Opening the Personal Gate between Technology and Health Care
M. Giacomini et al. (Eds.)

doi:10.3233/SHTI260407

DataCastle: A Pragmatic Approach for Research and Real-World Data Management

Jori KERN[a,b,c,d,e,1], Markus Katharina BRECHTEL[f,g], Tim SCHUMACHER[a,b,c,d,e], Martin LABLANS[a,b,c,d,e], and Tobias KUSSEL[a,b,c,d,e]
[a] *Federated Information Systems, German Cancer Research Center (DKFZ) Heidelberg, Germany*
[b] *Complex Medical Informatics, Mannheim Institute for Intelligent Systems in Medicine, Medical Faculty Mannheim of Heidelberg University, Mannheim, Germany*
[c] *German Cancer Consortium (DKTK), Core Center Heidelberg, Germany*
[d] *DKFZ Hector Cancer Institute at the University Medical Center Mannheim, Germany*
[e] *Helmholtz Institute for Translational Oncology Mainz (HI-TRON Mainz) – A Helmholtz Institute of the DKFZ, Mainz, Germany*
[f] *Goethe University Frankfurt, Faculty of Medicine, Institute for Digital Medicine and Clinical Data Science, Germany*
[g] *University of Cologne, Faculty of Medicine and University Hospital Cologne, Department I for Internal Medicine, Cologne, Germany*

Abstract. Effective research data management (RDM) is essential for ensuring transparency, reproducibility, and collaboration in biomedical research, yet heterogeneous clinical and experimental data challenges fragmented infrastructures. DataCastle is a modular open-source platform that bridges FAIR data acquisition and FAIR data use by integrating enrollment-time pseudonymization, metadata extraction and background versioning of data and a processing environment for data analysis. Structured data are captured via an EDC system, unstructured data are stored within a filesystem-based data lake and metadata are mapped to Health DCAT-AP to support findability and EHDS alignment. DataCastle connects managed data to analysis and visualization tools to enable reproducible and version-linked workflows.

Keywords. Research Data Management (RDM), FAIR Principles, Interoperability

1. Introduction

Research Data Management (RDM) plays a crucial role in biomedical research by systematically collecting, organizing, documenting, and preserving data for long-term use. Thus, effective RDM is of importance for scientific progress by ensuring transparency, reproducibility, and collaboration. In biomedical research, heterogeneous and multimodal data (e.g. images, tables, documents, …), are required to answer research questions, hence integrating and managing this variety of data poses significant challenges. Compliance with the FAIR Principles by making data Findable, Accessible, Interoperable, and Reusable (1) is essential to address those challenges and prepare these data for secondary use. The DataCastle aims to provide a modular and open approach to RDM that goes beyond managing data but also includes versioning, structured interfaces, and applications to process data. Instead of focusing solely on storage, it enables users

[1] Corresponding Author: Jori Kern, j.kern@dkfz.de.

to *work* with the managed data as a part of established workflows and scientific best practices, building on existing concepts of trusted research environment architectures. The scientific contribution of this work is the specification of an end-to-end workflow from the data collection, via pseudonymization and metadata provisioning, towards – finally – the exploration and the orchestrated (re)use of heterogeneous research data.

2. Background and Related Work, state of the art

RDM in biomedical research relies on a multitude of specialized tools, including:

- Electronic Data Capture (EDC) tools such as REDCap (2) or OpenClinica (3) to capture structured data and manage studies.
- Electronic Lab Books (4) like LabFolder, eLabFTW, and RSpace are used for laboratory documentation and experiment management.
- RDM Platforms (5), such as Dataverse, RDMO, OpenBIS, focus on data integration, publication, and repository management. However, their adaptation to clinical workflows often requires additional customization and they usually lack native components for pseudonymization or integrated environments for data visualization and analysis. Domain-specific platforms, such as OMERO (6) and ABCD-J (7), integrate multiple steps in the data collection and processing workflow, but they are often highly optimized to their respective research domains and their data types.

Clinical data involving personally identifying information (PII, e.g., names, dates of birth) must be pseudonymized to be reused in scientific research, so an RDM focused on clinical usage should include pseudonymization in its data collection workflow. For the prototype implementation we decided to use Mainzelliste (8,9) as a pseudonymizationtool, since it is a freely available, open-source solution that is well established in multiple research networks, allowing reuse of existing pseudonymization workflows. The application and versioning layer are built on the previously published concept of the Data Science Orchestrator (10) to ensure that data in the DataCastle are not only managed but can be used directly within the applications to visualize or analyze the data, while continuously ensuring replicability. In the near future, additional requirements will be imposed on clinical data by the European Health Data Space (11), which mandates metadata to make data explorable and findable. The DataCastle can help meet these requirements.

