MEDINFO 2019: Health and Wellbeing e-Networks for All
L. Ohno-Machado and B. Séroussi (Eds.)

doi:10.3233/SHTI190231

Provenance Solutions for Medical Research in Heterogeneous IT-Infrastructure: An Implementation Roadmap

Marcel Parciak[a], Christian Bauer[a], Theresa Bender[a], Robert Lodahl[a], Björn Schreiweis[b], Erik Tute[c], Ulrich Sax[a]

[a] *Department of Medical Informatics, University Medical Center, Göttingen, Germany,*
[b] *Institute for Medical Informatics and Statistics, University Hospital Schleswig-Holstein and Kiel University, Kiel, Germany,*
[c] *Peter L. Reichertz Institute for Medical Informatics, Hannover Medical School, Hannover, Germany*

Abstract

Research data generated in large projects raise challenges about not only data analytics but also data quality assessments and data governance. The provenance of a data set – that is the history of data sets – holds information relevant to technicians and non-technicians and is able to answer questions regarding data quality, transparency, and more. We propose an implementation roadmap to extract, store, and utilize provenance records in order to make provenance available to data analysts, research subjects, privacy officers, and machines (machine readability). Each aspect is tackled separately, resulting in the implementation of a provenance toolbox. We aim to do so within the context of HiGHmed, a research consortium established within the medical informatics initiative in Germany. In this testbed of federated IT-infrastructures, the toolbox shall assist each stakeholder in answering domain-specific and domain-agnostic questions regarding the provenance of data sets. This way, we will improve data re-use, transparency, and reproducibility.

***Keywords*:**

Data Accuracy, Metadata, Research, Reproducibility of Results

Introduction

In addition to the variety of hospital information systems, university hospitals and academic medical centers operate specialized systems, each one serving specific use cases in a best-of-breed manner to support research [1]. Furthermore, a hospital runs multiple software systems in order to serve medical care, generating vast amounts of heterogeneous data. Using information that is contained in healthcare and research systems requires cleansing, harmonizing, and integrating the data sets – typically referred to as extract, transform, and load (ETL) processes. Finally, an integrated data pool can be generated for further analysis. As the numbers of data sets and systems grow, for example by federating systems from multiple sites, the complexity of performing ETL related tasks grows as well. This makes the assessment of data quality a tedious task. Moreover, data reuse is oftentimes difficult and then neglected, which results in recaptured and regenerated data sets per research project [2]. Thus, creating reusable data sets represents an important focus in current research [3]. The FAIR principles require data sets to be findable, accessible, interoperable, and reproducible in order to ensure high-quality data sharing and enable data reuse [4] both for humans and machines. Fine-grained documentation of processes, transformations, and influences on an object – also known as provenance – from data sources, data integration pipelines, and data repositories can be exploited to enable data reuse and improve data quality [5,6]. Provenance is used to enhance reproducibility, transparency, presentation, meta-analysis, and machine-readability [7].

Capturing, storing, and utilizing provenance have shown to provide valuable insights within biomedical research projects [8] – technical experts inspect provenance in order to assess data quality and gain insight into data processing across multiple sites. We aim to implement a generalized toolbox focused on provenance using our experience, providing instruments to capture, store, and utilize provenance for technicians and non-technicians alike. More discipline-specific experience on how to handle provenance in federated research projects is available from the geographic information systems (GIS) [9], bio-informatics [10] or the physics domain communities [11].

We will implement and test the provenance toolbox in the context of the HiGHmed platform, which aims to implement a federated platform of medical data sets in order to improve healthcare and medical research at multiple institutions [12]. Conforming to the HiGHmed principles, the toolbox needs to be scalable, compliant to local data safety and privacy regulations, sustainable, federated, and focused on improving patient healthcare. This toolbox should be applicable by different stakeholders in medical research projects.

We propose this implementation roadmap towards a provenance system architecture describing how we intend to develop solutions to capture provenance appropriately, store provenance securely embedded into a heterogeneous IT-infrastructure, and implement the means to utilize provenance for each relevant stakeholder in medical research.

Methods

A thorough literature review was conducted as a basis for our work [13]. Additionally, surveys of Moreau [14], Herschel et al. [15] as well as Pérez et al. [16] were considered. These works present background knowledge regarding provenance research as well as several solutions that implement provenance.

