Aus der Abteilung für Biomedizinische Informatik
am Zentrum für Präventivmedizin und Digitale Gesundheit
der Medizinischen Fakultät Mannheim
(Direktor: Kommissarische Leitung Dr. med. Fabian Siegel)

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

Inauguraldissertation
zur Erlangung des
Doctor scientiarum humanarum (Dr. sc. hum.)
der
Medizinischen Fakultät Mannheim
der Ruprecht-Karls-Universität
zu
Heidelberg

vorgelegt von
Kerstin Gierend, geb. Welter

aus
Ottweiler
2024

Dekan:
Prof. Dr. med. Sergij Goerdt

Referent:
Prof. Dr. med. Thomas Ganslandt

TABLE OF CONTENTS

Page

LIST OF ABBREVIATIONS

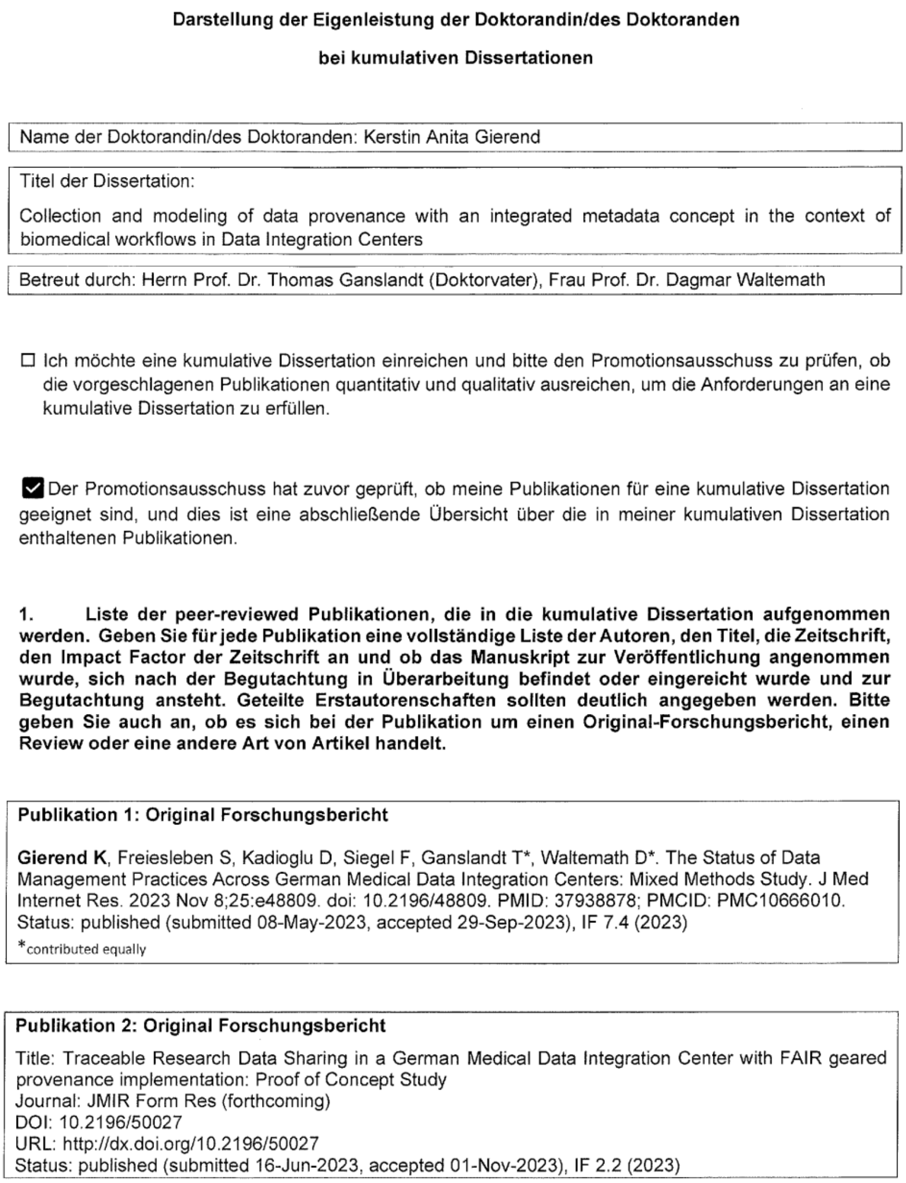
ALCOA	Attributable, Legible, Contemporaneous, Original, and Accurate
API	Application Programming Interface
CIS	Clinical Information System
CSV	Comma Separates Values
DIC	Data Integration Center
DIFUTURE	Data Integration for Future Medicine
ETL	Extract-Transform-Load
EHDS	European Health Data Space
FAIR	Findability, Accessibility, Interoperability, and Reusability
FDA	Food and Drug Administration
FDPG	Forschungsdatenportal für Gesundheit
HL7 FHIR	Health Level Seven Fast Health Interoperability Resources
HIGHMED	Heidelberg-Göttingen-Hannover Medizininformatik
JMIR	Journal of Medical Internet Research
MII	Medical Informatics Initiative
MIRACUM	Medical Informatics in Research and Care in University Medicine
MIRAPIE	Minimal Requirements for Automated Provenance Information Enrichment
PISA	Provenance Information System trAces
PROV	Provenance
PROV-DM	Provenance Data Model
PROV-O	Provenance Ontology
PubMed	Public Medicine
RDF	Resource Description Framework
REDCap	Research Electronic Data Capture
SFL	Software Framework Lifecycle
SPHN	Swiss Personalized Health Network
TRANSFoRm	Territories as response and accountable networks of S3 through new forms of open and responsible decision-making
UML	Unified Modeling Language
UMM	University Medicine Mannheim
W3C	World Wide Web Consortium
XML	Extensible Markup Language

LIST OF FIGURES

All other figures are available within the original publications.

1 PREFACE

The following publications are part of the cumulative dissertation. A detailed description of the personal contribution to each of the publications is provided according to the document “Presentation of the doctoral candidate’s personal contribution” (“Darstellung der Eigenleistung der Doktorandin/des Doktoranden”).



Darstellung der Eigenleistung der Doktorandin/des Doktoranden

bei kumulativen Dissertationen

Name der Doktorandin/des Doktoranden: Kerstin Anita Gierend

Titel der Dissertation:

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

Betreut durch: Herrn Prof. Dr. Thomas Ganslandt (Doktorvater), Frau Prof. Dr. Dagmar Waltemath

□ Ich möchte eine kumulative Dissertation einreichen und bitte den Promotionsausschuss zu prüfen, ob die vorgeschlagenen Publikationen quantitativ und qualitativ ausreichen, um die Anforderungen an eine kumulative Dissertation zu erfüllen.

☑ Der Promotionsausschuss hat zuvor geprüft, ob meine Publikationen für eine kumulative Dissertation geeignet sind, und dies ist eine abschließende Übersicht über die in meiner kumulativen Dissertation enthaltenen Publikationen.

1. Liste der peer-reviewed Publikationen, die in die kumulative Dissertation aufgenommen werden. Geben Sie für jede Publikation eine vollständige Liste der Autoren, den Titel, die Zeitschrift, den Impact Factor der Zeitschrift an und ob das Manuskript zur Veröffentlichung angenommen wurde, sich nach der Begutachtung in Überarbeitung befindet oder eingereicht wurde und zur Begutachtung ansteht. Geteilte Erstautorenschaften sollten deutlich angegeben werden. Bitte geben Sie auch an, ob es sich bei der Publikation um einen Original-Forschungsbericht, einen Review oder eine andere Art von Artikel handelt.

Publikation 1: Original Forschungsbericht

Gierend K, Freiesleben S, Kadioglu D, Siegel F, Ganslandt T*, Waltemath D*. The Status of Data Management Practices Across German Medical Data Integration Centers: Mixed Methods Study. J Med Internet Res. 2023 Nov 8;25:e48809. doi: 10.2196/48809. PMID: 37938878; PMCID: PMC10666010.
Status: published (submitted 08-May-2023, accepted 29-Sep-2023), IF 7.4 (2023)
*contributed equally

Publikation 2: Original Forschungsbericht

Title: Traceable Research Data Sharing in a German Medical Data Integration Center with FAIR geared provenance implementation: Proof of Concept Study
Journal: JMIR Form Res (forthcoming)
DOI: 10.2196/50027
URL: http://dx.doi.org/10.2196/50027
Status: published (submitted 16-Jun-2023, accepted 01-Nov-2023), IF 2.2 (2023)

Publikation 3: Review Protocol

Autoren: **Gierend K**, Krüger F, Waltemath D, Fünfgeld M, Ganslandt T, Zeleke AA.

Titel: Approaches and Criteria for Provenance in Biomedical Data Sets and Workflows: Protocol for a Scoping Review.

Zeitschrift: JMIR Res Protoc. 2021 Nov 22;10(11):e31750. doi: 10.2196/31750. PMID: 34813494; PMCID: PMC8663663.

Status: accepted, IF 1.7 (2023)

Publikation 4: Review

Autoren: **Kerstin Gierend**, Frank Krüger, Sascha Genehr, Francisca Hartmann, Fabian Siegel, Dagmar Waltemath, Thomas Ganslandt, Atinkut Alamirrew Zeleke

Titel: Capturing provenance information for biomedical data and workows: A scoping review.

Status: submitted 27-Jul- 2023, in peer-review

Publikation 5: Konferenzbeitrag

Autoren: **Kerstin Gierend***, Judith A.H. Wodke*, Sascha Genehr, Robert Gött, Ron Henkel, Frank Krüger, Markus Mandalka, Lea Michaelis, Alexander Scheuerlein, Max Schröder, Atinkut Zeleke, Dagmar Waltemath.

Titel: TAPP: Defining standard provenance information for clinical research data and workflows - Obstacles and opportunities.

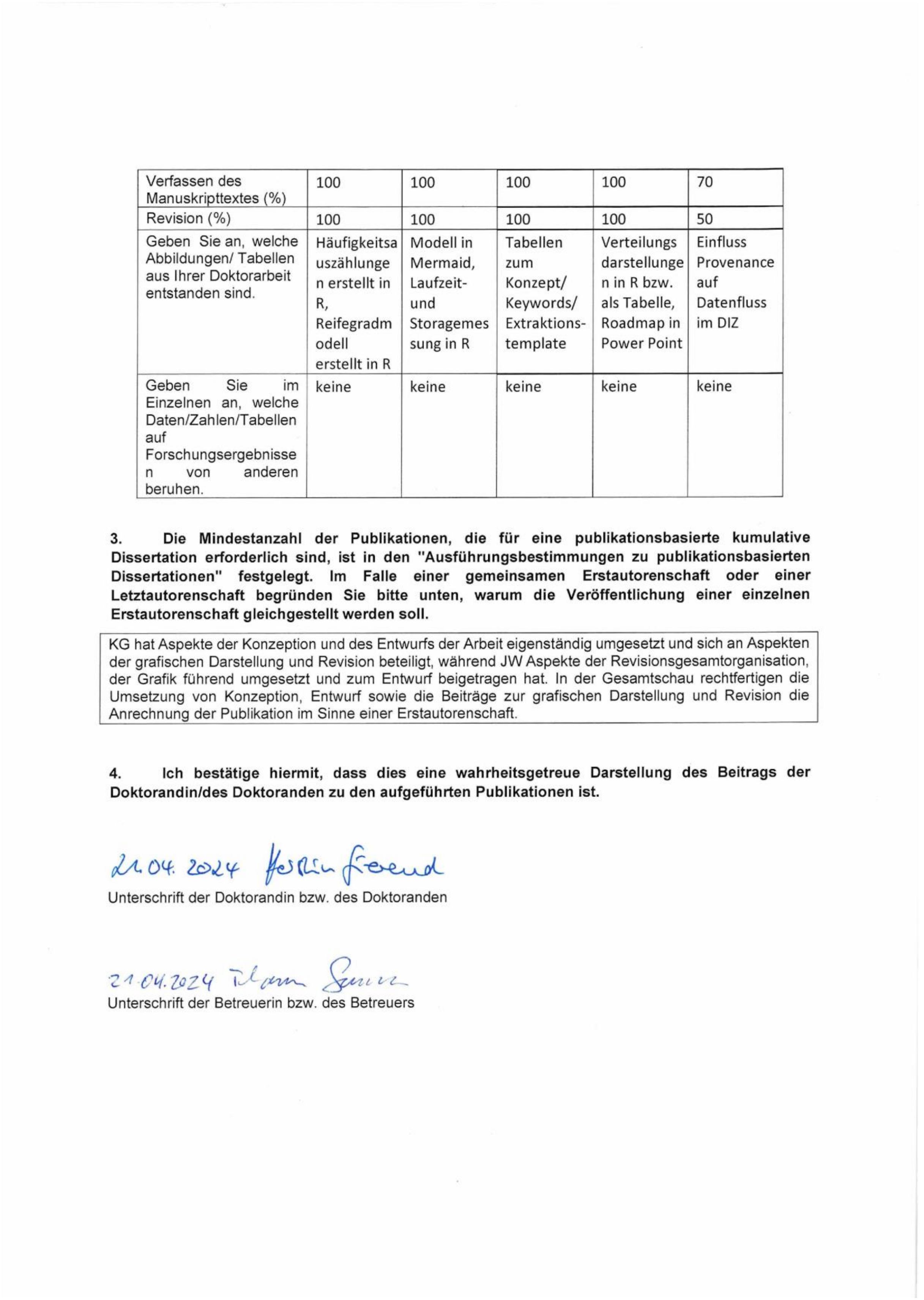
Konferenzbeitrag in Texas/Austin:

In Companion Proceedings of the ACM Web Conference 2023 (WWW '23 Companion). Association for Computing Machinery, New York, NY, USA, 1551–1554. https://doi.org/10.1145/3543873.3587562

* geteilte Erstautorenschaft

2. Zusammenfassung des Beitrags der Doktorandin/des Doktoranden zu der in jedem Manuskript berichteten Arbeit

Arbeitsschritte	**Publikation 1**	**Publikation 2**	**Publikation 3**	**Publikation 4**	**Publikation 5**
Konzeption (%)	100	100	80	100	70
Literaturrecherche (%)	100	100	100	100	50
Ethikantrag (%)	100 (entfällt nach Anfrage)	entfällt	entfällt	entfällt	entfällt
Tierversuchsantrag (%)	entfällt	entfällt	entfällt	entfällt	entfällt
Datenerhebung (%)	100	90	100	90	entfällt
Datenauswertung (%)	90	100	100	90	entfällt
Ergebnisinterpretation (%)	100	90	100	100	entfällt



Verfassen des Manuskripttextes (%)	100	100	100	100	70
Revision (%)	100	100	100	100	50
Geben Sie an, welche Abbildungen/ Tabellen aus Ihrer Doktorarbeit entstanden sind.	Häufigkeitsauszählungen erstellt in R, Reifegradmodell erstellt in R	Modell in Mermaid, Laufzeit- und Storagemessung in R	Tabellen zum Konzept/ Keywords/ Extraktions-template	Verteilungsdarstellungen in R bzw. als Tabelle, Roadmap in Power Point	Einfluss Provenance auf Datenfluss im DIZ
Geben Sie im Einzelnen an, welche Daten/Zahlen/Tabellen auf Forschungsergebnissen von anderen beruhen.	keine	keine	keine	keine	keine

3. Die Mindestanzahl der Publikationen, die für eine publikationsbasierte kumulative Dissertation erforderlich sind, ist in den "Ausführungsbestimmungen zu publikationsbasierten Dissertationen" festgelegt. Im Falle einer gemeinsamen Erstautorenschaft oder einer Letztautorenschaft begründen Sie bitte unten, warum die Veröffentlichung einer einzelnen Erstautorenschaft gleichgestellt werden soll.

KG hat Aspekte der Konzeption und des Entwurfs der Arbeit eigenständig umgesetzt und sich an Aspekten der grafischen Darstellung und Revision beteiligt, während JW Aspekte der Revisionsgesamtorganisation, der Grafik führend umgesetzt und zum Entwurf beigetragen hat. In der Gesamtschau rechtfertigen die Umsetzung von Konzeption, Entwurf sowie die Beiträge zur grafischen Darstellung und Revision die Anrechnung der Publikation im Sinne einer Erstautorenschaft.

4. Ich bestätige hiermit, dass dies eine wahrheitsgetreue Darstellung des Beitrags der Doktorandin/des Doktoranden zu den aufgeführten Publikationen ist.

21.04.2024 Kerstin Gierend

Unterschrift der Doktorandin bzw. des Doktoranden

21.04.2024

Unterschrift der Betreuerin bzw. des Betreuers

2 INTRODUCTION

2.1 Background of research work

The reuse of clinical routine data offers enormous potential for the clinical research. Clinical information systems (CIS), primarily developed for patient care in hospitals, store this huge data treasure. The use of this data in cross-hospital and transnational research projects demands extensive coordination of the technical and nontechnical tasks associated to data management. National initiatives have already been established in different countries e.g. the French Health Data Hub[1], the Swiss Personalized Health Network (SPHN)[2] or the Health-RI (Health-Research Infrastructure)[3]. In Germany, in 2018, the 'Federal Ministry of Education and Research' launched a national funding program, the German Medical Informatics Initiative (MII), to implement the reuse of clinical routine data for research at large scale[4]. Along with the collateral digitalization aspiration in the healthcare sector, the four MII funded consortia (MIRACUM, DIFUTURE, SMITH, HIGHMED) aim to strengthen medical research and improve medical treatment for patients. The associated university hospitals, like the University Medicine Mannheim (UMM) in the MIRACUM consortium, play an essential role in the establishment and networking of medical data integration centers (DIC)[5]. DICs are at the heart of the MII since they make medical data from care and research accessible. They create and implement the technical and organizational prerequisites for a cross-hospital data exchange between patient care and clinical/biomedical research. The data to be exchanged are jointly defined by the consortia members and implemented as the MI-I core data set[6]. The core data set consists of interdisciplinary basic modules like person, fall, consent, procedure, laboratory, medication, and extension modules which include data from specific medical fields or applications like oncology, biospecimen data, pathology findings. Data flows, starting from the routine healthcare systems into the DIC, are established in strict compliance with the generic data protection concept and the respective patient consents, in order to fill internal research data repositories of the hospital which are tailored to specialized or generic storage platforms. In addition, the DICs contribute to a cross-location repository, the German Research Data Portal for Health ("Forschungsdatenportal für Gesundheit" (FDPG)) which integrates data protection compliant data of about 25 university hospitals[7]. This data pool currently commands data of over 9.5 million patients with more than 40 million data items of diagnoses and more than 300 million laboratory data points. The FDPG serves as a central point of contact for scientists who want to carry out a research project with clinical routine data or biosamples. It handles the submitted feasibility requests, contractual regulations for the use and coordinates the data provision.

Any secondary data use presumes trustworthy and high-quality data management in general as a precondition for guaranteeing a precise and sustainable preparation of all related digital information[8,9,10]. However, managing digital data pipelines is a complex task, in particular in the context of clinical data use and under the associated general and hospital-specific rules for the protection of legal and ethical constraints for data access[11]. It is critical to have access to an unbroken and transparent chain of data transformation to minimize the risk to the legacy of the medical datasets[12]. It is equally challenging and extensive to ensure data correctness when data, processes and software code for data processing are constantly changing under dynamical infrastructural environmental conditions[13].

Figure 1 illustrates a generalized overview of the data managing tasks in the data life cycle of clinical research data. Essentially, the three layers shown characterize the entire research data life cycle, from data planning, data acquisition, data processing and data analysis to the publication and archiving of research data and results. The layers are composed of input and output data and data elements, of technical and nontechnical organizational artefacts like researchers, policies, guidelines, the infrastructure, external organizational units as for example a trusted third party, and of activities responsible for managing and processing the data. Thereby, it is important to ensure a controlled interplay, both within and between the individual layers in order to generate traceable, and at best, reproducible data and results which can be reused in future research approaches.

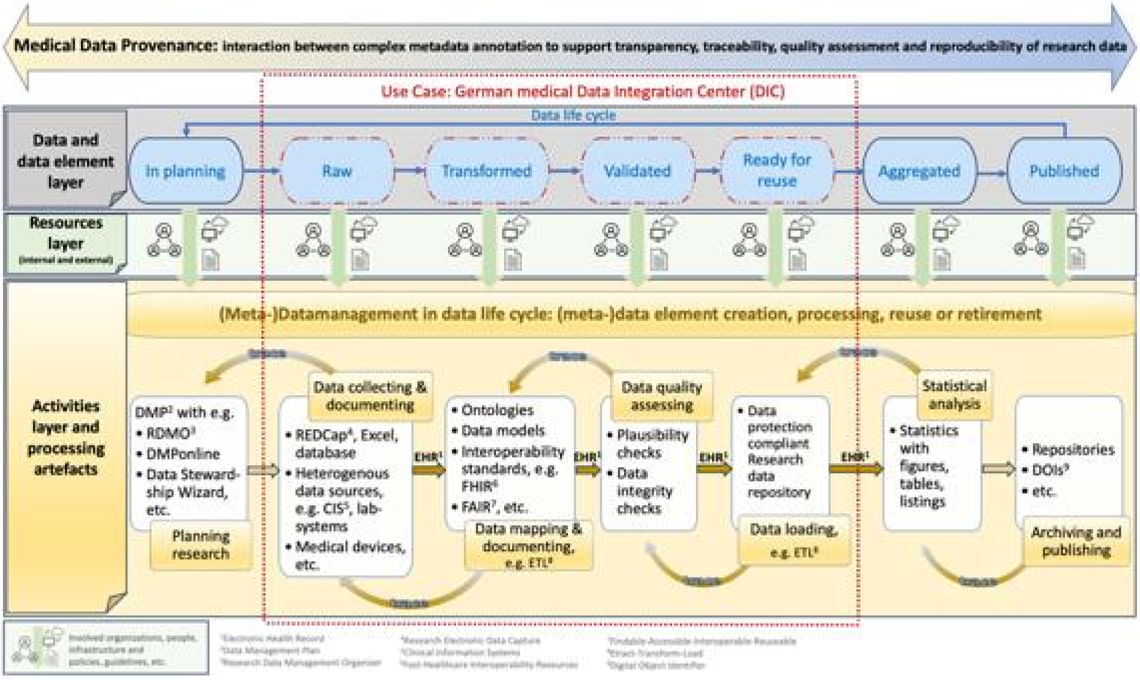
This illustration also depicts the core processing and artefacts in the DIC (red dotted line). The core data flows within a DIC and between associated CISs are subjected to complex data processing pipelines which require interdisciplinary medical and informatics stakeholder's knowledge. Extensive data protection compliant data extraction, transformation, and load (ETL) processes tackle the different data elements, which are also attributed to the MI-I core data sets, on their journey from arbitrary sources like the laboratory information system up to the target research data repository (see also Figure 1[14]). These complex ETL-routines are developed to transfer and store electronic health records (EHRs) from different clinical source systems into centralized data warehouse types while implementing approaches for healthcare data standardizing.

The implementation of sustainable interoperable data structures demands joint agreements at different four interoperability levels, (a) semantic level, (b) syntactic level, (c) structural level and (d) organizational level. In practice, this means scrutinizing and employing of related ontologies and healthcare specific interface standards, like the Health Level Seven Fast Health Interoperability Resources (HL7 FHIR)[15,16]. FHIR provides a generic definition of common health care concepts (e.g., patient, observation, practitioner, device, condition) and offers application programming interfaces (APIs) to access and reuse these resources while having a common understanding of the medical data.

Going beyond these measures, additional accompanying information about the data and processes, often referred to as metadata, contains valuable information. Key contextual metadata possess knowledge about the data and unlocks hidden data treasure. The crucial role of metadata and good (meta)data management is emphasized by the Guiding principles for FAIR data stewardship (Findability-Accessibility-Interoperability-Reusability)[17,18]. The FAIR principles are used to evaluate how well data comply with current standards in open and reproducible science[19]. Metadata contain substantial characteristics to express information for any kind of artifacts during data processing and managing[20].

Notably, the FAIR principles explicitly mention provenance as one component of metadata. The corresponding reuse principle (R1.2) is based on the provision of detailed provenance information. This information is required to identify data sources, input and output data sets and elements, and linked data transformation steps[21].

Figure 1: Different layers in Medical Data Provenance: dataflow and data life cycle with accompanying (meta-) data and information management in the DIC (red dotted line)


Medical Data Provenance: interaction between complex metadata annotation to support transparency, traceability, quality assessment and reproducibility of research data
Use Case: German medical Data Integration Center (DIC)
Data life cycle
Data and data element layer
In planning
Raw
Transformed
Validated
Ready for reuse
Aggregated
Published
Resources layer (internal and external)
Activities layer and processing artefacts
(Meta-)Datamanagement in data life cycle: (meta-)data element creation, processing, reuse or retirement
DMP² with e.g.
• RDMO³
• DMPonline
• Data Stewardship Wizard, etc.
Planning research
Data collecting & documenting
• REDCap⁴, Excel, database
• Heterogenous data sources, e.g. CIS⁵, lab-systems
• Medical devices, etc.
EHR¹
• Ontologies
• Data models
• Interoperability standards, e.g. FHIR⁶
• FAIR⁷, etc.
Data mapping & documenting, e.g. ETL⁸
EHR¹
Data quality assessing
• Plausibility checks
• Data integrity checks
EHR¹
• Data protection compliant Research data repository
Data loading, e.g. ETL⁸
EHR¹
Statistical analysis
• Statistics with figures, tables, listings
• Repositories
• DOIs⁹
• etc.
Archiving and publishing
Involved organizations, people, infrastructure and policies, guidelines, etc.
¹Electronic Health Record
²Data Management Plan
³Research Data Management Organiser
⁴Research Electronic Data Capture
⁵Clinical Information Systems
⁶Fast Healthcare Interoperability Resources
⁷Findable Accessible Interoperable Reusable
⁸Extract-Transform-Load
⁹Digital Object Identifier

A basic understanding of the term provenance is given with a description of what happened to the data[22]. The World Wide Web Consortium provenance (W3C-PROV) working group defines provenance as “information about entities, activities and people involved in producing a piece of data or thing, which can be used to form assessments about its quality, reliability or trustworthiness standard”[23]. The PROV standard defines a domain-agnostic conceptual data model (PROV-DM) with clear extensibility points, with core and extended structures to embed these alongside the data it belongs to. Targeted data provenance information tracking and its lawful compliant managing during the whole data life cycle are critically important properties for understandability and reproducibility of scientific results[24,25]. Advantages and opportunities of data provenance have been demonstrated, for instance, in the EU-Horizon 2020 TRANSFoRm project[22]. Researchers not considering the origin of data run into the hazard of systematically incomplete or wrong data. However, a provenance-oriented approach requires thorough planning, execution, and evaluation of data management processes in the respective application domain[26].

“Black box” processing and reporting of findings based on clinical routine data should no longer be acceptable since it may lead to loss of data and contextual knowledge about the data[17,27]. However, complex data transformation processes within DICs, as described above, demand rich provenance associated to these data elements in all data integration pipelines to gauge the quality of individual data elements.

Insufficient information about data, formation processes and metadata results in traceability issues which poses validity risks and can impede the quality assessment of extracted clinical data and related processes. In this vein, data documentation is highly relevant to enable traceability and essential to ensure data integrity[14,28]. This is one reason why the DICs face an increasing pressure to implement thorough data management and quality concepts. To address this shortcoming, the current status of provenance in clinical routine data seeks clarification.

The main objective of this dissertation is to build up a concept for relevant medical provenance capture at German medical Data Integration Centers. This dissertation aims to propose a solution that enhances the reusability of clinical routine data in medical DIC by meaningful provenance traces for their reliable and trustful secondary use of data in clinical research and patient care.

Secondly, this dissertation aims to serve as a preparatory data provenance contribution to the envisioned European Health Data Space (EHDS)[29].

In addition, the presented framework aims to foster the implementation of improved data managing concepts, leading to clear transparency, traceability and thus better provenance tracing through its lifecycle. Concurrently, this advances the accountability of a data integration center by reducing risks of the reuse of weak data in clinical research. It offers access to quality-assured and traceable data elements thereby boosting reliable and credible FAIR sharing of clinical research data.

2.2 Material and Methods in research work

The concept of two successive published studies forms the key to this dissertation.

The first study used a mixed-method study approach to examine the (meta-) data management practices for medical data elements throughout the data life cycle within German medical Data Integration Centers established in the Medical Informatics in Research and Care in University Medicine (MIRACUM) consortium (see publication 1)[14].

The second study followed-up on previous findings from publication 1, notably the data management maturity framework dedicated to empowering data provenance. Recommendations were picked up to develop a provenance gathering strategy. This study implemented and presented provenance traces as a proof-of-concept.
Metadata is specifically intended to strengthen the expressiveness of provenance traces which mirror both, the traceability chain, and the quality status of processed clinical data elements (see publication 2[30]).

Current literature was monitored continuously from the beginning and thoroughly examined based on a previously published scoping review protocol (see publication 3[31]) and the reported outcome (see publication 4[32]).

2.2.1 Mixed-method-study

Insights from a MIRACUM workshop on FAIR data management and discussions with data experts led to a mixed-method study which combines qualitative and quantitative research work (see publication 1[14]). First, this study aimed to obtain information about the current traceability and verifiability of processed patient data and metadata from heterogeneous clinical data sources in the DICs. In a second step, the development of a data management maturity framework should support the implementation of improved data management practices. It was hypothesized that a better provenance tracking would be feasible with a higher degree of transparency and traceability.
The study was performed as a semistructured interview. The interviews based on a survey using questionnaires covering clinical data processing and provenance practice within the DIC. Discussions with data experts from a MIRACUM FAIR data management workshop led to the development of these questionnaires (see Multimedia Appendix 2[14]).
A total of 22 experts and stakeholders from 10 DICs participated in the interviews, which were conducted remotely and individually with each DIC. All qualitative and quantitative data were concurrently entered into a REDCap database by the interviewing person while screensharing. Thematic analysis was conducted on the collected qualitative data without identifying the DIC. Coding was performed to identify relevant concepts or patterns within the data on the 4-eyes-principle. Qualitative results were integrated with the corresponding quantitative results. The categorical variables were characterized using counts and percentages, and represented in corresponding tables and figures, if applicable. The figures were created with R (version 4.2.0; The R Foundation)[33].
Results from this study comprising a DIC maturity model for provenance readiness were published (see Figure 6[14]) and reported compliant to the Good Reporting of a Mixed Methods Study (GRAMMS) checklist[34] (see Multimedia Appendix 1[14]).

2.2.2 Proof-of-concept

The proof-of-concept (see Figure 4[30]) combined a multi-stage approach to investigate the feasibility of automated generation of data provenance. This approach has been applied to different data integration pipelines at the University Medicine Mannheim (UMM-DIC).

The proof-of-concept entails a thorough requirements analysis (an interdisciplinary team of internal stakeholders in the UMM-DIC (lead, medical experts, computer scientists, technical staff, process owner of the ETL process)) to acquire the different aspects in the system border and system context of the planned provenance tracking system (Provenance Information System trAces (PISA)). In addition, the pressing requirement details from the preceding scoping review and the mixed-method study were considered for a precise definition of the functionalities. The resulting requirements served subsequently to develop the logical data model. The logical data model was designed using a unified-modeling-language (UML) class diagram which displays the structure of data elements and the relationships between them and describes how the data needs to be implemented.

The characterization of the associated data elements was achieved by determining the necessary metadata, preliminary value sets, W3C ontology mapping, and annotation. Documentary efforts were subjected to the good documentation practices like the ALCOA(+) principles[35]. An exemplary instantiation of the provenance information model on the described approach is given (see Figure 5[30]).

Finally, the provenance class was developed using Python, pewee, and a relational database. Test data element definitions were generated to develop and test the provenance class. Test data elements with comprehensive annotation were chosen to reflect the composition of a typical data integration repository. Seven data element types were defined, 100.000 data elements for each data element type were generated to produce a total of 700.000 provenance records using a Python (Python Foundation) script.

Extraction of provenance traces to any format like a csv, W3C RDF/XML file or HL7 FHIR resource “provenance” has been enabled. Finally, execution times for generating provenance traces were measured and evaluated.

2.2.3 Literature Review

A scoping review was conducted to present the current state of research on the provenance in the biomedical context. This scoping review followed the methodological framework by Arksey and O’Malley[36] and investigated evidence regarding approaches and criteria for provenance tracking in the biomedical domain. The corresponding research questions in this review were also targeted to the potential value of provenance information, the guidelines, demands and challenges during accomplishment of provenance and the completeness evaluation of provenance.

Based on the search strategy, the databases of PubMed and Web of Science were queried for articles published between 2006 and 2022. Title abstract screening with Rayyan[37], full text reading and screening, information extraction in pre-tested templates have been performed independently. The protocol of this scoping review has previously been published in JMIR Research Protocols[31], the report is currently undergoing a peer-review process and exists as pre-print[32].

This scoping review provides an extensive summary of current approaches and criteria for provenance tracking in the biomedical research domain. It discloses technical, implementation, and knowledge gaps with a focus on modeling and metadata frameworks for (sensitive) scientific biomedical data and provides a roadmap for a tailor-made provenance software-framework-lifecycle (Provenance-SFL) with many additional results (see Figure 6[32]).

3 PUBLICATIONS

3.1 Publication 1: The Status of Data Management Practices across German Medical Data Integration: Mixed-Method Study

This section delves into the status of data management practices across German Medical Data integration centers, as originally published in the Journal of Medical Internet Research (JMIR), an international, peer-reviewed, and open access journal.
The original publication and appendices are available at JMIR
(https://doi.org/10.2196/48809).

Original Paper

The Status of Data Management Practices Across German Medical Data Integration Centers: Mixed Methods Study

Kerstin Gierend[1], Dipl Inf (FH); Sherry Freiesleben[2], MSc; Dennis Kadioglu[3,4], MSc; Fabian Siegel[1], MD; Thomas Ganslandt[5*], MD; Dagmar Waltemath[2*], Dr -Ing

[1] Department of Biomedical Informatics at the Center for Preventive Medicine and Digital Health, Medical Faculty Mannheim, Heidelberg University, Mannheim, Germany

[2] Core Unit Data Integration Center and Medical Informatics Laboratory, University Medicine Greifswald, Greifswald, Germany

[3] Institute for Medical Informatics (IMI), Goethe University Frankfurt, University Hospital, Frankfurt am Main, Germany

[4] Department for Information and Communication Technology (DICT), Data Integration Center (DIC), Goethe University Frankfurt, University Hospital, Frankfurt am Main, Germany

[5] Chair of Medical Informatics, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany

[*] these authors contributed equally

Corresponding Author:
Kerstin Gierend, Dipl Inf (FH)
Department of Biomedical Informatics at the Center for Preventive Medicine and Digital Health
Medical Faculty Mannheim
Heidelberg University
Theodor-Kutzer-Ufer 1-3
Mannheim, 68167
Germany
Phone: 49 621383 ext 8087
Email: kerstin.gierend@medma.uni-heidelberg.de

Abstract

Background: In the context of the Medical Informatics Initiative, medical data integration centers (DICs) have implemented complex data flows to transfer routine health care data into research data repositories for secondary use. Data management practices are of importance throughout these processes, and special attention should be given to provenance aspects. Insufficient knowledge can lead to validity risks and reduce the confidence and quality of the processed data. The need to implement maintainable data management practices is undisputed, but there is a great lack of clarity on the status.

Objective: Our study examines the current data management practices throughout the data life cycle within the Medical Informatics in Research and Care in University Medicine (MIRACUM) consortium. We present a framework for the maturity status of data management practices and present recommendations to enable a trustful dissemination and reuse of routine health care data.

Methods: In this mixed methods study, we conducted semistructured interviews with stakeholders from 10 DICs between July and September 2021. We used a self-designed questionnaire that we tailored to the MIRACUM DICs, to collect qualitative and quantitative data. Our study method is compliant with the Good Reporting of a Mixed Methods Study (GRAMMS) checklist.

Results: Our study provides insights into the data management practices at the MIRACUM DICs. We identify several traceability issues that can be partially explained with a lack of contextual information within nonharmonized workflow steps, unclear responsibilities, missing or incomplete data elements, and incomplete information about the computational environment information. Based on the identified shortcomings, we suggest a data management maturity framework to reach more clarity and to help define enhanced data management strategies.

Conclusions: The data management maturity framework supports the production and dissemination of accurate and provenance-enriched data for secondary use. Our work serves as a catalyst for the derivation of an overarching data management strategy, abiding data integrity and provenance characteristics as key factors. We envision that this work will lead to the generation of fairer and maintained health research data of high quality.

(J Med Internet Res 2023;25:e48809) doi: 10.2196/48809

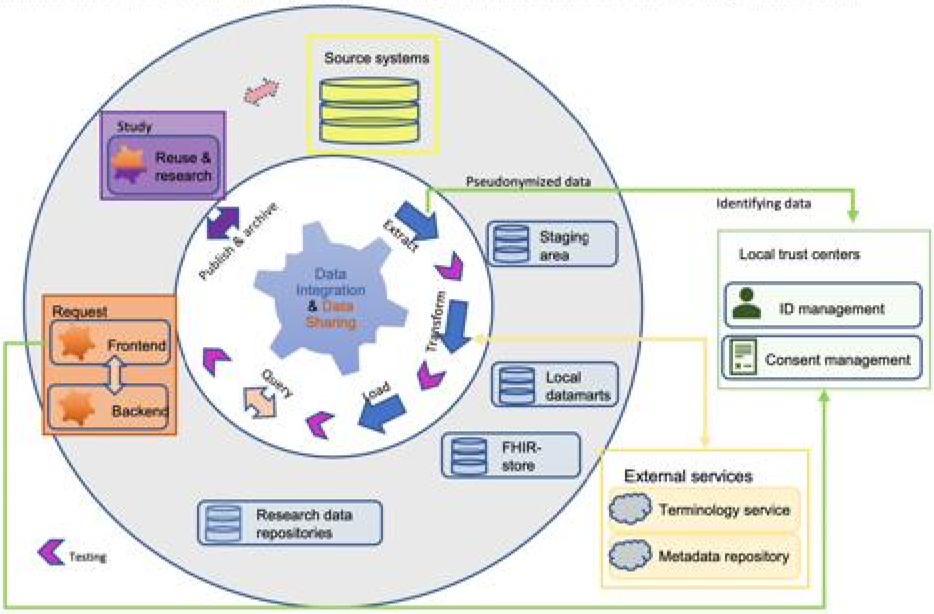
KEYWORDS

data management; provenance; traceability; metadata; data integration center; maturity model

Introduction

Data integration centers (DICs) within the German Medical Informatics Initiative (MII) have evolved rapidly in the past years [1-4]. DICs process and provide digital medical data for the secondary use in research. The foundation of data sharing (DS) and interoperability within the MII is an agreed-upon common core data set (CDS). The basic modules are generic and include data items encoding laboratory results, diagnosis, procedures, or medication data. The extension modules contain domain-specific data such as oncology or microbiology data [5]. The CDS data items are processed using a standardized extract-transform-load (ETL) development process that follows the data life cycle (Figure 1). Specific testing measures throughout the data processing chain are implemented to ensure accuracy and high quality. The architecture of every Medical Informatics in Research and Care in University Medicine (MIRACUM) DIC (see also Figure 1) is built upon the medical informatics reusable ecosystem of open source linkable and interoperable software tools [6]. Data requests by researchers are limited to and based on generic institutional policies and a defined legal framework. The concrete status of the DICs with respect to enabling the findable, accessible, interoperable, reusable (FAIR) principles still needs to be determined [7]. However, several initiatives have already outlined the importance of applying the FAIR principles for both input and output data [8-10].

Figure 1. Data life cycle and data management processes. An overview of core processes and artifacts from data management practice in a Medical Informatics in Research and Care in University Medicine data integration center. FHIR: fast health care interoperable resources.



The data life cycle describes the journey of biomedical data from data collection to final analysis and publication (Figure 1). Particularly when working with (sensitive) patient data, the understanding of the data's origin and the relationship between an element and its predecessors, also called traceability (see Textbox 1), is highly relevant for legal requirements and a fundamental prerequisite of data quality. "Black box" processing and reporting of findings based on routine data should no longer be acceptable [11] since it may lead to loss of data and contextual knowledge about the data [12]. This is a reason why the DICs faces an increasing pressure to implement thorough data management concepts, in particular provenance. An option is the adoption of generic provenance concepts from the World Wide Web Consortium (W3C) [13]. However, the application of these concepts requires insights and understanding of the data management tasks in the given context.

Insufficient information about data formation processes and metadata (see Textbox 1) pose validity risks and can impede the quality assessment of extracted clinical data and related processes. Data with unknown provenance and lack of traceability endure from a confidence deficiency and therefore minimize the acceptance for secondary use.

Textbox 1. Related terminologies.

- Provenance (World Wide Web Consortium [W3C] working definition): "Provenance of a resource is a record that describes entities and processes involved in producing and delivering or otherwise influencing that resource. Provenance provides a critical foundation for assessing authenticity, enabling trust, and allowing reproducibility. Provenance assertions are a form of contextual metadata and can themselves become important records with their own provenance" [13].
- W3C provenance: is a family of specifications for provenance with a generic concept to express specific meta-information (or metadata) about data and its related artifacts. Provenance records contain the agents (eg, people and institutions), entities (eg, data sources and data elements), and activities (eg, extract, load, and transform), involved in producing, influencing, or delivering a piece of data or a thing. The granularity of the W3C provenance concepts influences the level of traceable data management activities [13]. Provenance can be distinguished as data and workflow provenance [14].
- Meta-information (or metadata): machine understandable information for the web [15]. Metadata contain substantial characteristics to express (provenance) information for any kind of artifacts during data managing and play a crucial role in the implementation of the findable, accessible, interoperable, reusable (FAIR) principles [7].
- Traceability: ability to retain the identity of the product and its origin [11]. Traceability is essential to ensure data integrity and trust in the data [16]. In our study traceability is the ability to trace (identify and measure) all the steps that led to a particular point in a data transformation process. Traceability assumes enrichment of data with proper meta-information.

In this work, we seek clarification about the data management processes in German DICs. We aim to facilitate a comprehensive understanding and transparency of these processes to boost data reliability and integrity. We therefore ran a mixed method study across all MIRACUM DICs to get a picture of current traceability and verifiability of patient data and metadata processed from heterogeneous clinical data sources. We expect that DICs would benefit from an increased focus on governance of data management practices rather than random or only partly managed data processing. To support the change, we offer a maturity framework which can be implemented in DICs for self-evaluation. We hypothesize that the framework will foster the implementation of improved data management processes, transparency, traceability, and better provenance tracing.

Methods

Study Design

This study uses a mixed methods design [17] and associated best practices [18]. A mixed methods design leads to more plausible and comprehensible quantitative outcomes if combined with qualitative statements. The design involved the collection of qualitative and quantitative data in a single interview and subsequent analysis to strengthen the study's conclusions. The collection of quantitative and qualitative data was performed concurrently on the same survey and with the same priority. The study has been reported according to the Good Reporting of a Mixed Methods Study (GRAMMS) checklist [19] (Multimedia Appendix 1). Based on the survey results and discussions among the authors, a maturity framework was developed, following the capability maturity model (CMM) [20].

Study Settings and Participants

The study was performed as a semistructured interview. The interview questions cover clinical data processing and provenance practices within the DICs. The results from a MIRACUM workshop on FAIR data management and discussions with data experts from different DICs contributed to the design of the questionnaires. In addition, we build the questions upon insights from a survey on the research field of provenance [14].

For this work, we distinguished data management operations that concern the data integration (DI) phase (blue items in Figure 1) from operations concerning the DS phase (orange items in Figure 1). Thus, the interview questions were split into 2 separate questionnaires, containing 16 questions (DI) and 38 questions (DS), respectively. The DI questions covered data management activities during the extraction, transformation, and loading of electronic health records. DS questions comprised available documentation of resources, activities for DS processes, and organizational information. The interview does not cover the management of patients' consent since it is a precondition for data processing and release from the DICs [3].

The questions were numbered and grouped by subject. A mixture of open and closed questions was chosen to get a more comprehensive insight into the respective fields. The questionnaires were created in German language and pilot-tested internally with data experts.

Stakeholders from each MIRACUM site participated in the interview. We provided the questionnaires in advance with the option to delegate the task to accountable staff members. This kept both the interviewer and the participant in line, avoided distractions, and encouraged an open communication. Participants consent was obtained in written form ahead of the actual interview.

Sample

A total of 10 DICs (all MIRACUM sites) were invited to participate. We subsequently collected data from all sites with 22 participants, thereof 4 women and 18 men contributors. Due to the COVID-19 pandemic, all interviews were conducted virtually. The interviewing person shared the screen with the questionnaire displayed on it while the interview was conducted. Qualitative and quantitative data were collected in German language based on the participant's answers during the interview phase. All data were concurrently entered into a database (Research Electronic Data Capture [REDCap; Vanderbilt University]) by the interviewing person during the interview [21]. The data collection took between 1.5 hours and 4 hours per DIC. Overall, the data collection period lasted over 3 months. Due to the interview technique no missing data occurred.

Data Collection

The data collection method relied on asking questions within the predetermined thematic framework. Even for closed questions, there was always the option to ask additional questions and to store the answers.

The data collection included quantitative and qualitative data with equal emphasis (Multimedia Appendix 2). The qualitative data were collected in free-text fields and during the interviews with stakeholder professionals (Multimedia Appendix 2). The data collection took between 1.5 hours and 4 hours per DIC. Before starting the data analysis, all collected data were translated into English and covalidated.

Data Analysis

After performing the semistructured interview, we conducted a thematic analysis. We converted or transformed qualitative data into quantitative scores or constructs by "coding" the qualitative responses into different groups. We identified common topics or patterns and ensured that these patterns appropriately represent the participants' responses using the 4-eyes-principle.

The analyses were conducted anonymously without identifying the respective DIC. The tables and figures outline the individual characteristics and frequency counts were calculated. The categorical variables are described using counts and percentages, if applicable. The data were described using median and range for the continuous variables, if applicable. The figures were created with R (version 4.2.0; The R Foundation) [22]. Qualitative, free-text data were read, analyzed, and coded, if necessary. The narratives representing the coded themes were produced from the data material. The data analysis was reviewed by all authors.

Integration

Qualitative data were combined with quantitative data whenever possible. Thus, the qualitative results were integrated with the corresponding quantitative results and then presented numerically. The outcome was reported as descriptive statistical results. Whenever integration was not possible, we reported qualitative results instead. After analysis of the qualitative and quantitative data, the preliminary findings were presented and discussed among the authors.

Ethical Considerations

The ethics approval was waived by the University of Heidelberg or Mannheim University Medicine Ethics Committee II. Informed consent was obtained from all subjects (the stakeholders) to participate in the interview about the status of their data processing pipelines. All study data are deidentified. The participants did not receive any compensation.

Results

Overview

In our study, we seek clarification about the data management processes in German DICs. We aim to facilitate a comprehensive understanding and transparency of the prevailing practices for data extractions, data transformations, data storage, and data provision to boost data reliability and integrity. We first present the main survey outcomes, and then we introduce a maturity framework.

Results Overview

All 10 DICs of the MIRACUM consortium participated in the survey between July and October 2021. All 22 participants, either the head of a DIC or a member of the technical staff, responded to a total of 66 questions, thereof 16 questions about the DI phase and 12 questions about the locally used data elements and catalogs from the MII CDS. A total of 9 DICs answered the 38 DS specific question (Table 1); data from the Core Unit Data Integration Center at the University Medicine Greifswald is missing.

Table 1. The number of data integration center participants (Medical Informatics in Research and Care in University Medicine) in the 3 survey sections.

	Questions of data integration (n=16), n	Questions of status Medical Informatics Initiative data elements and catalogs (n=12), n	Questions of data sharing (n=38), n
University Medicine Dresden	1	1	1
University Medicine Erlangen	3	3	3
Goethe University Frankfurt	2	2	2
University Hospital Freiburg	1	1	1
University Hospital Giessen	2	2	2
University Medicine Greifswald	4	4	–
University Medicine Magdeburg	2	2	2
University Medicine Mannheim	3	3	3
University Medical Center Mainz	1	1	1
Philipps-University Marburg	3	3	3

General and Organizational Matters

Expectation Regarding Provenance

The interview revealed considerable expectations regarding the collection and use of provenance and metadata information, also beyond the W3C provenance definition (Table 2). Interestingly, the most common expectations were associated with the assessment of data quality (n=7), with traceability and information capability (n=7), and with the transparency in processing steps, workflows, or data sets (n=2). Other frequently named expectations were linked to technical reasons (n=4) such as debugging or performance evaluation. Less frequent terms included compliance with regulations (n=2), reproducibility, support of scientific usage process, or increased confidence in data. Expectations like clear regulation of responsible parties, interoperability, and increased acceptance were mentioned once (n=1). In this, 1 DIC stated no usage of provenance information at all.

