# Notes on resubmitting revision

---

## Reviewer A

**Reviewer Comment A.1** — The paper describes an IT system set up for general use with an example implementation based on soil and sediment information from an archaeological project in Romania. The system consists of an OLTP and OLAP component, tied together with pipelines and each having a (relational) data schema although the OLAP schema is de-normalised.

The work is interesting and not atypical of many IT systems constructed in many scientific academic departments. It is well designed for the purpose. It may provide some guidance for others developing such systems.

The problem is the claim to generality of application, and to interoperability. On both claims, the difficulty is that the IT system is locally designed, clearly heavily influenced by an individual project and takes no account of wider developments in generality and interoperation aspects of IT systems (across a range of scientific disciplines). Thus, there is no motivation for this work to be seen as part of global information (see detailed comments).

There is no reference – for example – to the metadata schemas of the European Open Science Cloud which aims for the generality and interoperability this paper claims (but does not demonstrate). It does not mention more recent work on Scientific Knowledge Graphs. It does not mention the leading large geoscience IT systems in Europe (which provide generality and interoperability) nor those in North America, East Asia and Australasia. The data model and schemas are not compared with those of the aforementioned systems.

The pipelines (workflows) appear not to use CWL (Common Workflow Language) which is a basis for generality and interoperability. Similarly, in many scientific disciplines the use of RO-Crates (Research Object Crates) is widely encouraged for generality and interoperability (including provenance and reproducibility).

The difficult problem of semantic consistency and formalisation is mentioned in lines 375-380. However, this is consigned to future work and there is no roadmap or plan of how such semantic consistency – for generality and interoperation – is to be achieved except a mention of discipline-related thesauri and a nod to ontologies. The key research question is how to bridge across heterogeneous thesauri – it requires semantic relationships and is likely to include also probabilistic measures.

**Reply**: We would like to thank you very much for your careful review of our manuscript, and for the detailed, insightful feedback you provided. Your comments have been extremely helpful in enabling us

to critically reflect on and refine the positioning and conceptual contribution of our work. We believe that your helpful and critical review has improved our revised manuscript, making it much clearer and more effective in communicating the ideas and added value of our framework, and positioning it better within the scientific landscape.
Your assessment that our approach is 'not atypical of many IT systems constructed in scientific academic departments' has made it clear to us that we have not emphasised the strategic and conceptual novelty of our work enough in its current form. Our primary concern is not to develop another self-contained software package. Instead, we are presenting a modular research data framework as an architectural approach which would hopefully inspire cultural and organisational change in academic data management. Unlike many typical systems, which focus on static storage or final applications, our approach is holistic and supports the entire research data lifecycle.
We acknowledge the validity of your criticism of the claim to universal validity and interoperability. It has made us realise that we need to distinguish more clearly between our bottom-up approach, which is specifically aimed at small, resource-limited projects, and global, top-down infrastructures. In our view, criticising the system as 'locally designed' confuses the specific implementation with the general concept.
The fact that we do not discuss the metadata schemas of the European Open Science Cloud, similar structures, or research on scientific knowledge graphs is due to the level at which our framework is situated and the purpose it serves. We explicitly mention this distinction (between laboratory information systems, repositories, and data catalogues) in the discussion, which is now located in the revised chapter "Positioning the framework within the research data ecosystem". Through the revision and now clearer description and discussion of our framework, as well as the adaptation of the abstract, we believe that it is now clear that a discussion of the aforementioned structures and concepts would go beyond the scope of our manuscript.
To address these issues and improve the clarity of the manuscript, we implemented the following major revisions based on your feedback. We hope these revisions address the criticisms you raised and clarify the contribution of our manuscript, which provides a pragmatic architectural blueprint for data management in research:

1. We have completely restructured the discussion. We now first position our work within the greater *research data ecosystem* before discussing how our framework addresses the challenges in data management, as outlined in the similarly revised chapter on challenges. We now discuss reusability and interoperability together before discussing workflow complexity, data heterogeneity and preservation. We now dedicate interoperability and the contribution and limitations of our approach in much greater detail. The discussion concludes with a discussion of the framework's applicability, which we have edited to remove redundancies.