We will use the provenance definition by the W3C PROV-DM specification: “provenance is defined as a record that describes the people, institutions, entities, and activities involved in producing, influencing, or delivering a piece of data or a thing.” [6] Provenance is captured in several granularities, ranging from provenance meta-data to data provenance as depicted in Figure 1 [15]. Provenance meta-data refers to a generic, hardly utilizable type of provenance,

e.g. a free-text description of a method in a research paper. Workflow and information system provenance refer to more specific types of machine-usable data, for example, a definition and execution of a workflow defined in a workflow management system. Distinguishing between these two types of provenance is not always possible. Data provenance refers to the most specific type of provenance, which is usually tailored to the precise context the data describes. Dublin Core, for example, describes several elements of the provenance of digital resources [17].

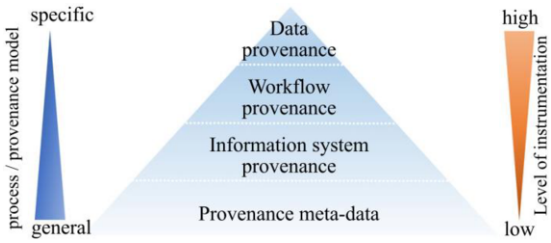

Figure 1 – Provenance Types Hierarchically Ordered by the Provenance Model and the Level of Instrumentation [15]

Capturing provenance manually is considered ineffective [18]. Workflow management systems enable capturing data based on predefined workflows, resulting in workflow-based provenance containing domain-specific insights on processing steps. Process-based (or activity-based) methods require a system or application itself to collect relevant provenance data. Operating-system-based (OS-based) methods allow to capture provenance based on OS functions, like file reads [13]. Similarly, captured provenance based on service-oriented architectures results in service-based provenance. Finally, provenance capturing in the context of relational database systems is considered as a specialized approach to collect provenance [16].

Tracing provenance is performed by two approaches: lazy and eager [16]. Lazy describes collecting provenance after the data processing took place, for example by scraping logs or reverse engineering database queries. The eager approach aims to collect provenance data during data processing (or immediately after), for example by wrapping data processing applications.

PROV-DM is a data model for provenance on the web, defined by the W3C in an effort to summarize and consolidate existing provenance models [6]. It models provenance as an acyclic directed graph, consisting of nodes and edges which represent provenance. Entities, Activities, and Agents represent objects, processes, and influences respectively.

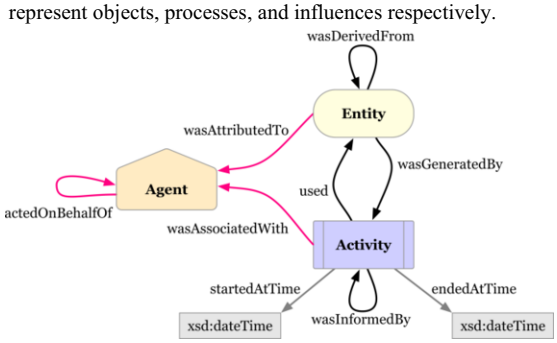

Figure 2 – W3C PROV-DM Depicted as an Acyclic Directed Graph, Showing the Three Starting Point Classes Entity, Activity, Agent in Yellow, Blue, Orange Respectively, Connected by the Relations Defined by PROV [21] and Two Activity-specific Properties are Visualized in Grey

These three nodes are connected by different relations as depicted in a general overview in Figure 2. PROV-DM is highly extensible, allowing to tailor general concepts to more specific constructs. The W3C also provides different serializations for PROV like XML [19] as well as general guidelines on how to use PROV-DM data, like access and query mechanisms [20].

Results

Using provenance from heterogeneous medical research information systems requires considering three aspects: extraction, storage, and utilization. Extraction (or capturing) of provenance is the documentation of data processing steps. Storage of provenance involves a data store, a data format and the means to query the stored provenance data. Finally, utilization generates knowledge from stored provenance data, enabling possible uses from provenance like meta-analysis or reproducibility of computing steps.