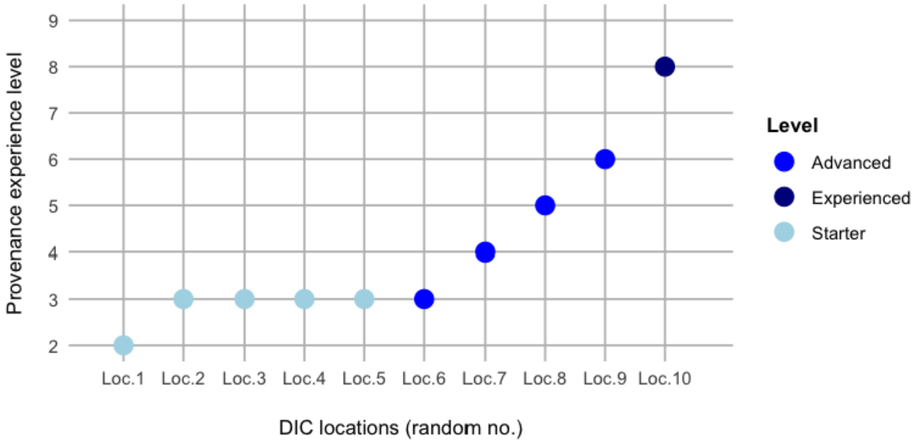
Table 2. Expectation regarding provenance, a summary of all reported expectations by 10 data integration centers.

	Frequency of expectations regarding provenance, n
Traceability and information capability	7
Data quality assessment	7
Technical reasons	4
Transparency of processing steps	2
Support of scientific process	2
Reproducibility of data flow	2
Proof of compliance	2
Increased confidence	2
Interoperability	1
Internal evaluation about changes in data elements	1
Increased acceptance	1
Clear regulation responsibilities	1
Concurrently no use	1

Self-Assessment of Provenance Experience

When analyzing the data in Figure 2, we observed a low provenance experience. More than half of the DICs ranked their provenance experience as a starter level with a score 0-3 (n=6). The 3 sites reported an advanced level with a score 4-7. Just 1 site rated their experience with a score of 8 (corresponding to expert level).

Figure 2. Self-assessment of provenance experience level. All reported self-assessments by the 10 participating data integration centers. DIC: data integration center.



Organizational Structure

Consistent with the W3C provenance model [13] and Herschel et al [14], the organizational component of a DIC represents a core unit at many German medical faculties. When asked for the organizational prerequisites, all DICs reported that specifications of the manufacturer systems and standard operating procedures (SOPs) were available. However, the degree of maturity varies across DICs.

At the time of the interview, all DICs (n=10, 100%) were in a continuous development process with drafted SOPs at different levels. However, some DICs (n=3) already reported gaps in their SOPs, preventing the full coverage of process flows for DI and DS. Nearly half of the sites (n=4, 40%) used already approved SOPs. Roles and responsibilities, as central parts of the SOP, had been defined in most DICs (n=8, 80%). Only a few DICs (n=3, 30%) had a dedicated role concept (Figure S1 in Multimedia Appendix 3).

Availability of Metadata and Related Tool Usage

No consequent and targeted practice for provenance capture could be determined. We hypothesize that it might be difficult to develop a standardized, structured, and machine-readable metadata schema across all German university hospitals. Similar results regarding insufficient availability of (semantic) metadata for provenance were observed [23]. Detailed results for the individual questions are given in Figure S2 in Multimedia Appendix 3.

Metadata Exploitation During Data Management

Overview

The development of metadata schemata is an important factor for high traceability [16]. Hence, we were interested in learning how organizational and document resources might help to generate metadata and to embed metadata within the digital object itself. This analysis section targeted the annotation status such as labeling of data elements, data sets, or tagging of files. Detailed results are available in Figures S3-S5 in Multimedia Appendix 3.

Documentation Matters

All interviewees declared that data management activities were not subjected to specific data management planning or tools. Any planning or preparational documentation was collectively performed using tools such as JIRA (Atlassian) [24] or Confluence (Atlassian) [25]. During the DS phase, most DICs (n=8) follow internal SOPs for the documentation of methods, or data management plans, respectively. All other DICs reported that internal, project-specific tools were applied. Processes were partially under construction.

Documentation Artifacts From Data Elements and Coding in Data Integration Phase

Appreciably all sites (n=10) reported about their level of documentation for accessing the source systems, for the maintenance of the data elements, for code development and execution, as well as the content of log files as part of their ETL-process as described below.

Annotation of Data Elements

As expected, and in line with the literature [26], preliminary attempts for data annotation exist. However, these attempts do not yet cover the whole processing pipeline in all DICs. The applied annotation approaches vary, too. It is noteworthy that the best, and partially automatic annotation was yielded on the joint segment Fast Healthcare Interoperability Resources (FHIR) to the research data repository (RDR; n=10), since this pipeline is part of the MIRACUM standard ETL process [3]. Detailed results are available in Figure S3 in Multimedia Appendix 3.

Log Files for Improved Traceability

Log files are text-based files, which include timestamps, store events, processes, and transactions. Thus, log files provide valuable provenance information. However, a direct access to log files was not possible during this study due to the risk of disclosing critical or sensitive information.

Most DICs (n=9) already established log files to trace environment and execution information, particularly during DI. In most cases, the log files contain configurable parameters and elements, mostly generated within the respective infrastructure framework. Some frameworks comprised self-defined information and messages for error, warning, and execution statistics. Depending on the actual process, short- or long-term retention could be observed. Long-term retention was applied for data transfer logs and short-term for application logs, for example, throughout the ETL life cycle. Half of the DICs (n=5) reported that the access to source systems is automatically logged with user information and time stamps, but without relationship to the particular data elements. In general, access to the source-application itself is not possible. More than two-thirds of the centers (n=4) have manual logging features in place. Only 1 DIC does not perform any logging (n=1).

Only sparse information was provided about the computational environment and execution workflows during runtime of scripts in productive operation (Figure S4 in Multimedia Appendix 3). A small number of sites (n=2) reported that automated and collaboratively accessible information were created. The 3 centers (n=3) said that no such information was generated. All other survey participants (up to n=7) asserted that the logging protocols were either compiled manually or generated automatically. Based on the survey data no systematic approach was deducible.

However, half of the DICs reported that scripts which are executed during the data requesting phase often do not produce log files. If log files were produced by the scripts, they contained information about execution and error history (n=4). Many DICs emphasized their capability of access-logging to data pools, computational environments, and execution history. The recorded information includes details about Docker containers such as the software status of the environment. However, some data seems to be missing in the logs, including the date of execution or the user account. Logs from the RDR Informatics for Integrating Biology and the Bedside indicated access logins (who and when) and querying of data elements. Extraction protocols were produced for some source systems.

Versioning Information Status

Version information, an important element for reproducible research, creates a history for each file. Based on the annotation of the source code and artifacts, for example, the used programming language, provenance data can create relationships between individual elements or documents. Figure S5 in Multimedia Appendix 3 illustrates that the generated code was in general subjected to textual documentation, and code was mostly versioned in a DI pipeline (n=10). GitHub is mainly used as version control tool. Importantly, the DI segment FHIR to the RDR provided full versioning capability in all DICs (n=10). The reason is that MIRACUM developed and delivered a centralized component for this workflow step [3]. Also, code for the ETL segment for data processing from staging servers to a FHIR repository is highly version controlled. Lower coverage was observed on the initial stretch source system to staging (n=7). In this, 1 interviewee explained that this circumstance was due the code being in the responsibility of the manufacturer. Another expert said that code was managed manually (n=1). In 1 case, no version control was implemented at all. The situation is similar when data is queried by scripts for research purposes since code versioning was tool-guided by the most DICs (n=7). Overall, the results suggest that version information is available, but needs to be prepared in more detail to be useful for provenance processing.

Documentation Artifacts of Testing Procedures and Script Validation

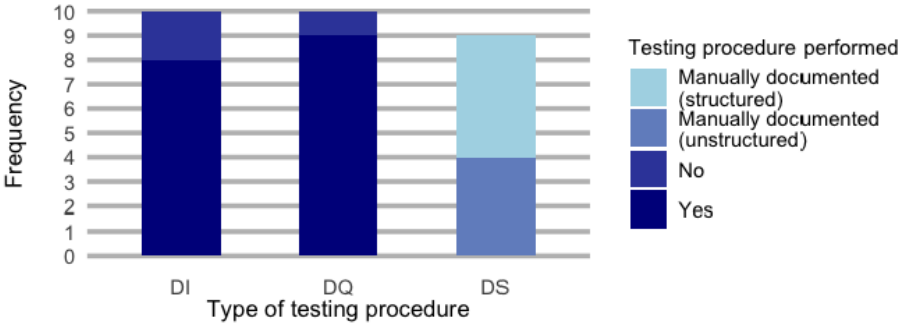
A considerably high number of DICs confirmed the implementation of test procedures (n=8) and data quality measures (n=9; Figure 3).

Notably, different test documentation strategies were reported by the stakeholders (Figure S6 in Multimedia Appendix 3). Most sites (n=4) mentioned the provision of an automated testing documentation which was collaboratively accessible for the authorized staff during the data integrity measurements pipeline.

Data quality is mainly assessed using the Data Quality Assessment reporting tool, which has been developed within MIRACUM (n=6) [27]. A total of 2 DICs used self-developed assessment and documentation of data quality.

All DICs validate their data querying scripts. Evidence for validation is provided by manually documenting the queries in a structured and permanently accessible way on GitLab (GitLab Inc) [28] or JIRA [24] (n=5). Unstructured evidence such as the four-eyes principle was practiced in 4 DICs.

Figure 3. Testing or validation procedures. A summary of all reported types of testing procedures during the scripting phase. Data integration centers reporting about their testing procedures. DI: data integration phase; DS: data sharing phase; DQ: data querying phase.



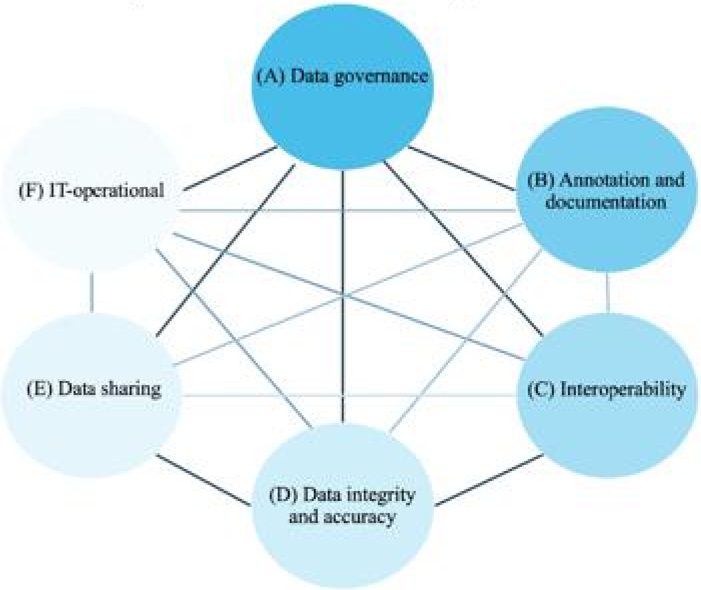
Documentation Artifacts From Final Review and Facts About Research Result Objects

All previous processes and individual outcomes contribute to the history of the so-called result object. We anticipate that the result object should contain all provenance-related metadata. As shown in Figure S7 in Multimedia Appendix 3, most participants (n=5) examine all the documentation and artifacts for traceability. Applied examination methods included the 4-eyes-principle, random sample checks, or ETL checklists with defined examination criteria. Approximately one-third of the respondents (n=3) indicated that the traceability of documentation and related artifacts was not checked. Only 1 DIC has plans to check traceability systematically and automated. Remarkably, examination of the result object regarding adherence to FAIR principles and provenance assessment was not performed in any DIC. These findings indicate a lack of awareness for FAIR data management, as has also been observed in a recent survey within the German Network University Medicine [29].

Derivation of a Maturity Framework

On the basis of our study results, we derived a data integration center toward maturity framework (DIC2MF), which incorporates the specific needs and metadata items of German DICs (Figure 4). The DIC2MF indicates a DIC's readiness status for provenance tracking ("provenance power") and can be used as a benchmarking tool.

Figure 4. Dimensions and their relationships within the framework for provenance tracking.



Dimensions and Categories of the Framework

The DIC2MF concept is based on the CMM. Unlike the already published maturity model for provenance management, which was established in the hydro- and geoscientists' field [30], our approach comprises 6 dimensions and related categories (Figure 4) which together constitute provenance characteristics that a DIC requires to be effective in delivering traceable and reliable patient data for secondary use. The dimensions and categories were influenced by the grouping of key interview findings from (1) related organizational, legal, and technical conditions, (2) the metadata exploitation based on data annotation and documentation degree and associated operations, and (3) including the measures to ensure quality during the different operations. Textbox 2 elucidates the proposed framework and the associated characteristics.

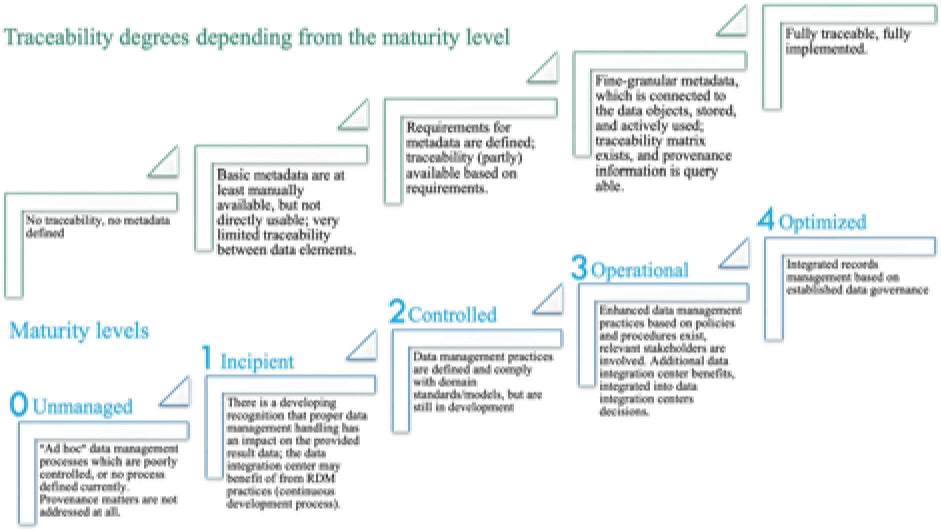
Each dimension is represented by a specific ability level. Figure 5 depicts the different gradations of the 5 ability levels "unmanaged," "incipient," "controlled," "operational," and "optimized." Each level describes a degree of traceability fulfillment and is an indicator for the provenance power in the DIC. The completeness and quality of traceability goes hand in hand with the levels of maturity. An instantiation of the framework is shown in Figure 6.

Textbox 2. Components of the framework for provenance tracking (data integration center toward maturity framework).

- Data management dimensions and categories
  - (A) Implement “Data governance” which explores the availability of important legislation, guidelines, or rules that directly relate to the scope of a data integration center
    - Roles and responsibilities (staff, roles, and training)
    - Standard operating procedures (quality management)
    - Regulations (eg, general data protection regulation and patient consent)
    - Risk management (controlling risks)
  - Build multiple data management dimensions (B up to E) for data processing and data analysis
    - (B) Addresses the practices to “Annotation and Documentation” of data and the related processes
      - Considers metadata about the management of the (automated) documentation and annotation steps of the individual data and process elements, including the provenance of any processed element
        - Access
        - Input sources
        - Output sets
        - Data elements
        - Scripts
        - Execution
        - Versioning
      - Considers information from log files created during data conversion, for example, to cover the facets of provenance according to the World Wide Web Consortium provenance recommendation (see Textbox 1)
    - (C) Enforces the transformation and processing of data into interoperable formats to enable translational research with patient data
      - Includes metadata about the usage of standard data models and catalogs
        - Common data model
        - Domain specific catalogs
    - (D) Examines the implementation of quality standards to ensure “Data Integrity and Accuracy” of the processed patient data
      - Comprises metadata about all methods for examining and maintaining the data quality
        - Testing procedures
        - Validation approach
    - (E) Data sharing
      - Includes metadata about the service of organization and reporting of the data request and analysis result as well as taking care of long-term archiving aspects
        - Organization and reporting
        - Long-term archiving
    - (F) IT-operational
      - Comprises metadata about
        - Data security of patient data
        - Data accessibility of patient data
        - Infrastructure and computation environment
      - Tools and software

- Relationship
  - It should be mentioned that relationships exist between the dimensions, for example, data processing must adhere to given data governance rules

Figure 5. The 5 maturity levels in the framework (data integration center toward maturity framework) and defined degrees of traceability.



No traceability, no metadata defined

Basic metadata are at least manually available, but not directly usable; very limited traceability between data elements.

Requirements for metadata are defined; traceability (partly) available based on requirements.

Fine-granular metadata, which is connected to the data objects, stored, and actively used; traceability matrix exists, and provenance information is query able.

Fully traceable, fully implemented.

"Ad hoc" data management processes which are poorly controlled, or no process defined currently. Provenance matters are not addressed at all.

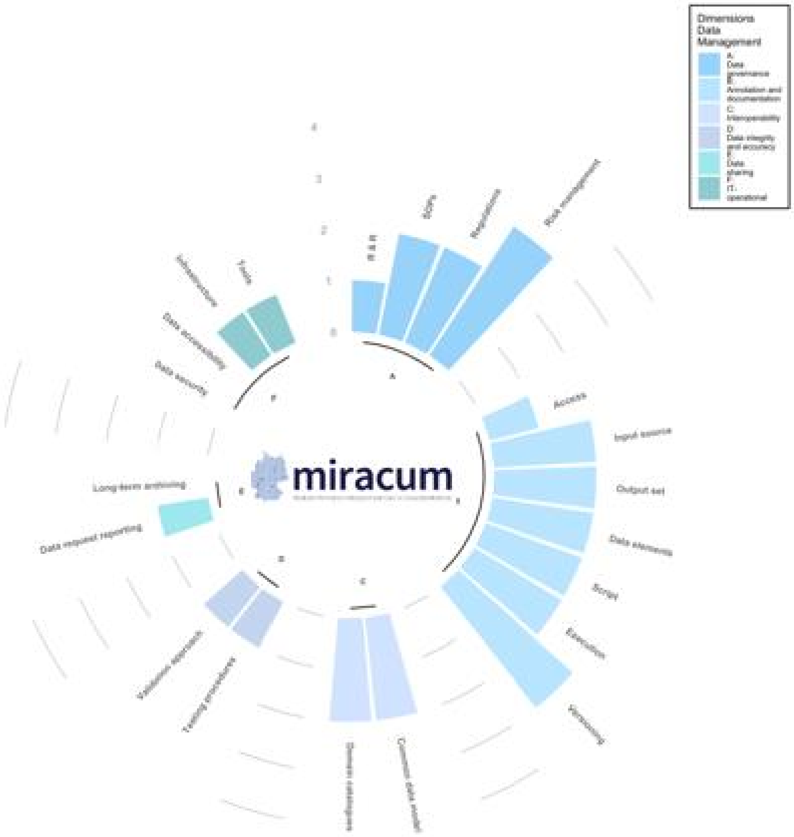
There is a developing recognition that proper data management handling has an impact on the provided result data; the data integration center may benefit of from RDM practices (continuous development process).

Data management practices are defined and comply with domain standards/models, but are still in development

Enhanced data management practices based on policies and procedures exist, relevant stakeholders are involved. Additional data integration center benefits, integrated into data integration centers decisions.

Integrated records management based on established data governance

Figure 6. DIC2MF—provenance power as part of the data management maturity framework. The DIC2MF indicates the DIC's readiness status for provenance ("provenance power"). Logo used with permission from the Medical Informatics in Research and Care in University Medicine (MIRACUM) Consortium [2]. DIC2MF: data integration center toward maturity framework; SOP: standard operating procedure.



Instantiation of the Framework

The inner circle (Figure 6) represents the grouped 6 data management dimensions for provenance tracking (following the specification in Figure 4). Each dimension contains multiple categories and each category reflects the substantial characteristics for the expression of provenance. The quality of provenance expression can be derived from the ability scale (between 0 and 4) which defines how reliably and maintainably the implemented practices within a DIC can produce the required outcomes. The higher the bar the more provenance information is available. Thus, the height of the bar is an indicator for the need to improve data management practices given on the description of the ability level. For example, progress from 1 maturity level to the next one may be reached by adding fine-granular metadata in compliance with the W3C provenance components agent, activity, and entity in a second step. The presented concepts are a first step toward identifying the requirements for traceability within a DIC.

Discussion

Principal Findings

We successfully performed a mixed method study and gained deep insight into the status of data management processes in the German medical DICs. Our work facilitates understanding and traceability and will potentially boost the reliability and integrity of data for secondary use. We derived a maturity framework and applied it as a benchmark to measure the degree of traceability and deriving from this the provenance power of individual data elements in MIRACUM German DICs. The proposed maturity framework for provenance readiness helps DICs to identify their conceptual bottlenecks in provenance tracking and increases trustful dissemination of clinical data.

We hypothesize that our work could serve as a catalyst for an overarching data management strategy for DICs. The beneficial approach presented here could be implemented widely as a common assessment tool, within the MII structure and in the medical research field itself.

Evaluation

Framework Applicability

The framework can be used for critical systematic self-evaluation. It can guide the identification of relevant components for provenance tracking and thus facilitate traceability of patient's data processing. The information obtained from the framework dimensions A to F help to develop the necessary metadata, and consequently enhance traceability on process and element level.

Establishing Traceability and Best Practices

Establishing traceability is one of the biggest challenges associated with any data conversion. A combination of several aspects may lead to the condition that traceability has not been implemented effectively at the DICs. Predominantly, a lack of awareness and provenance expertise could be a key finding from the self-assessment of provenance experience (Figure 2) and indicates a subordinate role of provenance to date. A lack of technological framework may furthermore hinder the uptake of provenance in the data processing pipelines. Here, the traceability issue can be linked to a lack of granularity including details about workflow steps and about the processed data elements themselves. ETL pipelines are mostly implemented individually by the DICs. Practices in the highly ranked centers for provenance expertise revealed that these include annotation and metadata documentation, even if it is not always machine-readable and automatically recorded.

A tentative explanation is that there is no systematic approach for gathering provenance data of individual data items (Table 2). The procedure of tracking data set or data items is neither formalized nor sufficiently standardized. Consequently, no targeted provenance collection and metadata concept has been established as of now. In addition, sparsely developed traceability decreases the reliability and thus the quality of single data elements for secondary use (Figure 3). Even if general testing procedures are available in the DI pipelines, there is a lack in quality traceability.

The following examples showcase how DICs may increase their maturity level by using the proposed framework dimensions and categories while connecting metadata to the associated artifact: (1) dimension A foresees (a) guidance on data managing activities, like define operations by SOPs, introduce data management plans, and consider legal restrictions and (b) regular data management training for the responsible staff. Connect both topics at least on data and process level. (2) The challenge of dimension B could be passed step by step (a) while gaining and deriving targeted annotation from log files for building and filling the maturity framework on a data element level, log files are configurable and enable the traceable storage of events so that these can be analyzed and optimized. In this way, log files thus help to track data and their processes, and to reconstruct transactions. Elements of log files could be selected as in the proposed framework, for example, source and target system, information about type of event or logged action, version or actor; (b) by having appropriate, clear, and complete documentation for all measured data in place, if possible, in machine-actionable way and connect this information to the data; (c) by making metadata accessible and adding richer prospective and retrospective provenance metadata. These actions will allow for fine-grained versioning workflows linking to outputs produced during the distinct executions of ETL pipelines. The metadata approach should consider information derived from the W3C components agents (such as developer and data owner), activities (such as different programming scripts), entities (such as data sources or data elements). (3) Convert the extracted data into common and interoperable health care standards as defined in dimension C and connect the associated metadata information to your processed data as described in dimension B. (4) Testing and validation (dimension D) approaches add quality information to the processed data itself. Collect available metadata on applied activities to ensure data quality as given in dimension B. (5) Dimension E, dedicated to the DS phase enriches a data element with information from the data requesting, reporting, and archiving phase. (6) Dimension F intends to collect meta-information about the operational environment in which the data were collected and processed.

Related Work

Provenance tracking and granularity issues were addressed in different papers [31,32]. Gierend et al [33] performed a scoping review on provenance in biomedical research and offered comprehensive results concerning the practical application of provenance and the associated challenges, including aspects like completeness and validation and provenance granularity issues. Curcin et al [34] reported that both data and processes need provenance, gathered in consistent, interoperable manner to make research results verifiable and reproducible. These works directed our study approach to examine the traceability aspect. Johns et al [35] tried to figure out knowledge on provenance methods in a more general way. Regarding the term provenance, Herschel et al [14] pointed to the definition of provenance, which leaves room for many different interpretations of and approaches to provenance and investigated the question why capturing provenance is useful. This led us to clearly define the goal of our study and give clear expectations regarding provenance accomplishment. Furthermore, this might give clear expectations regarding provenance accomplishment and provide the framework for the scope and the extent of implementation measures. In the same way, the outcome of our study can be used by the recently launched community-driven project which aims to define a "MInimal Requirements for Automated Provenance Information Enrichment" guideline [36]. The projects' goal is to build a general data model and semantics for provenance in the biomedical community.

Training issues were addressed as a challenge of poor data management practice [26]. Better health informatics training and permanent data manager and software architect positions are demanded in health research groups. This indicates that our maturity framework needs an iterative and interdisciplinary approach to implement traceability in data processing pipelines.

Lessons Learnt

During the conduction of the semistructured interview and the implementation of our framework, we learnt that the extent of the complex processing steps requires interdisciplinary team work to come to a proper level of provenance granularity. We are convinced that the community will benefit from a consequent exchange with stakeholders from different areas of expertise, like medical experts, data owners, and computer scientist. In addition, we encountered a major increase of transparency and traceability since we started with a consequent application of the maturity framework approach in our DIC. Moreover, having data governance in place, would facilitate the FAIR oriented data management planning and as such boost the data asset to be more reliable and trustful for or in the research field. Another recommendation is to spend more time on training in this field.

Ongoing Processes

Changing conditions in clinical routine, in granularity of requirements (decision-making, identifier management, and legal matters) demand continuous adaptation of the framework. We foresee extensions for provenance representation and storage, provenance retrieval, and usability along discussion for risk and benefit.

There are recent advancements to transform the dimension and categories into the W3C provenance concepts. We introduce a first provenance implementation in our DIC in Mannheim (University Medicine Mannheim DIC) in a proof-of-concept study in peer review phase.

Limitations

Our investigation is limited to the MIRACUM DICs, to their current service profiles and development stages as well as to the experience of the involved staff. Since provenance data are sporadically available, we were not able to consider maintainability aspects of provenance. Derivation of qualitative and quantitative results to the framework levels was performed by means of an evaluative description of metadata availability and the ability of traceable data. Integration of pseudonymization and consent management are external processes and not in primary scope for this study.

Conclusions

Implementing traceable data life cycles and transparent data management processes are sophisticated and challenging tasks, not only for the MIRACUM DICs. Notwithstanding, sufficient traceability would enable data to be a trusted asset in the medical DIC. Our paper provides insights on how institutions (attempt to) implement data management principles to provide clinical routine data for secondary use. However, to implement traceability, explainability of the relationships and the order between the data and process elements are required. We discussed the extensive transformations, curations, and linked artifacts of collected data elements and workflows during the entire data life cycle. The obtained insights led us to identify possible improvements and actions. One such action is the introduction of a maturity framework which visualizes the specific traceability challenges on a technical and organizational level observed at each DIC. In future, we seek to derive a generic provenance model and common data provenance strategy based on the traceability findings. To this end, we will investigate how complete provenance, as part of a FAIR data management strategy, can be delivered and what the limitations are in this regard. We envision that this work will lead to FAIR and maintained health and research data of high quality.

Acknowledgments

We would like to thank Achim Michel-Backofen, Albert Vass, Christian Haverkamp, Christian Seidemann, Daniel Schmitz, Dennis Kadioglu, Detlef Kraska, Jakub Lidke, Jan Maluche, Jan Sparenberg, Jonathan Mang, Kim Hee Eun, Noemi Deppenwiese, Robert Gött, Sebastian Stöcker, Sherry Freiesleben, Stephan Wojke, Tim Herrmann, and Thomas Litschko, who are the participants in the interview. We thank the reviewers for their constructive comments and valuable suggestions. This work was part of the precondition for obtaining the degree Dr sc. hum. from the University Heidelberg. This research is funded by the German Federal Ministry of Education and Research within the German Medical Informatics Initiative (FKZ 01ZZ1801A, 01ZZ1801E and 01ZZ1801M) and within the Network University Medicine (FKZ 01KX2121). For the publication fee we acknowledge financial support by Deutsche Forschungsgemeinschaft within the funding programme „Open Access Publikationskosten“ as well as by Heidelberg University.

Authors' Contributions

KG developed the questionnaires and conducted the semistructured interview, prepared data for analysis, performed the graphical analysis for the maturity framework, drafted all sections of the paper, coordinated reviewing, incorporated the comments from the coauthors, and submitted the paper. SF did the graphical analysis of the survey data and worked on the paper. DK contributed to editorial revision, finalization of the paper, and content support, especially in general on the topic of metadata. FS contributed to the discussion of the maturity framework and to the interpretation of data. DW contributed to the discussion of the concept, advisory support during preparation of the interview, editorial revision, and finalization of the paper. TG contributed to the discussion of the concept and finalization of the paper. All authors reviewed and approved the final version of the paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Good reporting of a mixed methods study checklist [19].
[PDF File (Adobe PDF File), 63 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Data collection items (in English and German language).
[PDF File (Adobe PDF File), 1292 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Additional results from the survey.
[PDF File (Adobe PDF File), 460 KB-Multimedia Appendix 3]

References

1. Semler SC, Wissing F, Heyder R. German medical informatics initiative. Methods Inf Med 2018;57(S 01):e50-e56 [FREE Full text] [doi: 10.3414/ME18-03-0003] [Medline: 30016818]
2. Data Integration Centers. MIRACUM. URL: https://www.miracum.org/en/das-konsortium/datenintegrationszentren/ [accessed 2022-09-22]
3. Prokosch HU, Acker T, Bernarding J, Binder H, Boeker M, Boerries M, et al. MIRACUM: medical informatics in research and care in university medicine. Methods Inf Med 2018;57(S 01):e82-e91 [FREE Full text] [doi: 10.3414/ME17-02-0025] [Medline: 30016814]
4. Data integration centres. Medical Informatics Initiative Germany. URL: https://www.medizininformatik-initiative.de/en/consortia/data-integration-centres [accessed 2022-09-22]
5. The Medical Informatics Initiative's core data set. Medical Informatics Initiative Germany. URL: https://www.medizininformatik-initiative.de/index.php/en/medical-informatics-initiatives-core-data-set [accessed 2022-09-22]
6. MIRACOLIX-Tools. MIRACUM. URL: https://www.miracum.org/en/das-konsortium/datenintegrationszentren/miracolix-tools/ [accessed 2022-09-22]
7. Jacobsen A, de Miranda Azevedo R, Juty N, Batista D, Coles S, Cornet R, et al. FAIR principles: interpretations and implementation considerations. Data Intell 2020;2(1-2):10-29 [FREE Full text] [doi: 10.1162/dint_r_00024]
8. Vasilevsky NA, Brush MH, Paddock H, Ponting L, Tripathy SJ, Larocca GM, et al. On the reproducibility of science: unique identification of research resources in the biomedical literature. PeerJ 2013;1:e148 [FREE Full text] [doi: 10.7717/peerj.148] [Medline: 24032093]
9. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A survey on bias and fairness in machine learning. ACM Comput Surv 2021;54(6):1-35 [doi: 10.1145/3457607]
10. Hasselbring W, Carr L, Hettrick S, Packer H, Tiropanis T. From FAIR research data toward FAIR and open research software. IT 2020;62(1):39-47 [FREE Full text] [doi: 10.1515/itit-2019-0040]
11. de Lusignan S, Liaw ST, Krause P, Curcin V, Vicente MT, Michalakidis G, et al. Key concepts to assess the readiness of data for international research: data quality, lineage and provenance, extraction and processing errors, traceability, and curation: contribution of the IMIA primary health care informatics working group. Yearb Med Inform 2018;20(01):112-120 [FREE Full text] [doi: 10.1055/s-0038-1638748]
12. Weng C. Clinical data quality: a data life cycle perspective. Biostat Epidemiol 2020;4(1):6-14 [FREE Full text] [doi: 10.1080/24709360.2019.1572344] [Medline: 32258941]
13. PROV-overview. W3C Working Group Note. URL: https://www.w3.org/TR/prov-overview/ [accessed 2022-09-22]
14. Herschel M, Diestelkämper R, Lahmar HB. A survey on provenance: what for? What form? What from? VLDB J 2017;26(6):881-906 [doi: 10.1007/s00778-017-0486-1]
15. Metadata and resource description. W3C Technology and Society domain. URL: https://www.w3.org/Metadata/ [accessed 2023-09-01]
16. Hume S, Sarnikar S, Noteboom C. Enhancing traceability in clinical research data through a metadata framework. Methods Inf Med 2020;59(2-03):75-85 [doi: 10.1055/s-0040-1714393] [Medline: 32894879]
17. Kelley K, Clark B, Brown V, Sitzia J. Good practice in the conduct and reporting of survey research. Int J Qual Health Care 2003;15(3):261-266 [FREE Full text] [doi: 10.1093/intqhc/mzg031] [Medline: 12803354]
18. Dowding D. Best practices for mixed methods research in the health sciences John W. Creswell, Ann Carroll Klassen, Vicki L. Plano Clark, Katherine Clegg Smith for the office of behavioral and social sciences research; qualitative methods overview Jo Moriarty. Qual Soc Work 2013;12(4):541-545 [doi: 10.1177/1473325013493540a]
19. O'Cathain A, Murphy E, Nicholl J. The quality of mixed methods studies in health services research. J Health Serv Res Policy 2008;13(2):92-98 [doi: 10.1258/jhsrp.2007.007074] [Medline: 18416914]
20. Humphrey WS. Characterizing the software process: a maturity framework. IEEE Softw 1988;5(2):73-79 [doi: 10.1109/52.2014]
21. Harris PA, Taylor R, Minor BL, Elliott V, Fernandez M, O'Neal L, et al. The REDCap consortium: building an international community of software platform partners. J Biomed Inform 2019;95:103208 [FREE Full text] [doi: 10.1016/j.jbi.2019.103208] [Medline: 31078660]
22. The R project for statistical computing. R. 2020. URL: https://www.R-project.org/ [accessed 2023-10-18]

23. Razick S, Močnik R, Thomas LF, Ryeng E, Drabløs F, Sætrom P. The eGenVar data management system--cataloguing and sharing sensitive data and metadata for the life sciences. Database (Oxford) 2014;2014:bau027 [FREE Full text] [doi: 10.1093/database/bau027] [Medline: 24682735]
24. Jira software features. ATLASSIAN. URL: https://www.atlassian.com/software/jira/features [accessed 2022-09-22]
25. ATLASSIAN. URL: https://www.atlassian.com/software/confluence/features [accessed 2022-09-22]
26. Curcin V, Soljak M, Majeed A. Managing and exploiting routinely collected NHS data for research. Inform Prim Care 2012;20(4):225-231 [FREE Full text] [doi: 10.14236/jhi.v20i4.1] [Medline: 23890333]
27. Kapsner LA, Mang JM, Mate S, Seuchter SA, Vengadeswaran A, Bathelt F, et al. Linking a consortium-wide data quality assessment tool with the MIRACUM metadata repository. Appl Clin Inform 2021;12(4):826-835 [FREE Full text] [doi: 10.1055/s-0041-1733847] [Medline: 34433217]
28. Kamoun C, Roméjon J, de Soyres H, Gallois A, Girard E, Hupé P. Biogitflow: development workflow protocols for bioinformatics pipelines with git and GitLab. F1000Res 2020;9:632 [FREE Full text] [doi: 10.12688/f1000research.24714.3] [Medline: 33732441]
29. Michaelis L, Poyraz RA, Muzoora MR, Gierend K, Bartschke A, Dieterich C, et al. Insights into the FAIRness of the German network university medicine: a survey. Stud Health Technol Inform 2023;302:741-742 [doi: 10.3233/SHTI230251] [Medline: 37203481]
30. Taylor K, Woodcock R, Cuddy S, Thew P, Lemon D. A provenance maturity model. In: Denzer R, Argent RM, Schimak G, Hrebícek J, editors. Environmental Software Systems Infrastructures, Services and Applications. Germany: Springer International Publishing; 2015:1-18
31. Groth P, Moreau L. Representing distributed systems using the open provenance model. Future Gener Comput Syst 2011;27(6):757-765 [FREE Full text] [doi: 10.1016/j.future.2010.10.001]
32. Guedes T, Martins LB, Falci MLF, Silva V, Ocaña KACS, Mattoso M, et al. Capturing and analyzing provenance from spark-based scientific workflows with SAMbA-RaP. Future Gener Comput Syst 2020;112:658-669 [FREE Full text] [doi: 10.1016/j.future.2020.05.031]
33. Gierend K, Krüger F, Waltemath D, Fünfgeld M, Ganslandt T, Zeleke AA. Approaches and criteria for provenance in biomedical data sets and workflows: protocol for a scoping review. JMIR Res Protoc 2021;10(11):e31750 [FREE Full text] [doi: 10.2196/31750] [Medline: 34813494]
34. Curcin V, Miles S, Danger R, Chen Y, Bache R, Taweel A. Implementing interoperable provenance in biomedical research. Future Gener Comput Syst 2014;34:1-16 [FREE Full text] [doi: 10.1016/j.future.2013.12.001]
35. Johns M, Meurers T, Wirth FN, Haber AC, Müller A, Halilovic M, et al. Data provenance in biomedical research: scoping review. J Med Internet Res 2023;25:e42289 [FREE Full text] [doi: 10.2196/42289] [Medline: 36972116]
36. Gierend K, Wodke JAH, Genehr S, Gött R, Henkel R, Krüger F, et al. TAPP: Defining standard provenance information for clinical research data and workflows—Obstacles and opportunities. New York, NY, United States: Association for Computing Machinery; 2023 Presented at: WWW '23 Companion: Companion Proceedings of the ACM Web Conference 2023; 30 April 2023-4 May 2023; Austin TX USA p. 1551-1554 URL: https://dl.acm.org/doi/10.1145/3543873.3587562 [doi: 10.1145/3543873.3587562]

Abbreviations

CDS: core data set
CMM: capability maturity model
DI: data integration
DIC: data integration center
DIC2MF: data integration center toward maturity framework
DS: data sharing
ETL: extract-transform-load
FAIR: findable, accessible, interoperable, reusable
FHIR: Fast Healthcare Interoperability Resources
GRAMMS: Good Reporting of a Mixed Methods Study
MII: Medical Informatics Initiative
MIRACUM: Medical Informatics in Research and Care in University Medicine
RDR: Research Data Repository
REDCap: Research Electronic Data Capture
SOP: standard operating procedure
W3C: World Wide Web Consortium

Edited by A Mavragani; submitted 08.05.23; peer-reviewed by M Sicilia, T Miksa, N Howe; comments to author 19.08.23; revised version received 09.09.23; accepted 29.09.23; published 08.11.23

Please cite as:
Gierend K, Freiesleben S, Kadioglu D, Siegel F, Ganslandt T, Waltemath D
The Status of Data Management Practices Across German Medical Data Integration Centers: Mixed Methods Study
J Med Internet Res 2023;25:e48809
URL: https://www.jmir.org/2023/1/e48809
doi: 10.2196/48809
PMID:

3.2 Publication 2: Traceable Research Data Sharing in a German Medical Data Integration Center with FAIR geared provenance implementation: Proof-of-Concept Study

This section provides a feasibility study for the implementation of medical provenance traces, as originally published in the Journal of Medical Internet Research (JMIR) Formative Research, an international, peer-reviewed, and open access journal.
The original publication is available at JMIR Formative Research (https://doi.org/10.2196/50027).

Original Paper

Traceable Research Data Sharing in a German Medical Data Integration Center With FAIR (Findability, Accessibility, Interoperability, and Reusability)-Geared Provenance Implementation: Proof-of-Concept Study

Kerstin Gierend[1], Dipl Inf (FH); Dagmar Waltemath[2], Dr -Ing; Thomas Ganslandt[3], MD; Fabian Siegel[1], MD

[1]Department of Biomedical Informatics at the Center for Preventive Medicine and Digital Health, Medical Faculty Mannheim, Heidelberg University, Mannheim, Germany

[2]Core Unit Data Integration Center and Medical Informatics Laboratory, University Medicine Greifswald, Greifswald, Germany

[3]Chair of Medical Informatics, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany

Corresponding Author:
Kerstin Gierend, Dipl Inf (FH)
Department of Biomedical Informatics at the Center for Preventive Medicine and Digital Health
Medical Faculty Mannheim, Heidelberg University
Theodor-Kutzer-Ufer 1-3
Mannheim, 68167
Germany
Phone: 49 621383 ext 8087
Email: kerstin.gierend@medma.uni-heidelberg.de

Abstract

Background: Secondary investigations into digital health records, including electronic patient data from German medical data integration centers (DICs), pave the way for enhanced future patient care. However, only limited information is captured regarding the integrity, traceability, and quality of the (sensitive) data elements. This lack of detail diminishes trust in the validity of the collected data. From a technical standpoint, adhering to the widely accepted FAIR (Findability, Accessibility, Interoperability, and Reusability) principles for data stewardship necessitates enriching data with provenance-related metadata. Provenance offers insights into the readiness for the reuse of a data element and serves as a supplier of data governance.

Objective: The primary goal of this study is to augment the reusability of clinical routine data within a medical DIC for secondary utilization in clinical research. Our aim is to establish provenance traces that underpin the status of data integrity, reliability, and consequently, trust in electronic health records, thereby enhancing the accountability of the medical DIC. We present the implementation of a proof-of-concept provenance library integrating international standards as an initial step.

Methods: We adhered to a customized road map for a provenance framework, and examined the data integration steps across the ETL (extract, transform, and load) phases. Following a maturity model, we derived requirements for a provenance library. Using this research approach, we formulated a provenance model with associated metadata and implemented a proof-of-concept provenance class. Furthermore, we seamlessly incorporated the internationally recognized Word Wide Web Consortium (W3C) provenance standard, aligned the resultant provenance records with the interoperable health care standard Fast Healthcare Interoperability Resources, and presented them in various representation formats. Ultimately, we conducted a thorough assessment of provenance trace measurements.

Results: This study marks the inaugural implementation of integrated provenance traces at the data element level within a German medical DIC. We devised and executed a practical method that synergizes the robustness of quality- and health standard–guided (meta)data management practices. Our measurements indicate commendable pipeline execution times, attaining notable levels of accuracy and reliability in processing clinical routine data, thereby ensuring accountability in the medical DIC. These findings should inspire the development of additional tools aimed at providing evidence-based and reliable electronic health record services for secondary use.

Conclusions: The research method outlined for the proof-of-concept provenance class has been crafted to promote effective and reliable core data management practices. It aims to enhance biomedical data by imbuing it with meaningful provenance, thereby bolstering the benefits for both research and society. Additionally, it facilitates the streamlined reuse of biomedical data.

As a result, the system mitigates risks, as data analysis without knowledge of the origin and quality of all data elements is rendered futile. While the approach was initially developed for the medical DIC use case, these principles can be universally applied throughout the scientific domain.

(JMIR Form Res 2023;7:e50027) doi: 10.2196/50027

KEYWORDS

provenance; traceability; data management; metadata; data integrity; data integration center; medical informatics

Introduction

Provenance—a piece of metadata—is considered information that is fundamental in the data life cycle because it expresses the traceability of the processed data and facilitates the reproducibility of the results [1,2]. The availability of provenance throughout the data life cycle is deemed a crucial factor for maintaining trust in the data at all stages [3]. The data life cycle encompasses data generation, processing, validation, analysis, reporting, and application for decision-making in any context, culminating in storage within a specified retention period [4]. Medical data integration centers (DICs), particularly those established within the German Medical Informatics Initiative, must enhance accountability for their activities. This is particularly crucial for the methods used in extracting, transforming, and loading sensitive patient data from heterogeneous clinical routine systems into (standardized) research data repositories for subsequent secondary use [5]. In this given context, it is necessary to understand the limitations of the provided data [6]. Collecting comprehensive and pertinent contextual provenance information along these processing pipelines is one approach to enhance the accountability of the medical DIC (Textbox 1). Provenance and integrity must be systematically evaluated and documented in routinely collected data sets to facilitate their reuse in clinical trials [7].

Textbox 1. Accountability in a German medical data integration center.

Accountability means accepting responsibility for activities and in this context entails all procedures and processes for data managing pipelines [8]. This includes keeping the movement of data elements transparent and traceable. Provenance traces enable documentation of this movement and hence generate trust in the data integrity and reliability of the provided data for secondary use.

To achieve reproducibility [9] and integrity when exchanging data between academia and industry, researchers must adhere to essential research principles, particularly following good practice guidelines (eg, good clinical practice, good research/scientific practice, commonly referred to as GxP) [10]. Ensuring and evaluating data integrity and data provenance are anticipated to be prerequisites for clinical trial data [11]. For instance, the clinical research data quality standard ALCOA+ (Attributable, Legible, Contemporaneous, Original, and Accurate+) articulates enhanced data integrity properties and fundamentally contributes to provenance information [12]. These properties pertain to attributable, legible, contemporaneous, original, accurate, complete, consistent, enduring, and available data characteristics [10].

In addition to adhering to good scientific practice [13], heightened legal requirements such as compliance with the General Data Protection Regulation (GDPR) in the European Union, or contractual obligations, mandate evidence-based data processing for both deidentification and reidentification of data, encompassing the life cycle of the patient’s consent [14].

A crucial factor in advancing these objectives is the metadata acquired from the data transformation and integration process throughout the data life cycle. The field of biological research has already acknowledged the significance of metadata, as outlined in ISO norms such as ISO/CD 20961 [15] and ISO/TC 276/WG5 on data processing and integration [16]. ISO 20961, for example, specifies requirements for the consistent formatting and documentation of data and metadata.

Furthermore, the FAIR (Findability, Accessibility, Interoperability, and Reusability) guiding principles for data management and data stewardship emphasize the overall relevance of metadata for the data itself, including those used in infrastructures and services [17]. Aspects of the FAIR recommendations explicitly address provenance capture. As such, the “R1.2” FAIR principle demands machine-accessible and readable metadata, which include provenance information about the data creation or generation. Related metadata accumulate not only during the data transformation itself but also within the software used [18]. The principle “R1.3” expects metadata to be adhering to domain-relevant community standards such as the HL7 Fast Healthcare Interoperability Resources (FHIR) or Dublin core [1]. FHIR is an internationally recognized standard that supports the exchange of data between different software systems within the health care sector [19]. In this vein, the FHIR resource “provenance” records entities and processes involved in creating a specific resource. From a technical point of view, the FHIR Provenance resource is founded on the framework of the open W3C standard PROV-Data Model definition and ontology [20], the successor to the Open Provenance Model [21]. Here, the concepts of linked entities, activities, and agent resources enable the establishment of a provenance model. Such resources can be described with the W3C Resource Description Framework (RDF) method [22]. RDF is a data model, which is commonly stored in formats such as RDF/XML (.rd) or JSON-LD (.json). All formats represent a knowledge graph.

As of now, the capture of provenance in health care is not adequately or uniformly implemented in German medical DICs, as revealed in a recent study on their data management status [23]. The results demonstrated that provenance is indeed a factor strongly influenced by the maturity level of data management practices. Following complex transformations in the data integration process, the provenance of data elements is often lost, making it difficult to impossible to assess the (measurement) quality of a data element. This reduction in traceability diminishes trust in the validity of the collected data.