2. We are now addressing global standards in the discussion. We argue that, unlike large repositories or knowledge graphs, our framework has a different objective. Rather than aiming for final archiving, we facilitate the generation of FAIR-compliant data during active research by integrating standardised storage solutions and data orchestration. We lay the technical groundwork (internal standardisation) that will enable future connection to external standards. However, we agree with the reviewer's assessment and now explicitly mention in the discussion the possibility of generating RO crates and CWL workflows from our framework in the future.

3. The review has made it clear to us that the manuscript could be significantly improved by clarifying

the concepts of our data model, OLTP, OLAP, 'data orchestration', 'data pipeline' and how it differs from 'scientific workflows'. We have specifically restructured and revised the sections on 'design and implementation' to more clearly distinguish between the respective concepts, technical implementation, and application.

- We have grouped the user interface, initial data processing, and the database, including its model, in the OLTP chapter to illustrate their architectural unity.
- We have significantly expanded the description of the data model. We now describe in greater detail, how it builds on the preliminary work of Nordsiek and Halisch, and how we are expanding it according to relational concepts by storing data directly in the database.
- We have revised the chapter on OLAP to more clearly distinguish between the concept, technical implementation and the respective project-specific applications. The difference between our approach and 'data warehouses' is now also becoming clearer.
- To better highlight the difference between the conceptual idea, the technical platform, and the project-specific implementation of actual pipelines, we have renamed the chapter 'Data Pipelines' to 'Data Orchestration' to underline that our framework provides the platform rather than the project or environment-specific pipelines.
- A 'data pipeline' (implemented with Dagster in our framework) refers to the technical, automated process of data orchestration, involving the extraction, transformation and loading of data. A scientific workflow, for which CWL is an excellent tool, describes the conceptual steps at a higher level. We now distinguish both concepts more clearly by defining in the manuscript pipelines as the computational steps of workflows.

4. To better reflect the revised structure and objectives of our discussion and to better clarify the current scientific context of our work, we have reorganised the chapter on challenges in (geo)scientific research data management and clarified the wording to define the scientific context and objectives of our work.

5. We agree with the reviewer that achieving semantic consistency is a major challenge. We are now emphasising more strongly in the 'Challenges' and in the 'Discussion' that resolving these semantic differences requires community-driven efforts that extend beyond the scope of our framework, and that instead, we have established the technical prerequisite for doing so. By creating an internally consistent, provenance-based data environment, we are preparing the data for the semantic mapping and bridging that will be developed through collaborative efforts. The revised version of the manuscript explains this more clearly, emphasising that our work is a prerequisite for addressing the major challenge of semantic interoperability, rather than an alternative to it.

6. To reinforce the conceptual separation from implementation and increase transparency, we published the code of the OLTP module along with a boilerplate for data orchestration on Zenodo.

**Reviewer Comment A.2 — Line 19 - "Wilkinson et al. introduced the FAIR principles"** The FAIR principles were produced by FORCE11 `https://force11.org/info/the-fair-\ data-principles/`, although Wilkinson et al elaborated their interpretation and even more formally by the RDA Working Group on FAIR Data Maturity `https://zenodo.org/records/3909563#.YGRNnq8za70`

**Reply**: Thank you for this important clarification regarding the origins of the FAIR principles within the community. You are right that the principles emerged from a broad community process within FORCE11. We decided to cite Wilkinson et al. (2016) in line with standard academic practice, as this was the formal, peer-reviewed publication that introduced and defined the principles for the scientific record. As noted on the FORCE11 website that you provided, this article formally published the principles. To acknowledge this important nuance, we revised the sentence in the manuscript to better reflect the community origin.