Extraction of provenance

As a first step, coordination-points – central computing systems which process high loads of data – will be tackled in order to extract provenance [22]. PROV@TOS has been implemented to capture provenance from data integration jobs based on Talend Open Studio for Data Integration [23]. In combination with version control systems, full-featured provenance documents for data integration pipelines will be extracted [24]. As a part of HiGHmed, the Medical Data Integration Center at the University Medical Center in Göttingen (UMG MeDIC) will use an IHE compliant infrastructure that will use Audit Record Repositories defined by IHE ATNA [25]. These will serve as sources to extract provenance besides data integration pipelines. Upon enabling provenance-awareness for those two coordination-points, existing IT-systems are tackled. In addition to wrapping, log-scraping or creating provenance templates [26], a concept to categorize different provenance sources regarding the quality of produced provenance data will be defined and used to prioritize the implementation of capturing techniques per system. Although the impact of different capturing techniques on provenance quality has been investigated [18], we intend to tailor a comparison to the medical domain in order to prioritize the implementation of per system provenance capturing mechanisms.

Solution to store provenance

Storing, accessing and querying provenance are made possible at a centralized provenance store [27] utilizing the standardized W3C PROV data model [6]. A neo4j NoSQL database (https://neo4j.com/) with an appropriate W3C PROV connector (https://github.com/DLR-SC/provneo4j) will serve as the provenance store. It will be a part of an IHE-compliant IT-infrastructure, hard linking provenance records to the actual data they represent [28]. User access rights are defined via the IHE Consent Profiles (BPPC/APPC) where applicable. The definition of appropriate access rights needs to consider data privacy (i.e. General Data Protection Regulation (EU) 2016/679 (GDPR)). Concepts analogous to the "Cross-Community" profiles defined by IHE will be created and implemented. As a basis, the PROV-AQ [20] mechanisms can be used and extended to fit the needs within a medical context. Furthermore, FHIR Resources for PROV (http://hl7.org/fhir/provenance.html) are available for usage, enabling captured provenance to be used across a whole data sharing community.

HiGHmed's architectural framework relies on the openEHR standard for semantic modeling of clinical information [12]. This standard provides the Feeder Audit (https://www.openehr.org/releases/RM/latest/docs/common/common.html#_feeder_system_audit) which enables linking or directly storing provenance information on different granularity levels even for single data entries where desired.

Visualization and exploration of provenance

On a per stakeholder basis, tools to access provenance records are – if no ready-to-use solutions are available – developed and evaluated, each serving the specialized needs of a stakeholder regarding provenance questions. As a starting point, three stakeholders are taken into account: domain experts (medical informaticians) that perform data analytics, data privacy officers that need to assess regulatory questions and laymen (patients) that seek informational value from provenance records. As intended, use case and domain specific knowledge varies greatly among these three stakeholders, therefore stakeholder-specific access and interaction interfaces to provenance must be designed. A stakeholder analysis needs to be performed prior to implementing the tools followed by validation of the implemented tools against the requirements previously gathered [29]. The PrIMe methodology can serve as a starting point for this task [30]. Tools visualizing such metadata to researchers preparing data analysis can benefit from the availability of standardized provenance information. In order to assist in tasks like data analysis effectively, the planned tools should be embedded into other applications (for example, tranSMART (http://transmartfoundation.org/) for data analysis). As other research fields have already implemented solutions for their respective fields (for example, DataONE [9]), an analysis of those implementations will be performed. Additionally, gaining insight from provenance records is not limited to visualizations, as network analytics are currently evaluated within the provenance research community [31]. To contribute to this research, the metrics that are defined by the authors can be put to use within controlling and monitoring systems.

Discussion

Solutions to capture, store, and use provenance data are scarce. DataONE [9] implements provenance capturing, storing, and usage to recall processing steps related to the data objects that are stored in DataONE, enabling reproducibility. Similarly, platforms like Galaxy [10] or CRISTAL [11] aim to make processing workflows reproducible and recallable. In the Software Evolution domain, the means to utilize provenance data within source code versioning solutions for domain specific needs exist, e.g. defect prediction [32] or simulation of software evolution processes [33]. Our aim will be to learn from these solutions and create the possibility to use provenance to answer domain-specific questions regarding reproducibility of results, assessing data quality, presentation of results, and meta-analysis of all processes within the context of medical research as Curcin et al. suggest [8].

Extraction of provenance

Several software solutions to capture provenance are available, most of them based on workflow management systems [16]. This makes it difficult to tailor them to the needs within the HiGHmed MeDICs. Moreover, the client-server architecture defined by the IHE profiles [25] makes the use of OS-based provenance capture mechanisms less viable. The majority of data sources are only available by proprietary software, urging us to choose an approach similar to PLUS [22]: we will focus on provenance from coordination-points first and tackle single systems second, employing PrIMe [24] to wrap or extend systems. Data privacy needs to be considered when extracting provenance from medical data sets due to GDPR. Where applicable, fine-grained domain-specific provenance data will be extracted from source systems in order to represent the whole data capturing process within the provenance store [11].

