, *tags*
campaign	label, *project*, date_start, date_end, *destination_country*, *study_areas*, season
Transect	identifier, *study_area*, *campaign*, description, multiline
ExposureType	main_type, abbreviation, name_ger, name_en
Location	data_source, *campaign*, identifier, *project*, *reference*, date_of_record, easting, northing, srid, location, altitude, *study_site*, *transect*, *processor*, exposure_type, line, sampling, gradient_upslope, gradient_downslope, slope_aspect, relief_description, current_weather_conditions, past_weather_conditions, tags
Layer	*location*, identifier, token, description, depth_top, depth_bottom, structure, fine_soil_field, munsell_hue_value, munsell_hue, munsell_value, munsell_chroma, calcite, secondary_calcite, *tags*
Sample	identifier, igsn, *project*, date, *location*, *processor*, *parent*, description, material, *layer*, weight, depth_top, depth_bottom, *type*, *tags*

Table A5. Laboratory.

Table	Attributes
Manufacturer	name, website
Device	name, description, token, *manufacturer*
Accessory	*device*, name, description
AccessoryParamter	method, accessory, parameter_name, parameter_value, parameter_unit
Method	name, description, token, *device*, category, laboratory, available
calibration	*device*, date, *researcher*, remarks
Firmware	*device*, version, installation_date, changelog

Code availability. The Python source code for the OLTP module is publicly available on GitHub (https://github.com/Cologne-Geomorphological-Software-Lab/CGDB, last access: 14 May 2026) under a permissive MIT license. A persistent, citable version of the repository corresponding to this publication has been archived on Zenodo: https://doi.org/10.5281/zenodo.17869730 (Handy and van der Meij, 2026).

Data availability. The data presented in the case study serve an illustrative purpose of the database. These data are part of ongoing research projects and will be published alongside their corresponding research articles. Until then, requests for data access can be directed to the corresponding author.

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 contact author has declared that none of the authors has any competing interests.

Disclaimer. Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims made in the text, published maps, institutional affiliations, or any other geographical representation in this paper. The authors bear the ultimate responsibility for providing appropriate place names. Views expressed in the text are those of the authors and do not necessarily reflect the views of the publisher.

Acknowledgements. 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 and Julia Rauchalles 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.

To ensure a high standard of communication, we used AI tools to improve the grammar, style, clarity and readability of this manuscript.

Financial support. 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.

This open-access publication was funded by Universität zu Köln.

Review statement. This paper was edited by Lev Eppelbaum and reviewed by C. Kristina Rossavik and one anonymous referee.

References

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

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

Crusoe, M. R., Abeln, S., Iosup, A., Amstutz, P., Chilton, J., Tijanić, N., Ménager, H., Soiland-Reyes, S., Gavrilović, B., Goble, C., and Community, T. C.: Methods Included: Standardizing Computational Reuse and Portability with the Common Workflow Language, Commun. ACM, 65, 54–63, https://doi.org/10.1145/3486897, 2022.

Curdt, C., Hoffmeister, D., Kramm, T., Lang, U., and Bareth, G.: Etablierung von Forschungsdatenmanagement-Services in geowissenschaftlichen Sonderforschungsbereichen am Beispiel des SFB/Transregio 32, SFB 1211 und SFB/ Transregio 228, https://doi.org/10.17192/BFDM.2019.2.8103, 2019.

De Castro, P., Herb, U., Rothfritz, L., and Schöpfel, J.: IGSN – Building and Expanding a Community-Driven PID System, Tech. rep., Zenodo, https://doi.org/10.5281/zenodo.7330498, 2023.

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

Deutsche Forschungsgemeinschaft: Guidelines for Safeguarding Good Research Practice. Code of Conduct, Zenodo, 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, https://doi.org/10.2777/02999, 2018.

Faniel, I. M. and Jacobsen, T. E.: Reusing Scientific Data: How Earthquake Engineering Researchers Assess the Reusability of Colleagues' Data, Comp. Supp. Coop. Wor., 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, Int. J. Geogr. Inf. Sci., 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, Comput. Geosci., 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 (last access: 14 May 2026), 2020.

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

Handy, D. and van der Meij, M.: Cologne-Geomorphological-Software-Lab/CGDB: Implemented Dagster Boilerplate (v1.1.0), Zenodo [code], https://doi.org/10.5281/zenodo.17869730, 2026.