**Reviewer Comment A.3 — Line 38 - "For this reason, spatial data must have explicit locational information in its metadata"** For FAIR all datasets (and for that matter software services and workflows) need to have metadata. This is not discussed in the paper, nor is there any suggestion of adopting/improving widely used metadata standards such as DCAT `https://dcat.org/` and particularly extensions (APs: Application Profiles) that allow for domain specialisation while retaining interoperability and generality through the main entities of DCAT. Use of such standards improves greatly interoperability.

**Reply**: This passage does not deny that detailed metadata is required for a dataset to comply with the FAIR principles. Instead, we focus on the specific challenges associated with spatially explicit data, which we consider to be a unique feature of geoscientific data. While we agree with the importance of metadata, the demand for further discussion on metadata standards misses the scope and objective of the presented system. As mentioned previously, we are not introducing a system for sharing static scientific data. However, DCAT refers to "an RDF vocabulary for representing data catalogues" (https://www.w3.org/TR/vocab-dcat-3/#dcat-scope). Furthermore, DCAT is built around distinct datasets which "represent a collection of data published or curated by a single agent or identifiable community". Demanding its adoption, or even improvement, within the scope of our work misconstrues not only the nature of relational databases, but also the objectives and properties of our approach as a whole. If anything, the application of DCAT comes after our approach.
To clarify this, we rephrased the sentence. This suggests that the paper's aim is not to provide a comprehensive discussion of metadata's significance, but to address the inherent structural challenges of geoscientific data.

**Reviewer Comment A.4 — Line 46 - "a comprehensive, interdisciplinary approach has been missing"** See EPOS https://www.epos-eu.org/ and for workflows its extension through DT-GEO https://dtgeo.eu/ and/or EGDI https://www.europe-geology.eu/

**Reply**: Thank you for highlighting these significant European initiatives. You are right that large-scale infrastructures such as EPOS and EGDI offer comprehensive data integration platforms at a macro level. However, our work focuses on a different — and, we would argue, complementary — part of the research data lifecycle. These platforms operate at a macro-infrastructural level, focusing primarily on the aggregation, harmonisation and publication of data from various sources. In contrast, our framework is a 'bottom-up' solution operating at the micro-level of an individual research project, particularly those with limited resources. Our primary goal is not data publication, but rather the management of data throughout the active research process — from field collection to final analysis.
We believe that effective management of data at its point of origin is a crucial prerequisite for it to be FAIR enough to be ingested into a system like EPOS. Our framework enables smaller projects to achieve this. DT-GEO's focus on digital twins for geophysical modelling is conceptually and technically distinct

from our goal of providing a foundational data management architecture.
To clarify our position, we have strengthened the positioning of our work within the research data ecosystem to explicitly distinguish it from large-scale publication and integration infrastructures. This will better define the specific gap that our work aims to fill and demonstrate that our approach facilitates the preparation of data for inclusion in larger, domain-specific databases.

**Reviewer Comment A.5 — Line 46 - "a comprehensive, interdisciplinary approach has been missing"** The paper does not provide a solution that is comprehensive; it is limited to a particular area of geoscience (sediment and soil samples)

**Reply**: Thank you for raising this important point. You are right that our case study focuses on soil and sediment samples. This gives us a valuable opportunity to clarify the intended meaning of 'comprehensive' in our manuscript. The comprehensiveness we claim is architectural and procedural, rather than disciplinary. Our framework is comprehensive in that it covers the entire research data lifecycle — from initial data collection in the field, through processing and analysis, to preparation for reuse. While the case study is specific, it demonstrates how this general architectural blueprint can be implemented in a real-world research context. Although the specific data models and workflows would change, the core components are designed to be adaptable to other geoscientific domains.
We recognise that our original wording was ambiguous. Therefore, we revised the abstract, the 'Design and Implementation' chapter, and the discussion to explicitly state that the claimed comprehensiveness refers to coverage of the data lifecycle rather than universal disciplinary applicability.