Audit logs compliant to the Good Clinical Practice (GCP) showed to yield valuable information regarding provenance [5]. Hence, the audit logs captured based on the ATNA IHE-profile specification may yield comparable data from a coordination-point. Source systems that forward audit logs compliant with ATNA could also be addressed this way. Further investigation is needed to validate this assumption, being an integral part of establishing provenance-aware medical research IT-infrastructures.

Solutions to store provenance

As W3C PROV-DM is modeled as an acyclic directed graph [6], the neo4j graph database is a primary candidate to store PROV data in contrast to commonly used relational database storage systems [8]. Neo4j provides a reliable and scalable solution to store and query graph data [34] and is also used by established provenance recording systems [22,35].

Data privacy within provenance is oftentimes regarded as the need to ensure access rights to provenance data [8,14]. Albeit an important consideration, restricting access to provenance is insufficient to meet both, data privacy regulations and interests of data analysts. Chapman et al. introduce the use of surrogates to mask sensitive data to unprivileged users [22]. This enables the use of sensitive data without violating privacy rights. Under the umbrella term "secure (data) provenance", Torra et al. summarize four requirements [36]: distribution of provenance, integrity of provenance, availability of provenance and privacy and confidentiality of provenance. Tailoring these requirements to the medical research domain will be an important part of future research and the implementation in the UMG MeDIC and other Data Integration Centers.

Provenance data is considered essential within big data applications [37]. Although medical research IT-infrastructures are not necessarily big data applications [1], some characteristics (variety, veracity and, to some degree, volume of data) remain important in data integration in translational research. Within the GIS community, distributing data with its provenance is common practice [38]. Hence, federating provenance across multiple sites in ways similar to the federation of medical data sets is an important focus which needs to be addressed when implementing data federation infrastructures. FHIR, as a standard for medical data exchange, incorporates W3C PROV and provides a starting point to focus on provenance federation.

Visualization and exploration of provenance

Utilizing provenance usually comes in two steps: visualization and analysis. Several solutions to visualize provenance exist, primarily an acyclic directed graph defined by PROV-DM [6]. Prov-O-Viz by Hoekstra and Groth is another graph-based visualization that focuses on the data flow within activities and entities [39]. Schreiber and Struminski present a solution to make provenance understandable for laymen using a representation with comics [40]. The latter solution will serve as a starting point to implement provenance visualization solutions for all stakeholders. In combination with the data

itself, provenance is able to assist visualization of medical records in a similar approach to EVLIN [41].

Due to the nature of PROV-DM, analysis of provenance – that is, computations utilizing provenance without visualizing it – are closely related to network analysis techniques. Huynh et al. tailored several metrics established within network analysis to provenance data, demonstrating the usefulness of such techniques [31]. The usage of provenance templates may add insight in combination with network analysis [26].

Provenance insight serves different purposes [7]. Tackling all purposes in a single application would result in a highly complex tool that requires not only a deep understanding of provenance itself but also of the data it describes. To circumvent this challenge, insights from research regarding data quality of medical data sets can be exploited. Kahn et al. recommend utilizing provenance to improve data quality and data reuse of observational and administrative data [42]. Also, the term "fitness for purpose" was coined to describe that every use case requires a specialized view on data sets to gain optimal performance [43]. Following this paradigm, we aim to make provenance usable for each stakeholder by selecting relevant provenance data sets and by using a specialized applications, tailored to a small set of provenance questions.

As an example, the GCP-compliant data capture software secuTrial (https://www.secutrial.com) implements a function to review audit trails within the application itself, bringing provenance directly into the context of the data itself.

Conclusions

The means to extract, store, and utilize provenance have been tackled individually by several tools within the provenance community. Solutions that grant all of these features are scarce but show great potential to improve reproducibility, recall, insight, presentation, and enable meta-analysis of processing steps within heterogeneous IT-infrastructures. We aim to learn from successful implementations in other disciplines (GIS [9], bio-informatics [10] and software evolution [32,33]), and harness this knowledge to jump start provenance applications in medical research implemented in a provenance toolbox that pools all features related to extraction, storage, and utilization of provenance. This will improve reproducibility, re-use, interoperability, and overall quality in medical research [8,44].

