Title: Metadata traceability and reuse in research projects

Authors: Miguel Ángel Sicilia & Alberto Ballesteros Rodríguez

Date: March 2025

Institution: Biomedical Data Science and Engineering (IRYCIS)

Introduction

In research projects, the use of Data Management Plans (DMPs) is becoming increasingly common. These documents describe the data used and generated throughout a project, aligning with the FAIR principles [1] (Findable, Accessible, Interoperable, Reusable). Machine-actionable DMPs [2] (maDMPs) are structured, standardized, and encoded versions of DMPs that allow systems to automatically interpret, validate, and process them. This enables automation, dynamic updates, and seamless integration with other research tools, enhancing efficiency, reproducibility, and transparency. One of the most established standards for developing maDMPs is the RDA-DMP Common Standard [3], promoted by the Research Data Alliance (RDA). To support the creation and validation of maDMPs according to this standard, tools like the Python library madmpy[1] offer a programmatic interface that enables their seamless integration into data workflows and research applications.

In medical research, REDCap[2] [4] (Research Electronic Data Capture) is one of the most widely used tools for data collection and management, extensively adopted by hospitals and research institutions. However, REDCap stores data in isolated silos within each project, making it difficult to integrate with an organization-wide repository that spans an entire institution, such as a large hospital or a network of research centers. This data fragmentation poses challenges for traceability, governance, and reuse, as researchers struggle to determine what data was collected and utilized across multiple projects.

Additionally, there is currently no direct link between DMPs and the data stored in REDCap, making it impossible to verify whether the data used or generated in a project aligns with what was initially planned in the DMP. Without this connection, compliance validation and data standardization efforts become increasingly complex [5].

A metadata management system, such as OpenMetadata[3], offers a structured solution by centralizing and cataloging metadata across projects [6]. OpenMetadata enables abstraction, making it easier to organize, track, and reuse metadata while ensuring compliance with institutional and international data standards. Developing an automated integration pipeline between REDCap and OpenMetadata enables research institutions to bridge the gap between DMP planning and real-world data collection. This solution enhances data traceability, ensures adherence to standards, and improves metadata interoperability across various research initiatives.

Proposal

To address the challenges associated with data isolation in REDCap and the lack of integration with organization-wide repositories, we propose an automated pipeline that

[1] https://madmpy.readthedocs.io/en/latest/
[2] https://project-redcap.org/
[3] https://open-metadata.org/

links REDCap projects with a metadata management system like OpenMetadata. This ensures that each exported project is stored in a database with the same name, allowing for seamless integration in OpenMetadata. Also, mapping project metadata and enriching it with standardized ontology tags, such as OMOP[4] CDM codes, provide additional standardized information about each attribute, improving metadata discoverability and semantic consistency across projects

Technical implementation

The whole process to achieve complete integration consists of several key steps, as shown in Figure 1.

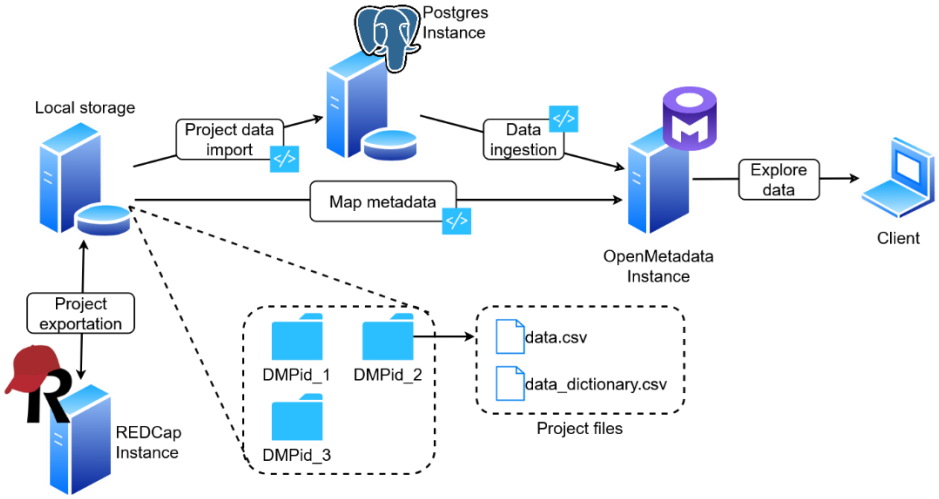

Figure 1 - Architecture

First, we propose using the REDCap project name as an identifier for the corresponding DMP. This ensures a structured relationship between the DMP and the REDCap project, allowing OpenMetadata to create a database entry with the same name when performing ingestion. This linkage establishes that project metadata remains clearly associated with its respective research context.