**Reviewer Comment A.6 — Fig. 2 - The OLTP/OLAP architecture is not suitable for real-time or event-driven workflows** Re-think the architecture to allow for real-time data ingestion and inline analysis to detect events if it is to be comprehensivee

**Reviewer Comment A.7 — Fig. 2** The figure does not include cardinality and optionality symbols.

**Reviewer Comment A.8 — Fig. 2 - The entity sample seems to imply a physical sample of sediment or soil** Consider a sample could also be e.g., a digital seismogram or a chemical analysis of a volcanic gas if it is to be comprehensive

**Reviewer Comment A.9 — Fig. 2** There is no clear indication of what is data and what is metadata, and the schema does not match well-known schemas used in geoscience or schemas used generally in research

**Reply**: Thank you for your detailed and constructive feedback on Figure 2. Your comments provide us with a valuable opportunity to clarify the purpose of the figure and the core design principles of our framework. The central point that addresses several of your comments is that Figure 2 is intended as a conceptual illustration, rather than a complete and prescriptive implementation schema. Its primary purpose is to contrast the structure of a normalised OLTP schema (Fig. 2a) with that of a denormalised OLAP star schema (Fig. 2b), thereby explaining the architectural trade-offs. It is not intended to propose a new universal standard for geoscientific data. To make this clearer, we will revise the figure and its caption to explicitly state that it is a conceptual model's purpose. With this in mind, we would like to address your specific points.

1. **Data vs. Metadata & Schema Standards:** You are right that the figure does not depict a specific schema, such as the International Generic Sample Number (IGSN), or label entitiy types explicitly as data or metadata. This is because our framework adheres to a fundamental principle of relational database design, whereby the distinction is logical rather than physical. Both data (e.g. a measurement value) and metadata (e.g. its unit) are stored as values in columns. The context is defined by the relationships between tables. This approach avoids the rigid separation of data and metadata into different systems, a problem that we address explicitly. We clarified the logical handling of metadata in the revised figure caption.

2. **Definition of Sample:** Thank you for highlighting this ambiguity. In our model, a sample is deliberately defined as a physical specimen (e.g. sediment or rock). Other items, such as chemical analyses or seismograms, are modelled as analyses or measurements associated with that sample. This conceptual separation is fundamental to the fidelity of the framework. To avoid ambiguity for a broader audience, we added a clarifying sentence to the text.

3. **Cardinality and Optionality:** We agree completely. Adding this notation will make the conceptual model more informative. We revised Figure 2 to include the correct symbols in the updated manuscript.

4. **Real-time vs. Batch Processing:** You are right to point out that our architecture is not a real-time stream processing engine. This is a deliberate design choice tailored to the dominant workflows in our target geoscientific domain, in which data is usually processed in batches. However, we would like to clarify that our architecture fully supports event-driven workflows. The decoupled orchestration layer (Dagster) can trigger pipelines based on events such as the arrival of a new data file. As the core message of the paper is an architectural blueprint for batch-oriented research projects, we feel that a detailed discussion of real-time architectures would be beyond the scope of the paper.

We are confident that these clarifications and revisions will address your concerns and significantly improve the clarity of our manuscript. However, we believe that the criticism stems from a misunderstanding of the objectives and architecture of our framework. We therefore think that the original presentation of our approach was insufficient to communicate our idea clearly. For this reason, we have fundamentally revised the chapter 'Design and implementation'. To further clarify how our model builds upon that provided by Nordsiek and Halisch (2024) and takes it further in terms of relational strictness, we revised chapter 3.1. We describe the data model now in greater detail. Furthermore, we added tables to the appendix, depicting in detail the entity types and their attributes. We now explain the concept of OLAP databases in greater detail and, in contrast to OLTP, highlight that they do not serve as a single central data storage but as project-specific analytical databases. We completely revised the chapter on data pipelines and renamed it 'data orchestration' to distinguish between the concept, the technical framework and the institution- and project-specific pipelines.