The primary objective of this study is to improve the reusability of clinical routine data within a medical DIC for its secondary application in clinical research. Our goal is to enhance processed clinical routine data by incorporating appropriate semantic metadata, a key requirement guided by the FAIR principles [17]. Furthermore, our intention is to bolster the accountability of our DIC by mitigating the risks associated with the reuse of compromised data in clinical research.

To our knowledge, this is the first demonstration of provenance integration within a medical DIC.

Methods

Materials

We used test data to develop and test our provenance class. Test data elements were chosen to reflect the composition of a typical data integration repository. We created exemplary dummy data element definitions with comprehensive annotation (Textbox 2). We defined 7 data element types and generated 100,000 data elements for each data element type to generate a total of 700,000 provenance records using a Python (Python Foundation) script.

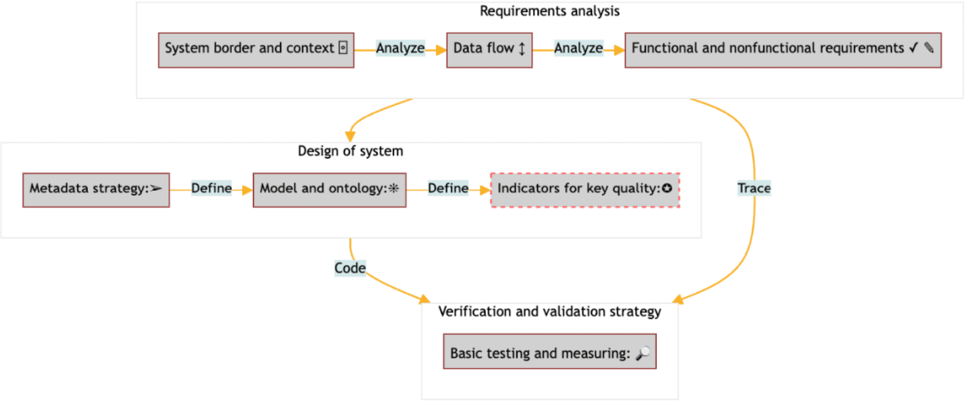
Textbox 2. Exemplary dummy text–based data element definition.

```
id='syst_blood_pressure',
name='syst_blood_pressure',
description='Systolic Blood Pressure',
source='stg_sap_vitalis',
source_variable='SysBP',
destination='dwh_vitalis',
destination_variable='SBP',
description_of_transformation='copy',
description_of_qualitycheck='range check 80-160',
status_log='passed date 12.May2022',
sop_name='SOP p'
sop_version='v1.5',
sop_status='approved',
steward_name='no name given')
```

Proof-of-Concept Solution

Following the tailor-made provenance framework [3], we developed a proof-of-concept provenance solution. This framework complements a standard software engineering cycle (requirements, design, coding, testing, and implementation) with insights from a comprehensive literature search and uses established works as a guide to the users of the framework. The expanded requirements analysis is substantiated by the topics identified through the literature search. Details are described in Figure 1.

Figure 1. Overview of the road map steps.


Requirements analysis
System border and context
Analyze
Data flow
Analyze
Functional and nonfunctional requirements ✓
Design of system
Metadata strategy:
Define
Model and ontology:
Define
Indicators for key quality:
Trace
Code
Verification and validation strategy
Basic testing and measuring:

Requirements Analysis

Overview

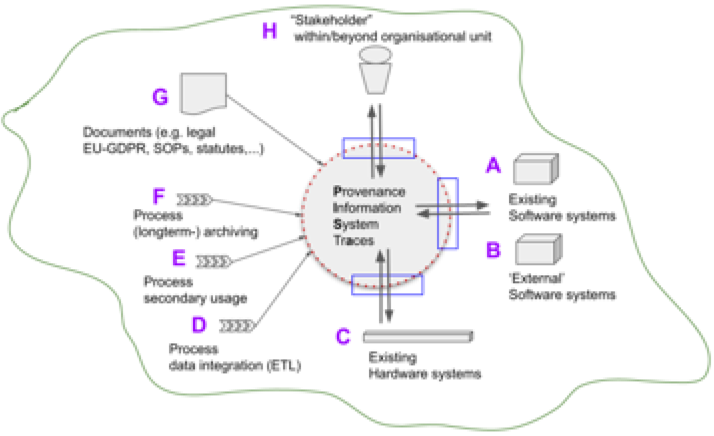
An interdisciplinary team of internal stakeholders in the University Medicine Mannheim-DIC (lead, medical experts, computer scientists, technical staff, and process owner of the ETL [extract, transform, and load] process) performed the requirements analysis for the research approach. Initially, we engaged in discussions, documented feedback, and obtained approval for our own data pipeline processes, based on the WH questions (what, when, where, who, why, how, which, whose). This was done to ensure accurate and risk-managed data processing pipelines. Our focus centered on questions related to data governance, annotation, documentation, interoperability, data integrity and accuracy, data sharing, and information technology operations. This emphasis aligns with a prior investigation on data management practices in German DICs [23], where these questions were identified as integral to tracing patient data through the DICs.

Building on the previous steps, we initiated the process by visualizing the scope definition (system border and context) of the planned provenance tracking systems. Using notation according to DeMarco [24], we generated a data flow diagram. Following this, we documented the resultant requirements, representing them in free text and as a unified modeling language (UML) class diagram to address various requirements perspectives [25].

System Border and Context

The context view (Figure 2) is used to delineate the scope of our system, establishing the boundary between functionalities that are considered in and out of scope. The system to be modeled, known as the Provenance Information System Traces (PISA), is depicted as a circle in the center (outlined by the dotted red line in Figure 2). At the conceptual level, we established the system border to encompass all aspects within the object scope. We delineated the system context (depicted in green as a freehand drawing) with aspects (A to H) that impact the planned provenance tracking system in our medical DIC. The processes that were modeled had been previously defined by local stakeholders and were influenced by the processes of the medical informatics initiative community [5]. The core process, the ETL process (D), includes valid documents (G) (eg, statutes, standard operating procedures, European Union-GDPR) and the involvement of stakeholders within and beyond the organizational unit (H), representing the primary focus of our development efforts. Existing software and hardware systems (A–C), as well as the processes of secondary usage for data request (E) and long-term archiving (F), are outside the scope of this study.

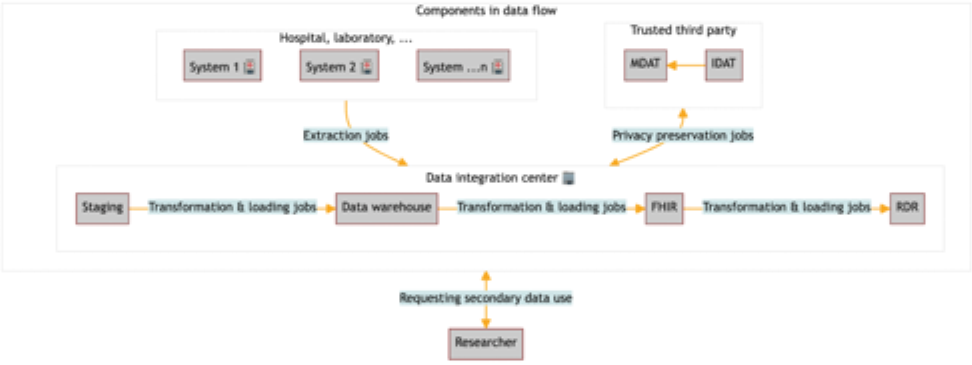
Figure 2. Aspects in the system context and border of the Provenance Information System Traces (PISA). EU: European Union; GDPR: General Data Protection Regulation; SOP: standard operating procedure.



Data Flow

Given the multitude of processes within a DIC, we confined our focus to the requirements related to the data integration process (Figure 2; ETL, letter D). We scrutinized the data flow and derived a data flow diagram, illustrating the functional requirements perspective (Figure 3). As part of the Medical Informatics Initiative, all DICs in Germany modeled a comparable, generic data flow. This data flow delineates the movement of data among processes (ETL), storage entities (staging area, data warehouse, FHIR server, and research data repository), and involved actors (staff in DIC, researcher, and trusted third party). Processes encapsulate functions responsible for transforming processing data. These processes consume input data from diverse systems, manage these data, and convey the results to an output. Storage ensures data persistence, allowing processes to access the storage in read or write modes. Actors actively engage in information exchange with the system.

Figure 3. Simplified general data flow diagram in the data integration center. The simplified general data flow diagram in the data integration center (DIC) provides information about components participating in data flow: different hospital or laboratory systems donating the data, the independent trust center (trusted third party) enabling the separate processing of identifying data (IDAT) and medical data (MDAT), the data integration center with the different integration phases staging, data warehouse, FHIR and the research data repository (RDR). Individual DIC may deviate from this general data flow. FHIR: Fast Healthcare Interoperability Resources.


Components in data flow
Hospital, laboratory, ...
System 1
System 2
System ...n
Trusted third party
MDAT
IDAT
Extraction jobs
Privacy preservation jobs
Data integration center
Staging
Transformation & loading jobs
Data warehouse
Transformation & loading jobs
FHIR
Transformation & loading jobs
RDR
Requesting secondary data use
Researcher

Requirements Description

In a previous publication, we conducted interviews with various German medical DICs [23]. Through these interviews, we identified the most crucial requirements, emphasizing assessments of data quality, traceability, and information capability. Additionally, transparency in processing steps, workflows, and data sets emerged as a significant consideration. Other identified requirements encompassed aspects such as debugging or performance evaluation. Additionally, there was a focus on compliance with regulations, reproducibility, support of the scientific utilization process, increased confidence in data, and clear regulation of responsible parties [23].

In alignment with this study, we established preconditions and requirements along the data flow for implementing the provenance tracking system. We identified the intended features for the implementation of the PISA and derived the system's requirements (Table 1). In general, PISA should have the capability to trace the complete production history of a data element while incorporating domain-specific characteristics of the data element. These provenance traces for an individual data element must be captured along the presented data flow.

Table 1. Requirements for the proof of concept for PISA[a].

Number	Requirements (functional and nonfunctional)	Explanation
1	PISA must have the capability to track the complete processing history of a data element, and the provenance information must be stored in a database. This encompasses all derivation steps performed on data elements during their processing steps.	It includes all the information (metadata) required for producing a specific data set or a data element while preserving its data integrity status. This encompasses details such as data source, data destination, method, tools, software, and versions used. The benchmark should align with the "entities" and "activities" components of the W3C model.
2	PISA must possess the capability to trace organizational responsibilities and the means used.	It includes information (metadata) about all the involved agents in producing a data set or data elements, such as staff, standard operating procedures, and guidance. The benchmark should align with the "agent" components of the W3C model.
3	PISA must be analyzable by an authorized user and capable of producing diverse representations and export formats for the provenance traces.	Detailed provenance traces are accessible and exportable to support evaluation by users, including formats such as log files, FHIR[b] provenance, W3C[c] RDF[d]/XML, and RDF/JSON-LD provenance.
4	PISA must be able to track the quality status and assessment of data elements.	The provenance information for a data element is expanded to include the quality status of the processed data element.
5	PISA must be able to track the status of the script execution.	At a minimum, the provenance information should encompass the verification status and time stamp of the processed scripts.
7	PISA must provide a high level of ease of use for ETL[e] programmers and should be usable without requiring in-depth knowledge of provenance terms and concepts.	PISA should facilitate easy integration into ETL pipelines with transfer interfaces, allowing seamless integration with established technologies. Moreover, it must be easy to install, for example, by supporting widely used and easily set up databases.
8	PISA must be time-efficient and capable of ensuring acceptable performance.	Time measurements per data element must take place and be evaluated to verify the feasibility of the proof-of-concept approach.
9	Verification by unit tests/code coverage >80%	Passed testing results.

[a] PISA: Provenance Information System Traces.

[b] FHIR: Fast Healthcare Interoperability Resources.

[c] W3C: Word Wide Web Consortium.

[d] RDF: Resource Description Framework.

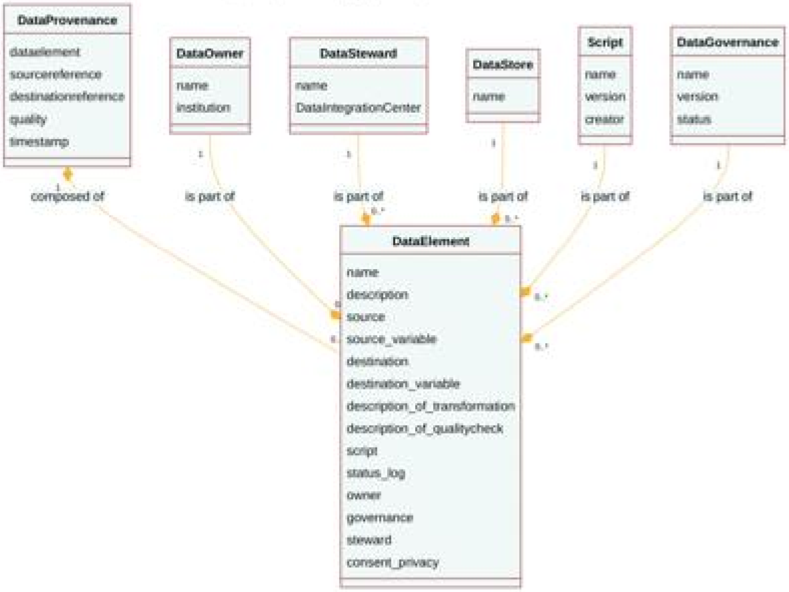
[e] ETL: extract, transform, and load.

Design and Architecture of the Provenance Class

Development of the Logical Data Model

Based on the aforementioned requirements (Table 1) and the DIC maturity model [23], we constructed the logical data model as a UML class diagram, identifying classes and their associations (Figure 4).

Figure 4. The logical data model as UML class diagram (technology-agnostic).



Metadata Strategy

Our metadata strategy centered on characterizing the data elements and their associated artifacts throughout their processing pipeline.

Aligned with the requirements and the logical model, we extracted the pertinent provenance metadata and aligned this provenance profile with the W3C components entity, agent, and activity. Simultaneously, we diligently enforced documentation efforts and annotation, guided by good documentation practices such as the ALCOA(+) principles for the identified components [10]. The annotation process we implemented enhanced the comprehension, increased understanding, and improved the traceability of the processed data elements.

The FAIR principles R1.2 and R1.3 guided us to enrich (R) data elements with meaningful (provenance) metadata. Consequently, we characterized data elements by collecting content-rich contextual and technical metadata that narrate the story of the entire data processing workflow and link to related artifacts (Table 2). During the transformation processes, we documented quality procedures and incorporated coding practices and versioning information.

Table 2. Levels of contextual and technical metadata and their related FHIR[a] mapping: a mapping example of our metadata to the FHIR Provenance resource. The FHIR Provenance elements are aligned with the W3C[b] PROV model elements.

Level[c]	Description[d]	Possible mapping[e]	Exemplified output[f]
Data Governance[g]	Name and version of the standard operating procedures or regulation (eg, “DIC_ETL-ST.pdf, v1, approved”)	.policy .agent.type	“policy” : [“http://example.org/policy/1234”], “location”: { “reference”: “DIC” },

Data Owner	Name of the (hospital) department and the responsible person owning the patient data (eg, physician or stakeholder name)	.authorization
		.agent
		.agent.type
		.agent.role
		.agent.who
		.agent.onBehalfOf
			“authorization”: {
			“coding”: [
			{
			“system”: “http://terminology.hl7.org/CodeSystem/v3-ActReason”,
			“code”: “TRANSRCH”
			}
			]
			},

Data Steward	Name of the responsible data steward (eg, person who takes care of data management)	.location
		.agent
		.agent.type
		.agent.role
		.agent.who
		.agent.onBehalfOf
			“agent”: {
			“who”: {
			“display”: “Hr. Koch”
			}
			}

Data Store	Used input or created output data file as part of the processing pipeline (eg, name original source system and name target system)	.entity
		.entity.role
		.entity.what
		.target (as mapping from entity)
			“entity”: {
			“what”: {
			“identifier”: [
			{
			“system”: “urn:ietf:rfc:3986”,
			“value”: “243c773b-8936-407e-9c23-270d0ea49cc4”,
			“display”: “”
			}
			]
			}
			}

Data Script	Scripts or programs developed to process the data with a description of script version and name and creator (eg, etl_st.py v1 MZ)	.activity
		.basedOn
		.agent.type
			“activity”: {
			“coding”: [
			{
			“system”: “http://terminology.hl7.org/CodeSystem/iso-21089-lifecycle”,
			“code”: “averaging”,
			“display”: “Transform”
			}
			]
			}
			“basedOn”: [
			{
			“reference” : “ServiceRequest”
			}
			]

Level[c]	Description[d]	Possible mapping[e]	Exemplified output[f]
Data Element	Individual characteristics per data element during a processing step such as ID, name, description, source and destination information from Data Store Level, description of the transformation approach, description of quality check (testing and validation approach), privacy and security status, and information from Script Level	.entitiy .entity.role .entity.what .entity.agent	Schema as in Data Store Level
Data Provenance	References to all other mentioned levels and testimony for quality (eg, “25, 3, 5, good, 2023-02-03 06:01:34”)	.id .occuredDateTime .recorded .patient .encounter .target	“id” : “id” ”occuredDateTime“: ”timestamp“, ”recorded“: ”timestamp“
Data Infrastructure[g]	Used hardware and software conditions during data processing	N/A[h]	N/A

[a] FHIR: Fast Healthcare Interoperability Resources.

[b] W3C: Word Wide Web Consortium.

[c] Level corresponds to the maturity level of the data integration center.

[d] Description of the possible content or annotation.

[e] Possible mapping to the Health Level 7 FHIR resource “Provenance.”

[f] One possible exemplified output extract as a serialization in FHIR JSON.

[g] Not yet or only partly implemented.

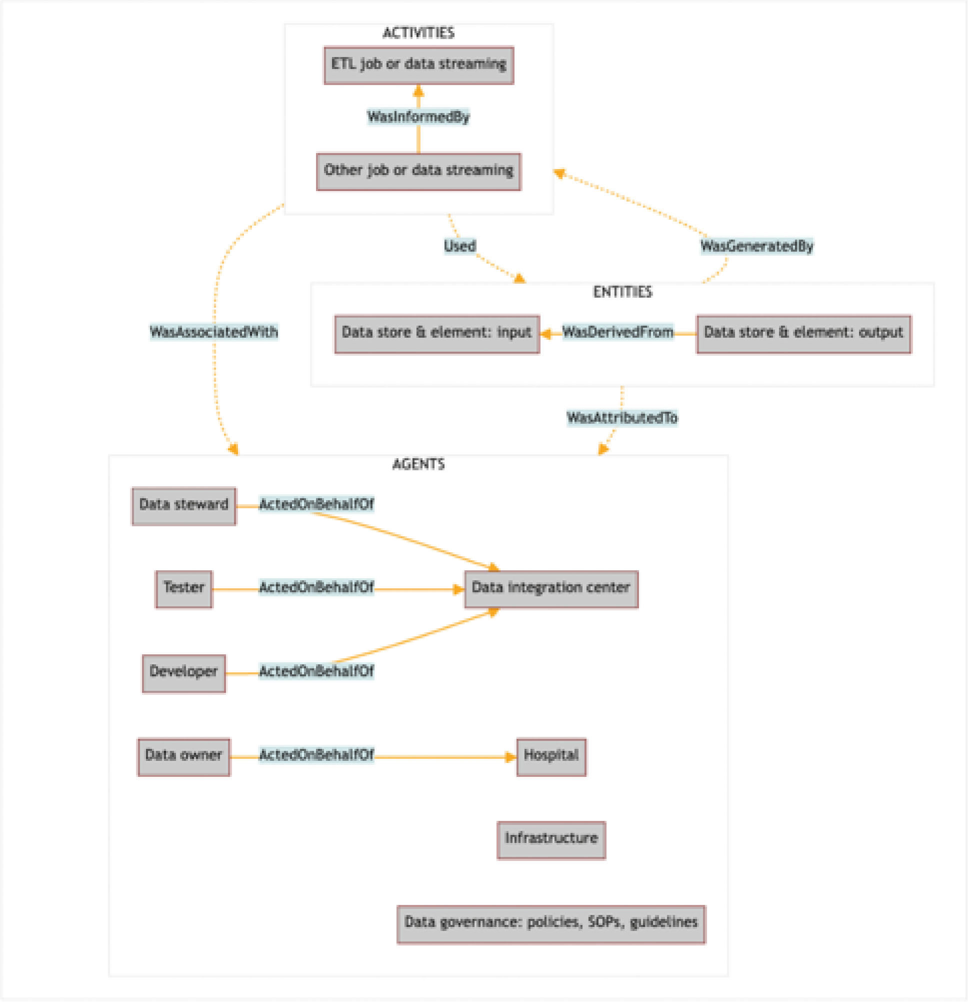
[h] N/A: not applicable.

Examples of expanded metadata elements are more detailed descriptions of the transformation, the quality check, and the status of the data element in scope, or the results of the used log files. The metadata gathering for provenance comprises both manual annotation and an automated collection process, representing a hybrid form of provenance [26].

Ontology

We organized, annotated, and represented information using WebProtégé 4.0.2 (Protege Team in the Biomedical Informatics Research Group at Stanford University), a tool designed for collaboratively creating complex ontologies [27]. The W3C PROV ontology and the fundamental relationships between entities, activities, and agents served as a framework for representing the provenance graph [20]. More specifically, we mapped processes onto activities, actors onto agents, and input/output data onto entities. The attributes of the provenance data model were aligned with the attributes of the data set. An instantiation of the provenance model, reflecting the W3C PROV vocabulary and layout convention, is illustrated in Figure 5. Additionally, the W3C PROV supports interoperable interchange of provenance in heterogeneous environments.

Figure 5. Exemplary instantiation of the provenance information model. SOP: standard operating procedure.



Implementation and Verification Approach

Finally, building on the preceding steps, we developed an open-source Python class "Data Provenance" with associated methods, and validated our approach in an exemplified data integration pipeline [28]. Provenance traces were mapped exemplarily onto the W3C RDF/XML and HL7 FHIR resource "Provenance" in its current maturity level (version R 5). We utilized peewee (version 3.15.4), a Python Object-Relational Mapping library that supports the binding of objects to relational databases such as SQLite, MySQL, or PostgreSQL [29]. To visualize the provenance traces, we used the Mermaid plotting framework [30].

The verification and validation approach for the developed provenance class involved an independent code review and unit tests to ensure that the code meets the requirements of the design. We assessed efficiency (storage space in kilobytes and computing time) and ensured the maintainability of the program (code structure, modularity, comments in code, currency, and comprehensibility of documentation).

While creating provenance records, we conducted a runtime experiment to measure the performance of our developed class. We recorded the time that the program took to run for proper execution. The runtime environment comprised the operating system Ubuntu 22.04.2 LTS (Canonical Ltd.), 32 GB memory, and an 8-core Intel Xeon Platinum 8276 CPU @ 2.20-GHz computer.

As a runtime environment, we used a virtual machine running on top of the machine. The runtime period was defined as the duration when the program was actively running.

We conducted measurements per data element and per provenance record on 9 virtual machines, each utilizing different data element block sizes (starting with 1, 10, 100, 1000, 10,000, and 100,000 up to 9, 90, 900, 9000, 90,000, and 900,000 data elements). For the analysis of runtime measurements, we used R version 4.2.0 (2022; R Foundation for Statistical Computing), and figures were generated using the ggplot2 package [31].

The code is available in a git repository under the Massachusetts Institute of Technology (MIT) license [32].

Ethical Considerations

Given the nature of the proof-of-concept study relying on dummy test data, ethics approval, informed consent, and deidentification were not applicable.

Results

Provenance Traces Representation

All the gathered provenance information is in a machine-readable format. Additionally, FHIR health care standards were used [33].

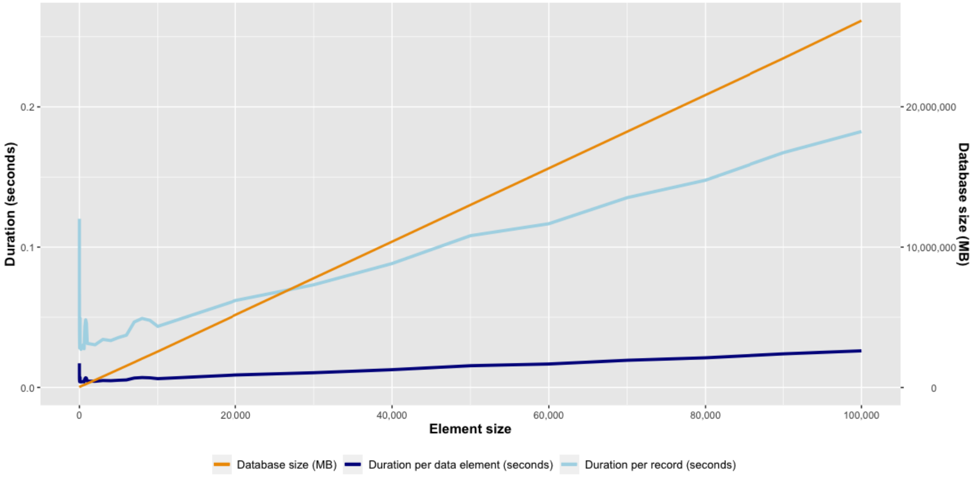
We developed an FHIR profile based on the "provenance" resource, resulting in a record that delineates the entities and processes involved in producing, delivering, or otherwise influencing that resource. This was accomplished by mapping the contextual and technical metadata to the corresponding resource provenance elements (Table 2).

Through the integration of all metadata levels, we facilitated the traceability of each data element. We illustrated the traceability using a data flow diagram and presented it in a human-readable text form. Additionally, the provenance information was exported into various formats such as FHIR-JSON, W3C-RDF/XML, W3C-RDF/JSON-LD, or a text-based log file. This approach aligns with data obtained in other studies [34].

Measurement of Provenance Traces

As anticipated, the specified provenance class successfully generated the database and the metadata tables according to the UML class diagram (illustrated in Table 2). Provenance records were automatically appended to the provenance table throughout the execution of the exemplified data integration pipeline. We recorded runtime measurements of the algorithm, displayed separately for the storage duration of a data element and for a record, as well as the corresponding increase in the database (Figure 6). As evident, the runtime complexity of the algorithm per data element indicates a nearly linear relationship with the size of the input data.

Figure 6. Provenance-Runtime-Experiment presenting storage duration per element and per record.



We observed an acceptable runtime duration ranging from 0.0039 to 0.02601 seconds per data element. However, when measuring the runtime for a provenance record, we encountered an increasing duration, ranging from 0.0271 to 0.1882 seconds. Given that our approach incorporates novel aspects, we were unable to find comparable studies for this measurement. Nevertheless, the data obtained here suggest that using this approach to establish provenance traces can yield accurate and timely information.

Verification and Validation

The validation status for our proof-of-concept provenance class is outlined in Table 3. We anticipate that our results can be readily adopted for additional metadata components and seamlessly transferred to decision-making applications.

Table 3. Validation status of requirements.

Requirement number	Validation result
1	Introduction of metadata for data elements and their processing collected automatically during ETL[a] job running in data flow. Relevant tables (DataProvenance, DataElement, and associated tables) in the provenance database were created and continuously updated during processing.
2	Organizational topics (DataGovernance, DataSteward, and DataOwner) were recorded in the provenance database and continuously updated during processing.
3	Provenance traces were created in different formats. Detailed provenance traces are accessible and exportable to support evaluation by users (eg, FHIR[b] provenance, W3C[c] RDF[d]/XML RDF/JSON-LD provenance).
4	The quality status of a processed data element is tracked and currently presented with a placeholder value in the DataProvenance table (see the “Future Work” section).
5	The verification status of used scripts and time stamps were recorded in the table DataElement. More specific content-related provenance information needs to be added in the second step. This compromises detailed annotation about the performed transactions and can be used for handling inconsistencies and rules for conflict resolution (see the “Future Work” section)
7	Easy integration into the ETL pipeline setup: only 3 lines of code, set up per data element: 1 line (see the “Future Work” section).
8	Time measurements confirmed satisfying results.
9	We achieved a code coverage of >90%, confirming that the code is comprehensively verified (quality aspect for software). We successfully verified the provenance with unit tests and validated all results against the defined requirements.

[a]ETL: extract, transform, and load.

[b]FHIR: Fast Healthcare Interoperability Resources.

[c]W3C: Word Wide Web Consortium.

[d]RDF: Resource Description Framework.

Discussion

Principal Findings

Our study introduces the first ready-to-use library designed to record provenance information from clinical data processing pipelines in a German medical DIC. This current research extends previous work in provenance by using an approach that systematically combines detailed insights from medical, data management, and information technology operational experts. This method aims to facilitate the reuse of enriched patient data with precision and rigor. We demonstrated that our research approach successfully facilitates the implementation of traceability in the processing of data elements. This, in turn, contributes to the promotion of good data management and documentation practices, ultimately ensuring sufficient provenance quality. Furthermore, these good practices pave the way for the (automated) generation of annotations [23] and prevent poor data integrity, thereby enhancing data quality [35]. Through this, we hypothesize that our work could contribute to the reliability and safety of quality-assured patient data for secondary use. Simultaneously, we mitigate the risks associated with the reuse of weak data in clinical research.

We fulfilled the requirement for FAIR (Findability, Accessibility, Interoperability, and Reusability) provenance information by adhering to standards for syntactic and semantic interoperability, including JSON, W3C PROV, and FHIR mapping. Compared with the FHIR resource Provenance, we noted that our metadata recording offers significantly more detailed contextual information for each data element. We suggest that improvements to the FHIR Provenance resource, particularly for data within medical DICs, be deliberated and harmonized with existing FHIR resources such as “AuditEvent” or the “FiveWs Pattern” [19].

The strengths of this study are (1) the provision of provenance information for data elements with export options to interchange standard formats such as FHIR-JSON or W3C RDF/XML; (2) the simplicity of integrating this provenance class into ETL and other data pipelines; and (3) the extensibility of metadata components along with acceptable runtime measurements.

Related Work

In general, research on provenance and related management has progressed significantly in recent years. Numerous studies have been conducted, both domain specific and domain independent, focusing on provenance. Recently submitted scoping review results on provenance tracking have yielded valuable insights and provided an extensive summary of current approaches and criteria [3]. The scoping review revealed technical, implementation, and knowledge gaps, with a specific emphasis on modeling and metadata frameworks for (sensitive) scientific biomedical data. Moreover, the primary focus of the research was centered on workflow provenance. This involved the utilization of models such as the Open Provenance Model or the W3C PROV data model across various semantic levels and tools in scientific workflows or experiments, as demonstrated in frameworks such as BioWorkbench or the OpenPREDICT use case [36,37]. Additionally, other work has delved into different yet more general approaches for metadata usage and harvesting [38,39]. A systematic literature analysis on functional requirements for medical data integration outlined general requirements for data traceability and metadata management [40].

While these prior efforts are crucial, they still lack the specific requirements and considerations tailored for a DIC use case. By contrast, our approach is finely tuned to the unique needs of a DIC, providing a comprehensive exploration of provenance that imparts medical meaning and understanding to the data elements, thereby enhancing their reusability.

Lessons Learned

We discovered that interdisciplinary competence profiles; fostering communication between medical experts, data stewards, and information technology developers; and establishing a common language were pivotal factors leading to significant progress in our specific DIC use case. Implementing proper data governance and comprehensive data management documentation, such as data management plans, would be instrumental in mitigating the risk of incorrect use of the data.

The lessons learned from our description could serve as motivation for other researchers aiming to establish FAIR-oriented provenance. This would not only advance the reuse of their research data and results but also underscore the importance of maintaining overall responsibility for the data, even after project funding concludes.

Future Work

Future work should also prioritize the development of a strategy for assessing data privacy, data integrity, and related quality of a data element. Integrating this information into the framework would enhance the expressiveness of the provenance information and enable the derivation of quality dimensions. For this reason, data elements may need to be accompanied by additional properties (refer to Table 2) that are significant for interpretability, helping determine limitations or detect duplications for use in similar research studies. Addressing the adequacy and relevance of the data element for upcoming research questions aids in supporting interpretation and, consequently, the reuse of a data element, as already highlighted in a draft Food and Drug Administration guidance [41]. To facilitate easy integration with other programming languages, we will provide an application programming interface.

Future studies should also explore ways to enhance the script for generating the provenance class in alignment with the FAIR for Research Software Principles [42]. Determining appropriate software metadata that accurately describe the specific characteristics of the software is an essential aspect to be addressed [18].

Before the future implementation and integration of the provenance class into real-world data integration processes, it is advisable to seek recommendations for risk measures. Factors such as the confidentiality level and security of provenance information, storage considerations, performance issues, and scalability should be carefully considered. In addition, it is crucial to consider experiences gained from maintaining metadata management and interoperable technologies, especially from professional data stewards. Ongoing exchanges with stakeholders and conducting usability evaluations are essential aspects that should be taken into account.

This work also contributes to a broader community project that seeks to establish the “Minimal Requirements for Automated Provenance Information Enrichment” (MIRAPIE) project [43].

Limitations

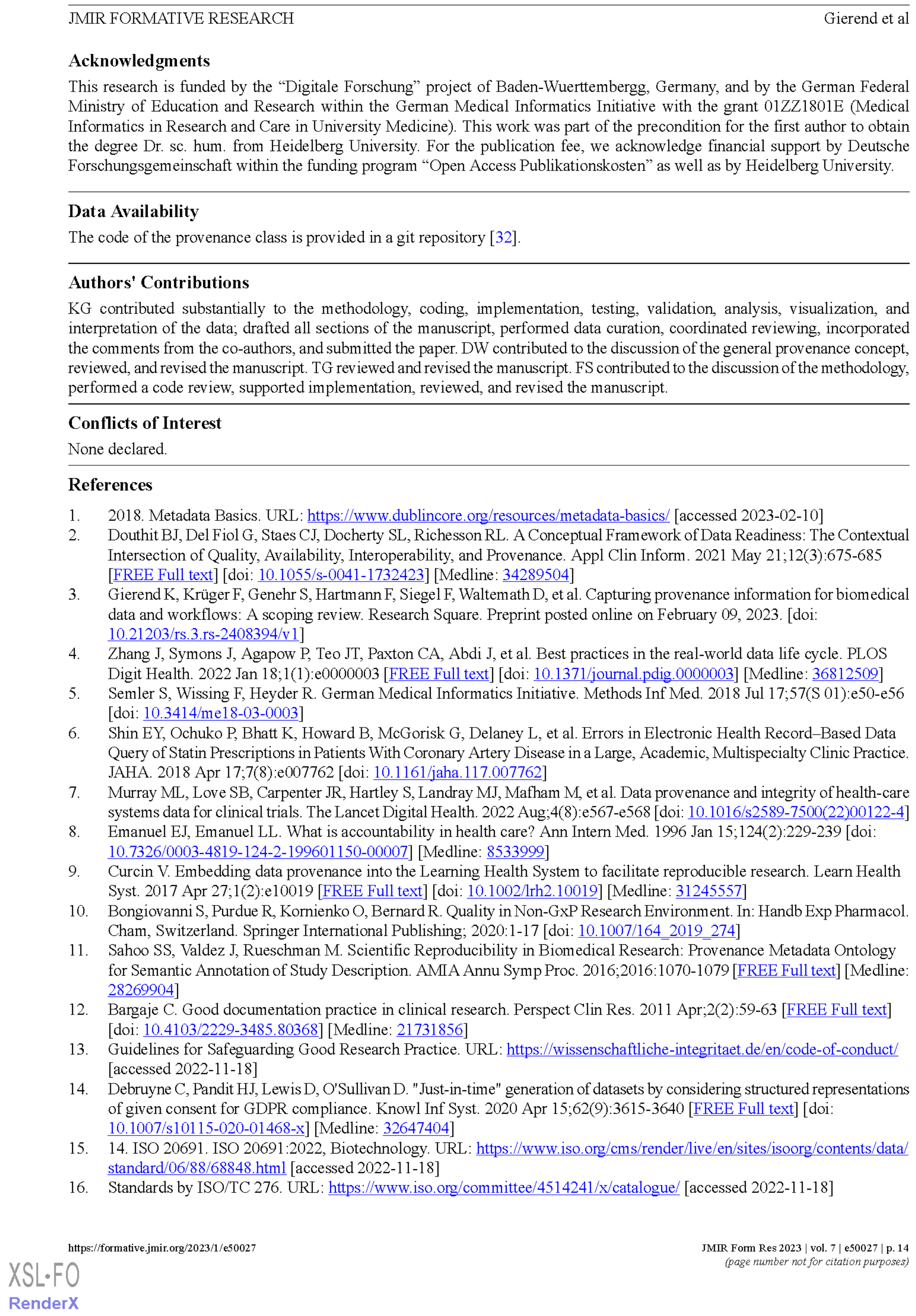
As the library has only been tested with simulated data, the next step—testing in a real environment—is currently in preparation. Despite the straightforward ETL integration approach, we will carefully assess the complexity and associated costs of implementation within the medical DIC. We recognize the need to bolster the overall qualification and validation concept. We believe it is crucial to expand the current provenance class to one that is inspection- or audit-ready, although accreditation demands additional measures and efforts. Additionally, further scalability analysis should be incorporated into the research approach.

Trust involves more than just the provenance of data elements; it also implies correctness and security against malicious users. This challenge can only be addressed through technical access limitations and organizational measures. Nevertheless, automated provenance traces can contribute to building trust in the transformation and movement of data within the DIC. Moreover, it empowers us to confidently assess the quality and validity of the original data points even after undergoing complex transformations within a data warehouse.

Conclusions

We have designed, developed, and implemented provenance traces at the data element level for a German medical DIC, with the potential for extension at the national level. The described research method for the proof-of-concept provenance class has been crafted to promote effective and reliable core data management practices, enriching biomedical data with meaningful provenance. This, in turn, strengthens the benefits for research and society while simplifying the reuse of biomedical data. While the approach was initially developed for the medical DIC use case, these principles can be applied universally throughout the scientific domain. The implementation and analysis of provenance traces play a crucial role in minimizing risks associated with undetected or unintended data integrity breaches. Hence, provenance traces significantly contribute to building trust in routine clinical data and enhancing the accountability of a medical DIC. We are confident that by adhering to this advanced practice, the existing gaps between industry (pharmaceutical companies), service providers, and academia can be mitigated. Consequently, this can lead to an increase in the secondary use of (sensitive) patient data in clinical investigations.

The outcomes of our research prompt additional questions, particularly regarding how in-depth exploration of further provenance analysis can predict the quality of data using machine learning methods. The limitations identified in our study indicate the need for further investigations into provenance theory, standards, and practices in the clinical field.



Acknowledgments

This research is funded by the “Digitale Forschung” project of Baden-Wuerttembergg, Germany, and by the German Federal Ministry of Education and Research within the German Medical Informatics Initiative with the grant 01ZZ1801E (Medical Informatics in Research and Care in University Medicine). This work was part of the precondition for the first author to obtain the degree Dr. sc. hum. from Heidelberg University. For the publication fee, we acknowledge financial support by Deutsche Forschungsgemeinschaft within the funding program “Open Access Publikationskosten” as well as by Heidelberg University.

Data Availability

The code of the provenance class is provided in a git repository [32].

Authors' Contributions

KG contributed substantially to the methodology, coding, implementation, testing, validation, analysis, visualization, and interpretation of the data; drafted all sections of the manuscript, performed data curation, coordinated reviewing, incorporated the comments from the co-authors, and submitted the paper. DW contributed to the discussion of the general provenance concept, reviewed, and revised the manuscript. TG reviewed and revised the manuscript. FS contributed to the discussion of the methodology, performed a code review, supported implementation, reviewed, and revised the manuscript.

Conflicts of Interest

None declared.

References

1. 2018. Metadata Basics. URL: https://www.dublincore.org/resources/metadata-basics/ [accessed 2023-02-10]
2. Douthit BJ, Del Fiol G, Staes CJ, Docherty SL, Richesson RL. A Conceptual Framework of Data Readiness: The Contextual Intersection of Quality, Availability, Interoperability, and Provenance. Appl Clin Inform. 2021 May 21;12(3):675-685 [FREE Full text] [doi: 10.1055/s-0041-1732423] [Medline: 34289504]
3. Gierend K, Krüger F, Genehr S, Hartmann F, Siegel F, Waltemath D, et al. Capturing provenance information for biomedical data and workflows: A scoping review. Research Square. Preprint posted online on February 09, 2023. [doi: 10.21203/rs.3.rs-2408394/v1]
4. Zhang J, Symons J, Agapow P, Teo JT, Paxton CA, Abdi J, et al. Best practices in the real-world data life cycle. PLOS Digit Health. 2022 Jan 18;1(1):e0000003 [FREE Full text] [doi: 10.1371/journal.pdig.0000003] [Medline: 36812509]
5. Semler S, Wissing F, Heyder R. German Medical Informatics Initiative. Methods Inf Med. 2018 Jul 17;57(S 01):e50-e56 [doi: 10.3414/me18-03-0003]
6. Shin EY, Ochuko P, Bhatt K, Howard B, McGorisk G, Delaney L, et al. Errors in Electronic Health Record–Based Data Query of Statin Prescriptions in Patients With Coronary Artery Disease in a Large, Academic, Multispecialty Clinic Practice. JAHA. 2018 Apr 17;7(8):e007762 [doi: 10.1161/jaha.117.007762]
7. Murray ML, Love SB, Carpenter JR, Hartley S, Landray MJ, Mafham M, et al. Data provenance and integrity of health-care systems data for clinical trials. The Lancet Digital Health. 2022 Aug;4(8):e567-e568 [doi: 10.1016/s2589-7500(22)00122-4]
8. Emanuel EJ, Emanuel LL. What is accountability in health care? Ann Intern Med. 1996 Jan 15;124(2):229-239 [doi: 10.7326/0003-4819-124-2-199601150-00007] [Medline: 8533999]
9. Curcin V. Embedding data provenance into the Learning Health System to facilitate reproducible research. Learn Health Syst. 2017 Apr 27;1(2):e10019 [FREE Full text] [doi: 10.1002/lrh2.10019] [Medline: 31245557]
10. Bongiovanni S, Purdue R, Kornienko O, Bernard R. Quality in Non-GxP Research Environment. In: Handb Exp Pharmacol. Cham, Switzerland. Springer International Publishing; 2020:1-17 [doi: 10.1007/164_2019_274]
11. Sahoo SS, Valdez J, Rueschman M. Scientific Reproducibility in Biomedical Research: Provenance Metadata Ontology for Semantic Annotation of Study Description. AMIA Annu Symp Proc. 2016;2016:1070-1079 [FREE Full text] [Medline: 28269904]
12. Bargaje C. Good documentation practice in clinical research. Perspect Clin Res. 2011 Apr;2(2):59-63 [FREE Full text] [doi: 10.4103/2229-3485.80368] [Medline: 21731856]
13. Guidelines for Safeguarding Good Research Practice. URL: https://wissenschaftliche-integritaet.de/en/code-of-conduct/ [accessed 2022-11-18]
14. Debruyne C, Pandit HJ, Lewis D, O'Sullivan D. "Just-in-time" generation of datasets by considering structured representations of given consent for GDPR compliance. Knowl Inf Syst. 2020 Apr 15;62(9):3615-3640 [FREE Full text] [doi: 10.1007/s10115-020-01468-x] [Medline: 32647404]
15. 14. ISO 20691. ISO 20691:2022, Biotechnology. URL: https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/06/88/68848.html [accessed 2022-11-18]
16. Standards by ISO/TC 276. URL: https://www.iso.org/committee/4514241/x/catalogue/ [accessed 2022-11-18]

17. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016 Mar 15;3(1):160018 [FREE Full text] [doi: 10.1038/sdata.2016.18] [Medline: 26978244]
18. Lamprecht A, Garcia L, Kuzak M, Martinez C, Arcila R, Martin Del Pico E, et al. Towards FAIR principles for research software. DS. 2020 Jun 12;3(1):37-59 [FREE Full text] [doi: 10.3233/DS-190026]
19. HL7 FHIR Foundation enabling healthcare interoperability through FHIR. URL: https://fhir.org/ [accessed 2023-02-10]
20. W3C PROV Overview. URL: https://www.w3.org/TR/prov-overview/ [accessed 2022-11-18]
21. Moreau L, Clifford B, Freire J, Futrelle J, Gil Y, Groth P, et al. The Open Provenance Model core specification (v1.1). Future Generation Computer Systems. 2011 Jun;27(6):743-756 [doi: 10.1016/j.future.2010.07.005]
22. Sikos LF, Philp D. Provenance-Aware Knowledge Representation: A Survey of Data Models and Contextualized Knowledge Graphs. Data Sci Eng. 2020 May 08;5(3):293-316 [doi: 10.1007/s41019-020-00118-0]
23. Gierend K, Freiesleben S, Kadioglu D, Siegel F, Ganslandt T, Waltemath D. The Status of Data Management Practices Across German Medical Data Integration Centers: Mixed Methods Study. J Med Internet Res. 2023 Nov 08;25:e48809 [FREE Full text] [doi: 10.2196/48809] [Medline: 37938878]
24. DeMarco T. Structure AnalysisSystem Specification. In: Broy M, Denert E. editors. Pioneers and Their Contributions to Software Engineering Berlin, Heidelberg. Springer Berlin Heidelberg; 1979:255
25. Bornberg-Bauer E, Paton NW. Conceptual data modelling for bioinformatics. Brief Bioinform. 2002 Jun 01;3(2):166-180 [doi: 10.1093/bib/3.2.166] [Medline: 12139436]
26. Lim C, Lu S, Chebotko A, Fotouhi F. Prospective and Retrospective Provenance Collection in Scientific Workflow Environments. In: ProspectiveRetrospective Provenance Collection in Scientific Workflow Environments IEEE International Conference on Services Computing Miami, FL. USA. IEEE; 2010 Presented at: Conference on Services Computing; 2010; Miami, FL, USA p. 449 [doi: 10.1109/SCC.2010.18]
27. Provenance in Data Integration Center, WebProtégé. URL: https://webprotege.stanford.edu/#login [accessed 2023-05-10]
28. Python. URL: https://www.python.org/ [accessed 2022-11-18]
29. 15. Peewee documentation. URL: https://docs.peewee-orm.com/en/latest/ [accessed 2022-11-18]
30. Woodward M. Include diagrams in your Markdown files with Mermaid. The GitHub Blog. URL: https://github.blog/2022-02-14-include-diagrams-markdown-files-mermaid/ [accessed 2022-12-02]
31. R: A language and environment for statistical computing. Vienna, Austria. R Foundation for Statistical Computing URL: https://www.R-project.org/ [accessed 2022-12-02]
32. GitHub: kegieKG/Provenance-in-Data-Integration-Center. URL: https://github.com/kegieKG/Provenance-in-Data-Integration-Center [accessed 2023-05-10]
33. Vorisek CN, Lehne M, Klopfenstein SAI, Mayer PJ, Bartschke A, Haese T, et al. Fast Healthcare Interoperability Resources (FHIR) for Interoperability in Health Research: Systematic Review. JMIR Med Inform. 2022 Jul 19;10(7):e35724 [FREE Full text] [doi: 10.2196/35724] [Medline: 35852842]
34. de Oliveira W, Braga R, David JMN, Stroele V, Campos F, Castro G. Visionary: a framework for analysis and visualization of provenance data. Knowl Inf Syst. 2022 Jan 04;64(2):381-413 [doi: 10.1007/s10115-021-01645-6]
35. Mitchell SN, Lahiff A, Cummings N, Hollocombe J, Boskamp B, Field R, et al. FAIR data pipeline: provenance-driven data management for traceable scientific workflows. Philos Trans A Math Phys Eng Sci. 2022 Oct 03;380(2233):20210300 [FREE Full text] [doi: 10.1098/rsta.2021.0300] [Medline: 35965468]
36. Mondelli M, Magalhães T, Loss G, Wilde M, Foster I, Mattoso M, et al. BioWorkbench: a high-performance framework for managing and analyzing bioinformatics experiments. PeerJ. 2018;6:e5551 [FREE Full text] [doi: 10.7717/peerj.5551] [Medline: 30186700]
37. Celebi R, Rebelo Moreira J, Hassan A, Ayyar S, Ridder L, Kuhn T, et al. Towards FAIR protocols and workflows: the OpenPREDICT use case. PeerJ Comput Sci. 2020;6:e281 [FREE Full text] [doi: 10.7717/peerj-cs.281] [Medline: 33816932]
38. Gonçalves RS, Musen MA. The variable quality of metadata about biological samples used in biomedical experiments. Sci Data. 2019 Feb 19;6(1):190021 [FREE Full text] [doi: 10.1038/sdata.2019.21] [Medline: 30778255]
39. Bönisch C, Kesztyüs D, Kesztyüs T. Harvesting metadata in clinical care: a crosswalk between FHIR, OMOP, CDISC and openEHR metadata. Sci Data. 2022 Oct 28;9(1):659 [FREE Full text] [doi: 10.1038/s41597-022-01792-7] [Medline: 36307424]
40. Kinast B, Ulrich H, Bergh B, Schreiweis B. Functional Requirements for Medical Data Integration into Knowledge Management Environments: Requirements Elicitation Approach Based on Systematic Literature Analysis. J Med Internet Res. 2023 Feb 09;25:e41344 [FREE Full text] [doi: 10.2196/41344] [Medline: 36757764]
41. Girman CJ, Ritchey ME, Lo Re V. Real-world data: Assessing electronic health records and medical claims data to support regulatory decision-making for drug and biological products. Pharmacoepidemiol Drug Saf. 2022 Jul 03;31(7):717-720 [FREE Full text] [doi: 10.1002/pds.5444] [Medline: 35471704]
42. Barker M, Chue Hong NP, Katz DS, Lamprecht A, Martinez-Ortiz C, Psomopoulos F, et al. Introducing the FAIR Principles for research software. Sci Data. 2022 Oct 14;9(1):622 [FREE Full text] [doi: 10.1038/s41597-022-01710-x] [Medline: 36241754]

43. Gierend K, Wodke J, Genehr S, Gött R, Henkel R, Krüger F, et al. TAPP: Defining standard provenance information for clinical research data and workflows - Obstacles and opportunities. Companion Proceedings of the ACM Web Conference 2023 Austin TX USA. Association for Computing Machinery, New York, NY, USA; 2023 Apr 30 Presented at: In Companion Proceedings of the ACM Web Conference 2023 (WWW '23 Companion); 2023-04-30; Austin, TX, USA p. 1551-1554 [doi: 10.1145/3543873.3587562]

Abbreviations

ALCOA: Attributable, Legible, Contemporaneous, Original, and Accurate
DIC: data integration center
ETL: extract, transform, and load
FAIR: Findability, Accessibility, Interoperability, and Reusability
FHIR: Fast Healthcare Interoperability Resources
GDPR: General Data Protection Regulation
MIRAPIE: Minimal Requirements for Automated Provenance Information Enrichment
MIT: Massachusetts Institute of Technology
PISA: Provenance Information System Traces
RDF: Resource Description Framework
UML: unified modeling language
W3C: Word Wide Web Consortium

Edited by A Mavragani; submitted 16.06.23; peer-reviewed by T Miksa; comments to author 05.10.23; revised version received 25.10.23; accepted 01.11.23; published 07.12.23

Please cite as:
Gierend K, Waltemath D, Ganslandt T, Siegel F
Traceable Research Data Sharing in a German Medical Data Integration Center With FAIR (Findability, Accessibility, Interoperability, and Reusability)-Geared Provenance Implementation: Proof-of-Concept Study
JMIR Form Res 2023;7:e50027
URL: https://formative.jmir.org/2023/1/e50027
doi: 10.2196/50027
PMID: 38060305

3.3 Publication 3: Approaches and Criteria for Provenance in Biomedical Data Sets and Workflows: Protocol for a Scoping Review

This section contains the scoping review protocol, as originally published in the Journal of Medical Internet Research (JMIR) Research Protocol, an international, peer-reviewed, and open access journal.
The original publication and appendices are available at JMIR (https://doi.org/10.2196/31750).