3. Methods

The DataCastle architecture is modular to enable the interchangeability of its components. It provides ready to use “data plumbing” to flexibly connect existing selected existing open-source solutions in a coherent integrated platform, provided as a Docker compose file. At its core, it consists of an EDC System that enables collection of structured clinical data by providing access to partners and patients. To enable pseudonymization in patient data acquisition and the linkage of records across multiple data sources, a local pseudonymization tool called Mainzelliste is used in the DataCastle, acting like a trusted third party (TTP). Since Mainzelliste supports privacy-preserving record linkage via secure multi-party computation (SMPC) (12), enabling multiple DataCastle instances to securely link their datasets among each other without the need of a trusted third party. Utilising a callback during the pseudonymization an automated script creates the patient in the EDC System and provides a path in the filesystem. To store unstructured data of heterogeneous nature, this filesystem functions as a data lake, capable of storing arbitrary document-based data regardless of format. In the prototype we used Nextcloud. To make the data within the DataCastle explorable and findable, a MetadataExtractor component integrating ExifTool (13) provides metadata to make the data visible and, if of clinical nature, aligning with the EHDS by using the Health DCAT-AP metadata format. This MetadataExtractor is extensible to enable metadata extraction for data that goes beyond the standard ExifTool capabilities, if needed. To ensure data reproducibility and reusability, the data must be versioned. To version data within Nextcloud, its WebDAV interface is used in combination with Git-annex WebDAV special remote, keeping versioning processes in the background to minimize impact on the user experience (14). The versioning information can also serve as additional metadata to show how datasets evolve over time. DataCastle also encompasses the data usage process. To achieve this, the architecture of the Data Science Orchestrator again serves as a model for how different applications can operate within the DataCastle. This includes well established tools like JupyterLab and R-Studio, and, considering hardware requirements, can be extended to tools for machine learning and other AI-based tasks. The DataCastle also includes visualization tools, such as a DICOM Viewer (implemented with Orthanc (15) as a DICOM server in the prototype), as well as UI tools that help principal investigators (PIs) manage data collection.

4. Result

To clarify the scope of the DataCastle approach and the processes it supports, the patient enrollment and FAIR-compliant clinical data acquisition and integration workflow for federated biomedical data driven projects is used to demonstrate the functionality. In this scenario, the multi-centric data collection begins with the enrollment of a patient in a research study. This process requires handling of PII and managing patients' consent across institutional boundaries. Clinical scientists enter the clinical data into the EDC system. Thus, the EDC serves as the entry point for research-specific parameters, clinical observations, and biosample information. Mainzelliste enables interconnecting and synchronizing the trust centers in all connected institutions. When PII are entered during enrollment, Mainzelliste generates pseudonyms for patients, which can be used throughout the whole process without revealing their original identities. This enables the linkage of EDC data with information collected from other clinical systems via automated network interfaces. Since real-world data is often fragmented or incomplete, the EDC system can also serve to augment additional data for research purposes. The enriched research data are stored in a research data lake and made available for integrated analysis environments. The DataCastle is designed to fully support this complex workflow. Utilizing Mainzelliste's REST API, DataCastle supports cross-institutional pseudonymization and consent management workflows, allowing multiple instances to operate as either standalone (centralized) trust centers or as federated nodes in a distributed architecture. Following initial structured data capture and enrollment using the DataCastle's EDC system and Mainzelliste, the associated files are stored within the filesystem layer. Via the Metadata Extractor, descriptive metadata for each file are extracted. The current prototype supports automated metadata extraction for text files, image formats, and DICOM series. Using unified metadata standards benefits project interoperability and should be considered for each project. Each dataset is versioned using Git-annex and Nextcloud WebDAV, allowing all changes to be tracked over time. The versioning mechanism also connects metadata and data content, generating a complete provenance record, which is essential for reproducibility and auditability in research. On the application layer, the managed datasets are connected with analysis and visualization environments. Researchers can explore the available data, initiate analysis and management workflows, and transfer datasets to third party tools, such as JupyterLab, Orthanc, and Rstudio, as provided by the Data Science Orchestrator. This application layer operationalizes the principles of FAIR data use by enabling researchers not only to find and access data but also to work with them directly in a reproducible environment. By establishing a tool that encompasses the entire data processing workflow in one turnkey solution, the DataCastle bridges FAIR data and FAIR use, implementing the philosophy of open-science, open-collaboration in a community centered way (16).