Next, data and metadata are exported from REDCap as structured files, including the dataset and its corresponding data dictionary. These files are then processed and imported into a PostgreSQL[5] database through a script adapted to the project's data, which creates the necessary tables and preserves the data structure for further processing. Following this, a data ingestion pipeline integrates automatically the PostgreSQL database into OpenMetadata. This process registers the metadata structure, including variables, data types, and relationships. While actual data values are included, sample data visibility is disabled to prevent unauthorized access to sensitive information.

After the ingestion phase, the tag mapping process is executed. Each attribute within the data dictionary is analyzed, and OMOP CDM codes stored within a specific annotation column in PostgreSQL—originally derived from REDCap—are assigned as metadata tags.

[4] https://www.ohdsi.org/data-standardization/
[5] https://www.postgresql.org/

Although OMOP CDM tagging is the primary focus, the approach can be extended to incorporate other tagging mechanisms as needed.

Finally, once the metadata is fully integrated and enriched, researchers and data stewards can explore the cataloged metadata within OpenMetadata, allowing them to analyze which metadata attributes have been collected across different projects.

Through this implementation, institutions reinforce metadata centralization, traceability, and enrichment, enhancing data governance and interoperability across research environments. Additionally, this approach supports data reusability, ensures compliance verification, and supports seamless integration with external systems, ultimately promoting a more structured and accessible research ecosystem.

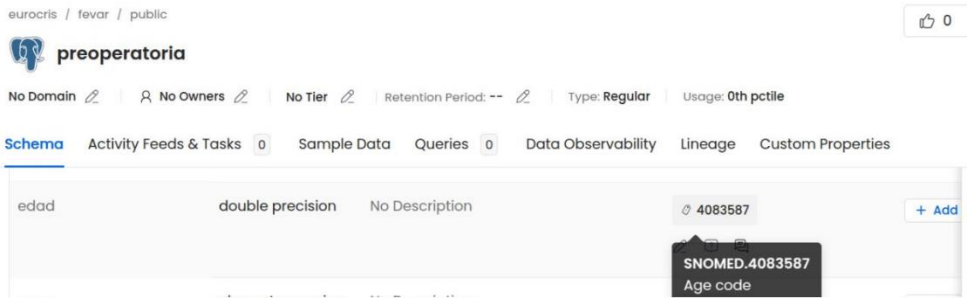

Figure 2 - Metadata Discovery

References

[1] Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al*. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

[2] Tomasz Miksa, Simon Oblasser, and Andreas Rauber. 2021. Automating Research Data Management Using Machine-Actionable Data Management Plans. ACM Trans. Manage. Inf. Syst. 13, 2, Article 18 (June 2022), 22 pages. https://doi.org/10.1145/3490396

[3] Miksa, T., Walk, P., & Neish, P. (2020). RDA DMP Common Standard for Machine-actionable Data Management Plans. Zenodo. https://doi.org/10.15497/rda00039

[4] Patridge EF, Bardyn TP. Research Electronic Data Capture (REDCap). J Med Libr Assoc. 2018 Jan;106(1):142–4. doi: 10.5195/jmla.2018.319. Epub 2018 Jan 2. PMCID: PMC5764586.

[5] Modly, L. A., & Smith, D. J. (2022). The need for data management standards in public health nursing: A narrative review and case study. Public Health Nursing, 39, 1027–1033. https://doi.org/10.1111/phn.13066

[6] Garnett, A., Leahey, A., Savard, D., Towell, B., & Wilson, L. (2017). Open Metadata for Research Data Discovery in Canada. Journal of Library Metadata, 17(3–4), 201–217. https://doi.org/10.1080/19386389.2018.1443698