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;<br>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:*