5. Discussion

Since the DataCastle's first implementation focus on specific use-cases, future work will focus on publishing the framework as an open-source platform and piloting it in more generic use cases. We will focus on evaluating the quality of the automatically generated metadata and performance evaluation for large datasets (e.g., imaging series, genomics, …), as seamless integration into clinical workflows requires both reliability and efficiency. To consolidate existing data management and data collection, collaboration with existing solutions will be sought out. Integrating the DataCastle and existing solutions like ABCD-J or OMERO could be beneficial. Further efforts will focus on aligning the DataCastle with established standards, including HL7, HL7 FHIR, and CDISC ODM, to promote interoperability among the research community.

6. Conclusion

DataCastle provides a pragmatic, modular architecture that links FAIR-oriented data acquisition with reproducible, version-linked data use across integrated tools. By combining enrollment-time pseudonymization, automated metadata extraction mapped to Health DCAT-AP, and background versioning of heterogeneous files, the prototype supports traceable dataset evolution and controlled reuse.

7. References

1. Wilkinson MD, Dumontier M, Aalbersberg IJJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3160018. doi:10.1038/sdata.2016.18 Cited in: PubMed; PMID 26978244.
2. Harris PA, Taylor R, Thielke R, et al. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform. 2009;42(2):377–81. doi:10.1016/j.jbi.2008.08.010 Cited in: PubMed; PMID 18929686.
3. Löbe M, Meineke F, Winter A. Scenarios for Using OpenClinica in Academic Clinical Trials. Stud Health Technol Inform. 2019;258211–5. Cited in: PubMed; PMID 30942748.
4. Murphy AJ, Kaifi J. Maintaining an Effective Lab Notebook and Data Integrity. In: Hunter CJ, Dudeja V, editors. Success in Academic Surgery: Basic Science.

Cham: Springer Nature Switzerland; 2025. p. 51–63. (Success in Academic Surgery; vol. 81).

5. Gieß A, Hutterer A. The Future of Data Management: A Delimitation of Data Platforms, Data Spaces, Data Meshes, and Data Fabrics. Inf Syst E-Bus Manage. 2025;32(1):3. doi:10.1007/s10257-025-00707-4
6. Allan C, Burel J-M, Moore J, et al. OMERO: flexible, model-driven data management for experimental biology. Nat Methods. 2012;9(3):245–53. doi:10.1038/nmeth.1896 Cited in: PubMed; PMID 22373911.
7. ABCD-J: Accessing Behavior for Clinical Data and Joint Usage [Internet]. 2025 [cited 2025 Nov 3]. Available from: https://www.abcd-j.de
8. Lablans M, Borg A, Ückert F. A RESTful interface to pseudonymization services in modern web applications. BMC Med Inform Decis Mak. 2015;152. doi:10.1186/s12911-014-0123-5 Cited in: PubMed; PMID 25656224.
9. Tremper G, Brenner T, Ben Amor M, et al. Mainzelliste – 10 Years of Pseudonymization, Record Linkage and Informed Consent Management (accepted); 2025. (vol. 1).
10. Kern J, Brechtel M, Kaluza P, et al. Data Science Orchestrator: A Containerized Trusted Research Environment for Flexible and Secure Analytical Pipelines. Stud Health Technol Inform. 2025;327698–702. doi:10.3233/SHTI250438 Cited in: PubMed; PMID 40380547.
11. Hussein R, Scherdel L, Nicolet F, et al. Towards the European Health Data Space (EHDS) ecosystem: A survey research on future health data scenarios. Int J Med Inform. 2023;170104949. doi:10.1016/j.ijmedinf.2022.104949 Cited in: PubMed; PMID 36521422.
12. Stammler S, Kussel T, Schoppmann P, et al. Mainzelliste SecureEpiLinker (MainSEL): Privacy-Preserving Record Linkage using Secure Multi-Party Computation. Bioinformatics. 2020. doi:10.1093/bioinformatics/btaa764 Cited in: PubMed; PMID 32871006.
13. Toevs B. Processing of Metadata on Multimedia Using ExifTool: A Programming Approach in Python. In: 2015 Annual Global Online Conference on Information and Computer Technology (GOCICT): IEEE; 04.11.2015 - 06.11.2015. p. 26–30.
14. Michal Szczepanik. Putting new git-annex features to use with Nextcloud [Internet]. 2025 [cited 2025 Nov 3]. Available from: https://blog.datalad.org/posts/annex-nextcloud
15. Jodogne S, Bernard C, Devillers M, et al. Orthanc - A lightweight, restful DICOM server for healthcare and medical research. In: 2013 IEEE 10th International Symposium on Biomedical Imaging: IEEE; 07.04.2013 - 11.04.2013. p. 190–3.
16. The Turing Way Community. The Turing Way: A handbook for reproducible, ethical and collaborative research: Zenodo; 2025.