**Reviewer Comment A.10 — Line 161 - relationships into smaller tables that are linked by foreign keys** More modern relational database work has relationships themselves as entities (tables) linking e.g., sample and device thus a tuple in each of the base tables is related by the relationship table and can be read as e.g., sample X - was collected by - device Y (ideally with added temporal and spatial information in the relationship table)

**Reply**: Thank you for this precise suggestion. You are right that the most robust way to model many-to-many relationships is to use a dedicated relationship table. This is a well-established principle of relational database theory (e.g. Codd, 1970) and our OLTP module strictly adheres to it. The reason this level of detail is not immediately apparent in Figure 2 is that the diagram is a conceptual simplification designed to illustrate the high-level differences between OLTP and OLAP structures. As mentioned in our previous response, we revised chapter 3.1 with a detailed description of the data model and added a tabular description of the model's entity types to the appendix to address your point fully and make our implementation explicit.

**Reviewer Comment A.11 — Line 191** In this section there is no discussion of using CWL (Common Workflow Language)https://www.commonwl.org/ which would allow for interoperability / reusability, nor the use of 'research object crates' https://www.researchobject.org/ro-crate/ to enable interoperability / reusability: reasons for rejecting these solutions should be explained

**Reply**: Thank you for raising this important point. As detailed in our main response to your general critique, we addressed this issue by revising the chapter on 'Data Pipelines' completely to clarify the distinction between data pipelines and scientific workflows, such as CWL. Moreover, the discussion section emphasises that our framework provides a foundation for, rather than replaces, these important standards.

**Reviewer Comment A.12 — Line 310 - geochemical fingerprints** No mention of geochemical measurements in the schema (Fig 2)

**Reply**: You are correct that geochemical measurements are not included in Figure 2. This is because, as detailed in our previous responses, the figure is a high-level conceptual illustration. In our actual data model, however, a geochemical fingerprint is handled exactly as one would expect in a robust relational system: it is modelled as a specific type of analysis linked to a physical specimen via a relationship. This flexible approach is a core strength of our design. It is an excellent example of the implementation-level detail omitted from the conceptual figure for clarity, which is now made explicit in the detailed description of our data model in both the manuscript and its appendix.

**Reviewer Comment A.13 — Line 369 - integrating metadata and measurements into a unified data model, ensuring that all information is stored consistently and is easily accessible.** Even for small projects it is possible that datasets are stored on different servers in discipline-based laboratories (e.g., geochemistry on one server, grain size data on another) and so the model of metadata in a database on one server (possibly replicated) pointing to files on many different servers may be more generally applicable

**Reply**: The situation you describe, of datasets being stored on different servers in discipline-based laboratories, is precisely the core problem of data fragmentation and siloing that our framework is designed to solve.
Your suggestion of a metadata model pointing to files on different servers leaves the data physically separated by design, which severely limits data integration and complex cross-dataset analysis. It is not possible to run a single query across geochemical data on one server and grain size data on another.
Our framework takes a different, more powerful approach. Rather than merely pointing to distributed files, the combination of a relational database and data orchestration is designed to ingest data from

various sources (e.g., geochemistry and grain-size servers) and integrate them into the unified system. The purpose of our architecture is to move beyond a simple catalogue and create a single, consistent, queryable environment where geochemical and grain-size data reside together. This enables advanced integrated analyses that are impossible in the federated model you describe.
We acknowledge that our manuscript did not make this fundamental design choice and its rationale explicit enough. The fact that our solution can be mistaken for the problem it solves is a clear sign that we need to improve our explanation. Therefore, we revised both the description of the OLTP and the OLAP modules, explaining the trade-offs and justifying why our deep integration is essential for achieving the analytical goals of modern, data-driven geoscientific research.

