Ruprecht-Karls-Universität Heidelberg
Medizinische Fakultät Mannheim
Dissertations-Kurzfassung

Collection and modeling of data provenance with an integrated metadata concept in the context of biomedical workflows in Data Integration Centers

Autor: Kerstin Gierend

Institut / Klinik: Medizinische Statistik, Biomathematik und Informationsverarbeitung

Doktorvater: Prof. Dr. T. Ganslandt

In the context of the Medical Informatics Initiative funded by the German government, medical data integration centers have implemented complex data flows to load routine health care data into research data repositories for secondary use. Data management practices to (sensitive) medical data elements are of key importance throughout these processes, but less scientific work has so far been undertaken to examine and enforce the data provenance aspects in this specific medical use case. Insufficient knowledge about these medical data and processes can lead to validity risks and weaken the quality of the extracted data.

This cumulative dissertation presents the combination of a two-stage methodological approach to facilitate extensive provenance information enrichment in the data integration pipelines. A MIRACUM wide mixed-method study investigated both, the data management maturity status and provenance readiness and presented recommendations. The subsequent proof-of-concept study took up this outcome to model and implement an algorithm gathering, storing and extracting continuously relevant provenance information on medical data element level and achieved satisfying pipeline execution times. Overall, the implemented provenance tracking solution indicates a high degree of traceability, accuracy, and reliability of the transformed medical data elements, with which a data integration center can meet any accountability obligations. In addition, this dissertation serves as a catalyst for the derivation of an overarching data management strategy, abiding data integrity and provenance characteristics as a key factor for quality and FAIR sustained health and research data.

This thesis enabled for the first-time extensive provenance information enrichment in the data integration pipelines in a German medical data integration center. The dissertation anticipates recommendations enforce quality of patient data dissemination and guide the implementation of auditable and measurable provenance approaches. This development has a potentially broad application since it contributes as initial work to the envisioned European Health Data Space.