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. International Soil Classification System for Naming Soils and Creating Legends for Soil Maps, 4th edn., International Union of Soil Sciences (IUSS), Vienna, Austria, 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, Comput. Geosci., 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, Geochim. Cosmochim. Ac., 351, 192–205, https://doi.org/10.1016/j.gca.2023.04.024, 2023.

Klump, J., Lehnert, K., Ulbricht, D., Devaraju, A., Elger, K., Fleischer, D., Ramdeen, S., and Wyborn, L.: Towards Globally Unique Identification of Physical Samples: Governance and Technical Implementation of the IGSN Global Sample Number, Data Science Journal, 20, 33, https://doi.org/10.5334/dsj-2021-033, 2021.

Kottek, M., Grieser, J., Beck, C., Rudolf, B., and Rubel, F.: World Map of the Köppen-Geiger Climate Classification Updated, Meteorol. Z., 15, 259–263, https://doi.org/10.1127/0941-2948/2006/0130, 2006.

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.

Leo, S., Crusoe, M. R., Rodríguez-Navas, L., Sirvent, R., Kanitz, A., De Geest, P., Wittner, R., Pireddu, L., Garijo, D., Fernández, J. M., Colonnelli, I., Gallo, M., Ohta, T., Suetake, H., Capella-Gutierrez, S., De Wit, R., Kinoshita, B. P., and Soiland-Reyes, S.: Recording Provenance of Workflow Runs with RO-Crate, PLOS ONE, 19, e0309210, https://doi.org/10.1371/journal.pone.0309210, 2024.

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, e0116781, 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, Philos. T. Roy. Soc. A, 380, 20210300, 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, 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 Int. J. Geo-Inform., 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, Geosci. Instrum. Method. Data Syst., 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, arXiv [preprint], 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, 1981–1984, ACM, Amsterdam, the 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), 13–20, IEEE, Portoroz, Slovenia, ISBN 978-1-7281-9532-2, https://doi.org/10.1109/SEAA51224.2020.00014, 2020.

Soiland-Reyes, S., Sefton, P., Crosas, M., Castro, L. J., Coppens, F., Fernández, J. M., Garijo, D., Grüning, B., La Rosa, M., Leo, S., Ó Carragáin, E., Portier, M., Trisovic, A., RO-Crate Community, Groth, P., and Goble, C.: Packaging Research Artefacts with RO-Crate, Data Science, 5, 97–138, https://doi.org/10.3233/DS-210053, 2022.

Soil Science Division Staff: Soil Survey Manual, no. 18 in USDA Handbook, Government Printing Office, Washington, D.C., ISBN 9780160937439, 2017.

Stocker, M., Paasonen, P., Fiebig, M., Zaidan, M. A., and Hardisty, A.: Curating Scientific Information in Knowledge Infrastructures, Data Science Journal, 17, 21, https://doi.org/10.5334/dsj-2018-021, 2018.

Suter, F., Coleman, T., Altintaş, İ., Badia, R. M., Balis, B., Chard, K., Colonnelli, I., Deelman, E., Di Tommaso, P., Fahringer, T., Goble, C., Jha, S., Katz, D. S., Köster, J., Leser, U., Mehta, K., Oliver, H., Peterson, J.-L., Pizzi, G., Pottier, L., Sirvent, R., Suchyta, E., Thain, D., Wilkinson, S. R., Wozniak, J. M., and Ferreira Da Silva, R.: A Terminology for Scientific Workflow Systems, Future Gener. Comp. Sy., 174, 107974, https://doi.org/10.1016/j.future.2025.107974, 2026.

Tedersoo, L., Küngas, R., Oras, E., Köster, K., Eenmaa, H., Leijen, Ä., Pedaste, M., Raju, M., Astapova, A., Lukner, H., Kogermann, K., and Sepp, T.: Data Sharing Practices and Data Availability upon Request Differ across Scientific Disciplines, Scientific Data, 8, 192, https://doi.org/10.1038/s41597-021-00981-0, 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, 205395172093483, https://doi.org/10.1177/2053951720934838, 2020.

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.

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, Semant. Web, 10, 95–114, https://doi.org/10.3233/SW-180305, 2018.

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, 160018, 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, T. GIS, 18, 496–509, https://doi.org/10.1111/tgis.12060, 2014.