**Reviewer Comment A.14 — Line 461 - expanding interdisciplinary collaboration**
(Some) other disciplines have well established metadata schemas, workflows (pipelines): there is no explanation in the paper to support the assertion i.e. how would the system interact with extant systems in geoscience, soil science, climate science, environmental science, biodiversity... and even archaeology?
The very nature of the system is to be small, project-based, using locally-designed solutions, limited by the institutional IT infrastructure.

**Reply**: Thank you for this important question regarding the final sentence of our manuscript. It highlights a misunderstanding that we will address in two points.
1. Our framework is a scalable blueprint, not a 'small, local' solution. The comment asserts that our system is 'small, project-based and uses locally designed solutions'. This is incorrect. It stems from confusing an instance of the framework, such as the Toboliu case study, with the framework's standardised, project-independent design. Our concluding sentence is the central thesis of the paper. It is based on the framework's standardised design, which can be deployed for any number of projects. The architecture's entire purpose, especially that of the OLAP module for large-scale, cross-project analysis, is the opposite of a 'small, local' solution since it is used to 'manage and integrate datasets across multiple projects'.
2. Our final statement is not a vague assertion. It is based on the most important capability of our framework: the deep integration of heterogeneous data into a unified model. While the manuscript does not detail specific application programming interfaces (APIs) for external systems, it does explain the foundational mechanisms that facilitate interdisciplinary collaboration.

- **Overcoming Data Silos:** The main issue with interdisciplinary work is that data is fragmented. Our framework addresses this issue by consolidating data from various sources, such as geochemistry and sedimentology, into a single, unified system. This internal integration is the essential first step for any meaningful external collaboration. You cannot share what you have not first integrated.
- **Enabling a "Single Source of Truth":** Our framework transforms diverse data into a consistent structure, creating a single, queryable source of truth. This unified dataset provides the necessary foundation for building interoperable data products or APIs for other systems, such as those in soil and climate sciences.
- **Flexibility through Pipelines:** The 'transform' step in our ETL pipelines is an architectural component designed explicitly to handle data heterogeneity. Although we have only demonstrated this with internal data, the same mechanism could be used to implement transformations from external schemas, thereby making the framework extensible by design.

We concede that the manuscript does not make this strategic position explicit enough. The fact that our scalable blueprint could be mistaken for a small, local solution demonstrates the need for further elaboration. We therefore explicitly address this in the new section 'Positioning the Framework in the Research Data Ecosystem'. We clarify that our primary contribution is the creation of a deeply integrated, analysis-ready data asset. This asset is the foundational prerequisite for subsequent steps, such as interdisciplinary data sharing and interaction with larger, external systems.

**Reviewer Comment A.15** — **Line 462 - Code availability.** This paragraph indicates that the code is not open source. Hardly FAIR (which applies to software and workflows as well as data)

**Reply**: We agree that the goal of making software and workflows as open and accessible as possible is crucial. Your comment has made it clear that the 'Code availability' statement was inadequate and misleading. Your comment prompted us to reconsider how best to support the FAIR principles, which involves clearly separating these two components. The OLTP module is the core intellectual contribution and the most reusable part. To maximise the reusability and FAIRness of our work and encourage its adoption and further development by the community, we published the Python code for the OLTP module along with a boilerplate for data orchestration under a permissive open-source licence on Zenodo to make it findable, accessible and reusable for any research group wishing to adopt our approach. We revised the 'Code availability' section of the manuscript accordingly.

---

## Reviewer B

**Reviewer Comment B.1** — *General Comments*
A report that presents data organization structure for geoscientific databases with geospatial properties and data pipelines.
*Specific Comments*
May want to use a more updated citation related to (Zickel et al., 2025).

**Reply**: We would like to thank the reviewer for their accurate summary of our work and the hints for technical corrections. With regard to the specific comment on the Zickel et al. (2025) citation, the reviewer is right to draw attention to it. Since, contrary to our original expectation, the cited paper has not been published and will not be published in the foreseeable future, we have decided to reword the relevant paragraph so as not to present facts without a basis in the literature.