Protocol

Approaches and Criteria for Provenance in Biomedical Data Sets and Workflows: Protocol for a Scoping Review

Kerstin Gierend[1], Dipl Inf (FH); Frank Krüger[2], Dr Ing; Dagmar Waltemath[3], Prof Dr Ing; Maximilian Fünfgeld[1], Dr rer nat; Thomas Ganslandt[1], Prof Dr med; Atinkut Alamirrew Zeleke[3], Dr rer medic

[1]Department of Biomedical Informatics at the Center for Preventive Medicine and Digital Health, Medical Faculty Mannheim, Heidelberg University, Mannheim, Germany

[2]Department of Communications Engineering, University of Rostock, Rostock, Germany

[3]Department of Medical Informatics, Institute for Community Medicine, University Medicine Greifswald, Greifswald, Germany

Corresponding Author:
Kerstin Gierend, Dipl Inf (FH)
Department of Biomedical Informatics at the Center for Preventive Medicine and Digital Health
Medical Faculty Mannheim
Heidelberg University
Theodor-Kutzer-Ufer 1-3
Mannheim, 68167
Germany
Phone: 49 0621 383 ext 8087
Email: kerstin.gierend@medma.uni-heidelberg.de

Abstract

Background: Provenance supports the understanding of data genesis, and it is a key factor to ensure the trustworthiness of digital objects containing (sensitive) scientific data. Provenance information contributes to a better understanding of scientific results and fosters collaboration on existing data as well as data sharing. This encompasses defining comprehensive concepts and standards for transparency and traceability, reproducibility, validity, and quality assurance during clinical and scientific data workflows and research.

Objective: The aim of this scoping review is to investigate existing evidence regarding approaches and criteria for provenance tracking as well as disclosing current knowledge gaps in the biomedical domain. This review covers modeling aspects as well as metadata frameworks for meaningful and usable provenance information during creation, collection, and processing of (sensitive) scientific biomedical data. This review also covers the examination of quality aspects of provenance criteria.

Methods: This scoping review will follow the methodological framework by Arksey and O'Malley. Relevant publications will be obtained by querying PubMed and Web of Science. All papers in English language will be included, published between January 1, 2006 and March 23, 2021. Data retrieval will be accompanied by manual search for grey literature. Potential publications will then be exported into a reference management software, and duplicates will be removed. Afterwards, the obtained set of papers will be transferred into a systematic review management tool. All publications will be screened, extracted, and analyzed: title and abstract screening will be carried out by 4 independent reviewers. Majority vote is required for consent to eligibility of papers based on the defined inclusion and exclusion criteria. Full-text reading will be performed independently by 2 reviewers and in the last step, key information will be extracted on a pretested template. If agreement cannot be reached, the conflict will be resolved by a domain expert. Charted data will be analyzed by categorizing and summarizing the individual data items based on the research questions. Tabular or graphical overviews will be given, if applicable.

Results: The reporting follows the extension of the Preferred Reporting Items for Systematic reviews and Meta-Analyses statements for Scoping Reviews. Electronic database searches in PubMed and Web of Science resulted in 469 matches after deduplication. As of September 2021, the scoping review is in the full-text screening stage. The data extraction using the pretested charting template will follow the full-text screening stage. We expect the scoping review report to be completed by February 2022.

Conclusions: Information about the origin of healthcare data has a major impact on the quality and the reusability of scientific results as well as follow-up activities. This protocol outlines plans for a scoping review that will provide information about current approaches, challenges, or knowledge gaps with provenance tracking in biomedical sciences.

International Registered Report Identifier (IRRID): DERR1-10.2196/31750

(JMIR Res Protoc 2021;10(11):e31750) doi: 10.2196/31750

KEYWORDS

provenance; biomedical; workflow; data sharing; lineage; scoping review; data genesis; scientific data; digital objects; healthcare data

Introduction

The (re-)use of electronic medical and patient-related data offers enormous potential for further investigations in clinical research [1,2]. Different national initiatives such as the French Health Data Hub initiative or the German Medical Informatics Initiatives are committed to better knowledge discovery and data sharing in the health care domain [3]. Resulting outcomes enable patients and physicians a safe and rapid access to therapies or treatment options. Subsequently, treatment costs can be reduced. In this context, the access to quality-assured, traceable, and hence, credible shared data is essential. Providing information about the origin of data demands concepts for traceability to gain understanding for the relationships between results and source data. There is an increasing interest and need to ensure traceability throughout scientific practice. Consequently, a systematic knowledge compilation regarding provenance and potential gaps is needed.

Provenance describes the origin of data. A basic understanding of the term “provenance” is given with the description “what happened” to the data [4]. Several different models exist to formally express provenance information, for instance, the World Wide Web Consortium PROV standard or CWLProv [5,6]. Advantages and opportunities of providing data provenance have been demonstrated, for instance, from the experiences in the EU-Horizon 2020 TRANSFoRm project [4]. Moreover, the importance of provenance and the relation to provenance within electronic health records is pointed out in the study of Johnson et al [7]. A previously published systematic review of provenance systems already investigated tools and systems [8]. However, our own work aims to understand current approaches and criteria as well as knowledge gaps for provenance in biomedical as well as domain-independent research.

The fields of research data management and FAIR (findable-accessible-interoperable-reusable) data principles consider provenance as one of the research pillars [9]. As such, a provenance-oriented approach requires thorough planning, execution, and evaluation of data management processes in the respective application domain [1]. While capturing provenance information in the research, adherence to criteria such as consistency, interoperability, and confidentiality are required across all software tools [2]. Furthermore, data privacy issues have to be respected during modeling to keep compliance with national and international requirements such as the European General Data Protection Regulation [10,11].

Process quality with the associated workflow quality can be achieved by monitoring and troubleshooting in applications or in data integration scenarios such as Extract-Transform-Load jobs. This implies workflow requirements to be established on a fine- or coarse-grained provenance level for troubleshooting [12]. Addressing data quality issues should support in reaching completeness, accuracy, and timeliness of the data and creates trust in it. However, heterogeneous data sources, dynamic infrastructures, data exchange across boundaries, and lack of standards for quality measures characterize the current state of electronic health record data sets [13]. Contrarily, provenance information strengthens the credibility of the data and proves that data have not been intentionally or unintentionally changed in its life cycle [14]. The concept and implementation of provenance is essential in most scientific domains such as environmental fields (geoprocessing workflows or climate assessments), in fusion engineering, or material sciences [15,16]. Since the use of machine learning techniques within the scope of decision support is becoming increasingly popular for medical researchers, they are under the obligation to prove their reproducibility [17]. Therefore, systematic knowledge about the “what happened” and about reproducibility metrics such as data sets and code accessibility is indispensable and is in need of further investigation to provide provenance [18].

The aim of this scoping review is to investigate existing evidence regarding approaches and criteria for provenance tracking as well as disclosing current knowledge gaps in the biomedical domain. This comprises modeling aspects as well as metadata frameworks for meaningful and usable provenance information during creation, collection, and processing of (sensitive) scientific biomedical data. The review also covers the examination of quality aspects of provenance criteria.

Methods

Design

The individual elements from the framework of Arksey and O’Malley [19] will be used as a roadmap for this scoping review. Essential methodological steps will cover the stages (1) identification of the research questions, (2) identification of relevant studies, (3) study selection, (4) data extraction and charting, and (5) collating, summarizing, and reporting the results. Any subsequent deviations of the final report from the scoping review protocol will be clearly highlighted and explained in the scoping review report.

Ethics

Ethical approval was not required because only literature will be evaluated without processing sensitive patient data.

Stage 1: Identification of the Research Questions

At first, an informal prescreening of relevant literature in PubMed and Web of Science as well as grey literature from conferences or organizations was carried out to determine the keywords in scope. Relevant literature was identified with the support of a librarian. PubMed was searched using the keywords “provenance” and “tracking.” The reviewer team explored, studied, and scrutinized additional literature based on search combinations of terms linked to the topic “provenance.” Ten publications were selected and reviewed by the team in an iterative process to guide the implementation of the research questions. During this step, keywords from titles and abstracts were gathered and analyzed by implementing the search strategy based on them. The following research questions were generated to meet the objective of this scoping review before study conduction: to investigate existing evidence regarding approaches and criteria for provenance tracking as well as disclosing current knowledge gaps in the biomedical domain. This review covers modeling aspects as well as metadata frameworks for meaningful and usable provenance information during creation, collection, and processing of (sensitive) scientific biomedical data. This review also covers the examination of quality aspects of provenance criteria.

Research question 1: Which potential (methodological) approaches exist for the classification and tracking of provenance criteria and methods in a biomedical or domain-independent context?

Research question 2: How can the potential value of provenance information be harnessed and by whom? How can usability be provided?

Research question 3: What are the challenges and potential problems or bottlenecks for the accomplishment of provenance?

Research question 4: Which guidelines or demands for the consideration of provenance criteria in a biomedical or domain-independent context have to be followed?

Research question 5: How completely can provenance be mapped in the data lifecycle or during data management?

Stage 2: Identification of Relevant Studies

Relevant publications will be retrieved using concepts together with their associated keywords as selected from “Stage 1: Identification of the research questions.” Concepts are categorized into 4 groups: target domain, provenance, provenance properties, and objective. Target domain refers to the context of the research topic and includes studies with a biomedical, health care, clinical, or scientific background. Scientific background is limited to domain-independent studies and excludes all other domain-specific studies. The concept “provenance” concerns the information about the genesis of a given object while the concept “provenance properties” covers specific requirements tied to the term “provenance” or describes selected characteristics in this context. The concept “objective” embraces the range of purpose or the intention of provenance. Table 1 provides an overview of the eligibility criteria derived from the categorization of the concepts together with the defined terms and their matching keywords.

Table 1. Concepts and matching keywords (eligibility criteria).

Concepts	Matching keywords (inclusion criteria)
Target domain	biomed*[a], EHR, electronic health record, healthcare, clinical, scientific[b]
Provenance	provenance, prov, lineage
Provenance properties	interop*, (data NEAR/2 [flow, quality, transformation]), metadata, workflow, semantic, framework, annotat*, ontolog*, management, document*, (model NEAR/2 provenance)
Objective	audit*, decision support, ETL, Extract-Transform-Load, FHIR, record linking, machine learning, reproducib*, transparen*, track*, implement*

[a] The * symbol (wildcard character) replaces or represents one or more characters.

[b] Will be used in a domain-independent context only.

A comprehensive search strategy for identifying the relevant literature, based on the given table, was implemented in PubMed and Web of Science. Medical subject headings were applied in PubMed. Additionally, the Boolean operators AND OR were used within the search strategy for combining the individual concepts and their associated keywords.

The inclusion criteria comprised all papers in the English language and published between January 1, 2006 and March 23, 2021. The concepts and their related keywords, as shown in Table 1, are considered during the selection of the papers within the biomedical or domain-independent area. The start date for inclusion of literature was chosen owing to the initiation of the Open Provenance Model in 2006 as a result of the Provenance Challenge series [20]. Grey literature from relevant project reports and proceedings were searched and reviewed for eligibility. All search results were exported to a reference management tool to eliminate duplications. Unique results were exported to the web-based screening tool Rayyan (Qatar Computing Research Institute) [21]. The PRISMA-ScR (Preferred Reporting Items for Systematic reviews and Meta-analyses extension for Scoping Reviews) will be used for reporting of this scoping review [22].

Stage 3: Study Selection

During the scoping review process, decisions to select or eliminate studies are tracked using Rayyan. That way, independent screening by the reviewers is enabled. Rayyan allows citation sharing and blinded comparison of decisions for inclusion and exclusion of selected studies. All imported publications will be screened by reading the title and abstract by all 4 reviewers. Title-abstract screening is the process of reviewing the references for inclusion based solely upon their title and abstract. Reviewers will screen out irrelevant references whereby the inclusion and exclusion criteria serve as the basis for their eligibility decision. Conflicts will be resolved since at least 3 unified classifications are necessary for inclusion or exclusion of a publication in an unblinded modus. The included (=eligible) publications will be examined in a full-text screening phase to determine the extent to which they can answer the research questions. Each publication must be read by 2 researchers to determine the relevance to the research questions. If there is no joint agreement, an independent researcher will be consulted. A description and a PRISMA flow chart of the selection process with frequencies for references considered in the different databases will be provided as well as counting in the subsequent title-abstract screening process based on the eligibility criteria.

Stage 4: Data Extraction and Charting

The data collection process will be documented by the reviewers while using the collectively developed template as provided in Table 2. The approach to data extraction needs to be consistent with the research question and purpose. This charting form will be pretested and will be used after closed alignment between the reviewers. "Pretested" means that 2 reviewers will independently complete the template for 5 studies ahead of the main study. They will compare the result with regard to a consistent approach and agree on necessary updates in the template, if necessary. Reviewers will diligently extract and update the study data from the identified papers in scope during their full-text review in an iterative process.

Table 2. Data charting template for key information from eligible papers.

Metadata publication	Characteristic extraction and specification
Title[a]	Title
Citation details[a]	Author (1st), journal, DOI
Year of publication[a]	For example, YYYY
Publication type[a]	Journal or website or conference, etc
Study type[a]	Use case or development or evaluation
Continent of study	For example, Australia
Institute[a]	Contributing institute (corresponding author or—if not provided—1st author)
Corresponding author's discipline	For example, data architect
Funding source	Public or industry or none or missing
Objective[a]	Aim of the publication
Methods	Strategies, processes, or techniques utilized in the collection or analyzing of data, how is the validity of the study judged
Summary results[a]	Short description of results
Conclusion	Short description of conclusion
Target domain[a]	Name specific domain or domain independent
Keywords	List keywords from abstract
Metadata to key findings related to research questions	Characteristic extraction and specification
Research question 1: Approaches for classification and tracking of provenance criteria and methods in biomedical or domain-independent context	Provide description in the domain for data suitability or data availability and other requirements or factors on data or systems regarding the trace of the data history (eg, role of provenance in terms of domain standards, ie, interoperability standards, FAIR [findable-accessible-interoperable-reusable] data, relation to metadata and model use, representation formalisms, etc), check definition of provenance
Research question 2: Potential value of provenance information	Provide possible use case description and types of data sources included, usability including effect on target domain and by whom it can be used and who will be the stakeholders; problems, if provenance is not available
Research question 3: Potential problems or bottlenecks for the accomplishment of provenance	Describe any challenges (eg, legal, organizational, or technical conditions) or problems that occurred during implementation phase of provenance
Research question 4: Guidelines or demands for the consideration of provenance to be adhered to	Describe any valid domain standard requirement, for example, legal, guidelines, rules
Research question 5: Completeness of provenance information during data management process or data life cycle	Describe any measurement or outcome available for completeness of provenance information

[a]Obligatory input.

Stage 5: Collating, Summarizing, and Reporting the Results

The charting results from stage 4 will be presented in the following steps [19]. Analysis will be given by a qualitative evaluation and by summary statistics, charts, or equivalent appraisal. The reporting of the results and outcome will be aligned to the research questions. The meaning of the findings and their relation to the overall objectives will be discussed. Implications for future research, practice, and policy will be outlined. The reporting of the results will be aligned with the PRISMA-ScR reporting guidelines [22].

Results

Schedule

The scoping review started with a tentative search of the databases in PubMed and Web of Science in early 2021 (see stages 1-3) and resulted in 469 matches. These papers will be subjected to title-abstract screening in an interactive selection process for eligibility, followed by a full-text screening stage. These papers will be examined within an iterative selection process for inclusion into data charting (see stage 4). Data extraction will be finalized during the 4th quarter of 2021. The scoping review will be completed by summarizing and synthesizing the results by February 2022 (see stage 5).

Anticipated Outcomes

The scoping review will identify potentially relevant initiatives on provenance, and it will provide an overview of the evidence, gaps, and limitations for provenance criteria. All the evidence will be elaborated on the basis of the research questions. As such, the review can serve as preparatory work for achieving a comprehensive usable result on approaches and criteria for provenance. Based on the review results, the quality of the provenance criteria will be examined for a potential demarcation regarding minimum requirements for structuredness and completeness of provenance. We believe that this investigation supports provenance research with respect to the implementation of provenance in secondary use projects such as the German Medical Informatics Initiative. Within the Medical Informatics in Research and Care in University Medicine consortium, as part of the Medical Informatics Initiative, provenance has an important meaning to bioinformaticians and researchers [23].

Discussion

Implications for future work will be derived from the current status of research activities and their underlying concepts. We anticipate that implications will encompass conceptual and modeling approaches up to the generation of provenance-aware data as well as gaps in the current practices within the health care domain. We believe that our results will support the further development of guidelines, thereby overcoming the identified challenges and disclosing new opportunities for the classification and tracking of provenance criteria. Evidence will assist in recognizing and defining the preconditions for data sharing. It will further characterize data suitability and categories (eg, data governance, relevance, quality) at a fitness for purpose level in the health domain, considering the interests of different stakeholders. Finally, the scoping review will provide insights into whether a further assessment of the results is useful within a full systematic review.

Acknowledgments

This research is funded by the German Federal Ministry of Education and Research within the German Medical Informatics Initiative with the grant 01ZZ1801E (Medical Informatics in Research and Care in University Medicine), by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) SFB 1270/1-99150580, and by the National Research Data Infrastructure for Personal Health Data (NFDI4Health) DFG-funded project (Project 442326535).

Conflicts of Interest

None declared.

References

1. Jayapandian CP, Zhao M, Ewing RM, Zhang G, Sahoo SS. A semantic proteomics dashboard (SemPoD) for data management in translational research. BMC Syst Biol 2012;6 Suppl 3:S20 [FREE Full text] [doi: 10.1186/1752-0509-6-S3-S20] [Medline: 23282161]
2. Curcin V, Miles S, Danger R, Chen Y, Bache R, Taweel A. Implementing interoperable provenance in biomedical research. Future Generation Computer Systems 2014 May;34:1-16. [doi: 10.1016/j.future.2013.12.001]
3. Cuggia M, Combes S. The French Health Data Hub and the German Medical Informatics Initiatives: Two national projects to promote data sharing in healthcare. Yearb Med Inform 2019 Aug;28(1):195-202 [FREE Full text] [doi: 10.1055/s-0039-1677917] [Medline: 31419832]
4. Curcin V. Embedding data provenance into the Learning Health System to facilitate reproducible research. Learn Health Syst 2017 Apr;1(2):e10019 [FREE Full text] [doi: 10.1002/lrh2.10019] [Medline: 31245557]
5. Groth P, Moreau L. PROV-overview. W3C. URL: https://www.w3.org/TR/prov-overview/ [accessed 2021-06-10]
6. Khan FZ, Soiland-Reyes S, Sinnott RO, Lonie A, Goble C, Crusoe MR. Sharing interoperable workflow provenance: A review of best practices and their practical application in CWLProv. Gigascience 2019 Nov 01;8(11):1-27 [FREE Full text] [doi: 10.1093/gigascience/giz095] [Medline: 31675414]

7. Johnson KE, Kamineni A, Fuller S, Olmstead D, Wernli KJ. How the provenance of electronic health record data matters for research: a case example using system mapping. EGEMS (Wash DC) 2014;2(1):1058 [FREE Full text] [doi: 10.13063/2327-9214.1058] [Medline: 25821838]
8. Pérez B, Rubio J, Sáenz-Adán C. A systematic review of provenance systems. Knowl Inf Syst 2018 Feb 17;57(3):495-543. [doi: 10.1007/s10115-018-1164-3]
9. Jauer M, Deserno T. Data provenance standards and recommendations for FAIR data. Stud Health Technol Inform 2020 Jun 16;270:1237-1238. [doi: 10.3233/SHTI200380] [Medline: 32570597]
10. Hume S, Sarnikar S, Noteboom C. Enhancing traceability in clinical research data through a metadata framework. Methods Inf Med 2020 May;59(2-03):75-85. [doi: 10.1055/s-0040-1714393] [Medline: 32894879]
11. Sahoo SS, Nguyen V, Bodenreider O, Parikh P, Minning T, Sheth AP. A unified framework for managing provenance information in translational research. BMC Bioinformatics 2011 Nov 29;12:461 [FREE Full text] [doi: 10.1186/1471-2105-12-461] [Medline: 22126369]
12. Zheng N, Alawini A, Ives Z. 2019 Apr Presented at: 35th International Conference on Data Engineering (ICDE); 2019; Macao, China p. 184-195 URL: http://europepmc.org/abstract/MED/31595143 [doi: 10.1109/ICDE.2019.00025]
13. Margheri A, Masi M, Miladi A, Sassone V, Rosenzweig J. Decentralised provenance for healthcare data. Int J Med Inform 2020 Sep;141:104197. [doi: 10.1016/j.ijmedinf.2020.104197] [Medline: 32540775]
14. Wing JM. The data life cycle. Harvard Data Science Review 2019 Jun 23:1-6. [doi: 10.1162/99608f92.e26845b4]
15. Schissel D, Abla G, Flanagan S, Greenwald M, Lee X, Romosan A, et al. Automated metadata, provenance cataloging and navigable interfaces: Ensuring the usefulness of extreme-scale data. Fusion Engineering and Design 2014 May;89(5):745-749. [doi: 10.1016/j.fusengdes.2014.01.053]
16. Yakutovich A, Eimre K, Schütt O, Talirz L, Adorf C, Andersen C, et al. AiiDAlab – an ecosystem for developing, executing, and sharing scientific workflows. Computational Materials Science 2021 Feb;188:110165 [FREE Full text] [doi: 10.1016/j.commatsci.2020.110165]
17. Samuel S, Löffler F, König-Ries B. Machine learning pipelines: provenance, reproducibility and FAIR data principles. arXiv.org. URL: http://arxiv.org/abs/2006.12117 [accessed 2021-05-16]
18. McDermott MBA, Wang S, Marinsek N, Ranganath R, Foschini L, Ghassemi M. Reproducibility in machine learning for health research: Still a ways to go. Sci Transl Med 2021 Mar 24;13(586):eabb1655. [doi: 10.1126/scitranslmed.abb1655] [Medline: 33762434]
19. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. International Journal of Social Research Methodology 2005 Feb;8(1):19-32. [doi: 10.1080/1364557032000119616]
20. Moreau L, Ludäscher B, Altintas I, Barga R, Bowers S, Callahan S, et al. Special issue: the first Provenance Challenge. Concurrency Computat.: Pract. Exper 2008 Apr 10;20(5):409-418. [doi: 10.1002/cpe.1233]
21. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. Syst Rev 2016 Dec 05;5(1):210 [FREE Full text] [doi: 10.1186/s13643-016-0384-4] [Medline: 27919275]
22. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 02;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
23. Pugliese P, Knell C, Christoph J. Exchange of clinical and omics data according to FAIR principles: a review of open source solutions. Methods Inf Med 2020 Jun;59(S 01):e13-e20 [FREE Full text] [doi: 10.1055/s-0040-1712968] [Medline: 32620018]

Abbreviations

PRISMA-ScR: Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews

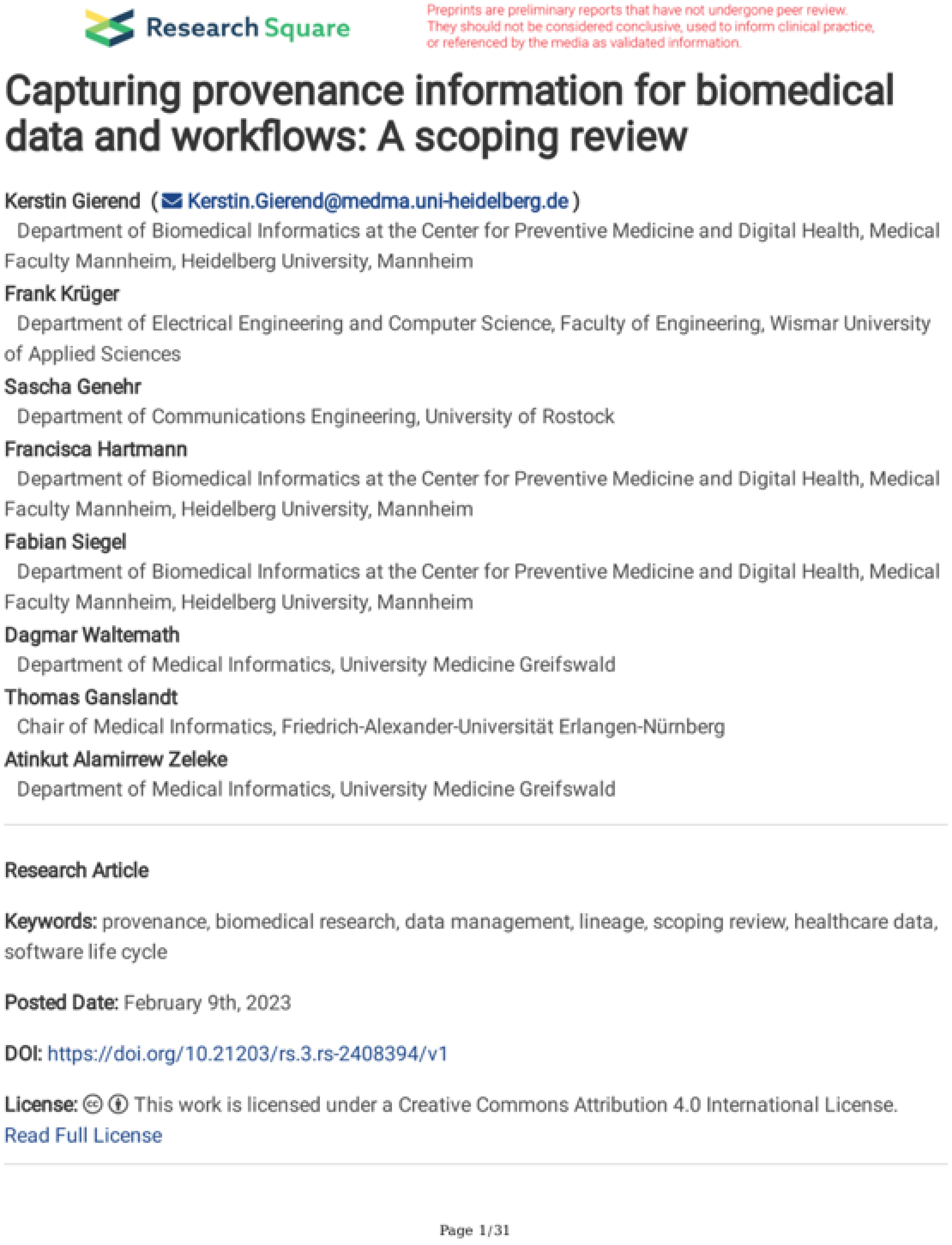
Edited by G Eysenbach; submitted 02.07.21; peer-reviewed by V Curcin, T Miksa; comments to author 11.08.21; revised version received 06.09.21; accepted 07.09.21; published 22.11.21

Please cite as:
Gierend K, Krüger F, Waltemath D, Fünfgeld M, Ganslandt T, Zeleke AA
Approaches and Criteria for Provenance in Biomedical Data Sets and Workflows: Protocol for a Scoping Review
JMIR Res Protoc 2021;10(11):e31750
URL: https://www.researchprotocols.org/2021/11/e31750
doi: 10.2196/31750
PMID:

3.4 Publication 4: Capturing provenance information for biomedical data and workflows: A scoping review (pre-print)

This section contains the scoping review, available as pre-print at research square. The manuscript is in peer-review at the Journal of Medical Internet Research (JMIR).

(https://doi.org/10.21203/rs.3.rs-2408394/v1)



Capturing provenance information for biomedical data and workflows: A scoping review

Kerstin Gierend (Kerstin.Gierend@medma.uni-heidelberg.de)
Department of Biomedical Informatics at the Center for Preventive Medicine and Digital Health, Medical Faculty Mannheim, Heidelberg University, Mannheim

Frank Krüger
Department of Electrical Engineering and Computer Science, Faculty of Engineering, Wismar University of Applied Sciences

Sascha Genehr
Department of Communications Engineering, University of Rostock

Francisca Hartmann
Department of Biomedical Informatics at the Center for Preventive Medicine and Digital Health, Medical Faculty Mannheim, Heidelberg University, Mannheim

Fabian Siegel
Department of Biomedical Informatics at the Center for Preventive Medicine and Digital Health, Medical Faculty Mannheim, Heidelberg University, Mannheim

Dagmar Waltemath
Department of Medical Informatics, University Medicine Greifswald

Thomas Ganslandt
Chair of Medical Informatics, Friedrich-Alexander-Universität Erlangen-Nürnberg

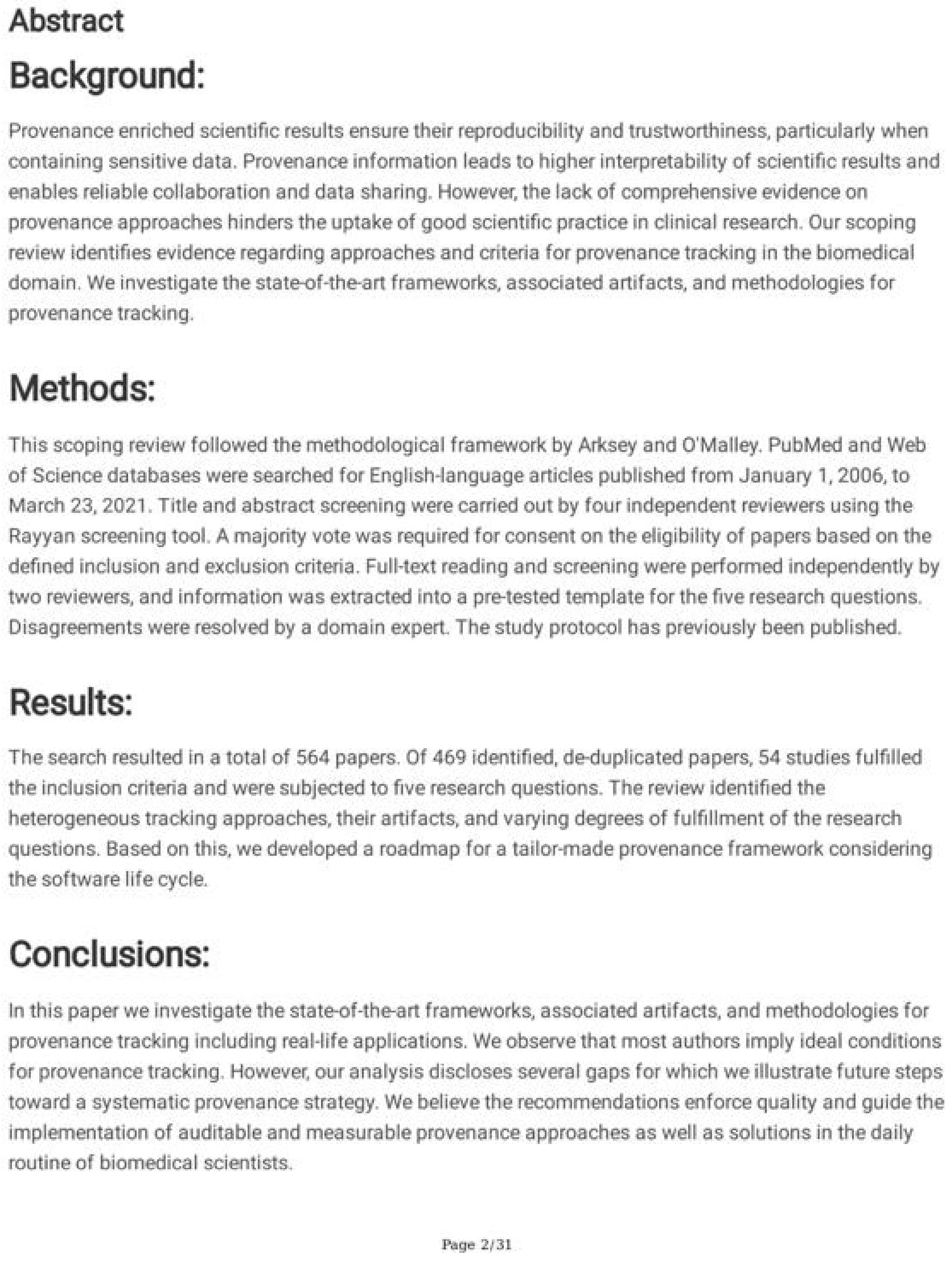
Atinkut Alamirrew Zeleke
Department of Medical Informatics, University Medicine Greifswald

Research Article

Keywords: provenance, biomedical research, data management, lineage, scoping review, healthcare data, software life cycle

Posted Date: February 9th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2408394/v1



Abstract

Background:

Provenance enriched scientific results ensure their reproducibility and trustworthiness, particularly when containing sensitive data. Provenance information leads to higher interpretability of scientific results and enables reliable collaboration and data sharing. However, the lack of comprehensive evidence on provenance approaches hinders the uptake of good scientific practice in clinical research. Our scoping review identifies evidence regarding approaches and criteria for provenance tracking in the biomedical domain. We investigate the state-of-the-art frameworks, associated artifacts, and methodologies for provenance tracking.

Methods:

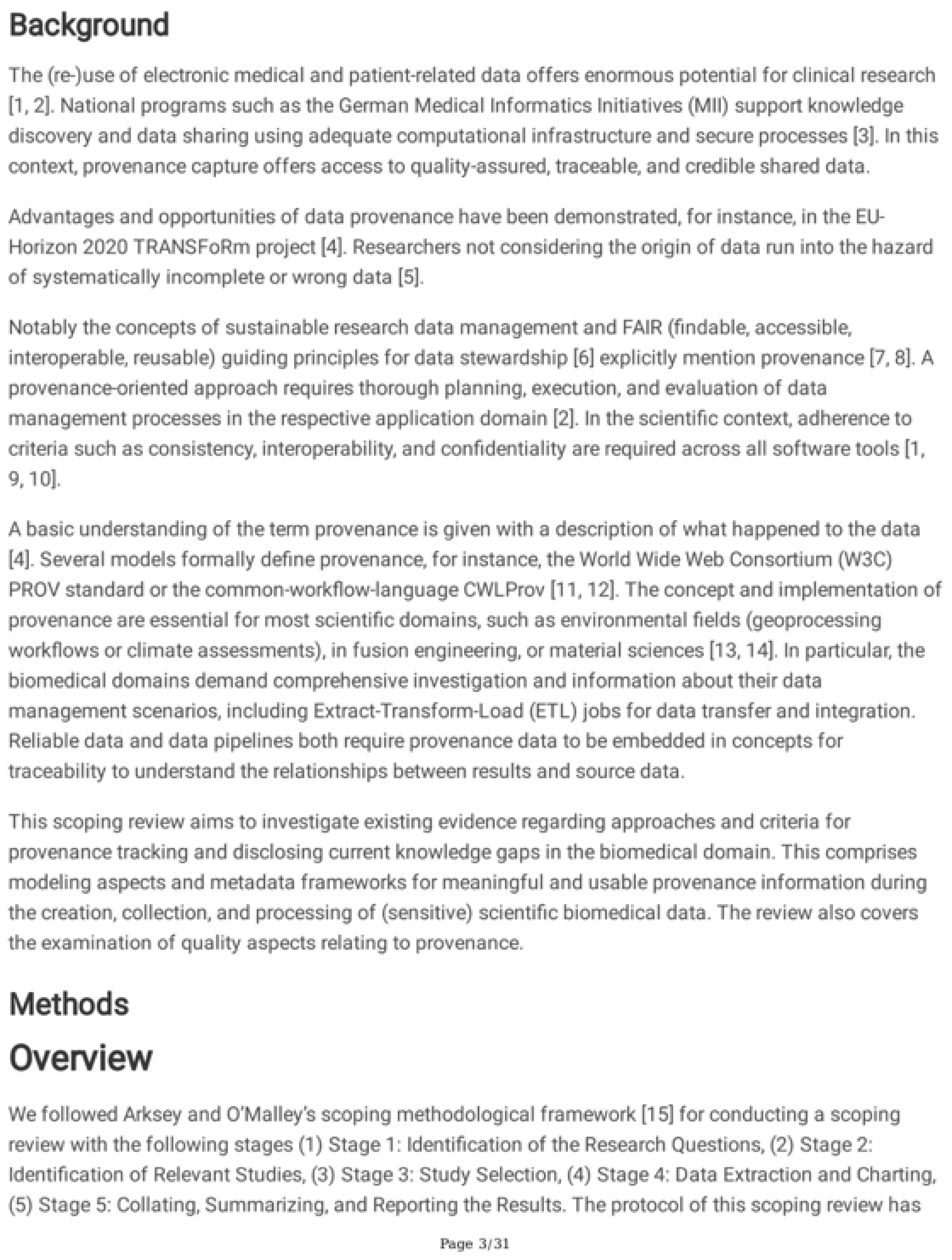
This scoping review followed the methodological framework by Arksey and O'Malley. PubMed and Web of Science databases were searched for English-language articles published from January 1, 2006, to March 23, 2021. Title and abstract screening were carried out by four independent reviewers using the Rayyan screening tool. A majority vote was required for consent on the eligibility of papers based on the defined inclusion and exclusion criteria. Full-text reading and screening were performed independently by two reviewers, and information was extracted into a pre-tested template for the five research questions. Disagreements were resolved by a domain expert. The study protocol has previously been published.

Results:

The search resulted in a total of 564 papers. Of 469 identified, de-duplicated papers, 54 studies fulfilled the inclusion criteria and were subjected to five research questions. The review identified the heterogeneous tracking approaches, their artifacts, and varying degrees of fulfillment of the research questions. Based on this, we developed a roadmap for a tailor-made provenance framework considering the software life cycle.

Conclusions:

In this paper we investigate the state-of-the-art frameworks, associated artifacts, and methodologies for provenance tracking including real-life applications. We observe that most authors imply ideal conditions for provenance tracking. However, our analysis discloses several gaps for which we illustrate future steps toward a systematic provenance strategy. We believe the recommendations enforce quality and guide the implementation of auditable and measurable provenance approaches as well as solutions in the daily routine of biomedical scientists.



Background

The (re-)use of electronic medical and patient-related data offers enormous potential for clinical research [1, 2]. National programs such as the German Medical Informatics Initiatives (MII) support knowledge discovery and data sharing using adequate computational infrastructure and secure processes [3]. In this context, provenance capture offers access to quality-assured, traceable, and credible shared data.

Advantages and opportunities of data provenance have been demonstrated, for instance, in the EU-Horizon 2020 TRANSFoRm project [4]. Researchers not considering the origin of data run into the hazard of systematically incomplete or wrong data [5].

Notably the concepts of sustainable research data management and FAIR (findable, accessible, interoperable, reusable) guiding principles for data stewardship [6] explicitly mention provenance [7, 8]. A provenance-oriented approach requires thorough planning, execution, and evaluation of data management processes in the respective application domain [2]. In the scientific context, adherence to criteria such as consistency, interoperability, and confidentiality are required across all software tools [1, 9, 10].

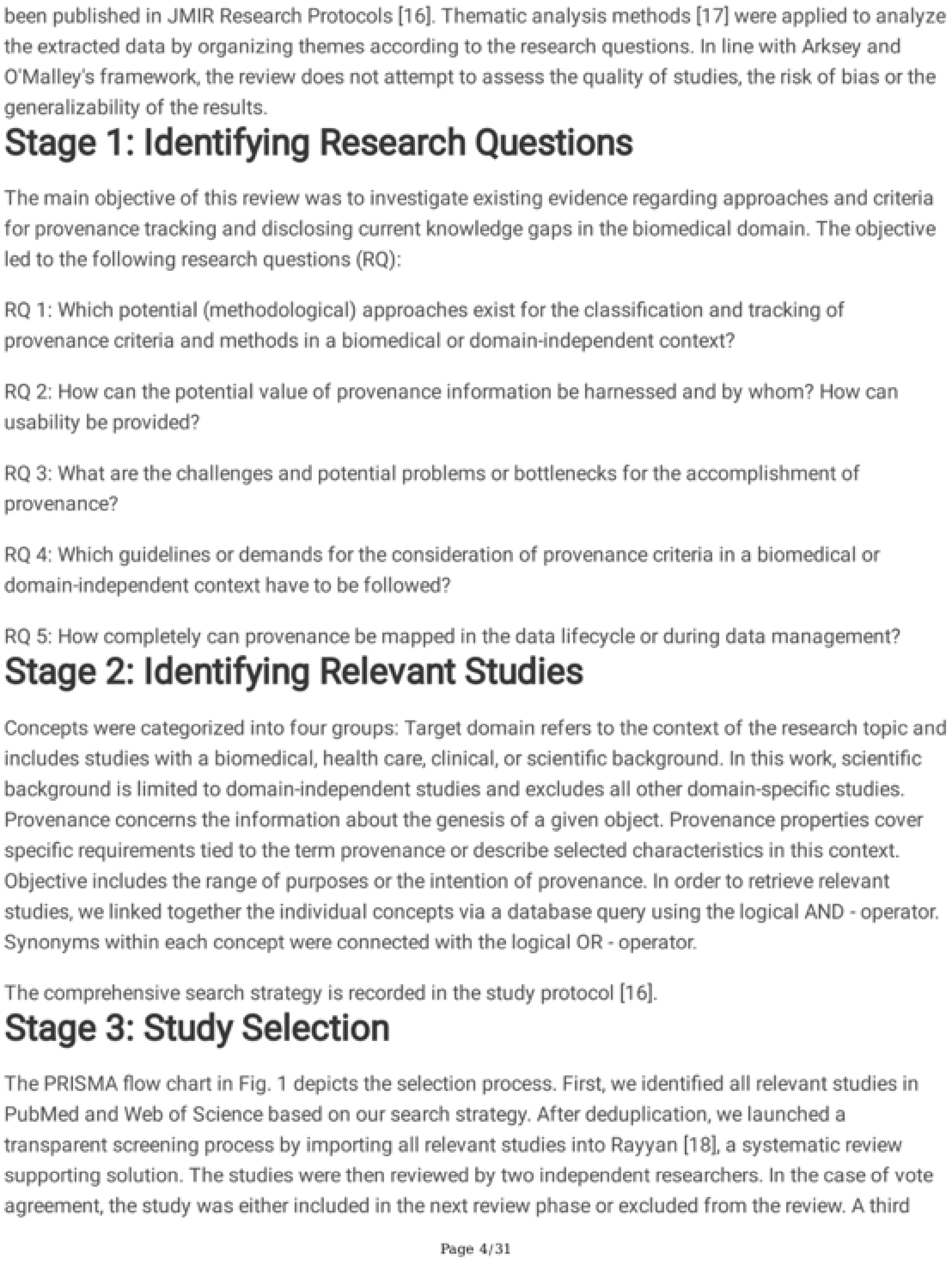
A basic understanding of the term provenance is given with a description of what happened to the data [4]. Several models formally define provenance, for instance, the World Wide Web Consortium (W3C) PROV standard or the common-workflow-language CWLProv [11, 12]. The concept and implementation of provenance are essential for most scientific domains, such as environmental fields (geoprocessing workflows or climate assessments), in fusion engineering, or material sciences [13, 14]. In particular, the biomedical domains demand comprehensive investigation and information about their data management scenarios, including Extract-Transform-Load (ETL) jobs for data transfer and integration. Reliable data and data pipelines both require provenance data to be embedded in concepts for traceability to understand the relationships between results and source data.

This scoping review aims to investigate existing evidence regarding approaches and criteria for provenance tracking and disclosing current knowledge gaps in the biomedical domain. This comprises modeling aspects and metadata frameworks for meaningful and usable provenance information during the creation, collection, and processing of (sensitive) scientific biomedical data. The review also covers the examination of quality aspects relating to provenance.

Methods

Overview

We followed Arksey and O'Malley's scoping methodological framework [15] for conducting a scoping review with the following stages (1) Stage 1: Identification of the Research Questions, (2) Stage 2: Identification of Relevant Studies, (3) Stage 3: Study Selection, (4) Stage 4: Data Extraction and Charting, (5) Stage 5: Collating, Summarizing, and Reporting the Results. The protocol of this scoping review has



been published in JMIR Research Protocols [16]. Thematic analysis methods [17] were applied to analyze the extracted data by organizing themes according to the research questions. In line with Arksey and O'Malley's framework, the review does not attempt to assess the quality of studies, the risk of bias or the generalizability of the results.

Stage 1: Identifying Research Questions

The main objective of this review was to investigate existing evidence regarding approaches and criteria for provenance tracking and disclosing current knowledge gaps in the biomedical domain. The objective led to the following research questions (RQ):

RQ 1: Which potential (methodological) approaches exist for the classification and tracking of provenance criteria and methods in a biomedical or domain-independent context?

RQ 2: How can the potential value of provenance information be harnessed and by whom? How can usability be provided?

RQ 3: What are the challenges and potential problems or bottlenecks for the accomplishment of provenance?

RQ 4: Which guidelines or demands for the consideration of provenance criteria in a biomedical or domain-independent context have to be followed?

RQ 5: How completely can provenance be mapped in the data lifecycle or during data management?