Acknowledgements

This work was supported by the German Federal Ministry of Education and Research (BMBF) within the framework of the research and funding concepts of the Medical Informatics Initiative (01ZZ1802B/HiGHmed) and the BMBF e:Med program (01ZX1606C/sysINFLAME, 031L0024A/myPathSem).

References

[1] C.R. Bauer, N. Umbach, B. Baum, K. Buckow, T. Franke, R. Grütz, L. Gusky, S.Y. Nussbeck, M. Quade, S. Rey, T. Rottmann, O. Rienhoff, and U. Sax, Architecture of a Biomedical Informatics Research Data Management Pipeline, *Stud Health Technol Inform*. **228** (2016) 262–266. doi:10.3233/978-1-61499-678-1-262.

[2] C. Safran, Update on Data Reuse in Health Care, *Yearb Med Inform*. **26** (2017) 24–27. doi:10.15265/IY-2017-013.

[3] S.M. Meystre, C. Lovis, T. Bürkle, G. Tognola, A. Budrionis, and C.U. Lehmann, Clinical Data Reuse or Secondary Use: Current Status and Potential Future Progress, *Yearb Med Inform*. **26** (2017) 38–52. doi:10.15265/IY-2017-007.

[4] M.D. Wilkinson, M. Dumontier, Ij.J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L.B. da S. Santos, P.E. Bourne, J. Bouwman, A.J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C.T. Evelo, R. Finkers, A. Gonzalez-Beltran, A.J.G. Gray, P. Groth, C. Goble, J.S. Grethe, J. Heringa, P.A.C. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S.J. Lusher, M.E. Martone, A. Mons, A.L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M.A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, and B. Mons, The FAIR Guiding Principles for scientific data management and stewardship, *Scientific Data*. **3** (2016). doi:10.1038/sdata.2016.18.

[5] M. Parciak, C.R. Bauer, B. Baum, H. Kusch, and U. Sax, Technical Aspects of Data Provenance in Clinical Trials, in: 62. Jahrestagung Der Deutschen Gesellschaft Für Medizinische Informatik, Biometrie Und Epidemiologie e. V. (GMDS), German Medical Science GMS Publishing House, Oldenburg, Germany, 2017. doi:10.3205/17gmds155.

[6] L. Moreau, and P. Missier, PROV-DM: The PROV Data Model, (2013). http://www.w3.org/TR/2013/REC-prov-dm-20130430/ (accessed August 27, 2018).

[7] E.D. Ragan, A. Endert, J. Sanyal, and J. Chen, Characterizing Provenance in Visualization and Data Analysis: An Organizational Framework of Provenance Types and Purposes, *IEEE Transactions on Visualization and Computer Graphics*. **22** (2016) 31–40. doi:10.1109/TVCG.2015.2467551.

[8] V. Curcin, S. Miles, R. Danger, Y. Chen, R. Bache, and A. Taweel, Implementing interoperable provenance in biomedical research, *Futur. Gener. Comp. Syst.* **34** (2014) 1–16. doi:10.1016/j.future.2013.12.001.

[9] Y. Cao, C. Jones, V. Cuevas-Vicenttín, M.B. Jones, B. Ludäscher, T. Mcphillips, P. Missier, C. Schwalm, P. Slaughter, D. Vieglais, L. Walker, and Y. Wei, DataONE: A Data Federation with Provenance Support, in: Provenance and Annotation of Data and Processes, Springer-Verlag New York, Inc., McLean, VA, USA, 2016: pp. 230–234. doi:10.1007/978-3-319-40593-3_28.

[10] J. Goecks, A. Nekrutenko, and J. Taylor, Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences, *Genome Biology*. **11** (2010). doi:10.1186/gb-2010-11-8-r86.

[11] A. Branson, R. McClatchey, J.-M. Le Goff, and J. Shamdasani, CRISTAL: A practical study in designing systems to cope with change, *Information Systems*. **42** (2014) 139–152. doi:10.1016/j.is.2013.12.009.

[12] B. Haarbrandt, B. Schreiweis, S. Rey, U. Sax, S. Scheithauer, O. Rienhoff, P. Knaup-Gregori, U. Bavendiek, C. Dieterich, B. Brors, I. Kraus, C. Thoms, D. Jäger, V. Ellenrieder, B. Bergh, R. Yahyapour, R. Eils, H. Consortium, and M. Marschollek, HiGHmed – An Open Platform Approach to Enhance Care and Research across Institutional Boundaries, *Methods of Information in Medicine*. **57** (2018) e66–e81. doi:10.3414/ME18-02-0002.

