1st Conference on Research Data Infrastructure

Harmonizing RDM

https://doi.org/10.52825/CoRDI.v1i.347

© Authors. This work is licensed under a Creative Commons Attribution 4.0 International License

Published: 07 Sept. 2023

Provenance Core Data Set

A Minimal Information Model for Data Provenance in Biomedical Research

Ulrich Sax[1,5] [https://orcid.org/0000-0002-8188-3495], Christian Henke[1] [https://orcid.org/0000-0002-4541-4018], Christian Draeger[2] [https://orcid.org/0000-0001-8835-4548], Theresa Bender[1] [https://orcid.org/0000-0001-6721-7034], Alessandra Kuntz[1,5] [https://orcid.org/0000-0002-8259-2577], Martin Golebiewski [3,5] [https://orcid.org/0000-0003-2039-8733], Hannes Ulrich [4] [https://orcid.org/0000-0002-8349-6798] and Matthias Löbe[2,5] [https://orcid.org/0000-0002-2344-0426]

[1] Department of Medical Informatics, University Medical Center Göttingen, Germany

[2] University Leipzig, Germany

[3] Heidelberg Institute of Theoretical Studies, Heidelberg, Germany

[4] Institute for Medical Informatics and Statistics, Kiel University, Germany

[5] part of the NFDI4Health Consortium

Abstract. The exchange, dissemination, and reuse of biological specimens and data have become essential for life sciences research. This requires standards that enable cross-organizational documentation, traceability, and tracking of data and its corresponding metadata. Thus, data provenance, or the lineage of data, is an important aspect of data management in any information system integrating data from different sources [1]. It provides crucial information about the origin, transformation, and accountability of data, which is essential for ensuring trustworthiness, transparency, and quality of healthcare data [2]. For biological material and derived data, a novel ISO standard was recently introduced that specifies a general concept for a provenance information model for biological material and data and requirements for provenance data interoperability and serialization [3,4]. However, a specific standard for health data provenance is currently missing. In recent years, there has been a growing need for developing a minimal core data set for representing provenance information in health information systems. This paper presents a Provenance Core Data Set (PCDS), a generalized data model that aims to provide a set of attributes for describing data provenance in health information systems and beyond.

Keywords: data provenance, lineage, Life Sciences, Harmonizing RDM, Linking RDM

1. Methods

The Provenance Core Data Set was developed based on inputs from various web conferences and discussions among experts in the field of health informatics organized by the NMDR2 project. Several data and metadata standards were examined on their ability to capture provenance metadata. The data model focuses on general attributes that are applicable to different scenarios and use cases, including data distribution, data transformation, and accountability.

2. Results

The Provenance Core Data Set provides a simple data model for representing data provenance in health information systems. It includes attributes that can capture important aspects

of provenance, such as the time of data creation, modification, and update, the source system information, the status of the data, accuracy assessment, and responsible parties. The data model is intended to support the trustworthiness, transparency, and accountability of data in health information systems, which are essential for ensuring data quality and integrity.

The data model includes attributes such as create date, change date, update date, source system type, source system name, source system URL, source system release, source system vendor name, source system vendor URL, status, accuracy, creator, interpretive comment, provider, frequency, depends on, measurement method, and measured by. These attributes are intended to provide comprehensive information about the provenance of data in health information systems.

3. Discussion

The Provenance Core Data Set can be applied to different scenarios and use cases in health information systems. It provides a common set of attributes that can be used to describe data provenance in a standardized and consistent manner. The data model can be used to capture important information about the origin, transformation, and accountability of data, which can be useful for various purposes such as data quality assessment, data integration, and data sharing. However, further research and validation are needed to evaluate the applicability and effectiveness of the Provenance Core Data Set in real-world health information systems.

4. Outlook

The Provenance Core Data Set is a promising approach for representing data provenance in health information systems and beyond. We need to discuss this approach with other communities especially in the NFDI context.

The data model has the potential to support the trustworthiness, transparency, and accountability of data in health information systems, which are crucial for ensuring data quality and integrity. Further research and validation are needed to evaluate the applicability and effectiveness of the Provenance Core Data Set in different health information systems and beyond. More standards integration has to be organized regarding huge initiatives like the German Medical Informatics Initiative [5] and their HL7 FHIR based core data set [6,7].

Data availability statement

-

Underlying and related material

-

Author contributions

All authors collaborate in the NFDI4health project. US prepared the manuscript, all authors reviewed and finalized.

Competing interests

The authors declare that they have no competing interests.

Funding

We greatly appreciate the funding from the Deutsche Forschungsgemeinschaft (DFG) through projects no. 442326535 (NFDI4health), 451265285 (NFDI4health TF COVID19), and 315072261 (NMDR2).

Acknowledgement

-

References

1. L. Moreau, and P. Missier, PROV-DM: The PROV Data Model, (2013). https://www.w3.org/TR/prov-dm/ (accessed April 24, 2023)
2. Parciak et al 2019: Provenance Solutions for Medical Research in Heterogeneous IT-Infrastructure: An Implementation Roadmap; Studies in Health Technology and Informatics Volume 264: MEDINFO 2019: Health and Wellbeing e-Networks for All; DOI 10.3233/SHTI190231[3] R. Wittner, P. Holub, C. Mascia, F. Frexia, H. Müller, M. Plass, C. Allocca, F. Betsou, T. Burdett, I. Cancio, A. Chapman, M. Chapman, M. Courtot, V. Curcin, J. Eder, M. Elliot, K. Exter, C. Goble, M. Golebiewski, B. Kisler, A. Kremer, S. Leo, S. Lin-Gibson, A. Marsano, M. Mattavelli, J. Moore, H. Nakae, I. Perseil, A. Salman, J. Sluka, S. Soiland-Reyes, C. Strambio-De-Castillia, M. Sussman, J.R. Swedlow, K. Zatloukal, J. Geiger, “Toward a common standard for data and specimen provenance in life sciences”, Learn Health Sys., e10365, April 2023, doi: https://doi.org/10.1002/lrh2.1036
4. “ISO/TS 23494-1:2023 Biotechnology — Provenance information model for biological material and data — Part 1: Design concepts and general requirements” https://www.iso.org/standard/80715.html (accessed 25 April 2023) [5] Semler et al 2018: German Medical Informatics Initiative; Methods Inf Med. 2018 Jul; 57(Suppl 1): e50–e56.PMID: 30016818
5. The Medical Informatics Initiative Core data set: https://www.medizininformatik-initiative.de/en/medical-informatics-initiatives-core-data-set (accessed April 23, 2023)7 FHIR Provenance: https://fhir-ru.github.io/provenance.html (accessed April 23, 2023)