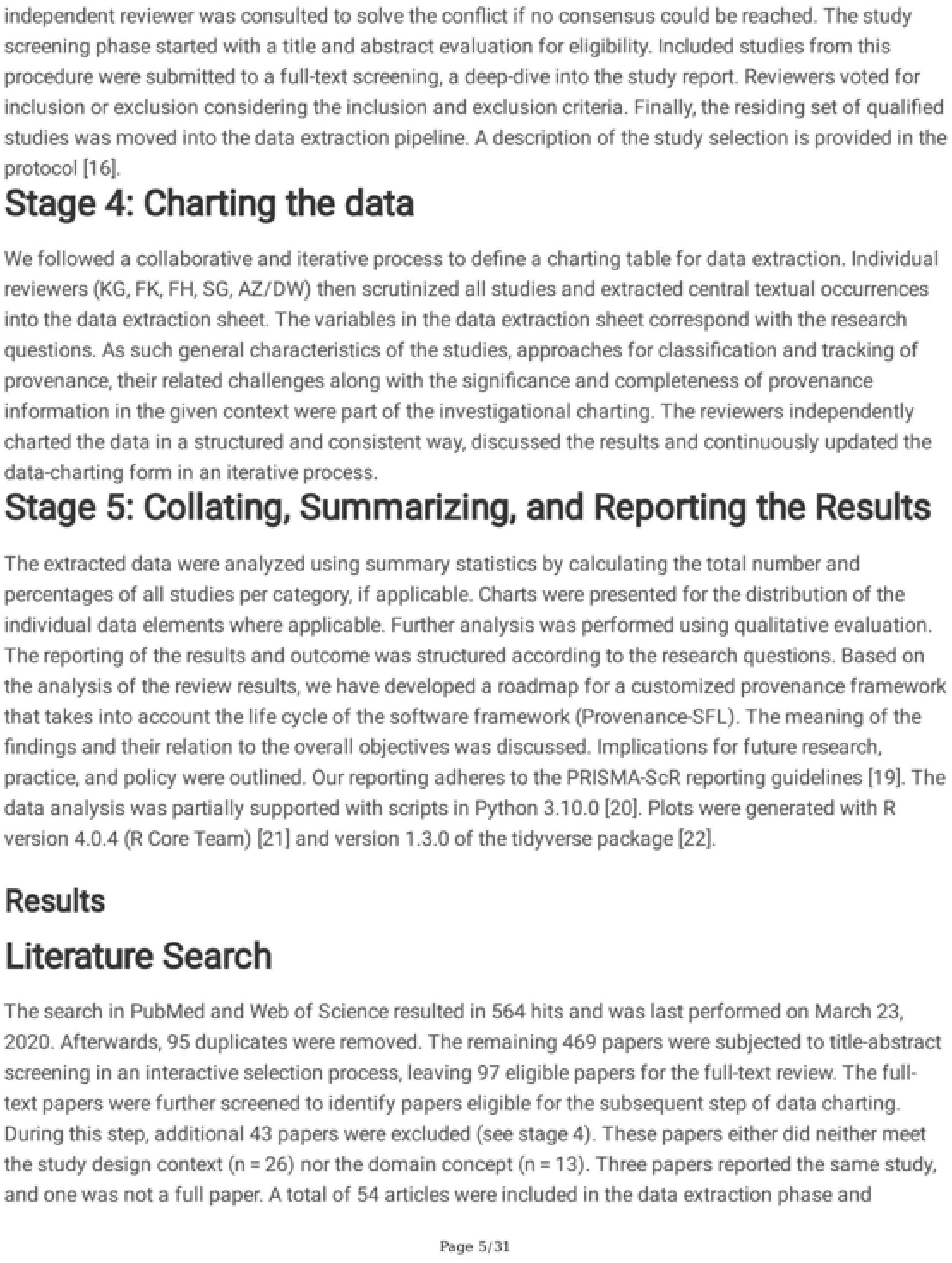
Stage 2: Identifying Relevant Studies

Concepts were categorized into four groups: Target domain refers to the context of the research topic and includes studies with a biomedical, health care, clinical, or scientific background. In this work, scientific background is limited to domain-independent studies and excludes all other domain-specific studies. Provenance concerns the information about the genesis of a given object. Provenance properties cover specific requirements tied to the term provenance or describe selected characteristics in this context. Objective includes the range of purposes or the intention of provenance. In order to retrieve relevant studies, we linked together the individual concepts via a database query using the logical AND - operator. Synonyms within each concept were connected with the logical OR - operator.

The comprehensive search strategy is recorded in the study protocol [16].

Stage 3: Study Selection

The PRISMA flow chart in Fig. 1 depicts the selection process. First, we identified all relevant studies in PubMed and Web of Science based on our search strategy. After deduplication, we launched a transparent screening process by importing all relevant studies into Rayyan [18], a systematic review supporting solution. The studies were then reviewed by two independent researchers. In the case of vote agreement, the study was either included in the next review phase or excluded from the review. A third



independent reviewer was consulted to solve the conflict if no consensus could be reached. The study screening phase started with a title and abstract evaluation for eligibility. Included studies from this procedure were submitted to a full-text screening, a deep-dive into the study report. Reviewers voted for inclusion or exclusion considering the inclusion and exclusion criteria. Finally, the residing set of qualified studies was moved into the data extraction pipeline. A description of the study selection is provided in the protocol [16].

Stage 4: Charting the data

We followed a collaborative and iterative process to define a charting table for data extraction. Individual reviewers (KG, FK, FH, SG, AZ/DW) then scrutinized all studies and extracted central textual occurrences into the data extraction sheet. The variables in the data extraction sheet correspond with the research questions. As such general characteristics of the studies, approaches for classification and tracking of provenance, their related challenges along with the significance and completeness of provenance information in the given context were part of the investigational charting. The reviewers independently charted the data in a structured and consistent way, discussed the results and continuously updated the data-charting form in an iterative process.

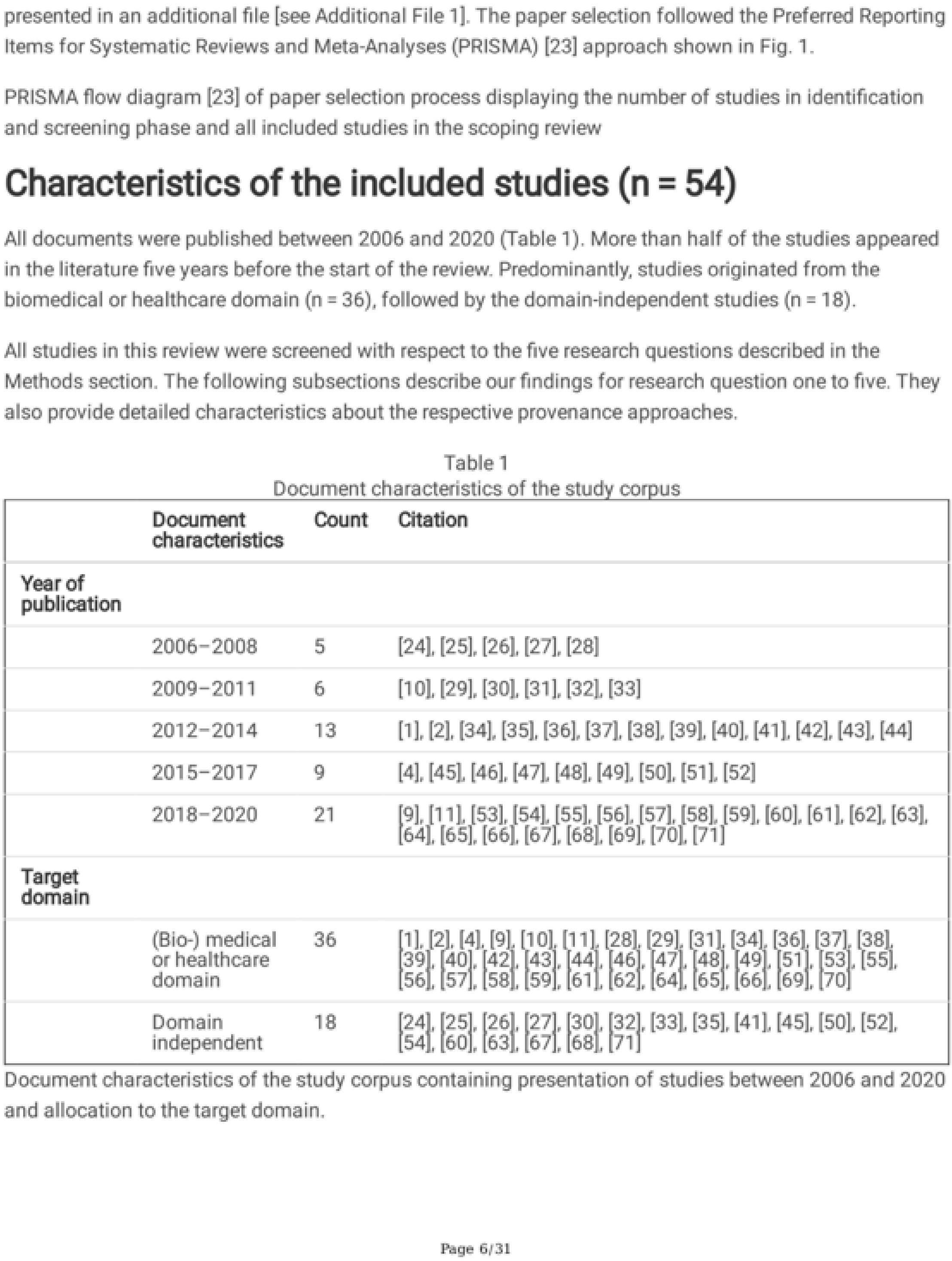
Stage 5: Collating, Summarizing, and Reporting the Results

The extracted data were analyzed using summary statistics by calculating the total number and percentages of all studies per category, if applicable. Charts were presented for the distribution of the individual data elements where applicable. Further analysis was performed using qualitative evaluation. The reporting of the results and outcome was structured according to the research questions. Based on the analysis of the review results, we have developed a roadmap for a customized provenance framework that takes into account the life cycle of the software framework (Provenance-SFL). The meaning of the findings and their relation to the overall objectives was discussed. Implications for future research, practice, and policy were outlined. Our reporting adheres to the PRISMA-ScR reporting guidelines [19]. The data analysis was partially supported with scripts in Python 3.10.0 [20]. Plots were generated with R version 4.0.4 (R Core Team) [21] and version 1.3.0 of the tidyverse package [22].

Results

Literature Search

The search in PubMed and Web of Science resulted in 564 hits and was last performed on March 23, 2020. Afterwards, 95 duplicates were removed. The remaining 469 papers were subjected to title-abstract screening in an interactive selection process, leaving 97 eligible papers for the full-text review. The full-text papers were further screened to identify papers eligible for the subsequent step of data charting. During this step, additional 43 papers were excluded (see stage 4). These papers either did neither meet the study design context (n = 26) nor the domain concept (n = 13). Three papers reported the same study, and one was not a full paper. A total of 54 articles were included in the data extraction phase and



presented in an additional file [see Additional File 1]. The paper selection followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) [23] approach shown in Fig. 1.

PRISMA flow diagram [23] of paper selection process displaying the number of studies in identification and screening phase and all included studies in the scoping review

Characteristics of the included studies (n = 54)

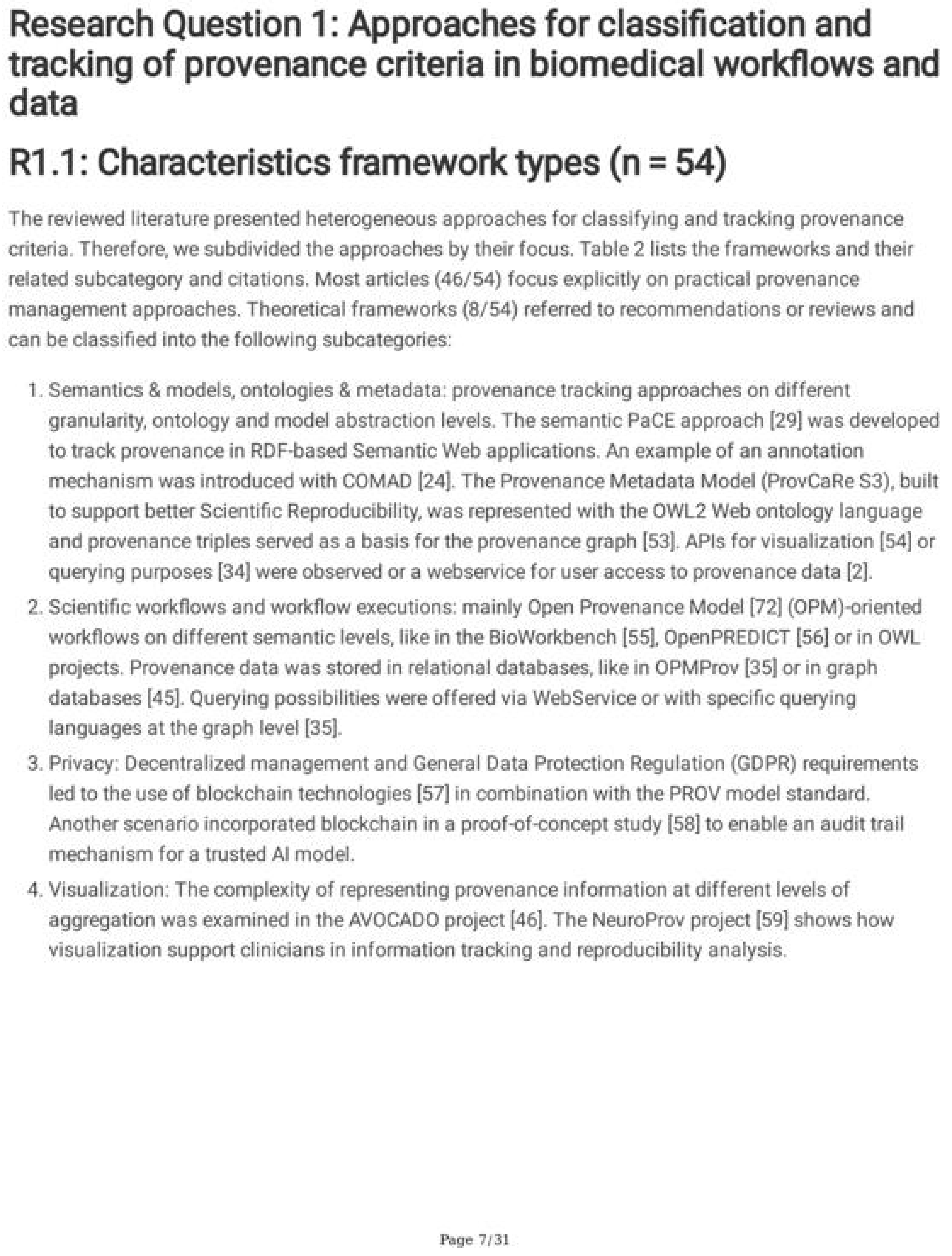
All documents were published between 2006 and 2020 (Table 1). More than half of the studies appeared in the literature five years before the start of the review. Predominantly, studies originated from the biomedical or healthcare domain (n = 36), followed by the domain-independent studies (n = 18).

All studies in this review were screened with respect to the five research questions described in the Methods section. The following subsections describe our findings for research question one to five. They also provide detailed characteristics about the respective provenance approaches.

Table 1
Document characteristics of the study corpus

	Document characteristics	Count	Citation
Year of publication			
	2006–2008	5	[24], [25], [26], [27], [28]
	2009–2011	6	[10], [29], [30], [31], [32], [33]
	2012–2014	13	[1], [2], [34], [35], [36], [37], [38], [39], [40], [41], [42], [43], [44]
	2015–2017	9	[4], [45], [46], [47], [48], [49], [50], [51], [52]
	2018–2020	21	[9], [11], [53], [54], [55], [56], [57], [58], [59], [60], [61], [62], [63], [64], [65], [66], [67], [68], [69], [70], [71]
Target domain			
	(Bio-) medical or healthcare domain	36	[1], [2], [4], [9], [10], [11], [28], [29], [31], [34], [36], [37], [38], [39], [40], [42], [43], [44], [46], [47], [48], [49], [51], [53], [55], [56], [57], [58], [59], [61], [62], [64], [65], [66], [69], [70]
	Domain independent	18	[24], [25], [26], [27], [30], [32], [33], [35], [41], [45], [50], [52], [54], [60], [63], [67], [68], [71]

Document characteristics of the study corpus containing presentation of studies between 2006 and 2020 and allocation to the target domain.



Research Question 1: Approaches for classification and tracking of provenance criteria in biomedical workflows and data

R1.1: Characteristics framework types (n = 54)

The reviewed literature presented heterogeneous approaches for classifying and tracking provenance criteria. Therefore, we subdivided the approaches by their focus. Table 2 lists the frameworks and their related subcategory and citations. Most articles (46/54) focus explicitly on practical provenance management approaches. Theoretical frameworks (8/54) referred to recommendations or reviews and can be classified into the following subcategories:

1. Semantics & models, ontologies & metadata: provenance tracking approaches on different granularity, ontology and model abstraction levels. The semantic PaCE approach [29] was developed to track provenance in RDF-based Semantic Web applications. An example of an annotation mechanism was introduced with COMAD [24]. The Provenance Metadata Model (ProvCaRe S3), built to support better Scientific Reproducibility, was represented with the OWL2 Web ontology language and provenance triples served as a basis for the provenance graph [53]. APIs for visualization [54] or querying purposes [34] were observed or a webservice for user access to provenance data [2].
2. Scientific workflows and workflow executions: mainly Open Provenance Model [72] (OPM)-oriented workflows on different semantic levels, like in the BioWorkbench [55], OpenPREDICT [56] or in OWL projects. Provenance data was stored in relational databases, like in OPMProv [35] or in graph databases [45]. Querying possibilities were offered via WebService or with specific querying languages at the graph level [35].
3. Privacy: Decentralized management and General Data Protection Regulation (GDPR) requirements led to the use of blockchain technologies [57] in combination with the PROV model standard. Another scenario incorporated blockchain in a proof-of-concept study [58] to enable an audit trail mechanism for a trusted AI model.
4. Visualization: The complexity of representing provenance information at different levels of aggregation was examined in the AVOCADO project [46]. The NeuroProv project [59] shows how visualization support clinicians in information tracking and reproducibility analysis.

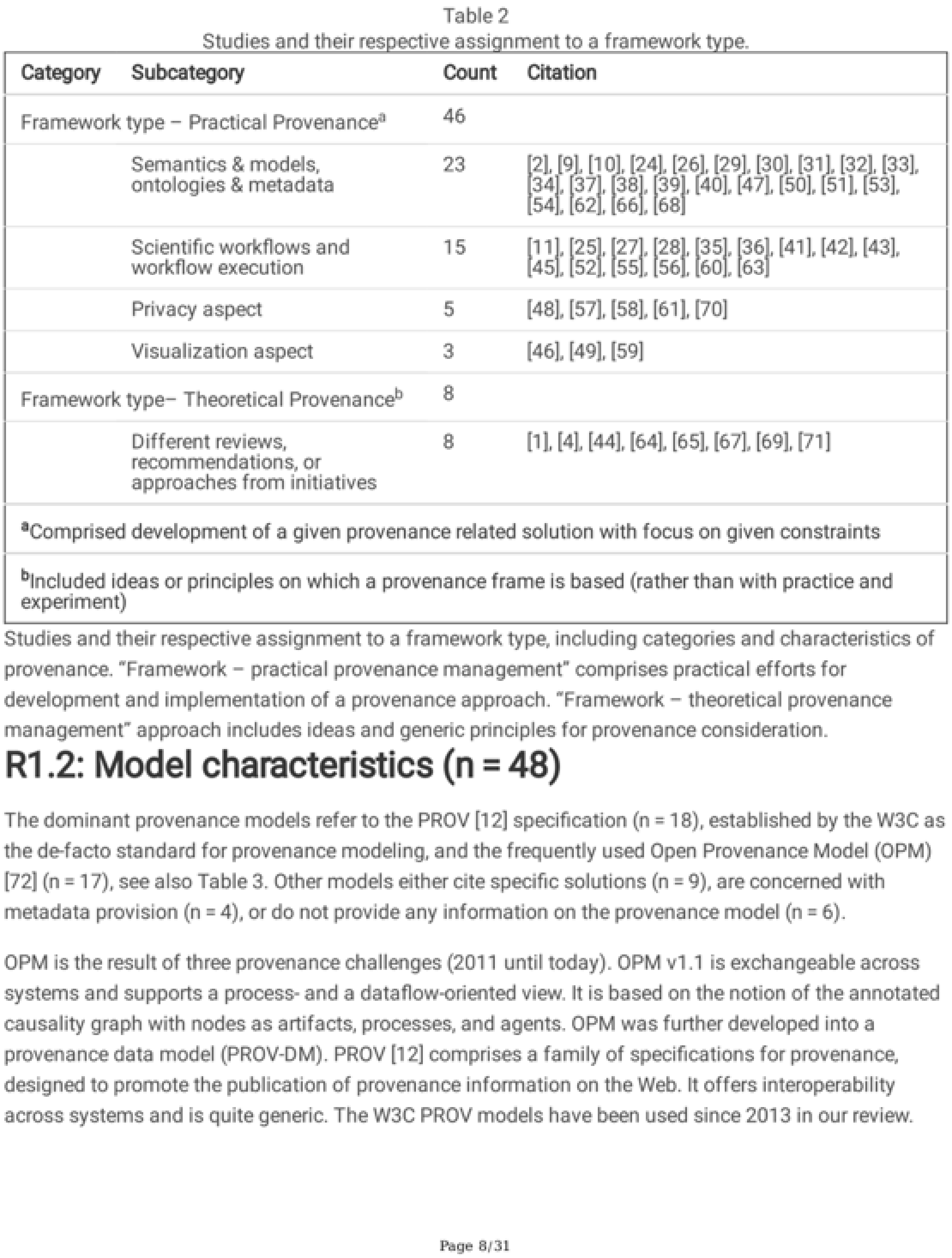


Table 2
Studies and their respective assignment to a framework type.

Category	Subcategory	Count	Citation
Framework type – Practical Provenance[a]		46	
	Semantics & models, ontologies & metadata	23	[2], [9], [10], [24], [26], [29], [30], [31], [32], [33], [34], [37], [38], [39], [40], [47], [50], [51], [53], [54], [62], [66], [68]
	Scientific workflows and workflow execution	15	[11], [25], [27], [28], [35], [36], [41], [42], [43], [45], [52], [55], [56], [60], [63]
	Privacy aspect	5	[48], [57], [58], [61], [70]
	Visualization aspect	3	[46], [49], [59]
Framework type– Theoretical Provenance[b]		8	
	Different reviews, recommendations, or approaches from initiatives	8	[1], [4], [44], [64], [65], [67], [69], [71]
[a]Comprised development of a given provenance related solution with focus on given constraints			
[b]Included ideas or principles on which a provenance frame is based (rather than with practice and experiment)			

Studies and their respective assignment to a framework type, including categories and characteristics of provenance. “Framework – practical provenance management” comprises practical efforts for development and implementation of a provenance approach. “Framework – theoretical provenance management” approach includes ideas and generic principles for provenance consideration.

R1.2: Model characteristics (n = 48)

The dominant provenance models refer to the PROV [12] specification (n = 18), established by the W3C as the de-facto standard for provenance modeling, and the frequently used Open Provenance Model (OPM) [72] (n = 17), see also Table 3. Other models either cite specific solutions (n = 9), are concerned with metadata provision (n = 4), or do not provide any information on the provenance model (n = 6).

OPM is the result of three provenance challenges (2011 until today). OPM v1.1 is exchangeable across systems and supports a process- and a dataflow-oriented view. It is based on the notion of the annotated causality graph with nodes as artifacts, processes, and agents. OPM was further developed into a provenance data model (PROV-DM). PROV [12] comprises a family of specifications for provenance, designed to promote the publication of provenance information on the Web. It offers interoperability across systems and is quite generic. The W3C PROV models have been used since 2013 in our review.

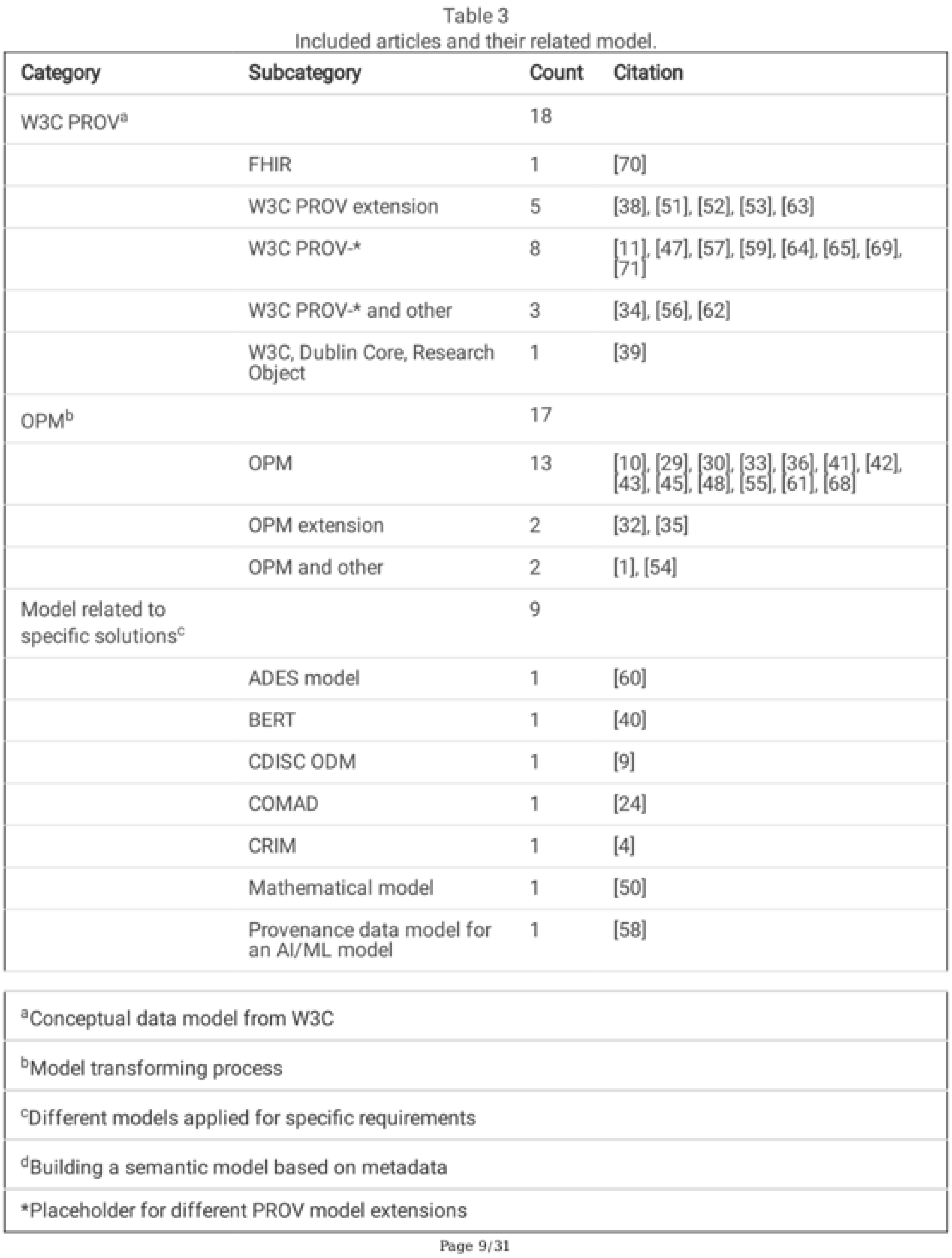


Table 3
Included articles and their related model.

Category	Subcategory	Count	Citation
W3C PROV[a]		18	
	FHIR	1	[70]
	W3C PROV extension	5	[38], [51], [52], [53], [63]
	W3C PROV-*	8	[11], [47], [57], [59], [64], [65], [69], [71]
	W3C PROV-* and other	3	[34], [56], [62]
	W3C, Dublin Core, Research Object	1	[39]
OPM[b]		17	
	OPM	13	[10], [29], [30], [33], [36], [41], [42], [43], [45], [48], [55], [61], [68]
	OPM extension	2	[32], [35]
	OPM and other	2	[1], [54]
Model related to specific solutions[c]		9	
	ADES model	1	[60]
	BERT	1	[40]
	CDISC ODM	1	[9]
	COMAD	1	[24]
	CRIM	1	[4]
	Mathematical model	1	[50]
	Provenance data model for an AI/ML model	1	[58]

[a] Conceptual data model from W3C

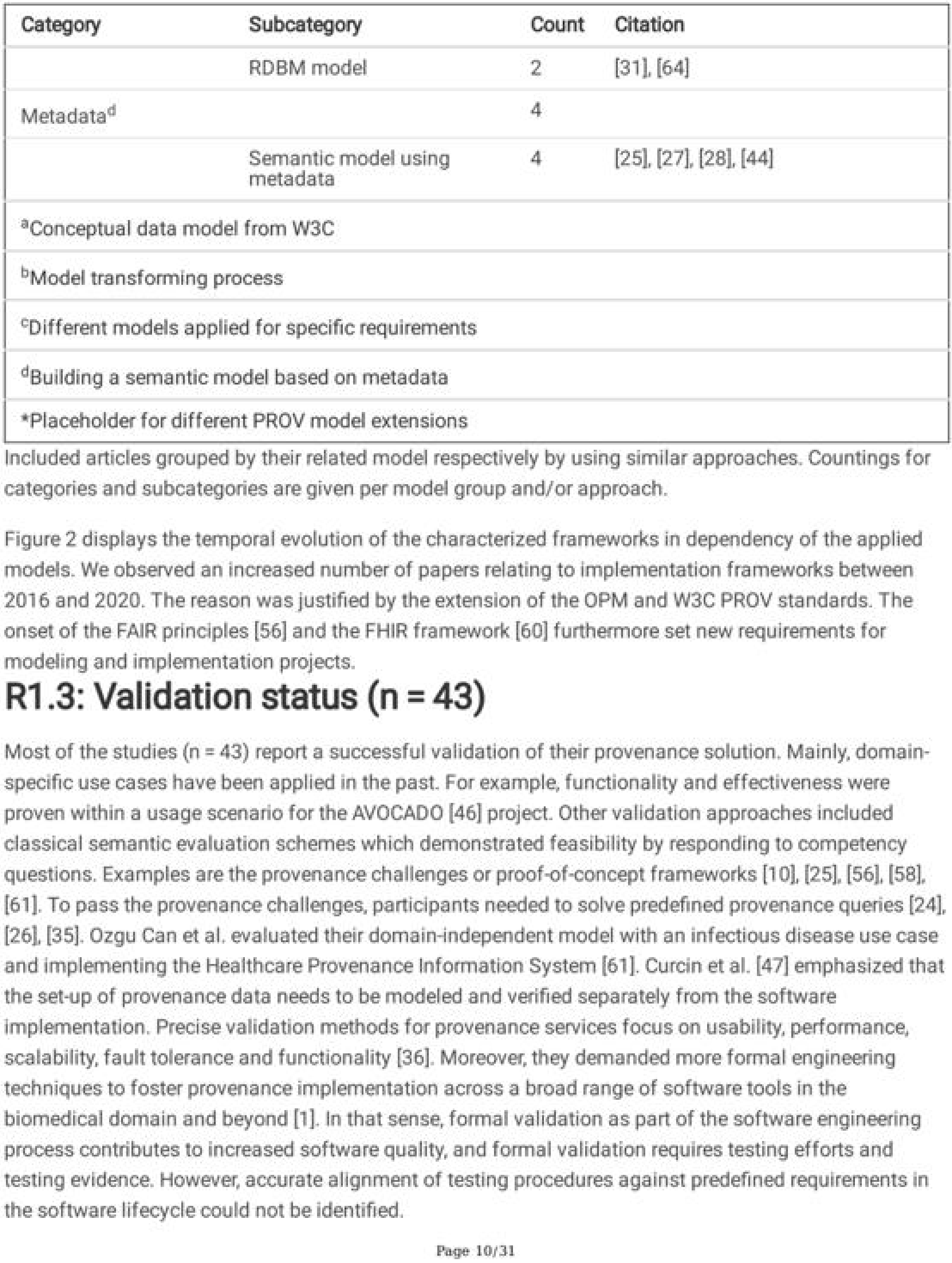
[b] Model transforming process

[c] Different models applied for specific requirements

[d] Building a semantic model based on metadata

*Placeholder for different PROV model extensions

Page 9/31



Category	Subcategory	Count	Citation
	RDBM model	2	[31], [64]
Metadata[d]		4	
	Semantic model using metadata	4	[25], [27], [28], [44]

[a]Conceptual data model from W3C

[b]Model transforming process

[c]Different models applied for specific requirements

[d]Building a semantic model based on metadata

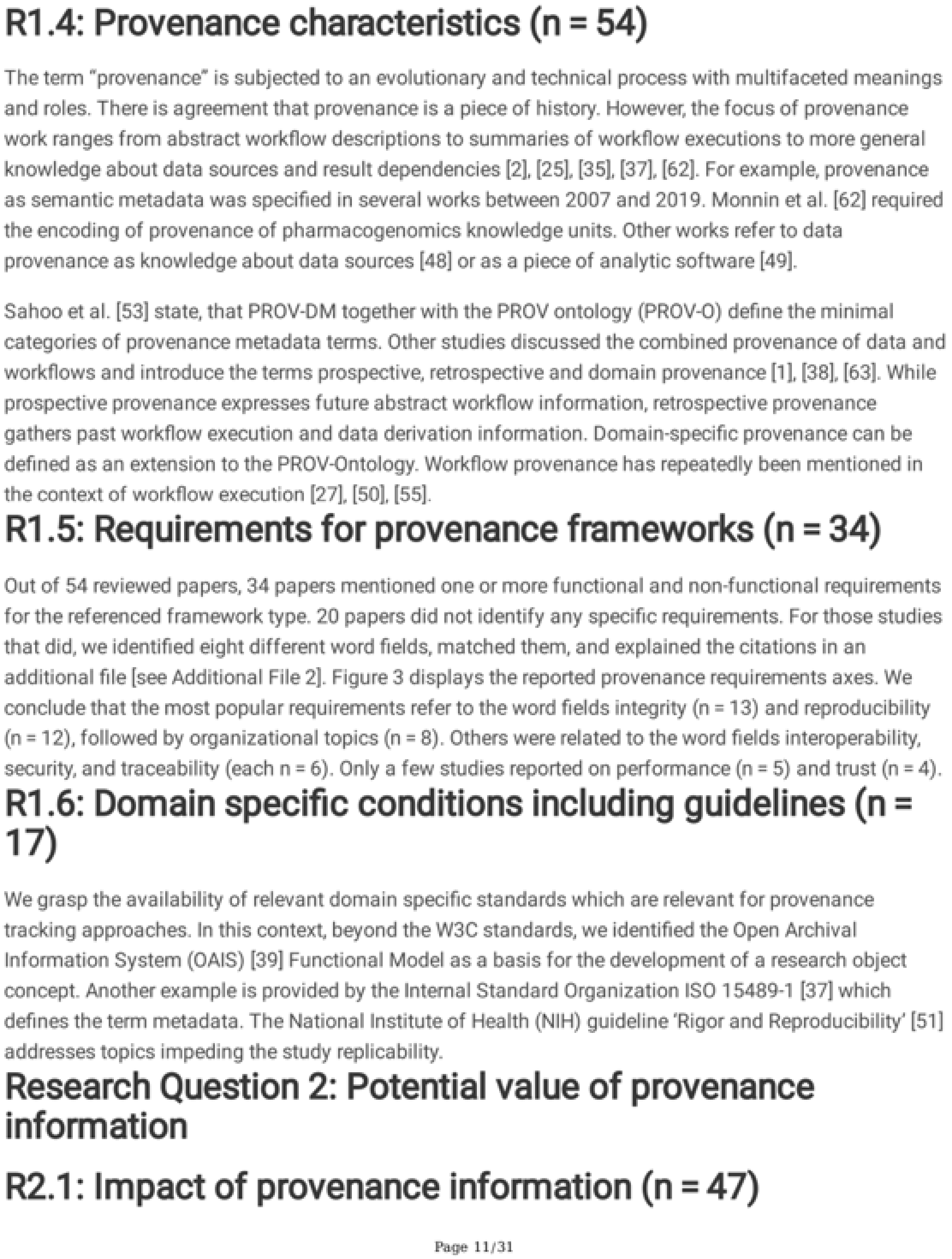
*Placeholder for different PROV model extensions

Included articles grouped by their related model respectively by using similar approaches. Countings for categories and subcategories are given per model group and/or approach.

Figure 2 displays the temporal evolution of the characterized frameworks in dependency of the applied models. We observed an increased number of papers relating to implementation frameworks between 2016 and 2020. The reason was justified by the extension of the OPM and W3C PROV standards. The onset of the FAIR principles [56] and the FHIR framework [60] furthermore set new requirements for modeling and implementation projects.

R1.3: Validation status (n = 43)

Most of the studies (n = 43) report a successful validation of their provenance solution. Mainly, domain-specific use cases have been applied in the past. For example, functionality and effectiveness were proven within a usage scenario for the AVOCADO [46] project. Other validation approaches included classical semantic evaluation schemes which demonstrated feasibility by responding to competency questions. Examples are the provenance challenges or proof-of-concept frameworks [10], [25], [56], [58], [61]. To pass the provenance challenges, participants needed to solve predefined provenance queries [24], [26], [35]. Ozgu Can et al. evaluated their domain-independent model with an infectious disease use case and implementing the Healthcare Provenance Information System [61]. Curcin et al. [47] emphasized that the set-up of provenance data needs to be modeled and verified separately from the software implementation. Precise validation methods for provenance services focus on usability, performance, scalability, fault tolerance and functionality [36]. Moreover, they demanded more formal engineering techniques to foster provenance implementation across a broad range of software tools in the biomedical domain and beyond [1]. In that sense, formal validation as part of the software engineering process contributes to increased software quality, and formal validation requires testing efforts and testing evidence. However, accurate alignment of testing procedures against predefined requirements in the software lifecycle could not be identified.



R1.4: Provenance characteristics (n = 54)

The term "provenance" is subjected to an evolutionary and technical process with multifaceted meanings and roles. There is agreement that provenance is a piece of history. However, the focus of provenance work ranges from abstract workflow descriptions to summaries of workflow executions to more general knowledge about data sources and result dependencies [2], [25], [35], [37], [62]. For example, provenance as semantic metadata was specified in several works between 2007 and 2019. Monnin et al. [62] required the encoding of provenance of pharmacogenomics knowledge units. Other works refer to data provenance as knowledge about data sources [48] or as a piece of analytic software [49].

Sahoo et al. [53] state, that PROV-DM together with the PROV ontology (PROV-O) define the minimal categories of provenance metadata terms. Other studies discussed the combined provenance of data and workflows and introduce the terms prospective, retrospective and domain provenance [1], [38], [63]. While prospective provenance expresses future abstract workflow information, retrospective provenance gathers past workflow execution and data derivation information. Domain-specific provenance can be defined as an extension to the PROV-Ontology. Workflow provenance has repeatedly been mentioned in the context of workflow execution [27], [50], [55].

R1.5: Requirements for provenance frameworks (n = 34)

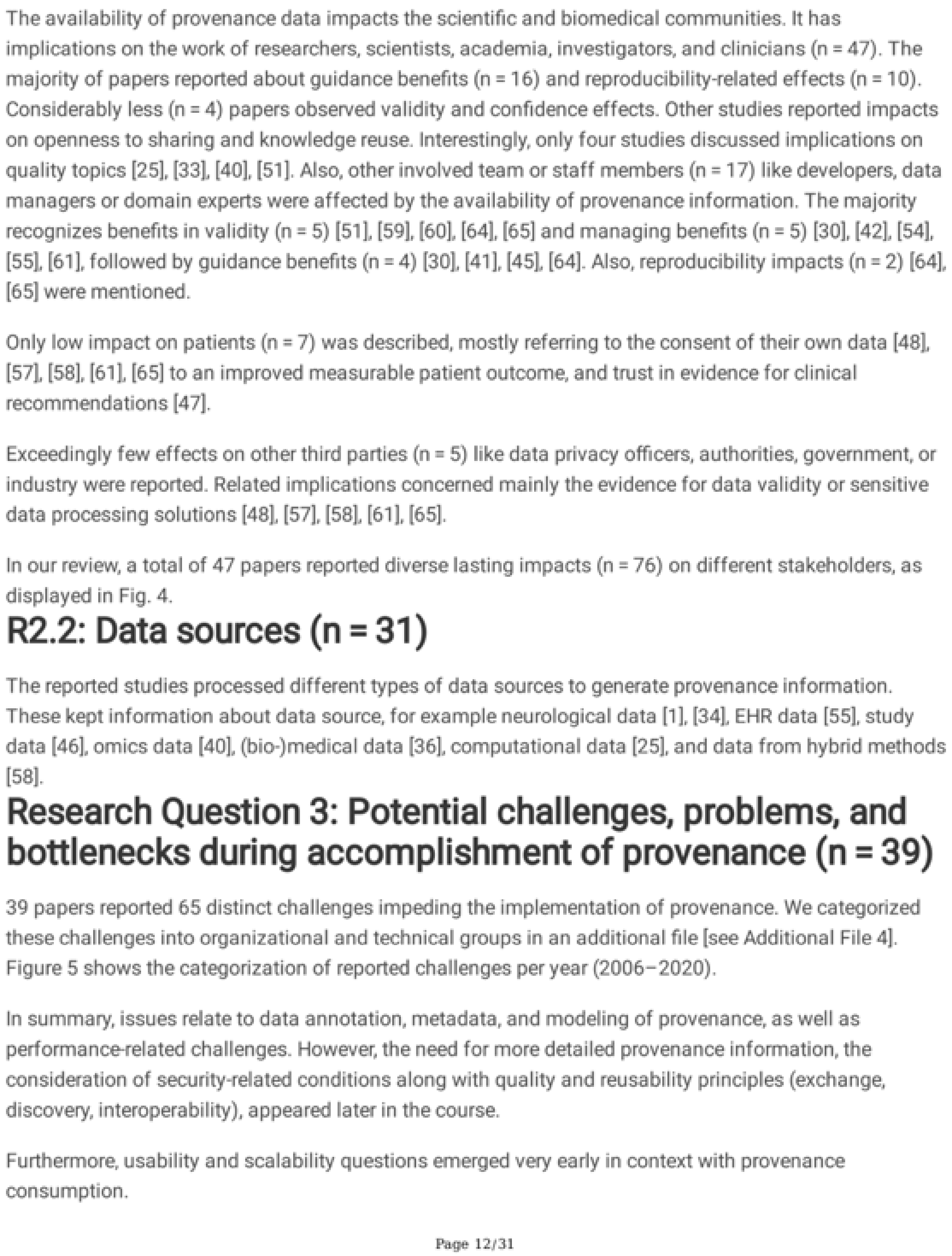
Out of 54 reviewed papers, 34 papers mentioned one or more functional and non-functional requirements for the referenced framework type. 20 papers did not identify any specific requirements. For those studies that did, we identified eight different word fields, matched them, and explained the citations in an additional file [see Additional File 2]. Figure 3 displays the reported provenance requirements axes. We conclude that the most popular requirements refer to the word fields integrity (n = 13) and reproducibility (n = 12), followed by organizational topics (n = 8). Others were related to the word fields interoperability, security, and traceability (each n = 6). Only a few studies reported on performance (n = 5) and trust (n = 4).

R1.6: Domain specific conditions including guidelines (n = 17)

We grasp the availability of relevant domain specific standards which are relevant for provenance tracking approaches. In this context, beyond the W3C standards, we identified the Open Archival Information System (OAIS) [39] Functional Model as a basis for the development of a research object concept. Another example is provided by the Internal Standard Organization ISO 15489-1 [37] which defines the term metadata. The National Institute of Health (NIH) guideline 'Rigor and Reproducibility' [51] addresses topics impeding the study replicability.

Research Question 2: Potential value of provenance information

R2.1: Impact of provenance information (n = 47)



The availability of provenance data impacts the scientific and biomedical communities. It has implications on the work of researchers, scientists, academia, investigators, and clinicians (n = 47). The majority of papers reported about guidance benefits (n = 16) and reproducibility-related effects (n = 10). Considerably less (n = 4) papers observed validity and confidence effects. Other studies reported impacts on openness to sharing and knowledge reuse. Interestingly, only four studies discussed implications on quality topics [25], [33], [40], [51]. Also, other involved team or staff members (n = 17) like developers, data managers or domain experts were affected by the availability of provenance information. The majority recognizes benefits in validity (n = 5) [51], [59], [60], [64], [65] and managing benefits (n = 5) [30], [42], [54], [55], [61], followed by guidance benefits (n = 4) [30], [41], [45], [64]. Also, reproducibility impacts (n = 2) [64], [65] were mentioned.

Only low impact on patients (n = 7) was described, mostly referring to the consent of their own data [48], [57], [58], [61], [65] to an improved measurable patient outcome, and trust in evidence for clinical recommendations [47].

Exceedingly few effects on other third parties (n = 5) like data privacy officers, authorities, government, or industry were reported. Related implications concerned mainly the evidence for data validity or sensitive data processing solutions [48], [57], [58], [61], [65].

In our review, a total of 47 papers reported diverse lasting impacts (n = 76) on different stakeholders, as displayed in Fig. 4.

R2.2: Data sources (n = 31)

The reported studies processed different types of data sources to generate provenance information. These kept information about data source, for example neurological data [1], [34], EHR data [55], study data [46], omics data [40], (bio-)medical data [36], computational data [25], and data from hybrid methods [58].

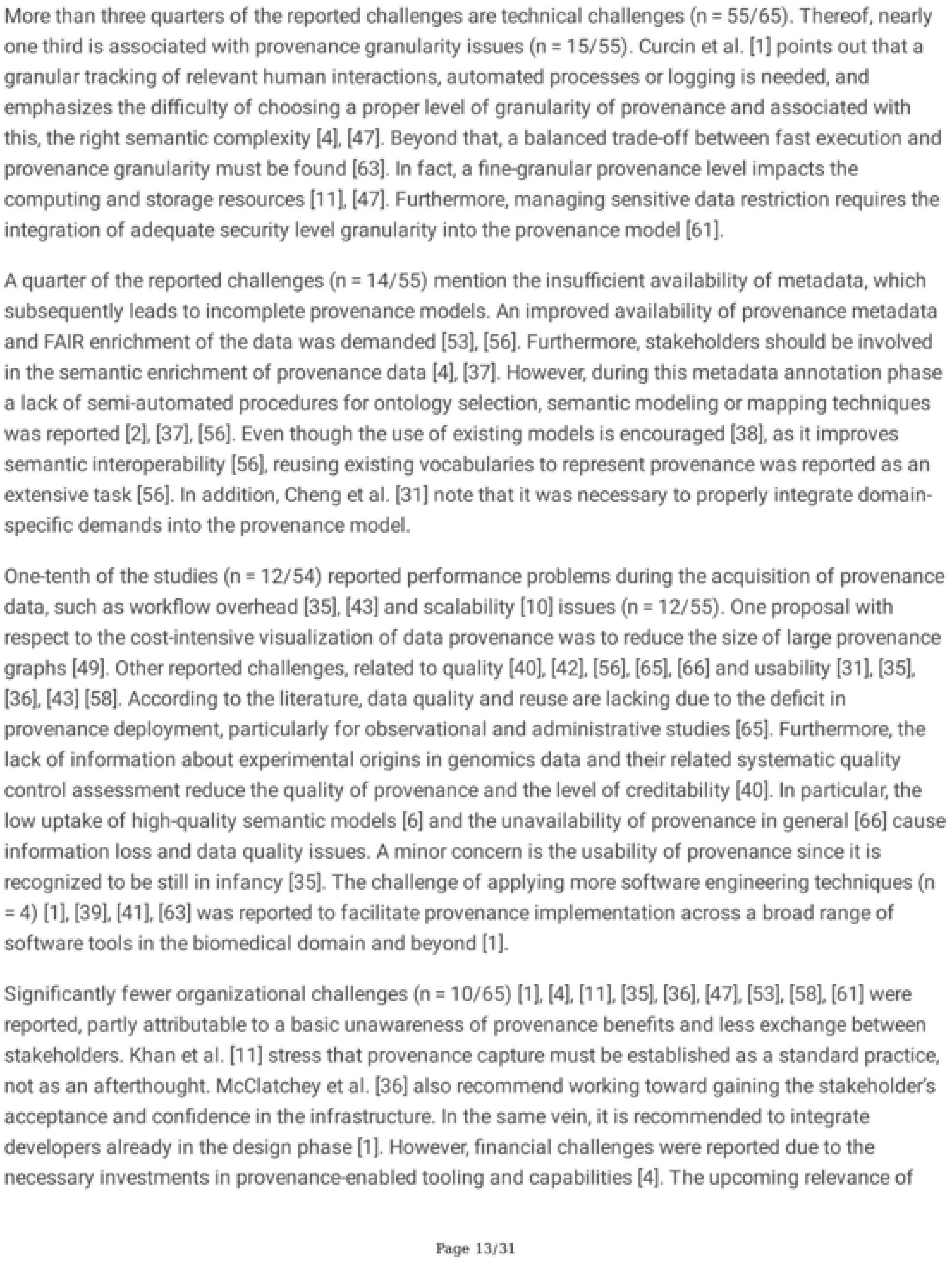
Research Question 3: Potential challenges, problems, and bottlenecks during accomplishment of provenance (n = 39)

39 papers reported 65 distinct challenges impeding the implementation of provenance. We categorized these challenges into organizational and technical groups in an additional file [see Additional File 4]. Figure 5 shows the categorization of reported challenges per year (2006–2020).

In summary, issues relate to data annotation, metadata, and modeling of provenance, as well as performance-related challenges. However, the need for more detailed provenance information, the consideration of security-related conditions along with quality and reusability principles (exchange, discovery, interoperability), appeared later in the course.

Furthermore, usability and scalability questions emerged very early in context with provenance consumption.

Page 12/31

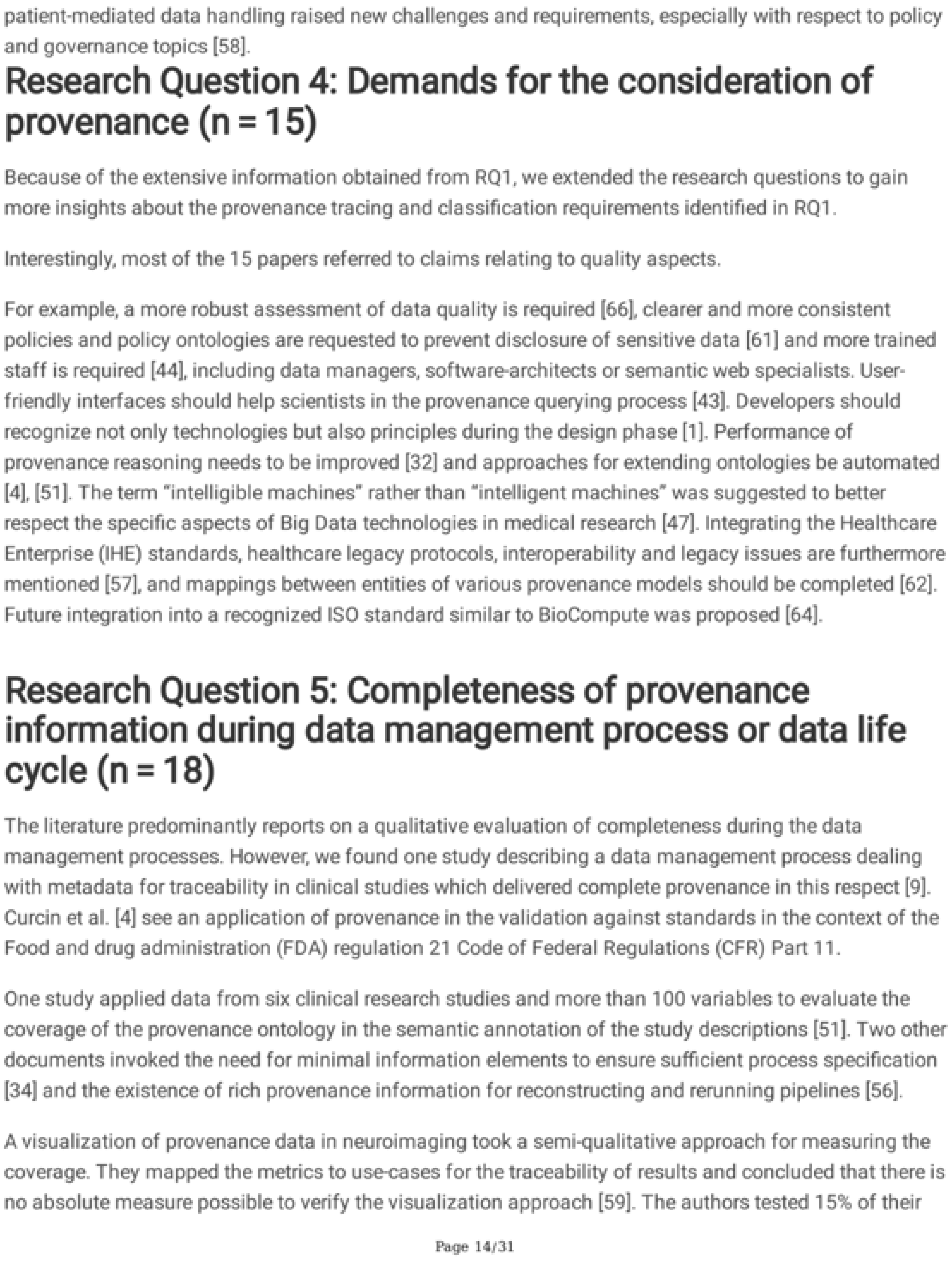


More than three quarters of the reported challenges are technical challenges (n = 55/65). Thereof, nearly one third is associated with provenance granularity issues (n = 15/55). Curcin et al. [1] points out that a granular tracking of relevant human interactions, automated processes or logging is needed, and emphasizes the difficulty of choosing a proper level of granularity of provenance and associated with this, the right semantic complexity [4], [47]. Beyond that, a balanced trade-off between fast execution and provenance granularity must be found [63]. In fact, a fine-granular provenance level impacts the computing and storage resources [11], [47]. Furthermore, managing sensitive data restriction requires the integration of adequate security level granularity into the provenance model [61].

A quarter of the reported challenges (n = 14/55) mention the insufficient availability of metadata, which subsequently leads to incomplete provenance models. An improved availability of provenance metadata and FAIR enrichment of the data was demanded [53], [56]. Furthermore, stakeholders should be involved in the semantic enrichment of provenance data [4], [37]. However, during this metadata annotation phase a lack of semi-automated procedures for ontology selection, semantic modeling or mapping techniques was reported [2], [37], [56]. Even though the use of existing models is encouraged [38], as it improves semantic interoperability [56], reusing existing vocabularies to represent provenance was reported as an extensive task [56]. In addition, Cheng et al. [31] note that it was necessary to properly integrate domain-specific demands into the provenance model.

One-tenth of the studies (n = 12/54) reported performance problems during the acquisition of provenance data, such as workflow overhead [35], [43] and scalability [10] issues (n = 12/55). One proposal with respect to the cost-intensive visualization of data provenance was to reduce the size of large provenance graphs [49]. Other reported challenges, related to quality [40], [42], [56], [65], [66] and usability [31], [35], [36], [43] [58]. According to the literature, data quality and reuse are lacking due to the deficit in provenance deployment, particularly for observational and administrative studies [65]. Furthermore, the lack of information about experimental origins in genomics data and their related systematic quality control assessment reduce the quality of provenance and the level of creditability [40]. In particular, the low uptake of high-quality semantic models [6] and the unavailability of provenance in general [66] cause information loss and data quality issues. A minor concern is the usability of provenance since it is recognized to be still in infancy [35]. The challenge of applying more software engineering techniques (n = 4) [1], [39], [41], [63] was reported to facilitate provenance implementation across a broad range of software tools in the biomedical domain and beyond [1].

Significantly fewer organizational challenges (n = 10/65) [1], [4], [11], [35], [36], [47], [53], [58], [61] were reported, partly attributable to a basic unawareness of provenance benefits and less exchange between stakeholders. Khan et al. [11] stress that provenance capture must be established as a standard practice, not as an afterthought. McClatchey et al. [36] also recommend working toward gaining the stakeholder's acceptance and confidence in the infrastructure. In the same vein, it is recommended to integrate developers already in the design phase [1]. However, financial challenges were reported due to the necessary investments in provenance-enabled tooling and capabilities [4]. The upcoming relevance of



patient-mediated data handling raised new challenges and requirements, especially with respect to policy and governance topics [58].

Research Question 4: Demands for the consideration of provenance (n = 15)

Because of the extensive information obtained from RQ1, we extended the research questions to gain more insights about the provenance tracing and classification requirements identified in RQ1.

Interestingly, most of the 15 papers referred to claims relating to quality aspects.

For example, a more robust assessment of data quality is required [66], clearer and more consistent policies and policy ontologies are requested to prevent disclosure of sensitive data [61] and more trained staff is required [44], including data managers, software-architects or semantic web specialists. User-friendly interfaces should help scientists in the provenance querying process [43]. Developers should recognize not only technologies but also principles during the design phase [1]. Performance of provenance reasoning needs to be improved [32] and approaches for extending ontologies be automated [4], [51]. The term "intelligible machines" rather than "intelligent machines" was suggested to better respect the specific aspects of Big Data technologies in medical research [47]. Integrating the Healthcare Enterprise (IHE) standards, healthcare legacy protocols, interoperability and legacy issues are furthermore mentioned [57], and mappings between entities of various provenance models should be completed [62]. Future integration into a recognized ISO standard similar to BioCompute was proposed [64].

Research Question 5: Completeness of provenance information during data management process or data life cycle (n = 18)

The literature predominantly reports on a qualitative evaluation of completeness during the data management processes. However, we found one study describing a data management process dealing with metadata for traceability in clinical studies which delivered complete provenance in this respect [9]. Curcin et al. [4] see an application of provenance in the validation against standards in the context of the Food and drug administration (FDA) regulation 21 Code of Federal Regulations (CFR) Part 11.

One study applied data from six clinical research studies and more than 100 variables to evaluate the coverage of the provenance ontology in the semantic annotation of the study descriptions [51]. Two other documents invoked the need for minimal information elements to ensure sufficient process specification [34] and the existence of rich provenance information for reconstructing and rerunning pipelines [56].

A visualization of provenance data in neuroimaging took a semi-qualitative approach for measuring the coverage. They mapped the metrics to use-cases for the traceability of results and concluded that there is no absolute measure possible to verify the visualization approach [59]. The authors tested 15% of their workflows for verifiability of results, comparability of workflows, progression of the data for the analysis and origin of results, and evolution to see how data products evolved during an experiment.

Furthermore, Sahoo et al. [53] examined the proportion of provenance metadata information across research articles using a qualitative hypothesis method. The method also provides a provenance ranking algorithm for the computation of a reproducibility rank for each article.

No numerical indication of completeness was not achieved in any of the other papers. However, the papers pointed out the advantages of provenance capture, for example, related to the longevity and accessibility of data after years [60].

Roadmap for a tailor-made provenance framework

Based on the insights obtained from the literature review, we developed a roadmap for the implementation of a tailor-made provenance framework based on the software-framework-lifecycle (Provenance-SFL). The heterogeneous tracking approaches, their artifacts, and varying degrees of fulfillment of the research questions are depicted in Fig. 6 and determine our main discussion points.

Discussion

This scoping review investigates evidence regarding approaches and criteria for provenance tracking. It discloses knowledge gaps in the biomedical domain with a focus on modeling and metadata frameworks for (sensitive) scientific biomedical data. Following the previously published scoping review protocol led us to include 54 full-text papers from initially 564 fetched papers found in PubMed and WoS databases. Using a structured and pre-tested data extraction sheet, contextual, but detailed enough, results were extracted to answer the outlined five research questions in the protocol.

Following the data extraction and analysis, the findings led us to define a Provenance-SLF roadmap elements. We essentially distinguished between the framework types and model characteristics, the validation status, and the requirement and provenance characteristics (see Fig. 6).

The provenance challenges, dealing with the need for provenance standardization, started in 2006 and gave rise to tailor-made models and metadata frameworks for the representation of provenance. These were later superseded by general-purpose standardized provenance models, which have more recently been combined with domain and application specific models or extensions such as the Provenance, Authoring and Versioning (PAV) ontology [38] or the ProvCaRe model [53]. The predominantly used models reported in this review referred to the W3C PROV and OPM standards. As shown in Fig. 2, an increased number of papers were related to the implementation frameworks that appeared between 2016 and 2020. One reason for the increase in implementations might be the substantiation to extend W3C PROV and OPM [11].

As of now, heterogeneous data sources, dynamic infrastructures, data exchange across boundaries, and a lack of standards for quality measures characterize the state of electronic health record data sets [57]. Additionally, various aspects of the term provenance [27], [37], [46], [53], [65] hamper the unique understanding and harmonization and engineering efforts for modeling, implementation, and validation interventions until now.

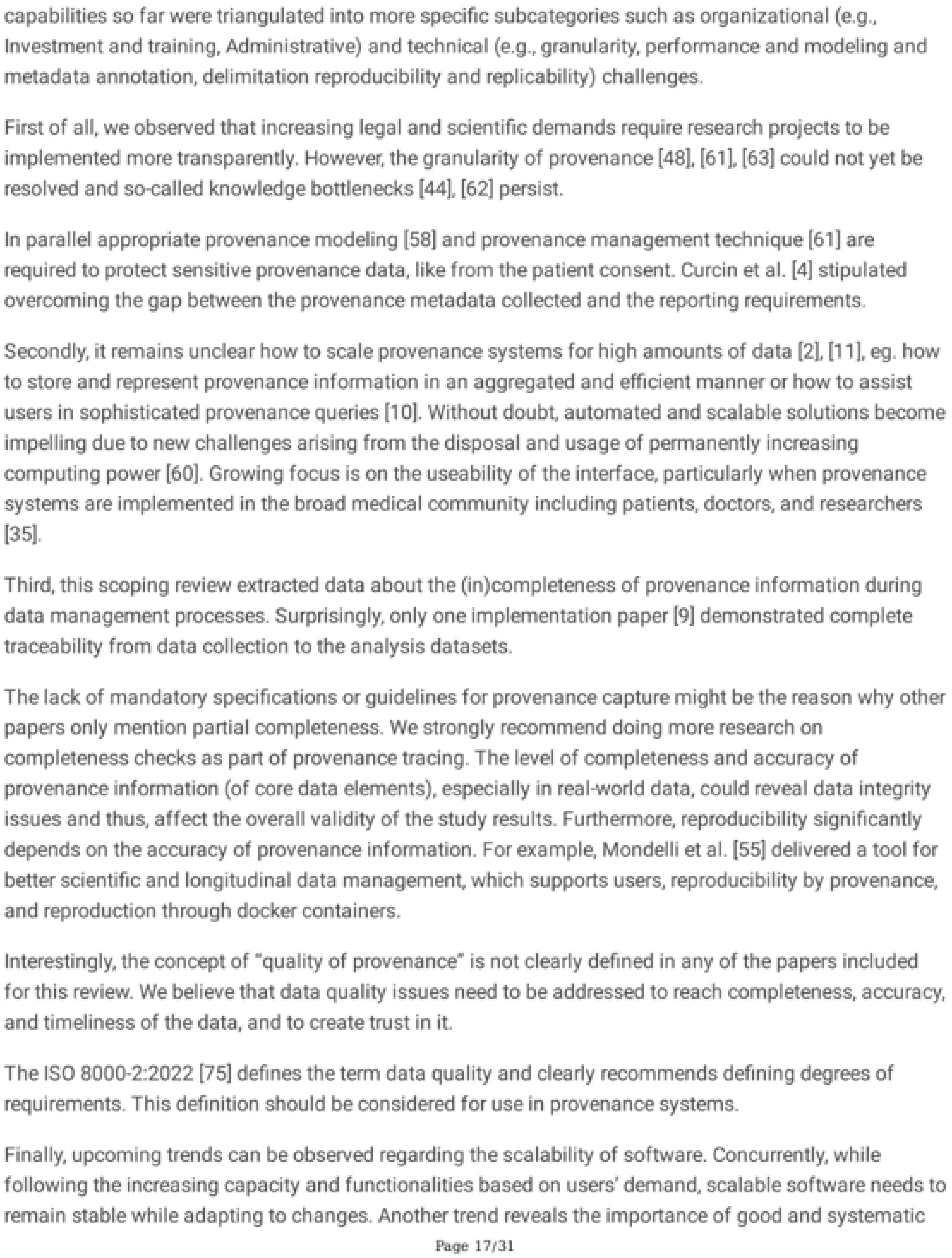
A provenance framework for today's demands must acknowledge the (semantic) complexity of the domain and its relevant facets and requirements [11] (see also Fig. 2). In addition to requirements analysis, a thorough strategy is necessary to plan the typical data management steps such as collecting, managing, and analyzing data (Pimentel et al. [67]). According to Curcin et al. [4], validation readiness can be achieved by separating modeling and verification of provenance data from the software implementation.

We agree that precise requirements analysis, as part of the software-life cycle, and the subsequent individual life-cycle steps, like testing and maintenance procedures, support the consequent temporal evolution and hence improve the quality of provenance frameworks and applications.

When incorporated in an official inspection, provenance information must be sufficient for a content-related validation against applicable and accepted standards [4]. Therefore precise validation methods for provenance services regarding usability and performance, scalability, fault tolerance, and functionality are needed [36]. We saw that validation approaches are linked to the evolution of provenance modeling and subsequent implementation attempts. Curcin et al. [1] argue that it was necessary to launch more formal software engineering techniques to foster provenance implementation across a broad range of software tools in the biomedical domain and beyond [1]. In that sense, formal validation as part of the software engineering process contributes to increased software and data quality. Formal validation requires testing efforts and testing evidence. Accurate alignment of testing procedures against predefined requirements in the software-lifecycle could not be identified in the included papers.

Provenance information is of high value for the scientific and biomedical community (eg. researchers), support staff (eg. developers), patients and other 3rd parties (eg. data privacy officer, authority) (see Fig. 4). It is interesting to see that despite the high impact of provenance [see Additional File 3] only some stakeholders provide sufficient provenance information. Rather, it appears that responsibility for overall provenance management is being shifted to the support staff [Gierend et al. (unpublished observations)]. We argue that available technology, IT knowledge and data management skills need to be paired with both domain-specific knowledge and combined with constraints of legal nature or guidance [4], [44]. This complexity indeed results in a very time-consuming business. However, automation and metadata collection can support this process [4], [73]. As a matter of fact, good provenance information strengthens the credibility of the data and proves that data have not been intentionally or unintentionally changed throughout the data life cycle [74].

Our review collects and summarizes the existing challenges during the accomplishment of provenance (Fig. 5). Challenges expressed in terms of missing, lacking, or hinderance on organizational and technical



capabilities so far were triangulated into more specific subcategories such as organizational (e.g., Investment and training, Administrative) and technical (e.g., granularity, performance and modeling and metadata annotation, delimitation reproducibility and replicability) challenges.

First of all, we observed that increasing legal and scientific demands require research projects to be implemented more transparently. However, the granularity of provenance [48], [61], [63] could not yet be resolved and so-called knowledge bottlenecks [44], [62] persist.

In parallel appropriate provenance modeling [58] and provenance management technique [61] are required to protect sensitive provenance data, like from the patient consent. Curcin et al. [4] stipulated overcoming the gap between the provenance metadata collected and the reporting requirements.

Secondly, it remains unclear how to scale provenance systems for high amounts of data [2], [11], eg. how to store and represent provenance information in an aggregated and efficient manner or how to assist users in sophisticated provenance queries [10]. Without doubt, automated and scalable solutions become impelling due to new challenges arising from the disposal and usage of permanently increasing computing power [60]. Growing focus is on the useability of the interface, particularly when provenance systems are implemented in the broad medical community including patients, doctors, and researchers [35].

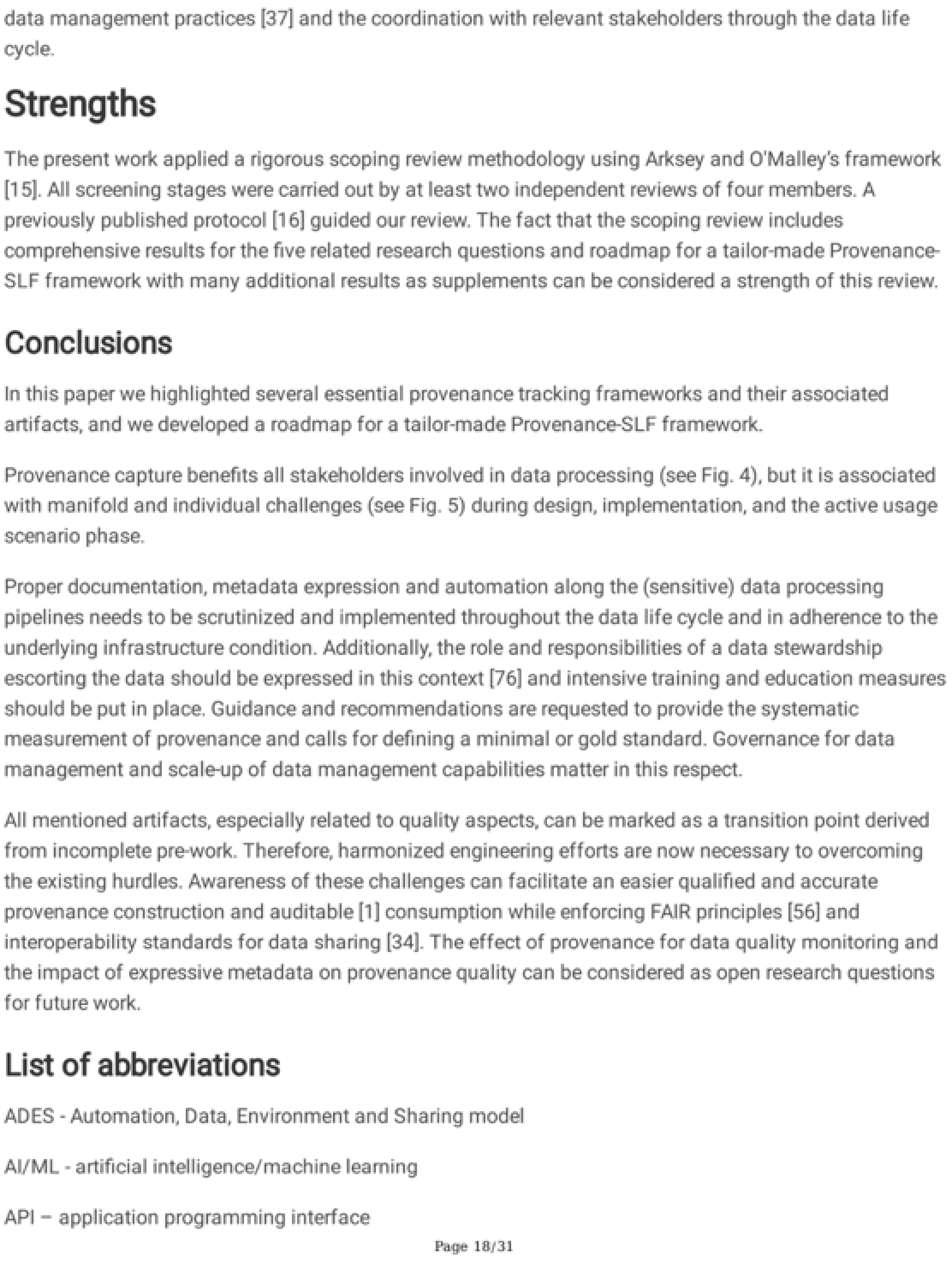
Third, this scoping review extracted data about the (in)completeness of provenance information during data management processes. Surprisingly, only one implementation paper [9] demonstrated complete traceability from data collection to the analysis datasets.

The lack of mandatory specifications or guidelines for provenance capture might be the reason why other papers only mention partial completeness. We strongly recommend doing more research on completeness checks as part of provenance tracing. The level of completeness and accuracy of provenance information (of core data elements), especially in real-world data, could reveal data integrity issues and thus, affect the overall validity of the study results. Furthermore, reproducibility significantly depends on the accuracy of provenance information. For example, Mondelli et al. [55] delivered a tool for better scientific and longitudinal data management, which supports users, reproducibility by provenance, and reproduction through docker containers.

Interestingly, the concept of "quality of provenance" is not clearly defined in any of the papers included for this review. We believe that data quality issues need to be addressed to reach completeness, accuracy, and timeliness of the data, and to create trust in it.

The ISO 8000-2:2022 [75] defines the term data quality and clearly recommends defining degrees of requirements. This definition should be considered for use in provenance systems.

Finally, upcoming trends can be observed regarding the scalability of software. Concurrently, while following the increasing capacity and functionalities based on users' demand, scalable software needs to remain stable while adapting to changes. Another trend reveals the importance of good and systematic



data management practices [37] and the coordination with relevant stakeholders through the data life cycle.

Strengths

The present work applied a rigorous scoping review methodology using Arksey and O'Malley's framework [15]. All screening stages were carried out by at least two independent reviews of four members. A previously published protocol [16] guided our review. The fact that the scoping review includes comprehensive results for the five related research questions and roadmap for a tailor-made Provenance-SLF framework with many additional results as supplements can be considered a strength of this review.

Conclusions

In this paper we highlighted several essential provenance tracking frameworks and their associated artifacts, and we developed a roadmap for a tailor-made Provenance-SLF framework.

Provenance capture benefits all stakeholders involved in data processing (see Fig. 4), but it is associated with manifold and individual challenges (see Fig. 5) during design, implementation, and the active usage scenario phase.

Proper documentation, metadata expression and automation along the (sensitive) data processing pipelines needs to be scrutinized and implemented throughout the data life cycle and in adherence to the underlying infrastructure condition. Additionally, the role and responsibilities of a data stewardship escorting the data should be expressed in this context [76] and intensive training and education measures should be put in place. Guidance and recommendations are requested to provide the systematic measurement of provenance and calls for defining a minimal or gold standard. Governance for data management and scale-up of data management capabilities matter in this respect.

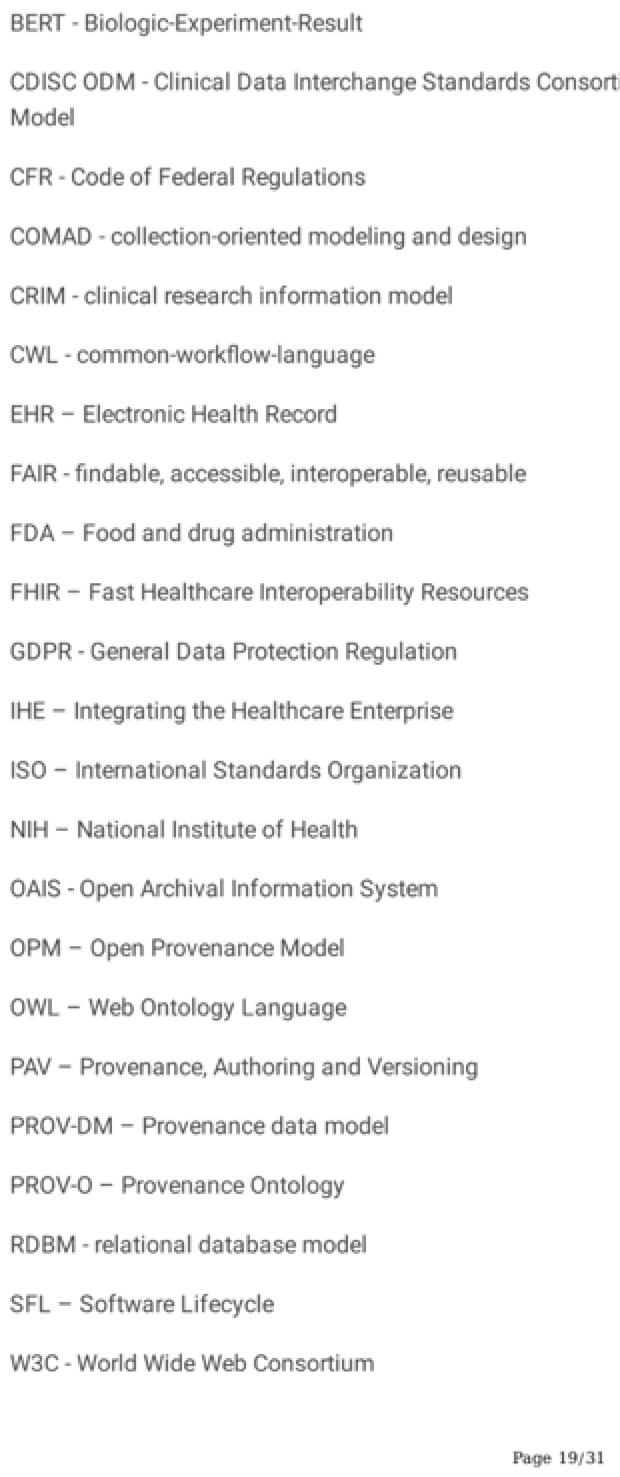
All mentioned artifacts, especially related to quality aspects, can be marked as a transition point derived from incomplete pre-work. Therefore, harmonized engineering efforts are now necessary to overcoming the existing hurdles. Awareness of these challenges can facilitate an easier qualified and accurate provenance construction and auditable [1] consumption while enforcing FAIR principles [56] and interoperability standards for data sharing [34]. The effect of provenance for data quality monitoring and the impact of expressive metadata on provenance quality can be considered as open research questions for future work.

List of abbreviations

ADES - Automation, Data, Environment and Sharing model

AI/ML - artificial intelligence/machine learning

API – application programming interface



BERT - Biologic-Experiment-Result

CDISC ODM - Clinical Data Interchange Standards Consortium standard, including Operational Data Model

CFR - Code of Federal Regulations

COMAD - collection-oriented modeling and design

CRIM - clinical research information model

CWL - common-workflow-language

EHR – Electronic Health Record

FAIR - findable, accessible, interoperable, reusable

FDA – Food and drug administration

FHIR – Fast Healthcare Interoperability Resources

GDPR - General Data Protection Regulation

IHE – Integrating the Healthcare Enterprise

ISO – International Standards Organization

NIH – National Institute of Health

OAIS - Open Archival Information System

OPM – Open Provenance Model

OWL – Web Ontology Language

PAV – Provenance, Authoring and Versioning

PROV-DM – Provenance data model

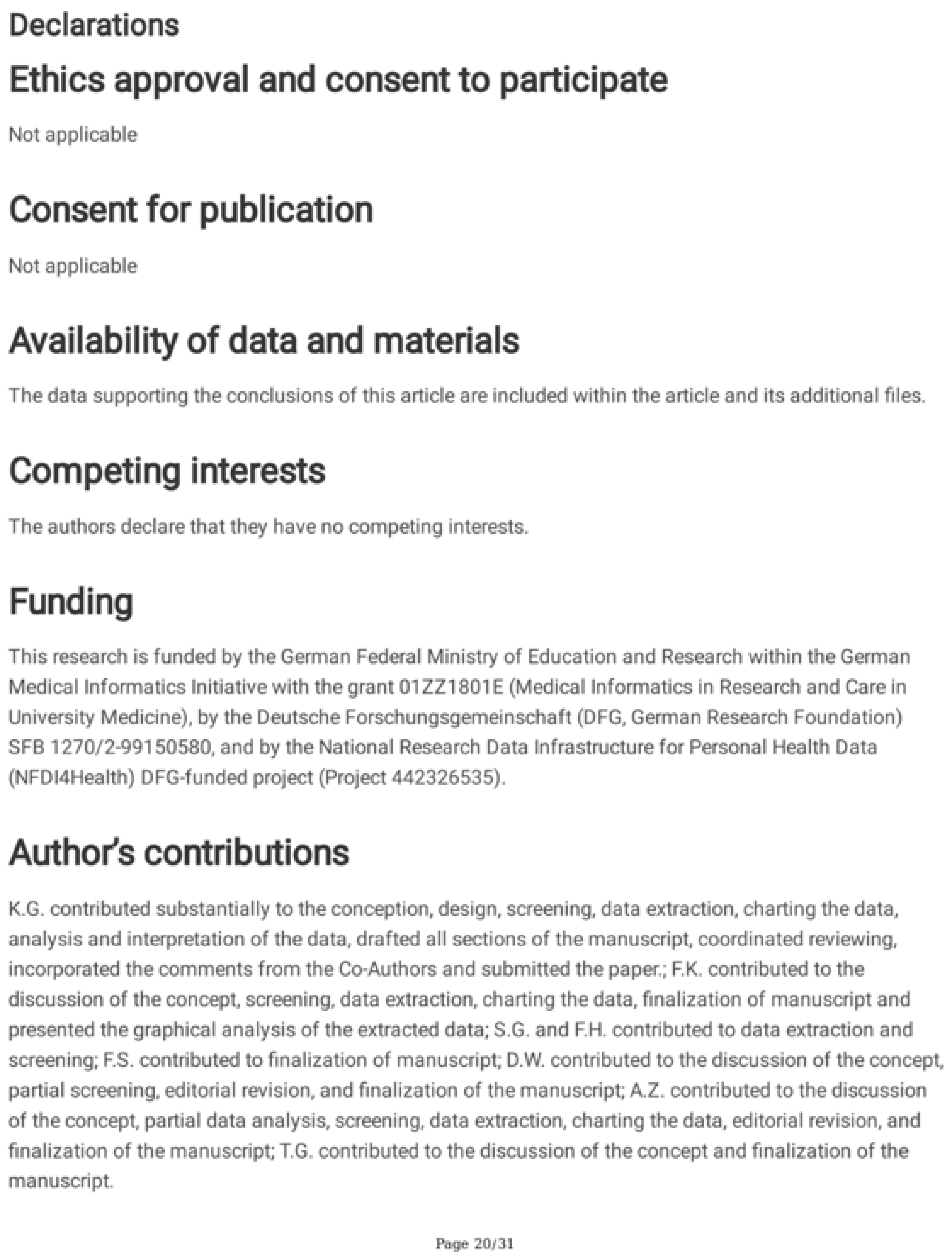
PROV-O – Provenance Ontology

RDBM - relational database model

SFL – Software Lifecycle

W3C - World Wide Web Consortium

Page 19/31



Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

The data supporting the conclusions of this article are included within the article and its additional files.

Competing interests

The authors declare that they have no competing interests.

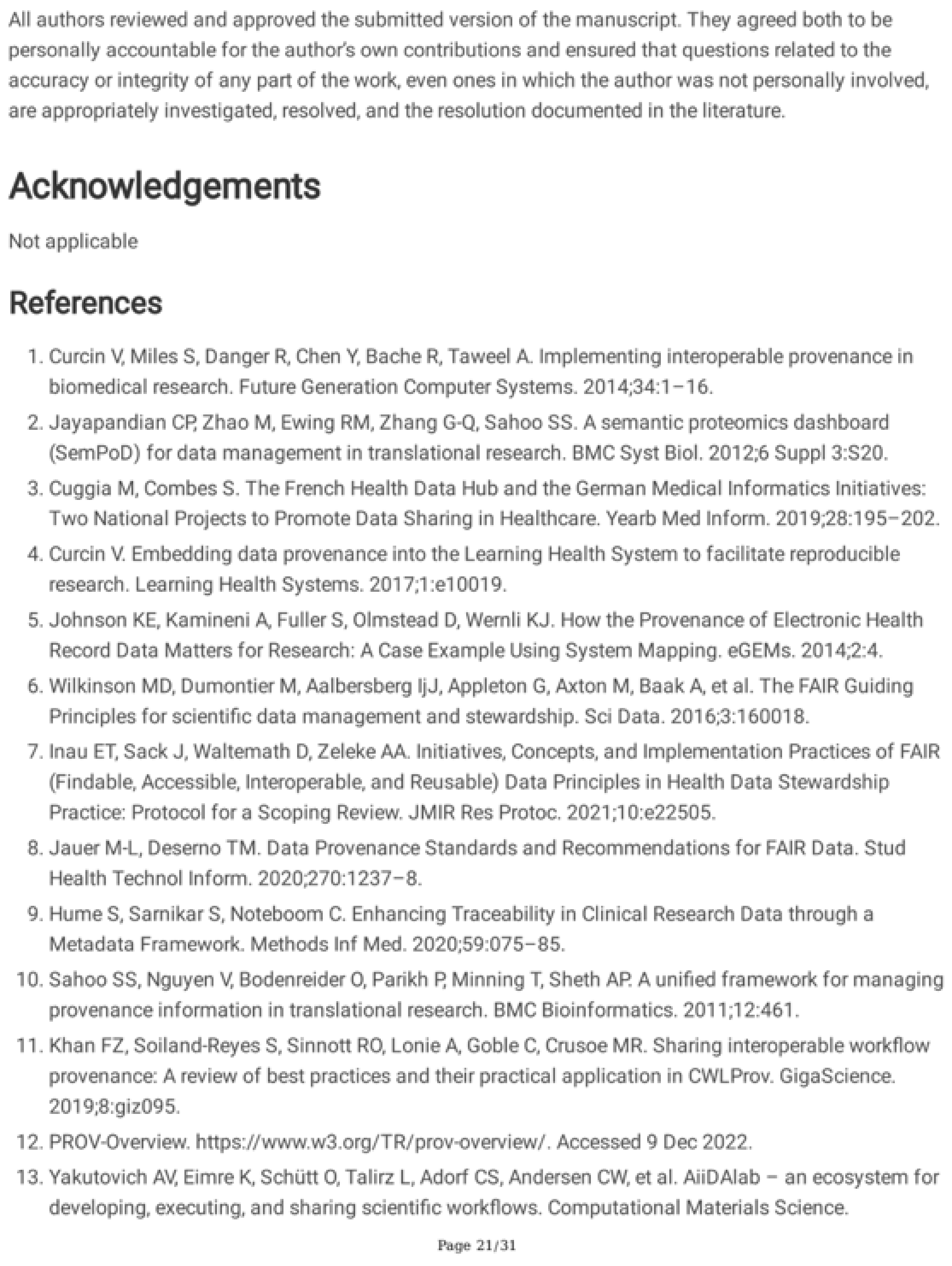
Funding

This research is funded by the German Federal Ministry of Education and Research within the German Medical Informatics Initiative with the grant 01ZZ1801E (Medical Informatics in Research and Care in University Medicine), by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) SFB 1270/2-99150580, and by the National Research Data Infrastructure for Personal Health Data (NFDI4Health) DFG-funded project (Project 442326535).

Author’s contributions

K.G. contributed substantially to the conception, design, screening, data extraction, charting the data, analysis and interpretation of the data, drafted all sections of the manuscript, coordinated reviewing, incorporated the comments from the Co-Authors and submitted the paper.; F.K. contributed to the discussion of the concept, screening, data extraction, charting the data, finalization of manuscript and presented the graphical analysis of the extracted data; S.G. and F.H. contributed to data extraction and screening; F.S. contributed to finalization of manuscript; D.W. contributed to the discussion of the concept, partial screening, editorial revision, and finalization of the manuscript; A.Z. contributed to the discussion of the concept, partial data analysis, screening, data extraction, charting the data, editorial revision, and finalization of the manuscript; T.G. contributed to the discussion of the concept and finalization of the manuscript.

Page 20/31



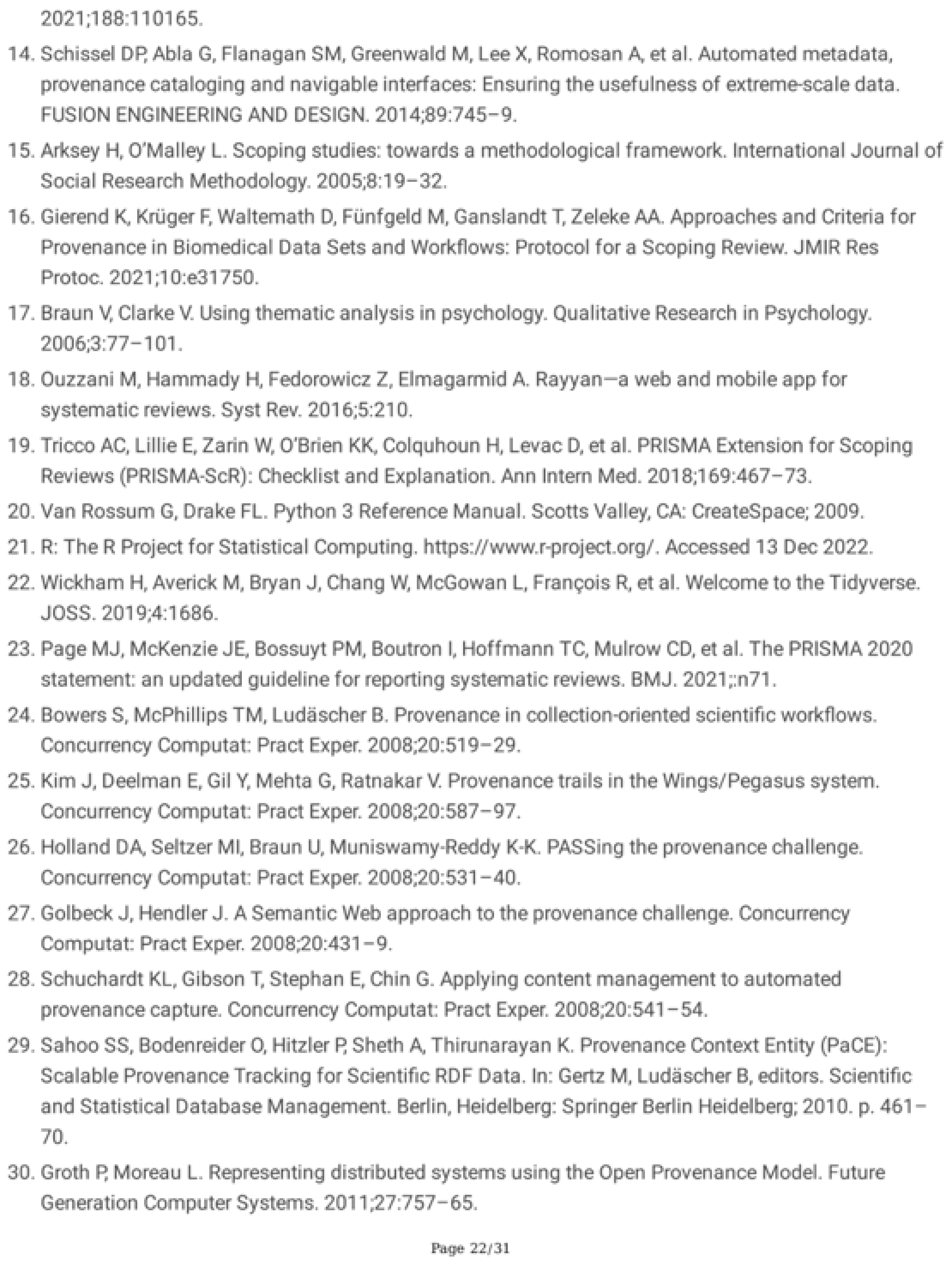
All authors reviewed and approved the submitted version of the manuscript. They agreed both to be personally accountable for the author's own contributions and ensured that questions related to the accuracy or integrity of any part of the work, even ones in which the author was not personally involved, are appropriately investigated, resolved, and the resolution documented in the literature.

Acknowledgements

Not applicable

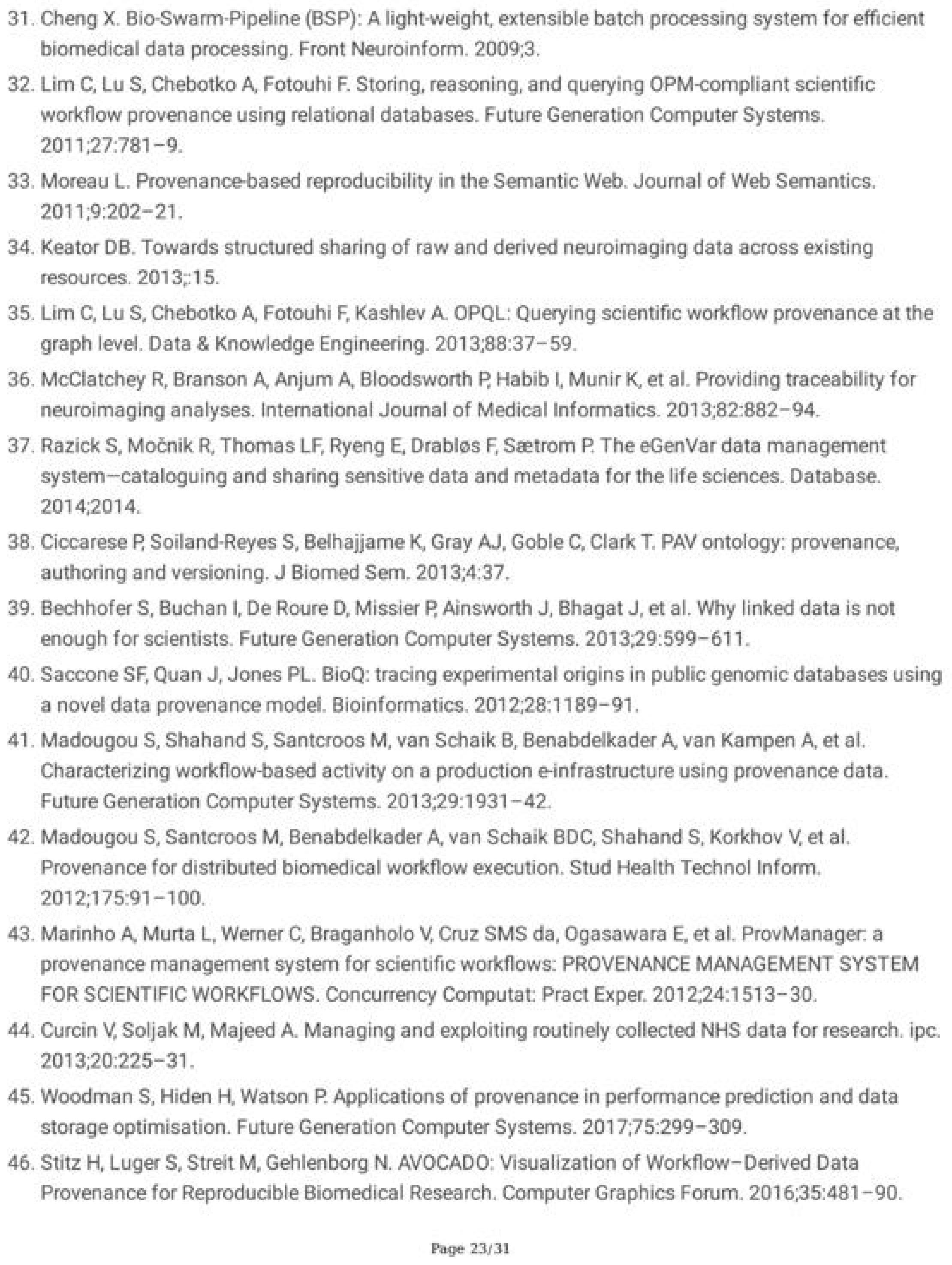
References

1. Curcin V, Miles S, Danger R, Chen Y, Bache R, Taweel A. Implementing interoperable provenance in biomedical research. Future Generation Computer Systems. 2014;34:1–16.
2. Jayapandian CP, Zhao M, Ewing RM, Zhang G-Q, Sahoo SS. A semantic proteomics dashboard (SemPoD) for data management in translational research. BMC Syst Biol. 2012;6 Suppl 3:S20.
3. Cuggia M, Combes S. The French Health Data Hub and the German Medical Informatics Initiatives: Two National Projects to Promote Data Sharing in Healthcare. Yearb Med Inform. 2019;28:195–202.
4. Curcin V. Embedding data provenance into the Learning Health System to facilitate reproducible research. Learning Health Systems. 2017;1:e10019.
5. Johnson KE, Kamineni A, Fuller S, Olmstead D, Wernli KJ. How the Provenance of Electronic Health Record Data Matters for Research: A Case Example Using System Mapping. eGEMs. 2014;2:4.
6. Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3:160018.
7. Inau ET, Sack J, Waltemath D, Zeleke AA. Initiatives, Concepts, and Implementation Practices of FAIR (Findable, Accessible, Interoperable, and Reusable) Data Principles in Health Data Stewardship Practice: Protocol for a Scoping Review. JMIR Res Protoc. 2021;10:e22505.
8. Jauer M-L, Deserno TM. Data Provenance Standards and Recommendations for FAIR Data. Stud Health Technol Inform. 2020;270:1237–8.
9. Hume S, Sarnikar S, Noteboom C. Enhancing Traceability in Clinical Research Data through a Metadata Framework. Methods Inf Med. 2020;59:075–85.
10. Sahoo SS, Nguyen V, Bodenreider O, Parikh P, Minning T, Sheth AP. A unified framework for managing provenance information in translational research. BMC Bioinformatics. 2011;12:461.
11. Khan FZ, Soiland-Reyes S, Sinnott RO, Lonie A, Goble C, Crusoe MR. Sharing interoperable workflow provenance: A review of best practices and their practical application in CWLProv. GigaScience. 2019;8:giz095.
12. PROV-Overview. https://www.w3.org/TR/prov-overview/. Accessed 9 Dec 2022.
13. Yakutovich AV, Eimre K, Schütt O, Talirz L, Adorf CS, Andersen CW, et al. AiiDAlab – an ecosystem for developing, executing, and sharing scientific workflows. Computational Materials Science.

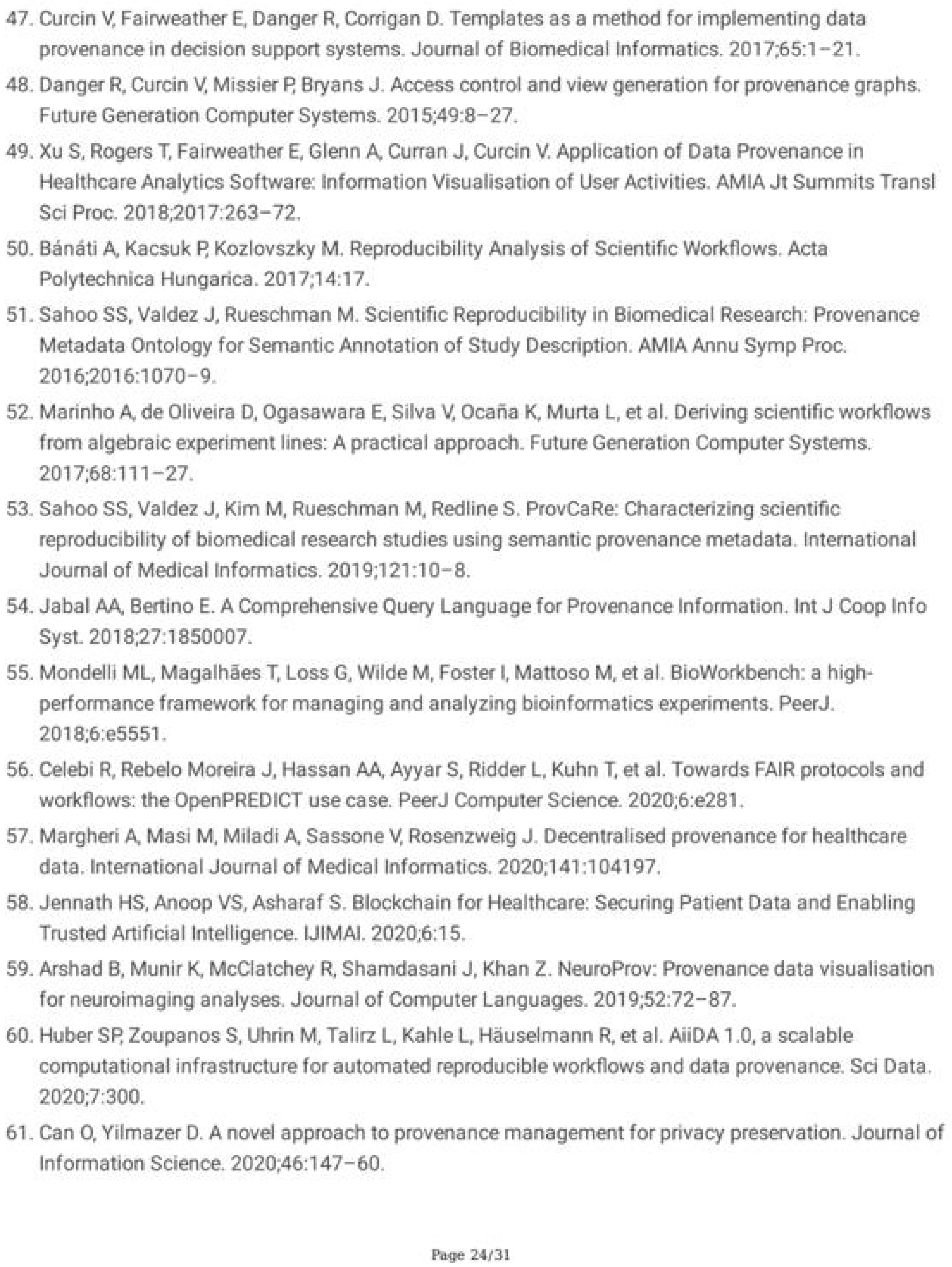


2021;188:110165.

14. Schissel DP, Abla G, Flanagan SM, Greenwald M, Lee X, Romosan A, et al. Automated metadata, provenance cataloging and navigable interfaces: Ensuring the usefulness of extreme-scale data. FUSION ENGINEERING AND DESIGN. 2014;89:745–9.
15. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. International Journal of Social Research Methodology. 2005;8:19–32.
16. Gierend K, Krüger F, Waltemath D, Fünfgeld M, Ganslandt T, Zeleke AA. Approaches and Criteria for Provenance in Biomedical Data Sets and Workflows: Protocol for a Scoping Review. JMIR Res Protoc. 2021;10:e31750.
17. Braun V, Clarke V. Using thematic analysis in psychology. Qualitative Research in Psychology. 2006;3:77–101.
18. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan—a web and mobile app for systematic reviews. Syst Rev. 2016;5:210.
19. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. Ann Intern Med. 2018;169:467–73.
20. Van Rossum G, Drake FL. Python 3 Reference Manual. Scotts Valley, CA: CreateSpace; 2009.
21. R: The R Project for Statistical Computing. https://www.r-project.org/. Accessed 13 Dec 2022.
22. Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, et al. Welcome to the Tidyverse. JOSS. 2019;4:1686.
23. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;:n71.
24. Bowers S, McPhillips TM, Ludäscher B. Provenance in collection-oriented scientific workflows. Concurrency Computat: Pract Exper. 2008;20:519–29.
25. Kim J, Deelman E, Gil Y, Mehta G, Ratnakar V. Provenance trails in the Wings/Pegasus system. Concurrency Computat: Pract Exper. 2008;20:587–97.
26. Holland DA, Seltzer MI, Braun U, Muniswamy-Reddy K-K. PASSing the provenance challenge. Concurrency Computat: Pract Exper. 2008;20:531–40.
27. Golbeck J, Hendler J. A Semantic Web approach to the provenance challenge. Concurrency Computat: Pract Exper. 2008;20:431–9.
28. Schuchardt KL, Gibson T, Stephan E, Chin G. Applying content management to automated provenance capture. Concurrency Computat: Pract Exper. 2008;20:541–54.
29. Sahoo SS, Bodenreider O, Hitzler P, Sheth A, Thirunarayan K. Provenance Context Entity (PaCE): Scalable Provenance Tracking for Scientific RDF Data. In: Gertz M, Ludäscher B, editors. Scientific and Statistical Database Management. Berlin, Heidelberg: Springer Berlin Heidelberg; 2010. p. 461–70.
30. Groth P, Moreau L. Representing distributed systems using the Open Provenance Model. Future Generation Computer Systems. 2011;27:757–65.



31. Cheng X. Bio-Swarm-Pipeline (BSP): A light-weight, extensible batch processing system for efficient biomedical data processing. Front Neuroinform. 2009;3.
32. Lim C, Lu S, Chebotko A, Fotouhi F. Storing, reasoning, and querying OPM-compliant scientific workflow provenance using relational databases. Future Generation Computer Systems. 2011;27:781–9.
33. Moreau L. Provenance-based reproducibility in the Semantic Web. Journal of Web Semantics. 2011;9:202–21.
34. Keator DB. Towards structured sharing of raw and derived neuroimaging data across existing resources. 2013;:15.
35. Lim C, Lu S, Chebotko A, Fotouhi F, Kashlev A. OPQL: Querying scientific workflow provenance at the graph level. Data & Knowledge Engineering. 2013;88:37–59.
36. McClatchey R, Branson A, Anjum A, Bloodsworth P, Habib I, Munir K, et al. Providing traceability for neuroimaging analyses. International Journal of Medical Informatics. 2013;82:882–94.
37. Razick S, Močnik R, Thomas LF, Ryeng E, Drabløs F, Sætrom P. The eGenVar data management system—cataloguing and sharing sensitive data and metadata for the life sciences. Database. 2014;2014.
38. Ciccarese P, Soiland-Reyes S, Belhajjame K, Gray AJ, Goble C, Clark T. PAV ontology: provenance, authoring and versioning. J Biomed Sem. 2013;4:37.
39. Bechhofer S, Buchan I, De Roure D, Missier P, Ainsworth J, Bhagat J, et al. Why linked data is not enough for scientists. Future Generation Computer Systems. 2013;29:599–611.
40. Saccone SF, Quan J, Jones PL. BioQ: tracing experimental origins in public genomic databases using a novel data provenance model. Bioinformatics. 2012;28:1189–91.
41. Madougou S, Shahand S, Santcroos M, van Schaik B, Benabdelkader A, van Kampen A, et al. Characterizing workflow-based activity on a production e-infrastructure using provenance data. Future Generation Computer Systems. 2013;29:1931–42.
42. Madougou S, Santcroos M, Benabdelkader A, van Schaik BDC, Shahand S, Korkhov V, et al. Provenance for distributed biomedical workflow execution. Stud Health Technol Inform. 2012;175:91–100.
43. Marinho A, Murta L, Werner C, Braganholo V, Cruz SMS da, Ogasawara E, et al. ProvManager: a provenance management system for scientific workflows: PROVENANCE MANAGEMENT SYSTEM FOR SCIENTIFIC WORKFLOWS. Concurrency Computat: Pract Exper. 2012;24:1513–30.
44. Curcin V, Soljak M, Majeed A. Managing and exploiting routinely collected NHS data for research. ipc. 2013;20:225–31.
45. Woodman S, Hiden H, Watson P. Applications of provenance in performance prediction and data storage optimisation. Future Generation Computer Systems. 2017;75:299–309.
46. Stitz H, Luger S, Streit M, Gehlenborg N. AVOCADO: Visualization of Workflow–Derived Data Provenance for Reproducible Biomedical Research. Computer Graphics Forum. 2016;35:481–90.



47. Curcin V, Fairweather E, Danger R, Corrigan D. Templates as a method for implementing data provenance in decision support systems. Journal of Biomedical Informatics. 2017;65:1–21.
48. Danger R, Curcin V, Missier P, Bryans J. Access control and view generation for provenance graphs. Future Generation Computer Systems. 2015;49:8–27.
49. Xu S, Rogers T, Fairweather E, Glenn A, Curran J, Curcin V. Application of Data Provenance in Healthcare Analytics Software: Information Visualisation of User Activities. AMIA Jt Summits Transl Sci Proc. 2018;2017:263–72.
50. Bánáti A, Kacsuk P, Kozlovszky M. Reproducibility Analysis of Scientific Workflows. Acta Polytechnica Hungarica. 2017;14:17.
51. Sahoo SS, Valdez J, Rueschman M. Scientific Reproducibility in Biomedical Research: Provenance Metadata Ontology for Semantic Annotation of Study Description. AMIA Annu Symp Proc. 2016;2016:1070–9.
52. Marinho A, de Oliveira D, Ogasawara E, Silva V, Ocaña K, Murta L, et al. Deriving scientific workflows from algebraic experiment lines: A practical approach. Future Generation Computer Systems. 2017;68:111–27.
53. Sahoo SS, Valdez J, Kim M, Rueschman M, Redline S. ProvCaRe: Characterizing scientific reproducibility of biomedical research studies using semantic provenance metadata. International Journal of Medical Informatics. 2019;121:10–8.
54. Jabal AA, Bertino E. A Comprehensive Query Language for Provenance Information. Int J Coop Info Syst. 2018;27:1850007.
55. Mondelli ML, Magalhães T, Loss G, Wilde M, Foster I, Mattoso M, et al. BioWorkbench: a high-performance framework for managing and analyzing bioinformatics experiments. PeerJ. 2018;6:e5551.
56. Celebi R, Rebelo Moreira J, Hassan AA, Ayyar S, Ridder L, Kuhn T, et al. Towards FAIR protocols and workflows: the OpenPREDICT use case. PeerJ Computer Science. 2020;6:e281.
57. Margheri A, Masi M, Miladi A, Sassone V, Rosenzweig J. Decentralised provenance for healthcare data. International Journal of Medical Informatics. 2020;141:104197.
58. Jennath HS, Anoop VS, Asharaf S. Blockchain for Healthcare: Securing Patient Data and Enabling Trusted Artificial Intelligence. IJIMAI. 2020;6:15.
59. Arshad B, Munir K, McClatchey R, Shamdasani J, Khan Z. NeuroProv: Provenance data visualisation for neuroimaging analyses. Journal of Computer Languages. 2019;52:72–87.
60. Huber SP, Zoupanos S, Uhrin M, Talirz L, Kahle L, Häuselmann R, et al. AiiDA 1.0, a scalable computational infrastructure for automated reproducible workflows and data provenance. Sci Data. 2020;7:300.
61. Can O, Yilmazer D. A novel approach to provenance management for privacy preservation. Journal of Information Science. 2020;46:147–60.

62. Monnin P, Legrand J, Husson G, Ringot P, Tchechmedjiev A, Jonquet C, et al. PGxO and PGxLOD: a reconciliation of pharmacogenomic knowledge of various provenances, enabling further comparison. BMC Bioinformatics. 2019;20:139.
63. Guedes T, Martins LB, Falci MLF, Silva V, Ocaña KACS, Mattoso M, et al. Capturing and Analyzing Provenance from Spark-based Scientific Workflows with SAMbA-RaP. Future Generation Computer Systems. 2020;112:658–69.
64. Alterovitz G, Dean D, Goble C, Crusoe MR, Soiland-Reyes S, Bell A, et al. Enabling precision medicine via standard communication of HTS provenance, analysis, and results. PLoS Biol. 2018;16:e3000099.
65. Parciak M, Bauer C, Bender T, Lodahl R, Schreiweis B, Tute E, et al. Provenance Solutions for Medical Research in Heterogeneous IT-Infrastructure: An Implementation Roadmap. Stud Health Technol Inform. 2019;264:298–302.
66. Danese MD, Halperin M, Duryea J, Duryea R. The Generalized Data Model for clinical research. BMC Med Inform Decis Mak. 2019;19:117.
67. Pimentel JF, Freire J, Murta L, Braganholo V. A Survey on Collecting, Managing, and Analyzing Provenance from Scripts. ACM Comput Surv. 2019;52:1–38.
68. Ornelas T, Braga R, David JMN, Campos F, Castro G. Provenance data discovery through Semantic Web resources. Concurrency Computat Pract Exper. 2018;30:e4366.
69. Daumke P, Heitmann KU, Heckmann S, Martínez-Costa C, Schulz S. Clinical Text Mining on FHIR. Stud Health Technol Inform. 2019;264:83–7.
70. Tyndall T, Tyndall A. FHIR Healthcare Directories: Adopting Shared Interfaces to Achieve Interoperable Medical Device Data Integration. Stud Health Technol Inform. 2018;249:181–4.
71. Thavasimani P, Cala J, Missier P. Why-Diff: Exploiting Provenance to Understand Outcome Differences From Non-Identical Reproduced Workflows. IEEE Access. 2019;7:34973–90.
72. Moreau L, Freire J, Futrelle J, McGrath RE, Myers J, Paulson P. The Open Provenance Model: An Overview. In: Freire J, Koop D, Moreau L, editors. Provenance and Annotation of Data and Processes. Berlin, Heidelberg: Springer Berlin Heidelberg; 2008. p. 323–6.
73. Schröder M, Staehlke S, Groth P, Nebe JB, Spors S, Krüger F. Structure-based knowledge acquisition from electronic lab notebooks for research data provenance documentation. J Biomed Semant. 2022;13:4.
74. Wing JM. The Data Life Cycle. Harvard Data Science Review. 2019. https://doi.org/10.1162/99608f92.e26845b4.
75. :00-17:00. ISO 8000-2:2022. ISO. https://www.iso.org/standard/85032.html. Accessed 13 Dec 2022.
76. Peng G. The State of Assessing Data Stewardship Maturity – An Overview. Data Science Journal. 2018;17:7.

Figures

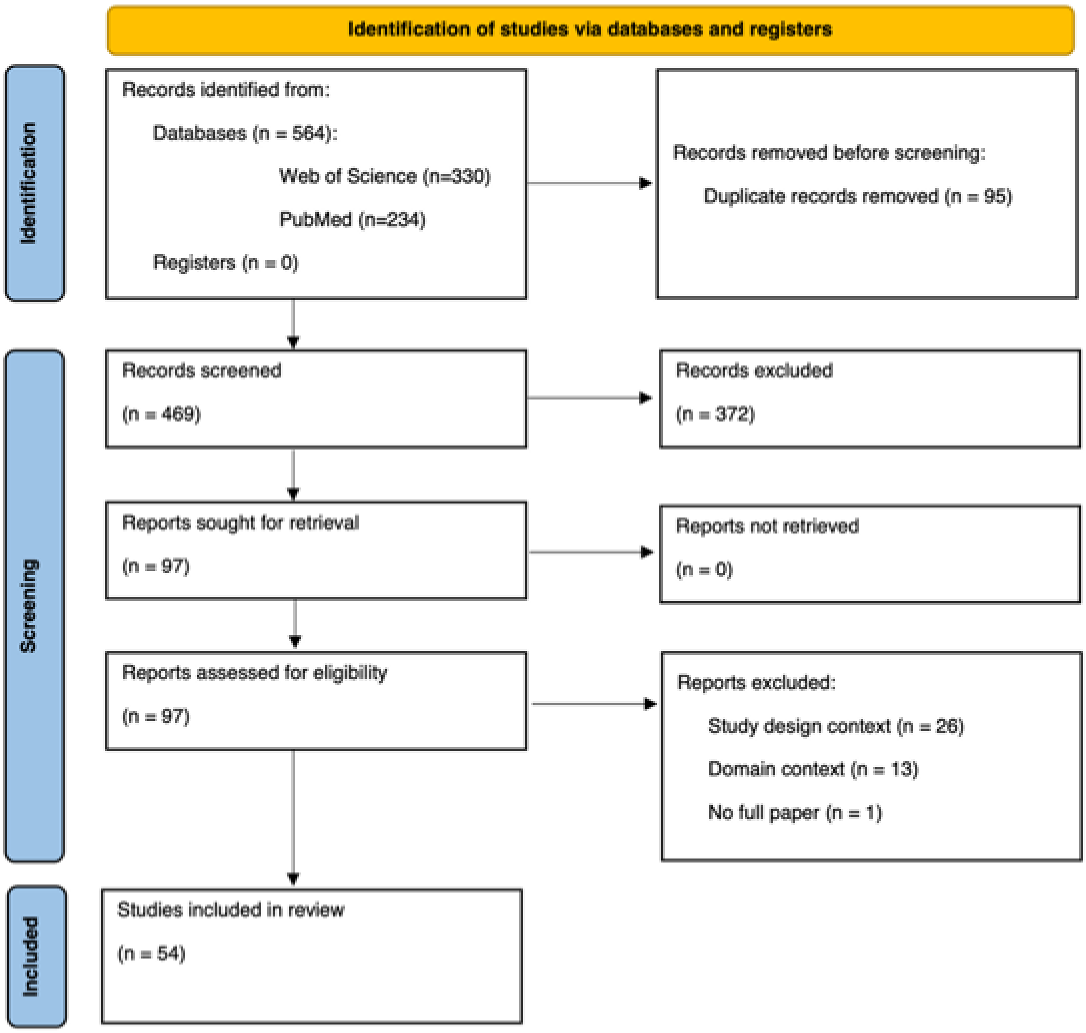


Figure 1

Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flow diagram

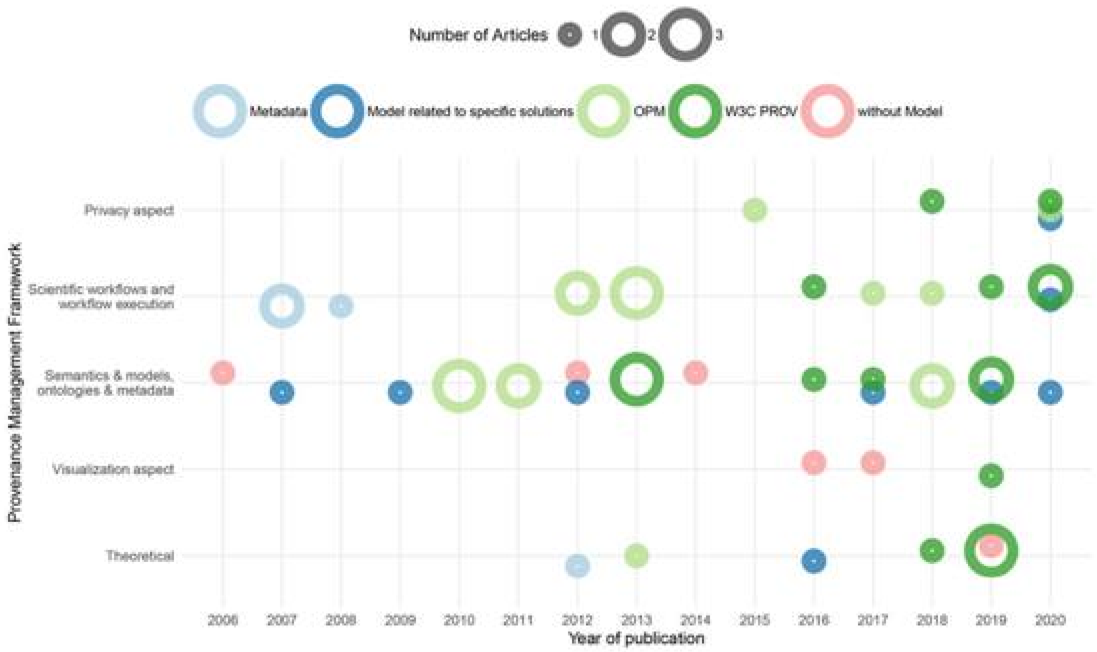


Figure 2

Reported provenance management frameworks per year.

The size of the ring corresponds with the number of articles per year that discuss a specific model (color-coded) in the context of the respective framework.

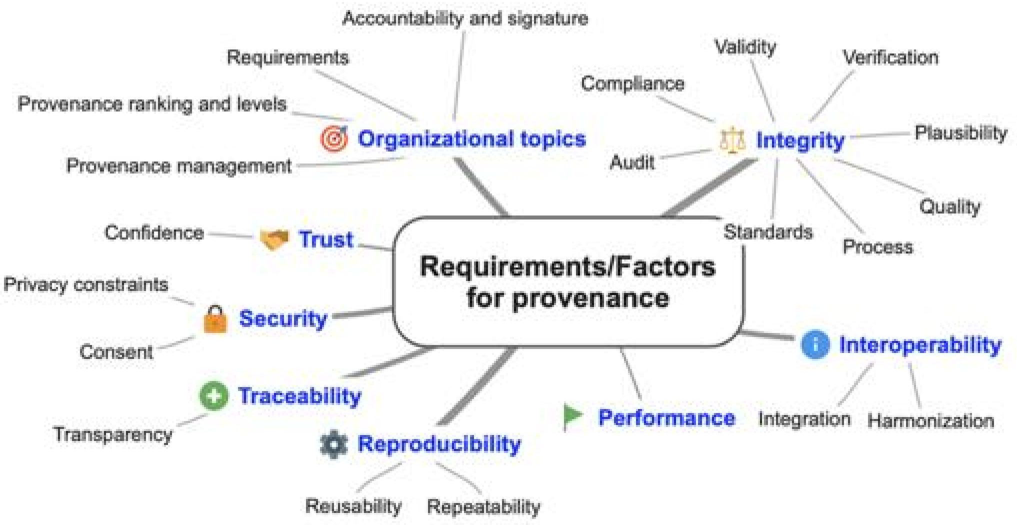


Figure 3

Reported provenance requirements/factors by word fields.

The line thickness in the first level proportionally reflects the respective characteristics count. The second level displays all occurred requirement classes.

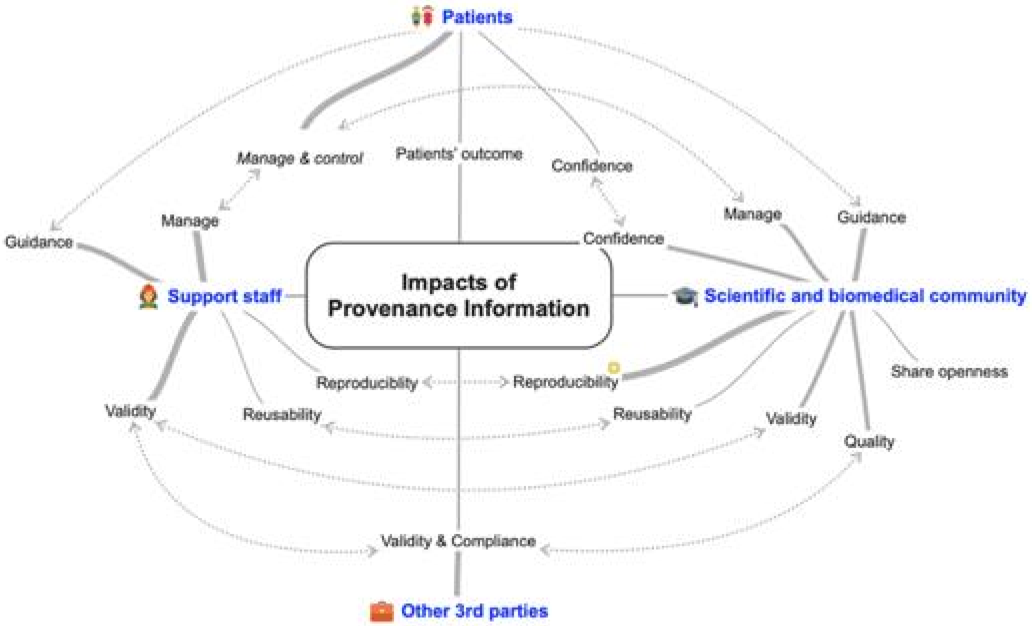


Figure 4

Reported impacts of provenance information.

Level one presents the stakeholder groups, level two the impacts on the stakeholders. The line thickness in the second level proportionally reflects the respective counts of the characteristics. An additional file provides details about the structure and relationship between the individual stakeholder groups and the reported impacts [see Additional File 3].

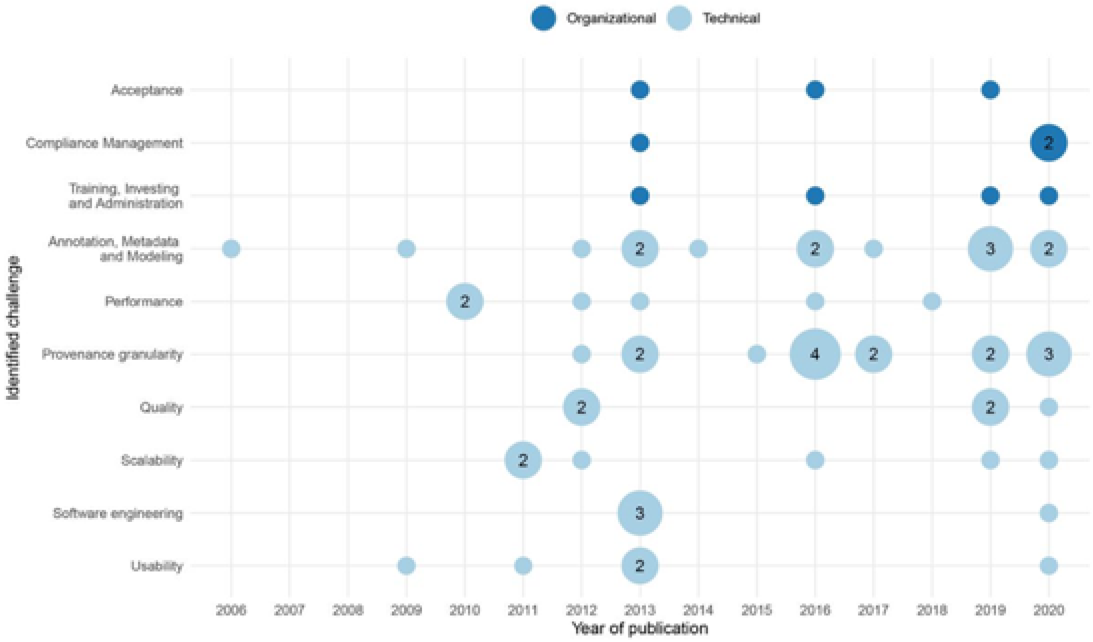


Figure 5

Challenges per year of publication.

The size of and the numbers in the circles represent the number of articles that reveal a challenge (color-coded). Note that numbers are omitted for single articles per category.

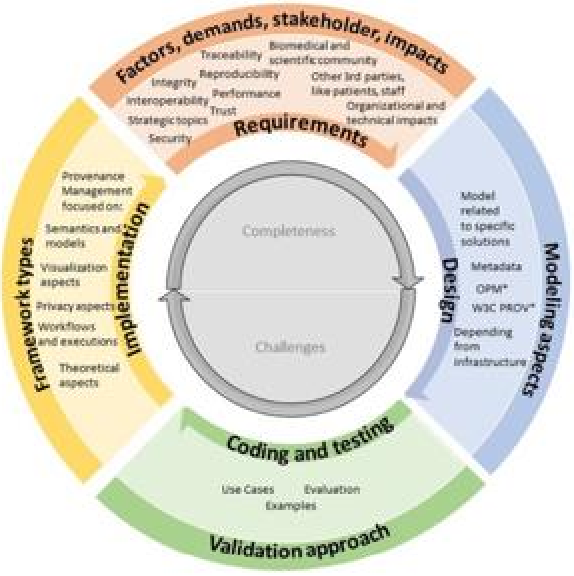


Figure 6

Roadmap for a tailor-made provenance framework (Provenance-SFL)

The roadmap for the tailor-made provenance framework (Provenance-SFL) shows the four major processing phases in the inner circle segments: starting with the requirements definition, set-up of the design based on the requirements, followed by coding and testing phase related to the given requirements and the implementation after successful testing. The outer and innermost circle present the mapped sections from our research questions approach to the Provenance-SFL.

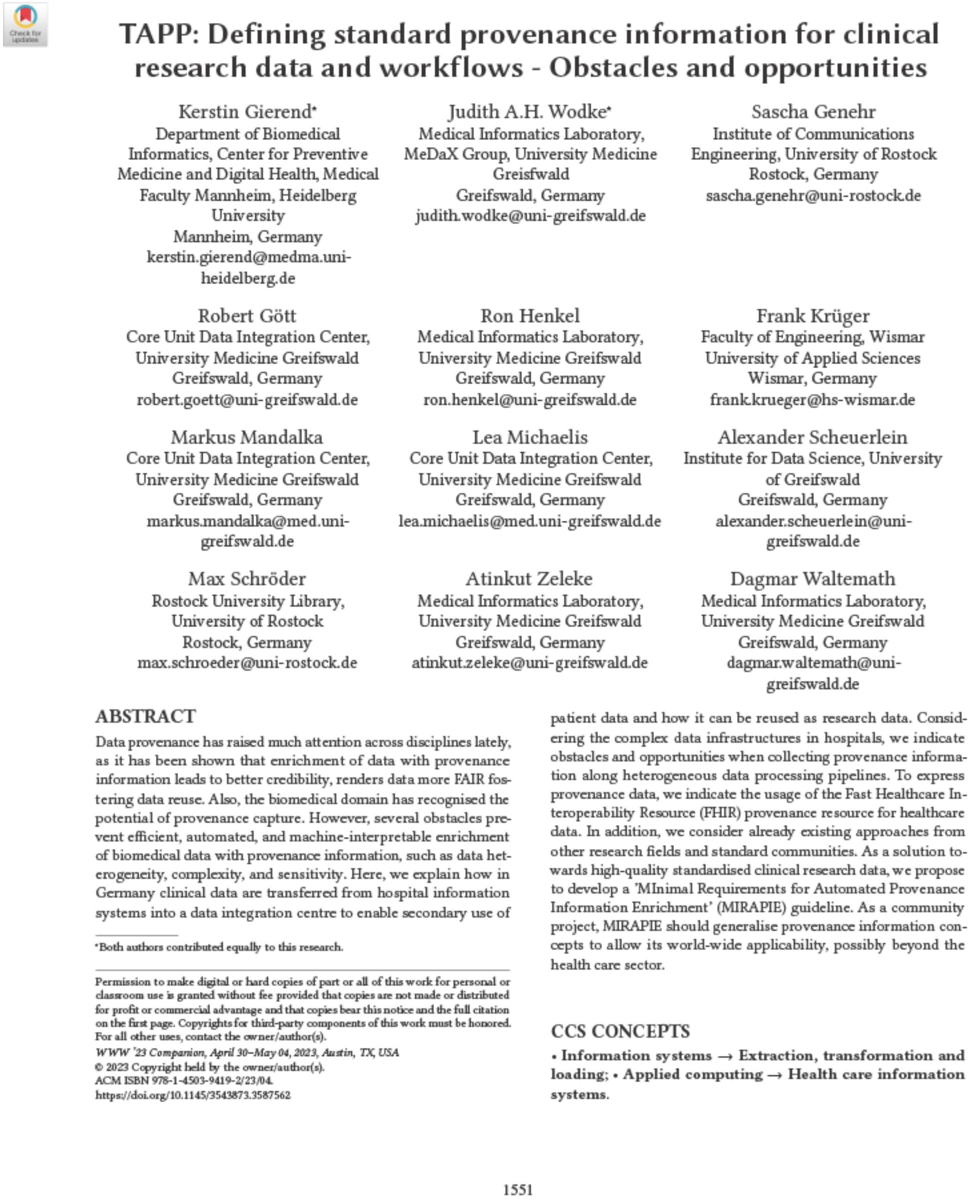
Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- AdditionalFile1.docx
- AdditionalFile2.docx
- AdditionalFile3.docx
- AdditionalFile4.docx

3.5 Publication 5: TAPP: Defining standard provenance information for clinical research data and workflows - Obstacles and opportunities

This section contains a conference (WWW ’23 Companion: Companion Proceedings of the ACM Web Conference 2023) contribution, as originally published in the Association for Computing Machinery Digital Library (ACM DL), an international, peer-reviewed, and open access publication.
The original publication is available at ACM Digital Library (https://doi.org/10.1145/3543873.3587562).



TAPP: Defining standard provenance information for clinical research data and workflows - Obstacles and opportunities

Kerstin Gierend*
Department of Biomedical Informatics, Center for Preventive Medicine and Digital Health, Medical Faculty Mannheim, Heidelberg University
Mannheim, Germany
kerstin.gierend@medma.uni-heidelberg.de

Judith A.H. Wodke*
Medical Informatics Laboratory, MeDaX Group, University Medicine Greisfwald
Greifswald, Germany
judith.wodke@uni-greifswald.de

Sascha Genehr
Institute of Communications Engineering, University of Rostock
Rostock, Germany
sascha.genehr@uni-rostock.de

Robert Gött
Core Unit Data Integration Center, University Medicine Greifswald
Greifswald, Germany
robert.goett@uni-greifswald.de

Ron Henkel
Medical Informatics Laboratory, University Medicine Greifswald
Greifswald, Germany
ron.henkel@uni-greifswald.de

Frank Krüger
Faculty of Engineering, Wismar University of Applied Sciences
Wismar, Germany
frank.krueger@hs-wismar.de

Markus Mandalka
Core Unit Data Integration Center, University Medicine Greifswald
Greifswald, Germany
markus.mandalka@med.uni-greifswald.de

Lea Michaelis
Core Unit Data Integration Center, University Medicine Greifswald
Greifswald, Germany
lea.michaelis@med.uni-greifswald.de

Alexander Scheuerlein
Institute for Data Science, University of Greifswald
Greifswald, Germany
alexander.scheuerlein@uni-greifswald.de

Max Schröder
Rostock University Library, University of Rostock
Rostock, Germany
max.schroeder@uni-rostock.de

Atinkut Zeleke
Medical Informatics Laboratory, University Medicine Greifswald
Greifswald, Germany
atinkut.zeleke@uni-greifswald.de

Dagmar Waltemath
Medical Informatics Laboratory, University Medicine Greifswald
Greifswald, Germany
dagmar.waltemath@uni-greifswald.de

ABSTRACT

Data provenance has raised much attention across disciplines lately, as it has been shown that enrichment of data with provenance information leads to better credibility, renders data more FAIR fostering data reuse. Also, the biomedical domain has recognised the potential of provenance capture. However, several obstacles prevent efficient, automated, and machine-interpretable enrichment of biomedical data with provenance information, such as data heterogeneity, complexity, and sensitivity. Here, we explain how in Germany clinical data are transferred from hospital information systems into a data integration centre to enable secondary use of patient data and how it can be reused as research data. Considering the complex data infrastructures in hospitals, we indicate obstacles and opportunities when collecting provenance information along heterogeneous data processing pipelines. To express provenance data, we indicate the usage of the Fast Healthcare Interoperability Resource (FHIR) provenance resource for healthcare data. In addition, we consider already existing approaches from other research fields and standard communities. As a solution towards high-quality standardised clinical research data, we propose to develop a 'MInimal Requirements for Automated Provenance Information Enrichment' (MIRAPIE) guideline. As a community project, MIRAPIE should generalise provenance information concepts to allow its world-wide applicability, possibly beyond the health care sector.

CCS CONCEPTS

• Information systems → Extraction, transformation and loading; • Applied computing → Health care information systems.

*Both authors contributed equally to this research.

WWW '23 Companion, April 30–May 04, 2023, Austin, TX, USA
© 2023 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9419-2/23/04.
https://doi.org/10.1145/3543873.3587562

KEYWORDS

Data Integration Center, provenance capture, biomedical data, Hospital Information System

ACM Reference Format:
Kerstin Gierend, Judith A.H. Wodke, Sascha Genehr, Robert Gött, Ron Henkel, Frank Krüger, Markus Mandalka, Lea Michaelis, Alexander Scheuerlein, Max Schröder, Atinkut Zeleke, and Dagmar Waltemath. 2023. TAPP: Defining standard provenance information for clinical research data and workflows - Obstacles and opportunities. In *Companion Proceedings of the ACM Web Conference 2023 (WWW '23 Companion), April 30–May 04, 2023, Austin, TX, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3543873.3587562

1 INTRODUCTION

The Medical Informatics Initiative (MII) Germany pushes digitisation and interoperability of clinical routine data in Germany. Towards this ambitious goal, university clinics set up data integration centers (DIZ) to provide data management and data-related services [19]. A DIZ is responsible for i) data collection from clinical information systems, ii) establishing data warehouses for harmonised data storage, and iii) controlled release of data based on a MII-standardised Broad Consent, which is managed together with identities and pseudonyms by the DIZ-independent local Trusted Third Party (TTP). To achieve interoperability at the syntax level, the MII uses standardised HL7 FHIR resources [4]. Data comparability is warranted for the MII Core Data Set (CDS) [6], a minimum set of data items each DIZ should cover. Today, DIZ data is a valuable data pool for cross-site clinical research in Germany [10]. Considerable data processing is, however, necessary to transfer data from clinical source systems to the DIZ and subsequently, after successful application for data usage, to Transfer Office and researchers (Figure 1). Despite being provided in standardised HL7 FHIR format, data items of the same type may have undergone differing processing steps respective of the DIZ they were handled in. This results in traceability and reproducibility issues due to a lack of contextual information within non-harmonized workflow steps, unclear responsibilities, missing or incomplete data elements (DEs), and incomplete information on the computational environment. In this setting, data provenance information promotes transparency throughout data processing [7]. Domain-specific concepts for data quality assessment and assurance aim to improve the situation [13]. As an intermediate step towards standardisation of clinical source systems and data processing pipelines, we propose a community project to define a "MInimal Requirements for Automated Provenance Information Enrichment" (MIRAPIE) guideline. This guideline could help rendering heterogeneously processed biomedical data more credible, better interpretable, and comparable.

2 REQUIREMENTS FOR PROVENANCE CAPTURE IN A DIZ

Provenance information tracks and documents the origin, ownership, processing, and custody of data throughout their life cycle. It renders data more FAIR and trustworthy by providing measures for comparison after multiple, often divergent processing steps. A syntax for provenance expression is the HL7 FHIR *Provenance* resource [3]. To define provenance requirements, profound knowledge on the underlying data management processes and the DEs is needed to prevent loss of context. As clinical DEs undergo multiple transformation processes from data generation to authorised data release in a DIZ, it is essential to enrich these data with provenance information during the different process phases. Especially when working with pseudonymised patient data, where a loss of context possibly renders the entire data item not reusable, provenance enrichment is essential. A recently published study outlined the role and possible dimensions of the associated metadata [7]. Here we focus on the storage and the processing structures implemented at the DIZ of the University Medicine Greifswald (UMG-DIZ) and of the University Medicine Mannheim (UMM-DIZ). The structures require provenance annotation during three processing phases (Figure 1):

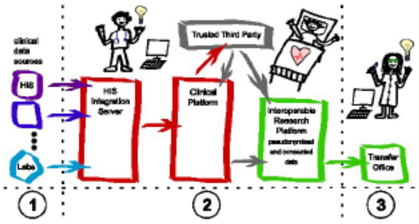
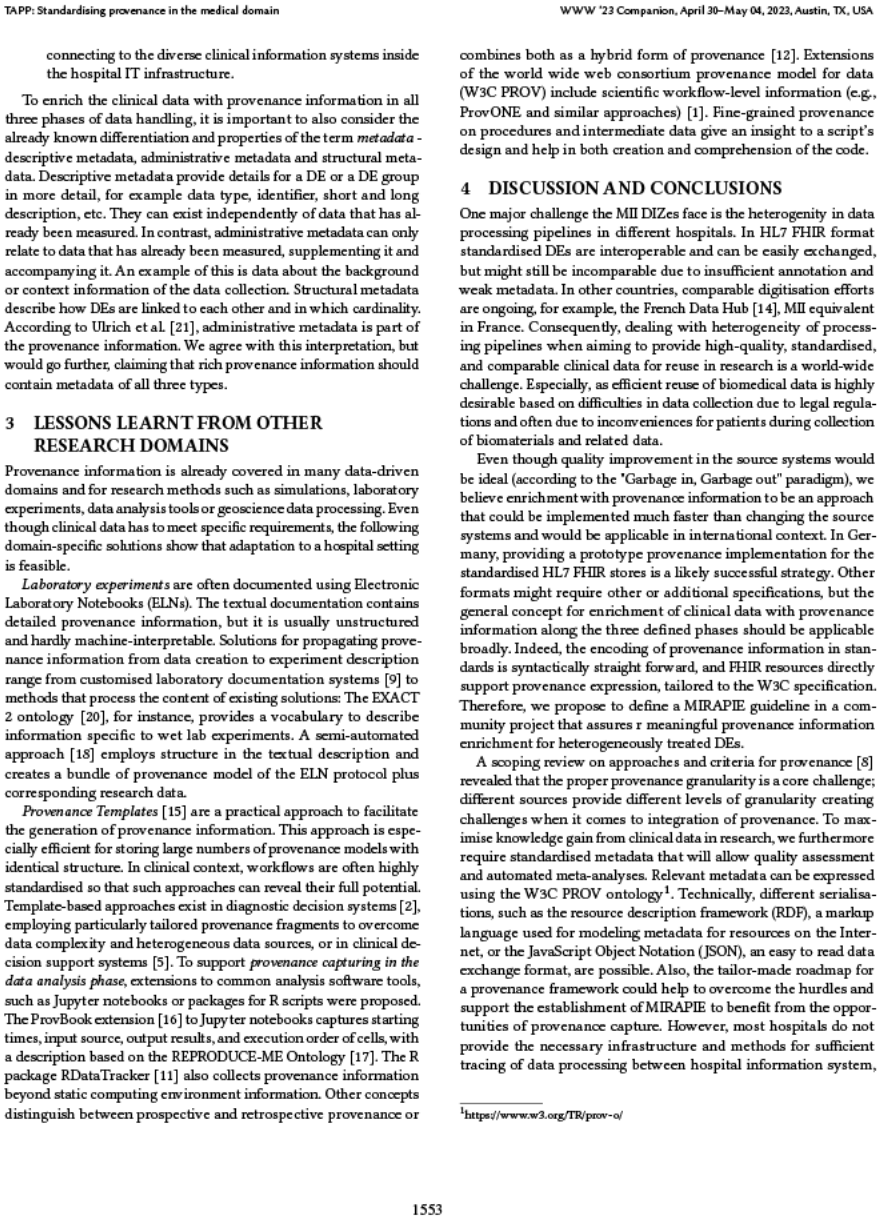


Figure 1: Three phases of provenance enrichment from capture in heterogeneous clinical source systems (1) via extract, transform, load (ETL) processing for storage and internal access in a data integration centre (DIZ) (2) to research data provision (3), exemplified for DIZes of University Medicine Greifswald (UMG-DIZ) and of University Medicine Mannheim (UMM-DIZ).

(1) **Data capturing in clinical source systems.** Heterogeneous source systems in clinical laboratories (from biological samples to digital testing results), Health electronic Case Report Form (eCRF; e.g., anamnesis, mostly structured), diagnosis, therapy, image generation processes, billing information etc. lead to different provenance granularity and availability.
(2) **Data storage and processing.** Transfer from hospital IT infrastructure into research data infrastructure via extract, transform, load (ETL) processes. Medical data generated by data sources are repeatedly rearranged during successive ETL steps. If necessary, DEs must be converted (data type compatibility), aggregated (calculated values) or split (semi-structured data). All DE processing should be described and made available by provenance data.
(3) **Data transfer.** Upon request and after permission by a *use and access committee*, data is transferred to the researcher. Further provenance data is generated, such as project application, timestamp of data provision, version of data release, data contained in the data package, etc. Provenance information in phase 3 can be captured inside the DIZ, without



connecting to the diverse clinical information systems inside the hospital IT infrastructure.

To enrich the clinical data with provenance information in all three phases of data handling, it is important to also consider the already known differentiation and properties of the term *metadata* - descriptive metadata, administrative metadata and structural metadata. Descriptive metadata provide details for a DE or a DE group in more detail, for example data type, identifier, short and long description, etc. They can exist independently of data that has already been measured. In contrast, administrative metadata can only relate to data that has already been measured, supplementing it and accompanying it. An example of this is data about the background or context information of the data collection. Structural metadata describe how DEs are linked to each other and in which cardinality. According to Ulrich et al. [21], administrative metadata is part of the provenance information. We agree with this interpretation, but would go further, claiming that rich provenance information should contain metadata of all three types.

3 LESSONS LEARNT FROM OTHER RESEARCH DOMAINS

Provenance information is already covered in many data-driven domains and for research methods such as simulations, laboratory experiments, data analysis tools or geoscience data processing. Even though clinical data has to meet specific requirements, the following domain-specific solutions show that adaptation to a hospital setting is feasible.

Laboratory experiments are often documented using Electronic Laboratory Notebooks (ELNs). The textual documentation contains detailed provenance information, but it is usually unstructured and hardly machine-interpretable. Solutions for propagating provenance information from data creation to experiment description range from customised laboratory documentation systems [9] to methods that process the content of existing solutions: The EXACT 2 ontology [20], for instance, provides a vocabulary to describe information specific to wet lab experiments. A semi-automated approach [18] employs structure in the textual description and creates a bundle of provenance model of the ELN protocol plus corresponding research data.

Provenance Templates [15] are a practical approach to facilitate the generation of provenance information. This approach is especially efficient for storing large numbers of provenance models with identical structure. In clinical context, workflows are often highly standardised so that such approaches can reveal their full potential. Template-based approaches exist in diagnostic decision systems [2], employing particularly tailored provenance fragments to overcome data complexity and heterogeneous data sources, or in clinical decision support systems [5]. To support *provenance capturing in the data analysis phase*, extensions to common analysis software tools, such as Jupyter notebooks or packages for R scripts were proposed. The ProvBook extension [16] to Jupyter notebooks captures starting times, input source, output results, and execution order of cells, with a description based on the REPRODUCE-ME Ontology [17]. The R package RDataTracker [11] also collects provenance information beyond static computing environment information. Other concepts distinguish between prospective and retrospective provenance or combines both as a hybrid form of provenance [12]. Extensions of the world wide web consortium provenance model for data (W3C PROV) include scientific workflow-level information (e.g., ProvONE and similar approaches) [1]. Fine-grained provenance on procedures and intermediate data give an insight to a script's design and help in both creation and comprehension of the code.

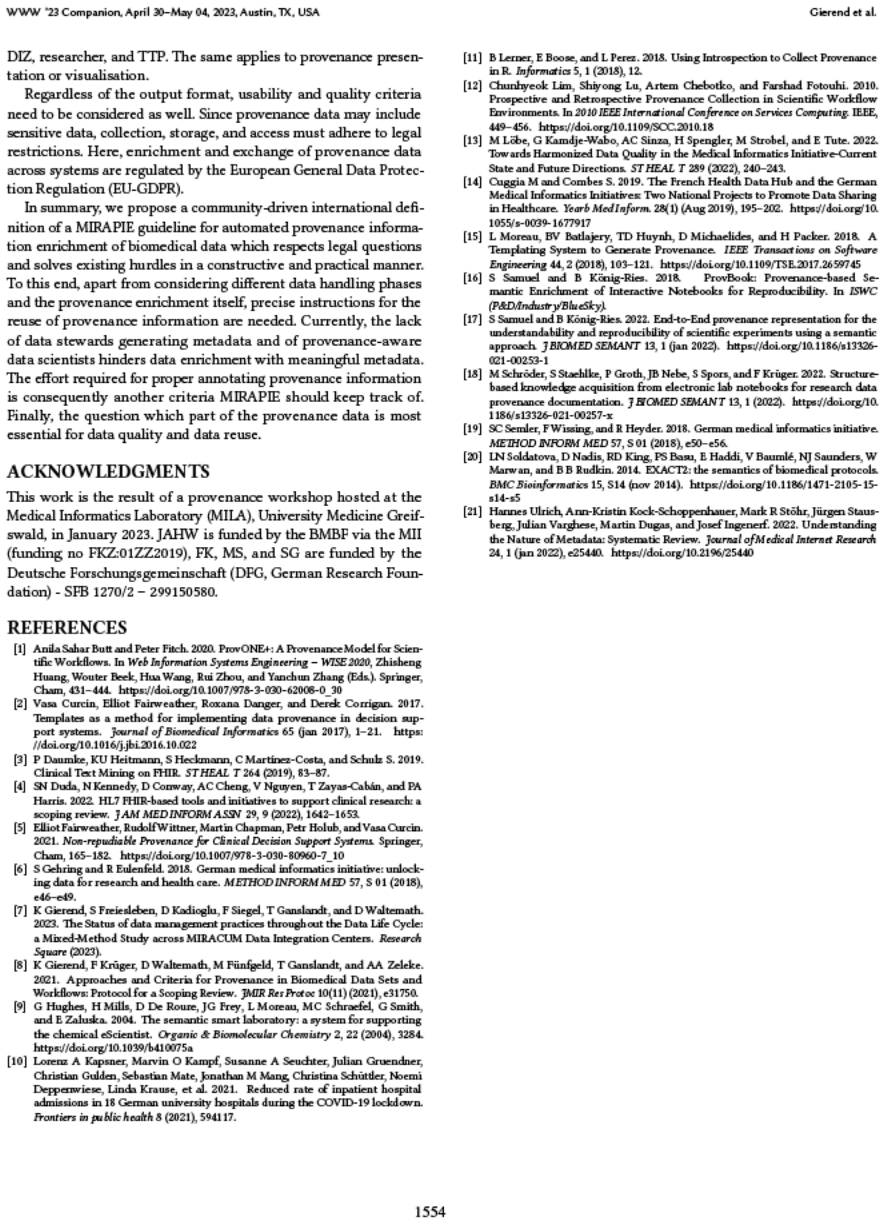
4 DISCUSSION AND CONCLUSIONS

One major challenge the MII DIZes face is the heterogenity in data processing pipelines in different hospitals. In HL7 FHIR format standardised DEs are interoperable and can be easily exchanged, but might still be incomparable due to insufficient annotation and weak metadata. In other countries, comparable digitisation efforts are ongoing, for example, the French Data Hub [14], MII equivalent in France. Consequently, dealing with heterogeneity of processing pipelines when aiming to provide high-quality, standardised, and comparable clinical data for reuse in research is a world-wide challenge. Especially, as efficient reuse of biomedical data is highly desirable based on difficulties in data collection due to legal regulations and often due to inconveniences for patients during collection of biomaterials and related data.

Even though quality improvement in the source systems would be ideal (according to the "Garbage in, Garbage out" paradigm), we believe enrichment with provenance information to be an approach that could be implemented much faster than changing the source systems and would be applicable in international context. In Germany, providing a prototype provenance implementation for the standardised HL7 FHIR stores is a likely successful strategy. Other formats might require other or additional specifications, but the general concept for enrichment of clinical data with provenance information along the three defined phases should be applicable broadly. Indeed, the encoding of provenance information in standards is syntactically straight forward, and FHIR resources directly support provenance expression, tailored to the W3C specification. Therefore, we propose to define a MIRAPIE guideline in a community project that assures r meaningful provenance information enrichment for heterogeneously treated DEs.

A scoping review on approaches and criteria for provenance [8] revealed that the proper provenance granularity is a core challenge; different sources provide different levels of granularity creating challenges when it comes to integration of provenance. To maximise knowledge gain from clinical data in research, we furthermore require standardised metadata that will allow quality assessment and automated meta-analyses. Relevant metadata can be expressed using the W3C PROV ontology[1]. Technically, different serialisations, such as the resource description framework (RDF), a markup language used for modeling metadata for resources on the Internet, or the JavaScript Object Notation (JSON), an easy to read data exchange format, are possible. Also, the tailor-made roadmap for a provenance framework could help to overcome the hurdles and support the establishment of MIRAPIE to benefit from the opportunities of provenance capture. However, most hospitals do not provide the necessary infrastructure and methods for sufficient tracing of data processing between hospital information system,

[1] https://www.w3.org/TR/prov-o/



DIZ, researcher, and TTP. The same applies to provenance presentation or visualisation.

Regardless of the output format, usability and quality criteria need to be considered as well. Since provenance data may include sensitive data, collection, storage, and access must adhere to legal restrictions. Here, enrichment and exchange of provenance data across systems are regulated by the European General Data Protection Regulation (EU-GDPR).

In summary, we propose a community-driven international definition of a MIRAPIE guideline for automated provenance information enrichment of biomedical data which respects legal questions and solves existing hurdles in a constructive and practical manner. To this end, apart from considering different data handling phases and the provenance enrichment itself, precise instructions for the reuse of provenance information are needed. Currently, the lack of data stewards generating metadata and of provenance-aware data scientists hinders data enrichment with meaningful metadata. The effort required for proper annotating provenance information is consequently another criteria MIRAPIE should keep track of. Finally, the question which part of the provenance data is most essential for data quality and data reuse.

ACKNOWLEDGMENTS

This work is the result of a provenance workshop hosted at the Medical Informatics Laboratory (MILA), University Medicine Greifswald, in January 2023. JAHW is funded by the BMBF via the MII (funding no FKZ:01ZZ2019), FK, MS, and SG are funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - SFB 1270/2 – 299150580.

REFERENCES

[1] Anila Sahar Butt and Peter Fitch. 2020. ProvONE+: A Provenance Model for Scientific Workflows. In *Web Information Systems Engineering – WISE 2020*, Zhisheng Huang, Wouter Beek, Hua Wang, Rui Zhou, and Yanchun Zhang (Eds.). Springer, Cham, 431–444. https://doi.org/10.1007/978-3-030-62008-0_30

[2] Vasa Curcin, Elliot Fairweather, Roxana Danger, and Derek Corrigan. 2017. Templates as a method for implementing data provenance in decision support systems. *Journal of Biomedical Informatics* 65 (jan 2017), 1–21. https://doi.org/10.1016/j.jbi.2016.10.022

[3] P Daumke, KU Heitmann, S Heckmann, C Martínez-Costa, and Schulz S. 2019. Clinical Text Mining on FHIR. *ST HEAL T* 264 (2019), 83–87.

[4] SN Duda, N Kennedy, D Conway, AC Cheng, V Nguyen, T Zayas-Cabán, and PA Harris. 2022. HL7 FHIR-based tools and initiatives to support clinical research: a scoping review. *J AM MED INFORM ASSN* 29, 9 (2022), 1642–1653.

[5] Elliot Fairweather, Rudolf Wittner, Martin Chapman, Petr Holub, and Vasa Curcin. 2021. *Non-repudiable Provenance for Clinical Decision Support Systems.* Springer, Cham, 165–182. https://doi.org/10.1007/978-3-030-80960-7_10

[6] S Gehring and R Eulenfeld. 2018. German medical informatics initiative: unlocking data for research and health care. *METHOD INFORM MED* 57, S 01 (2018), e46–e49.

[7] K Gierend, S Freiesleben, D Kadioglu, F Siegel, T Ganslandt, and D Waltemath. 2023. The Status of data management practices throughout the Data Life Cycle: a Mixed-Method Study across MIRACUM Data Integration Centers. *Research Square* (2023).

[8] K Gierend, F Krüger, D Waltemath, M Fünfgeld, T Ganslandt, and AA Zeleke. 2021. Approaches and Criteria for Provenance in Biomedical Data Sets and Workflows: Protocol for a Scoping Review. *JMIR Res Protoc* 10(11) (2021), e31750.

[9] G Hughes, H Mills, D De Roure, JG Frey, L Moreau, MC Schraefel, G Smith, and E Zaluska. 2004. The semantic smart laboratory: a system for supporting the chemical eScientist. *Organic & Biomolecular Chemistry* 2, 22 (2004), 3284. https://doi.org/10.1039/b410075a

[10] Lorenz A Kapsner, Marvin O Kampf, Susanne A Seuchter, Julian Gruendner, Christian Gulden, Sebastian Mate, Jonathan M Mang, Christina Schüttler, Noemi Deppenwiese, Linda Krause, et al. 2021. Reduced rate of inpatient hospital admissions in 18 German university hospitals during the COVID-19 lockdown. *Frontiers in public health* 8 (2021), 594117.

[11] B Lerner, E Boose, and L Perez. 2018. Using Introspection to Collect Provenance in R. *Informatics* 5, 1 (2018), 12.

[12] Chunhyeok Lim, Shiyong Lu, Artem Chebotko, and Farshad Fotouhi. 2010. Prospective and Retrospective Provenance Collection in Scientific Workflow Environments. In *2010 IEEE International Conference on Services Computing*. IEEE, 449–456. https://doi.org/10.1109/SCC.2010.18

[13] M Löbe, G Kamdje-Wabo, AC Sinza, H Spengler, M Strobel, and E Tute. 2022. Towards Harmonized Data Quality in the Medical Informatics Initiative-Current State and Future Directions. *ST HEAL T* 289 (2022), 240–243.

[14] Cuggia M and Combes S. 2019. The French Health Data Hub and the German Medical Informatics Initiatives: Two National Projects to Promote Data Sharing in Healthcare. *Yearb Med Inform.* 28(1) (Aug 2019), 195–202. https://doi.org/10.1055/s-0039-1677917

[15] L Moreau, BV Batlajery, TD Huynh, D Michaelides, and H Packer. 2018. A Templating System to Generate Provenance. *IEEE Transactions on Software Engineering* 44, 2 (2018), 103–121. https://doi.org/10.1109/TSE.2017.2659745

[16] S Samuel and B König-Ries. 2018. ProvBook: Provenance-based Semantic Enrichment of Interactive Notebooks for Reproducibility. In *ISWC (P&D/Industry/BlueSky)*.

[17] S Samuel and B König-Ries. 2022. End-to-End provenance representation for the understandability and reproducibility of scientific experiments using a semantic approach. *J BIOMED SEMANT* 13, 1 (jan 2022). https://doi.org/10.1186/s13326-021-00253-1

[18] M Schröder, S Staehlke, P Groth, JB Nebe, S Spors, and F Krüger. 2022. Structure-based knowledge acquisition from electronic lab notebooks for research data provenance documentation. *J BIOMED SEMANT* 13, 1 (2022). https://doi.org/10.1186/s13326-021-00257-x

[19] SC Semler, F Wissing, and R Heyder. 2018. German medical informatics initiative. *METHOD INFORM MED* 57, S 01 (2018), e50–e56.

[20] LN Soldatova, D Nadis, RD King, PS Basu, E Haddi, V Baumlé, NJ Saunders, W Marwan, and B B Rudkin. 2014. EXACT2: the semantics of biomedical protocols. *BMC Bioinformatics* 15, S14 (nov 2014). https://doi.org/10.1186/1471-2105-15-s14-s5

[21] Hannes Ulrich, Ann-Kristin Kock-Schoppenhauer, Mark R Stöhr, Jürgen Stausberg, Julian Varghese, Martin Dugas, and Josef Ingenerf. 2022. Understanding the Nature of Metadata: Systematic Review. *Journal of Medical Internet Research* 24, 1 (jan 2022), e25440. https://doi.org/10.2196/25440

4 DISCUSSION

The multilevel concept introduced in this cumulative dissertation marks the inaugural implementation of a solution for FAIR geared provenance tracking in a German medical DIC.

This dissertation extends previous provenance research from other research domains and applies it to the use case of clinical data integration in the German medical informatics community. To achieve this goal the dissertation integrated granular stakeholder knowledge from medical, data management and IT operational experts to facilitate reuse of contextual enriched patient data. The created provenance information on a data element level comprehensively increments value and limitations on secondary use of patient data disseminated through the DIC. Leveraging provenance consolidates data acceptance, the DIC position on an organizational level and strengthens its compelling accountability.

Using provenance will ultimately benefit researcher, patient's data safety and the DIC.

The studies reported here foresee a contribution to advances in sharing and reuse of biomedical and healthcare data for clinical research on an (inter-) national level in the forthcoming years. Particularly the envisioned EHDS, part of the European health data infrastructure, will benefit from provenance enriched real-world-data (RWD). Notably the challenges in healthcare, for example due to the increasing globalization, climate change and the SARS-CoV-2 pandemic, point out the high relevance in terms of health science and health policy for the establishment of high quality research databases incorporating provenance enriched data[38,39]. Careful data preparation and managing are essential for the success for the digitalization, sharing and access to quality health data in Europe[40].

A key finding from the underlying publications is that most provenance issues are related to the less governed data management practices in the DIC and to a lack of multistakeholder knowledge exchange. The studies showed that provenance capture in healthcare has not yet been implemented adequately nor homogeneously.

The mixed-method study (see publication 1), contributed significantly to the understanding and traceability of the processed data[14]. This study observed that a high degree on "black-box" processing hinders the proper uptake of provenance traces for the multiple transformed medical data elements. Furthermore, this study triggered both, the establishment, and the prototypic implementation of a maturity framework. It is conceivable that the launch of a maturity framework for data managing tailored to a DIC could be one pillar for provenance as it demonstrates that a supportive, striking, and effective approach is necessary to significantly improve steering of traceable processed data. This maturity framework indicates the provenance readiness and marks the importance of metadata management and possible bottlenecks in provenance tracking. Metadata are crucial to preserve access to the understanding and traceability of data. The generated data provenance itself displays the strong dependencies on properties of metadata, transparency, traceability, and trust and thus their data protection-compliant extraction for secondary use. Without disposing appropriate metadata, properties, and relationship on involved (1) agents, like data owner, responsible staff acting on behalf of organizations , software, (2) entities, like a data input sources or data output, (3) activities, like programs or scripts in a research process, transparency remains hidden, traceability is obscure and building trust in unreliable data and results is very difficult. For example, especially during essential semantic mappings of data, information about who mapped when what data and how is elementary to prove the correctness of the data. In addition, measuring points must be set up in the scripts and output needs evaluation to mark potential errors.

Overall, this study aims at catalyzing an overarching (meta-) data management strategy. On the one hand, an iterative approach is necessary to overcome poor data managing practices. Measures include the deployment of appropriate personnel like data stewards or increased training sections[41]. System's shortcoming could be properly identified and fixed by adequate data managing planning[42]. Beyond that, further major efforts are required to implement these developments on a broad scale like in the provenance workshop series "MInimal Requirements for Automated Provenance Information Enrichment" (MIRAPIE). First community-driven results to standardize provenance information enrichment for biomedical data and possibly beyond are disseminated (see publication 5[43]).

The second pillar towards a tailored data provenance implementation was built upon the proof-of-concept study. Here, the methodological and technical development led to a first ready-to-use python library (PISA), which enabled easy and fast automated establishment of provenance traces in the DIC (see publication 2[30]). The created provenance traces, stored in a relational database, mirror a benchmark for the suitability and readiness of clinical routine data as research data. In this context, data provenance helps to explain the traceability of individually processed data elements and improves its reusability. However, it must be considered that the collection of clinical routine data in the CIS is not primarily designed for medical research. Adapted CIS and enabling requirements for extended documentation specifications will be indispensable for data standardization and facilitating data comparability for the future use.

The PISA system displays extensible information about their source, destination, type of transformation, status of quality validation, and related ownership, valid governance documents and data stewardship. The possibility of extension offers the advantage that this provenance tracking solution can be further developed regarding an audit trail for provision in decision-supported systems. The generated provenance traces are format-neutral, so that any conversions, e.g. to FHIR, RDF, can be carried out at any time.

Both original works together, presented in this dissertation, provide a framework towards warranting a better data provenance integration and augmenting information quality on data elements, their transformation and movement. Moreover, patient's data safety benefits from this accomplishment[44].

The data issued to researchers show increased properties regarding reliability and integrity, reveal possible limitations, minimize risk for wrong use and ensure their trustful dissemination.

The aspects listed have various possible implications since trust entails more than provenance of data elements. Trust implies privacy preserving and security from malign users and measure not to corrupt or change provenance information unjustifiably. Since provenance information, probably scattered in repositories, may contain confidential information, provision, availability, and potential access require a granular concept. Compliance can only be solved by technical access limitations and organizational measure. Moreover, balance must be found to represent provenance information in a suitable way. Furthermore, scalability analysis should also be part of the research approach. Impact beyond the scope of this thesis is given in the context of artificial intelligence (AI) based systems in which transparent data provenance traces support decision-making purposes[28, 44].

Ongoing work determines appropriate methods and tools for provenance information management to achieve sustainability of the generated provenance and to uphold the provenance information properties. Further investigations into the structure of provenance information are relevant, for example why is certain kind of provenance information not available, was this information lost and when. Furthermore, the established concept and outcome will be transferred, integrated, and drive forward the MIRAPIE project.

Provenance traces used in clinical research, which is subjected to specific compliance requirements, e.g. from FDA side, need much more documentation effort during development and implementation for provenance traces. The approach and results of this dissertation take up the pivotal measures to adhere to new released technical specification ISO/TS 23494-1 for a provenance information model for biological material and data series[45]. Thus, it is important to extend the current provenance approach toward an inspection- or audit ready one. However, accreditation will require more measures and efforts.

Conclusion

With rising legal and scientific demands, there is an urgent need for greater transparency by implementing provenance systems in research projects.

This work identified the mandatory steps and synthesized underlying approaches and outcomes to implement fully automated provenance traces on medical data elements in German medical Data Integration Centers. The described provenance system PISA has been designed to minimize validity data risks and to produce a traceable pattern of data processing pipelines from their source. PISA thereby enriches biomedical data and pipelines with adequate FAIR maintained metadata on an (inter-) national level. The overall results pave the way for a reliable and trustful dissemination of FAIR and traceable data items for secondary use. Additionally, the crucial role and responsibility of a data steward escorting these data is highlighted.

Future research will investigate on how provenance traces impact data quality and how far this pushes the reproducibility of digital objects. Secondly, guidance and recommendations are requested to provide the systematic measurement of provenance and calls for defining a minimal or gold standard.

5 ABSTRACT

Summary

Provenance enriched scientific results contribute significantly to their trustworthiness, reliability, and possible reproducibility. Provenance information leads to a higher interpretability of scientific data and enables reliable collaboration and data sharing. As no standard system exists, provenance is often not tracked properly, leading to issues with data trust and data quality, ultimately blocking research on biomedical, and particularly clinical data. This is significant as it directly impacts the development of new treatments and tools for patient care. Moreover, the use of responsible AI in clinical settings demands a detailed reporting and transparency of data capture, transformations, and analysis.

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 data elements are of key importance throughout these processes, but no scientific work has so far been undertaken to examine the data provenance aspects. Insufficient knowledge about these data and processes can lead to validity risks and weaken the quality of the extracted data. The need to collect provenance data during the data life cycle is undisputed, but there is a great lack of clarity on the status.

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 continuously relevant provenance information during data integration pipelines.

The results of this dissertation demonstrate how to enable automated provenance gathering on a medical data element level in a data integration center by combining the strength of quality- and health standard guided (meta-) data management practices. The study analysis disclosed several gaps for which efficient steps were illustrated toward a systematic provenance strategy. The subsequent implementation of a novel provenance algorithm achieved satisfying pipeline execution times. Overall, this indicates a high degree of traceability, accuracy, and reliability of the transformed data, 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.

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 to the envisioned European Health Data Space.

Zusammenfassung

Mit Provenienz angereicherte wissenschaftliche Ergebnisse tragen wesentlich zur eigenen Vertrauenswürdigkeit, Zuverlässigkeit und möglichen Reproduzierbarkeit bei. Provenienz Informationen führen zu einer besseren Interpretierbarkeit wissenschaftlicher Daten und ermöglichen eine zuverlässige Zusammenarbeit und gemeinsame Nutzung von Daten. Da es kein Standardsystem gibt, wird die Provenienz oft nicht ordnungsgemäß verfolgt, was zu Problemen mit dem Vertrauen in die Daten und der Datenqualität führt und letztlich die Forschung an biomedizinischen und insbesondere klinischen Daten blockieren kann. Dies ist von großer Bedeutung, da es sich direkt auf die Entwicklung neuer Behandlungen und Instrumente für die Patientenversorgung auswirkt. Darüber hinaus erfordert der Einsatz verantwortungsvoller KI im klinischen Umfeld eine detaillierte Berichterstattung und Transparenz der Datenerfassung, -umwandlung und -analyse.

Im Rahmen der von der deutschen Regierung geförderten Medizininformatik-Initiative haben medizinische Datenintegrationszentren komplexe Datenflüsse implementiert, um Routinedaten aus der Gesundheitsversorgung für die Sekundärnutzung in Forschungsdatenrepositorien zu laden. Dabei ist das Datenmanagement zur Verarbeitung und Verwaltung der sensiblen Datenelemente von zentraler Bedeutung. Unzureichendes Wissen über diese Daten und Prozesse kann zu Validitätsrisiken führen und die Qualität der extrahierten Daten beeinträchtigen. Die Notwendigkeit der Erfassung von Provenienz Daten während ihres Datenlebenszyklus ist unbestritten, aber es besteht ein großer Mangel an Klarheit über den Status.

Diese kumulative Dissertation stellt die Kombination eines zweistufigen methodischen Ansatzes vor, um eine umfassende Anreicherung von Provenienz Informationen in den medizinischen Datenintegrationspipelines zu ermöglichen. Eine MIRACUM-weite „Mixed-Methods“ – Studie, eine Kombination aus qualitativer und quantitativer Forschungsmethode, untersuchte sowohl den Reifegrad des Datenmanagements als auch die Provenienz Bereitschaft und legte Empfehlungen vor. Die anschließende Machbarkeitsstudie griff diese Ergebnisse auf mit dem Ziel der Modellierung und Implementierung eines Provenienz Algorithmus.

Die Ergebnisse dieser Dissertation legen dar, wie in einem medizinischen Datenintegrationszentrum durch die Kombination von qualitäts- und gesundheitsstandardgeleiteten (Meta-)Datenmanagement-Praktiken erstmalig gezielt automatisch Provenienz Informationen auf Datenelementebene erfasst, gespeichert und ausgeleitet werden können. Die Studienanalyse deckte mehrere Lücken auf, für die effiziente Schritte hin zu einer systematischen Provenienz Strategie aufgezeigt wurden. Die anschließende Implementierung eines neuartigen Provenienz Algorithmus führte zu zufriedenstellenden Ausführungszeiten während der Datenintegrationspipelines. Damit konnte ein hohes Maß an Rückverfolgbarkeit, Genauigkeit und Zuverlässigkeit der transformierten Daten erreicht werden mit welcher ein Datenintegrationszentrum einer möglichen Rechenschaftspflicht nachkommen kann.

Darüber hinaus dient diese Dissertation als Katalysator für die Ableitung einer übergreifenden Datenmanagementstrategie, die Datenintegrität und Provenienz Merkmale als Schlüsselfaktor für qualitäts- und FAIR-gerechte Gesundheits- und Forschungsdaten beachtet.

Die Dissertation nimmt Empfehlungen vorweg, die die Qualität der Dissemination von Patientendaten verbessern und die Umsetzung von auditierbaren und messbaren Provenienz-Ansätzen anleiten sollen, und leistet damit einen Beitrag zum angestrebten Europäischen Gesundheitsdatenraum.

6 BIBLIOGRAPHY

1. Cuggia, M, Combes, S: The French Health Data Hub and the German Medical Informatics Initiatives: Two National Projects to Promote Data Sharing in Healthcare. *Yearb Med Inform,* 28**:** 195-202, 2019. https://doi.org/10.1055/s-0039-1677917
2. Coman Schmid, D, Crameri, K, Oesterle, S, Rinn, B, Sengstag, T, Stockinger, H, BioMed, ITnt: SPHN - The BioMedIT Network: A Secure IT Platform for Research with Sensitive Human Data. *Stud Health Technol Inform,* 270**:** 1170-1174, 2020. https://doi.org/10.3233/SHTI200348
3. Health-RI, S: health RI - enabling data driven health & life sciences, 2023. https://www.health-ri.nl/en. Retrieved 27-Dec-2023.
4. Semler, SC, Wissing, F, Heyder, R: German Medical Informatics Initiative. *Methods Inf Med,* 57**:** e50-e56, 2018. https://doi.org/10.3414/ME18-03-0003
5. Prokosch, HU, Acker, T, Bernarding, J, Binder, H, Boeker, M, Boerries, M, Daumke, P, Ganslandt, T, Hesser, J, Honing, G, Neumaier, M, Marquardt, K, Renz, H, Rothkotter, HJ, Schade-Brittinger, C, Schmucker, P, Schuttler, J, Sedlmayr, M, Serve, H, Sohrabi, K, Storf, H: MIRACUM: Medical Informatics in Research and Care in University Medicine. *Methods Inf Med,* 57**:** e82-e91, 2018. https://doi.org/10.3414/ME17-02-0025
6. Germany, MII: The Medical Informatics Initiative’s core data set, 2023. https://www.medizininformatik-initiative.de/en/medical-informatics-initiatives-core-data-set. Retrieved 28-Dec-2023.
7. Prokosch, HU, Gebhardt, M, Gruendner, J, Kleinert, P, Buckow, K, Rosenau, L, Semler, SC: Towards a National Portal for Medical Research Data (FDPG): Vision, Status, and Lessons Learned. *Stud Health Technol Inform,* 302**:** 307-311, 2023. https://doi.org/10.3233/SHTI230124
8. Kanza, S, Knight, NJ: Behind every great research project is great data management. *BMC Res Notes,* 15**:** 20, 2022. https://doi.org/10.1186/s13104-022-05908-5
9. Cooper, DR, Grabowski, M, Zimmerman, MD, Porebski, PJ, Shabalin, IG, Woinska, M, Domagalski, MJ, Zheng, H, Sroka, P, Cymborowski, M, Czub, MP, Niedzialkowska, E, Venkataramany, BS, Osinski, T, Fratczak, Z, Bajor, J, Gonera, J, MacLean, E, Wojciechowska, K, Konina, K, Wajerowicz, W, Chruszcz, M, Minor, W: State-of-the-Art Data Management: Improving the Reproducibility, Consistency, and Traceability of Structural Biology and in Vitro Biochemical Experiments. *Methods Mol Biol,* 2199**:** 209-236, 2021. https://doi.org/10.1007/978-1-0716-0892-0_13
10. Davis-Turak, J, Courtney, SM, Hazard, ES, Glen, WB, Jr., da Silveira, WA, Wesselman, T, Harbin, LP, Wolf, BJ, Chung, D, Hardiman, G: Genomics pipelines and data integration: challenges and opportunities in the research setting. *Expert Rev Mol Diagn,* 17**:** 225-237, 2017. https://doi.org/10.1080/14737159.2017.1282822
11. Steffens, S, Schroder, K, Kruger, M, Maack, C, Streckfuss-Bomeke, K, Backs, J, Backofen, R, Baessler, B, Devaux, Y, Gilsbach, R, Heijman, J, Knaus, J, Kramann, R, Linz, D, Lister, AL, Maatz, H, Maegdefessel, L, Mayr, M, Meder, B, Nussbeck, SY, Rog-Zielinska, EA, Schulz, MH, Sickmann, A, Yigit, G, Kohl, P: The challenges of research data management in cardiovascular science: a

DGK and DZHK position paper-executive summary. *Clin Res Cardiol*, 2023. https://doi.org/10.1007/s00392-023-02303-3

12. Carmona-Bayonas, A, Jimenez-Fonseca, P, Fernandez-Somoano, A, Alvarez-Mancenido, F, Castanon, E, Custodio, A, de la Pena, FA, Payo, RM, Valiente, LP: Top ten errors of statistical analysis in observational studies for cancer research. *Clin Transl Oncol,* 20**:** 954-965, 2018. https://doi.org/10.1007/s12094-017-1817-9
13. Bruha, P, Moucek, R, Salamon, J, Vacek, V: Workflow for health-related and brain data lifecycle. *Front Digit Health,* 4**:** 1025086, 2022. https://doi.org/10.3389/fdgth.2022.1025086
14. Gierend, K, Freiesleben, S, Kadioglu, D, Siegel, F, Ganslandt, T, Waltemath, D: The Status of Data Management Practices Across German Medical Data Integration Centers: Mixed Methods Study. *J Med Internet Res,* 25**:** e48809, 2023. https://doi.org/10.2196/48809
15. Stohr, MR, Gunther, A, Majeed, RW: Provenance for Biomedical Ontologies with RDF and Git. *Stud Health Technol Inform,* 267**:** 230-237, 2019. https://doi.org/10.3233/SHTI190832
16. Ayaz, M, Pasha, MF, Alzahrani, MY, Budiarto, R, Stiawan, D: The Fast Health Interoperability Resources (FHIR) Standard: Systematic Literature Review of Implementations, Applications, Challenges and Opportunities. *JMIR Med Inform,* 9**:** e21929, 2021. https://doi.org/10.2196/21929
17. Wilkinson, MD, Dumontier, M, Aalbersberg, IJ, Appleton, G, Axton, M, Baak, A, Blomberg, N, Boiten, JW, da Silva Santos, LB, Bourne, PE, Bouwman, J, Brookes, AJ, Clark, T, Crosas, M, Dillo, I, Dumon, O, Edmunds, S, Evelo, CT, Finkers, R, Gonzalez-Beltran, A, Gray, AJ, Groth, P, Goble, C, Grethe, JS, Heringa, J, t Hoen, PA, Hooft, R, Kuhn, T, Kok, R, Kok, J, Lusher, SJ, Martone, ME, Mons, A, Packer, AL, Persson, B, Rocca-Serra, P, Roos, M, van Schaik, R, Sansone, SA, Schultes, E, Sengstag, T, Slater, T, Strawn, G, Swertz, MA, Thompson, M, van der Lei, J, van Mulligen, E, Velterop, J, Waagmeester, A, Wittenburg, P, Wolstencroft, K, Zhao, J, Mons, B: The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data,* 3**:** 160018, 2016. https://doi.org/10.1038/sdata.2016.18
18. Jansen, P, van den Berg, L, van Overveld, P, Boiten, JW: Research Data Stewardship for Healthcare Professionals. In: *Fundamentals of Clinical Data Science.* edited by KUBBEN, P., DUMONTIER, M., DEKKER, A., Cham (CH), 2019, pp 37-53.
19. Michaelis, L, Poyraz, RA, Muzoora, MR, Gierend, K, Bartschke, A, Dieterich, C, Johann, T, Krefting, D, Waltemath, D, Thun, S: Insights into the FAIRness of the German Network University Medicine: A Survey. *CARING IS SHARING–EXPLOITING THE VALUE IN DATA FOR HEALTH AND INNOVATION***:** 741, 2023.
20. Jacobsen A, dMAR, Juty N, Batista D, Coles S, Cornet R, Courtot M, Crosas M, Dumontier M, Evelo CT, Goble C, Guizzardi G, Hansen KK, Hasnain A, Hettne K, Heringa J, Hooft RWW, Imming M, Jeffery KG, Kaliyaperumal R, Kersloot MG, Kirkpatrick CR, Kuhn T, Labastida I, Magagna B, McQuilton P, Meyers N, Montesanti A, van Reisen M, Rocca-Serra P, Pergl R, Sansone S-A, da Silva Santos LOB, Schneider J, Strawn G, Thompson M, Waagmeester A, Weigel T, Wilkinson MD, Willighagen EL, Wittenburg P, Roos M, Mons B, Schultes E.: FAIR Principles: Interpretations and Implementation Considerations. 2020. https://doi.org/10.1162/dint_r_00024

21. Gaignard, A, Rosnet, T, De Lamotte, F, Lefort, V, Devignes, MD: FAIR-Checker: supporting digital resource findability and reuse with Knowledge Graphs and Semantic Web standards. *J Biomed Semantics,* 14**:** 7, 2023. https://doi.org/10.1186/s13326-023-00289-5
22. Curcin, V: Embedding data provenance into the Learning Health System to facilitate reproducible research. *Learn Health Syst,* 1**:** e10019, 2017. https://doi.org/10.1002/lrh2.10019
23. group, WCw: An Overview of the PROV Family of Documents, 2023. https://www.w3.org/TR/prov-overview/
24. Samuel, S, Konig-Ries, B: End-to-End provenance representation for the understandability and reproducibility of scientific experiments using a semantic approach. *J Biomed Semantics,* 13**:** 1, 2022. https://doi.org/10.1186/s13326-021-00253-1
25. Martin, SJ: The FEBS Journal in 2022: trust the science and treasure the data. *FEBS J,* 289**:** 4-8, 2022. https://doi.org/10.1111/febs.16332
26. Jayapandian, CP, Zhao, M, Ewing, RM, Zhang, GQ, Sahoo, SS: A semantic proteomics dashboard (SemPoD) for data management in translational research. *BMC Syst Biol,* 6 Suppl 3**:** S20, 2012. https://doi.org/10.1186/1752-0509-6-S3-S20
27. Weng, C: Clinical data quality: a data life cycle perspective. *Biostat Epidemiol,* 4**:** 6-14, 2020. https://doi.org/10.1080/24709360.2019.1572344
28. Plass, M, Wittner, R, Holub, P, Frexia, F, Mascia, C, Gallo, M, Muller, H, Geiger, J: Provenance of specimen and data - A prerequisite for AI development in computational pathology. *N Biotechnol,* 78**:** 22-28, 2023. https://doi.org/10.1016/j.nbt.2023.09.006
29. Horgan, D, Hajduch, M, Vrana, M, Soderberg, J, Hughes, N, Omar, MI, Lal, JA, Kozaric, M, Cascini, F, Thaler, V, Sola-Morales, O, Romao, M, Destrebecq, F, Sky Gross, E: European Health Data Space-An Opportunity Now to Grasp the Future of Data-Driven Healthcare. *Healthcare (Basel),* 10, 2022. https://doi.org/10.3390/healthcare10091629
30. Gierend, K, Waltemath, D, Ganslandt, T, Siegel, F: Traceable Research Data Sharing in a German Medical Data Integration Center With FAIR (Findability, Accessibility, Interoperability, and Reusability)-Geared Provenance Implementation: Proof-of-Concept Study. *JMIR Form Res,* 7**:** e50027, 2023. https://doi.org/10.2196/50027
31. Gierend, K, Kruger, F, Waltemath, D, Funfgeld, M, Ganslandt, T, Zeleke, AA: Approaches and Criteria for Provenance in Biomedical Data Sets and Workflows: Protocol for a Scoping Review. *JMIR Res Protoc,* 10**:** e31750, 2021. https://doi.org/10.2196/31750
32. Gierend, K, Krüger, F, Genehr, S, Hartmann, F, Siegel, F, Waltemath, D, Ganslandt, T, Zeleke, AA: Capturing provenance information for biomedical data and workflows: A scoping review. Research Square, 2023.
33. Foundation, TR: The R Project for Statistical Computing, 2023. https://www.r-project.org. Retrieved 27 November 2023.
34. O'Cathain, A, Murphy, E, Nicholl, J: The quality of mixed methods studies in health services research. *J Health Serv Res Policy,* 13**:** 92-98, 2008. https://doi.org/10.1258/jhsrp.2007.007074
35. Bongiovanni, S, Purdue, R, Kornienko, O, Bernard, R: Quality in Non-GxP Research Environment. *Handb Exp Pharmacol,* 257**:** 1-17, 2020. https://doi.org/10.1007/164_2019_274

36. Arksey, H, O'Malley, L: Scoping studies: towards a methodological framework. *International journal of social research methodology,* 8**:** 19-32, 2005.
37. Ouzzani, M, Hammady, H, Fedorowicz, Z, Elmagarmid, A: Rayyan-a web and mobile app for systematic reviews. *Syst Rev,* 5**:** 210, 2016. https://doi.org/10.1186/s13643-016-0384-4
38. Morales, DR, Arlett, P: RCTs and real world evidence are complementary, not alternatives. *BMJ,* 381**:** 736, 2023. https://doi.org/10.1136/bmj.p736
39. Wicherski, J, Haenisch, B: [The application of real-world evidence in drug regulatory decision-making]. *Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz*, 2024. https://doi.org/10.1007/s00103-023-03830-0
40. Stellmach, C, Muzoora, MR, Thun, S: Digitalization of Health Data: Interoperability of the Proposed European Health Data Space. *Stud Health Technol Inform,* 298**:** 132-136, 2022. https://doi.org/10.3233/SHTI220922
41. Kohrs, FE, Auer, S, Bannach-Brown, A, Fiedler, S, Haven, TL, Heise, V, Holman, C, Azevedo, F, Bernard, R, Bleier, A, Bossel, N, Cahill, BP, Castro, LJ, Ehrenhofer, A, Eichel, K, Frank, M, Frick, C, Friese, M, Gartner, A, Gierend, K, Gruning, DJ, Hahn, L, Hulsemann, M, Ihle, M, Illius, S, Konig, L, Konig, M, Kulke, L, Kutlin, A, Lammers, F, Mehler, DMA, Miehl, C, Muller-Alcazar, A, Neuendorf, C, Niemeyer, H, Pargent, F, Peikert, A, Pfeuffer, CU, Reinecke, R, Roer, JP, Rohmann, JL, Sanchez-Tojar, A, Scherbaum, S, Sixtus, E, Spitzer, L, Strassburger, VM, Weber, M, Whitmire, CJ, Zerna, J, Zorbek, D, Zumstein, P, Weissgerber, TL: Eleven strategies for making reproducible research and open science training the norm at research institutions. *Elife,* 12, 2023. https://doi.org/10.7554/eLife.89736
42. Curcin, V, Soljak, M, Majeed, A: Managing and exploiting routinely collected NHS data for research. *Inform Prim Care,* 20**:** 225-231, 2012. https://doi.org/10.14236/jhi.v20i4.1
43. Gierend, K, Wodke, JAH, Genehr, S, Gött, R, Henkel, R, Krüger, F, Mandalka, M, Michaelis, L, Scheuerlein, A, Schröder, M, Zeleke, A, Waltemath, D: TAPP: Defining standard provenance information for clinical research data and workflows - Obstacles and opportunities. *Companion Proceedings of the ACM Web Conference 2023.* Austin, TX, USA, Association for Computing Machinery, 2023 pp 1551–1554.
44. Werder, K, Ramesh, B, Zhang, R: Establishing Data Provenance for Responsible Artificial Intelligence Systems. *ACM Trans Manage Inf Syst,* 13**:** Article 22, 2022. https://doi.org/10.1145/3503488
45. Wittner, R, Holub, P, Mascia, C, Frexia, F, Müller, H, Plass, M, Allocca, C, Betsou, F, Burdett, T, Cancio, I, Chapman, A, Chapman, M, Courtot, M, Curcin, V, Eder, J, Elliot, M, Exter, K, Goble, C, Golebiewski, M, Kisler, B, Kremer, A, Leo, S, Lin-Gibson, S, Marsano, A, Mattavelli, M, Moore, J, Nakae, H, Perseil, I, Salman, A, Sluka, J, Soiland-Reyes, S, Strambio-De-Castillia, C, Sussman, M, Swedlow, JR, Zatloukal, K, Geiger, J: Toward a common standard for data and specimen provenance in life sciences. *Learning Health Systems,* n/a**:** e10365, https://doi.org/https://doi.org/10.1002/lrh2.10365

7 APPENDICES

All appendices are available within the original publications.

8 CURRICULUM VITAE

PERSONAL DATA

Last and first name:	Gierend Kerstin Anita
Birth date:	30-Apr-1969
Birth place:	Ottweiler

SCHOOL BACKGROUND

08-Jun-1988	A-levels „Allgemeine Hochschulreife“, Cusanus Gymnasium St. Wendel

EDUCATION AND UNIVERSITARY BACKGROUND

1988 - 1991	School Medical Documentation, Gießen
08/1991	Graduation: „Staatliche Anerkennung“ *Topic „Facharbeit“:* „Durchführung und Auswertung einer humanpharmakologischen Prüfung anhand einer ausgewählten Phase I Studie“
1998 - 2000	Extra occupational study „Gesundheitsinformatik“, Start in 3rd semester, SRH Hochschule Heidelberg
27-Sep-2000	Graduation: Diplom-Informatikerin (FH) *Topic Diploma thesis:* „Konzeption und Implementierung eines zentralen Studien-Informations-Systems (SIS) zur Unterstützung des Projektmanagements im Bereich der klinischen Forschung“
24-Jul-2020	Admission to doctorate, Medical Faculty Mannheim, Heidelberg University

PROFESSIONAL CARRER

10/1989 - 12/1989	Practical training „Uniklinikum Gießen“
06/1990 - 09/1990	Practical training „Tumorregister des Landes Tirol, Uniklinikum Innsbruck“
03/1991 - 08/1991	Practical training „Institut für Klinische Pharmakologie, Prof. Dr. Lücker GmbH (IKP)“
09/1991 - 06/2013	CRS Clinical Research Mannheim GmbH (former IKP), Datamanagement & Biometrics
03/1996 - 01/2003	Authorized officer, Department “Datamanagement & Biostatistics”
07/2013 - 08/2019	Senior IT-Consultant (Computer System Validation), Roche Diagnostics GmbH (on behalf of “EXCO GmbH” and “Ingenieurbüro Prof. Dr. Gierend”)
since 09/2019	Department for Biomedical Informatics, Medical Faculty Mannheim, Heidelberg University
since 02/2023	Lecturer Research Data Management “Biomedizinische Informatik und Data Science“

ADDITIONAL PUBLICATIONS OR CONFERENCE CONTRIBUTIONS

Authors: Kohrs, FE, Auer, S, Bannach-Brown, A, Fiedler, S, Haven, TL, Heise, V, Holman, C, Azevedo, F, Bernard, R, Bleier, A, Bossel, N, Cahill, BP, Castro, LJ, Ehrenhofer, A, Eichel, K, Frank, M, Frick, C, Friese, M, Gartner, A, **Gierend, K**, Gruning, DJ, Hahn, L, Hulsemann, M, Ihle, M, Illius, S, Konig, L, Konig, M, Kulke, L, Kutlin, A, Lammers, F, Mehler, DMA, Miehl, C, Muller-Alcazar, A, Neuendorf, C, Niemeyer, H, Pargent, F, Peikert, A, Pfeuffer, CU, Reinecke, R, Roer, JP, Rohmann, JL, Sanchez-Tojar, A, Scherbaum, S, Sixtus, E, Spitzer, L, Strassburger, VM, Weber, M, Whitmire, CJ, Zerna, J, Zorbek, D, Zumstein, P, Weissgerber, TL
Title: Eleven strategies for making reproducible research and open science training the norm at research institutions.
Journal *Elife,* 12, 2023. https://doi.org/10.7554/eLife.89736
Status: accepted, IF 7.7 (2023)

Gierend K, Genehr S. MIRAPIE community project – a minimal biomedical data provenance model. Lightning talks, the combine HARMONY 2024, 08 April 2024

Michaelis L, Poyraz RA, Muzoora MR, **Gierend K**, Bartschke A, Dieterich C, Johann T, Krefting D, Waltemath D, Thun S.
Insights into the FAIRness of the German Network University Medicine: A Survey.
Stud Health Technol Inform. 2023 May 18;302:741-742. doi: 10.3233/SHTI230251. PMID: 37203481.

Kamdje Wabo G, Prasser F, **Gierend K**, Siegel F, Ganslandt T. Data Quality- and Utility-Compliant Anonymization of Common Data Model-Harmonized Electronic Health Record Data: Protocol for a Scoping Review. JMIR Res Protoc. 2023 Aug 11;12:e46471. doi: 10.2196/46471. PMID: 37566443; PMCID: PMC10457704.

Kerstin Gierend, Frank Krüger, Sascha Genehr, Francisca Hartmann, Thomas Ganslandt, Dagmar Waltemath, Atinkut Alamirrew Zeleke
Challenges and potential bottlenecks for the accomplishment of provenance in biomedical data sets and workflows
Deutsche Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie. 67. Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie e. V. (GMDS), 13. Jahreskongress der Technologie- und Methodenplattform für die vernetzte medizinische Forschung e.V. (TMF). sine loco [digital], 21.-25.08.2022. Düsseldorf: German Medical Science GMS Publishing House; 2022. DocAbstr. 163 (DOI: https://dx.doi.org/10.3205/22gmds062)

Erik Tute, Christian Draeger, **Kerstin Gierend**, Matthias Löbe, Julia Palm, Carsten Oliver Schmidt
A glimpse at representing data quality rules for their collaborative governance in the Medical Informatics Initiative
Deutsche Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie. 67. Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie e. V. (GMDS), 13. Jahreskongress der Technologie- und Methodenplattform für die vernetzte medizinische Forschung e.V. (TMF). sine loco [digital], 21.-25.08.2022. Düsseldorf: German Medical Science GMS Publishing House; 2022. DocAbstr. 166 (DOI: https://dx.doi.org/10.3205/22gmds018)

João Cardoso, Sarah Jones, Tomasz Miksa, Adil Hasan, Maria Praetzellis, Paulette Lieby, Elli Papadopoulou, **Kerstin Gierend.**
Mapping of maDMPs to Funder Templates. Zenodo; 2020. DOI: 10.5281/zenodo.3944458 (ohne peer-review)

REVIEW Activities

Deutsche Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie e. V. (GMDS), ongoing since 2022

Journal of Medical Internet Research and sister journals (JMIR),
ongoing since 2022

Invitation to join the Editorial Board of the Online Journal of Public Health Informatics, April 2024

9 ACKNOWLEDGEMENT

I would like to express my sincere gratitude to Prof. Dr.-Ing. Dagmar Waltemath for her invaluable guidance and expertise throughout this research project. Her profound insights, encouragement and dedication have shaped my academic work and have been essential in overcoming challenges and achieving milestones.

I am truly grateful to my supervisor Univ.-Prof. Dr. med. Thomas Ganslandt for his expertise, generosity and commitment to the realization of this dissertation project.

Additionally, I would like to extend my appreciation to Dr. rer. medic. Atinkut Alamirrew Zeleke and Dr. med. Fabian Siegel for their expertise and constructive feedback which have enriched my research work.

I acknowledge the funding support provided by the MIRACUM consortium as part of the Medical Informatics Initiative (MII), without which this research would not have been possible.

Finally, I am thankful to my beloved family for their unwavering encouragement and understanding during this challenging time. To my husband Christian, his endless support and patience and to my children Jens and Celine for being my greatest inspiration.