[13] M. Parciak, Provenancekonzept für Datenbestände aus einer heterogenen Forschungsinfrastuktur (am Beispiel einer klinischen Forschergruppe), Masters Thesis, Georg-August-Universität Göttingen, 2017.

[14] L. Moreau, The Foundations for Provenance on the Web, *WEB*. **2** (2010) 99–241. doi:10.1561/1800000010.

[15] M. Herschel, R. Diestelkämper, and H.B. Lahmar, A survey on provenance: What for? What form? What from?, *The VLDB Journal*. **26** (2017) 881–906. doi:10.1007/s00778-017-0486-1.

[16] B. Pérez, J. Rubio, and C. Sáenz-Adán, A systematic review of provenance systems, *Knowl Inf Syst*. (2018) 1–49. doi:10.1007/s10115-018-1164-3.

[17] J. Kunze, and T. Baker, The Dublin Core Metadata Element Set, *RFC 5013*. (2007). doi:10.17487/RFC5013.
[18] D. Gammack, S. Scott, and A.P. Chapman, Modelling Provenance Collection Points and Their Impact on Provenance Graphs, in: Provenance and Annotation of Data and Processes, Springer, Cham, 2016: pp. 146–157. doi:10.1007/978-3-319-40593-3_12.
[19] L. Moreau, H. Hua, C. Tilmes, and S. Zednik, PROV-XML: The PROV XML Schema, (2013). http://www.w3.org/TR/2013/NOTE-prov-xml-20130430/ (accessed October 25, 2018).
[20] L. Moreau, O. Hartig, Y. Simmhan, J. Myers, T. Lebo, K. Belhajjame, S. Miles, S. Soiland-Reyes, G. Klyne, and P. Groth, PROV-AQ: Provenance Access and Query, (2013). http://www.w3.org/TR/2013/NOTE-prov-aq-20130430/ (accessed October 25, 2018).
[21] T. Lebo, S. Sahoo, and D. McGuinness, PROV-O: The PROV Ontology, (2013). http://www.w3.org/TR/2013/REC-prov-o-20130430/ (accessed October 25, 2018).
[22] A. Chapman, B.T. Blaustein, L. Seligman, and M.D. Allen, PLUS: A provenance manager for integrated information, in: 2011 IEEE International Conference on Information Reuse Integration, IEEE, Las Vegas, NV, USA, 2011: pp. 269–275. doi:10.1109/IRI.2011.6009558.
[23] M. Parciak, C.R. Bauer, R. Lodahl, C. Thoms, H. Kusch, S. Rey, and U. Sax, PROV@TOS, a Java Wrapper to capture provenance for Talend Open Studio jobs, *63. Jahrestagung Der Deutschen Gesellschaft Für Medizinische Informatik, Biometrie Und Epidemiologie e.V. (GMDS)*. (2018). doi:10.3205/18gmds096.
[24] T. Bender, C.R. Bauer, M. Parciak, R. Lodahl, and U. Sax, FAIR conform ETL processing in translational research, *63. Jahrestagung Der Deutschen Gesellschaft Für Medizinische Informatik, Biometrie Und Epidemiologie e.V. (GMDS)*. (2018). doi:10.3205/18gmds095.
[25] IHE International, Inc., Integration Profiles, IHE International, Inc., 2017. https://www.ihe.net/uploadedFiles/Documents/ITI/IHE_ITI_TF_Vol1.pdf (accessed October 25, 2018).
[26] V. Curcin, E. Fairweather, R. Danger, and D. Corrigan, Templates as a method for implementing data provenance in decision support systems, *Journal of Biomedical Informatics*. **65** (2017) 1–21. doi:10.1016/j.jbi.2016.10.022.
[27] P. Groth, J. Sheng, S. Miles, S. Munroe, V. Tan, S. Meacham, and L. Moreau, An Architecture for Provenance Systems, University of Southampton, 2006.
[28] D. Koop, E. Santos, B. Bauer, M. Troyer, J. Freire, and C.T. Silva, Bridging Workflow and Data Provenance Using Strong Links, in: Scientific and Statistical Database Management, Springer, Berlin, Heidelberg, 2010: pp. 397–415. doi:10.1007/978-3-642-13818-8_28.
[29] R. Bjugn, and B. Casati, Stakeholder Analysis: A Useful Tool for Biobank Planning, *Biopreservation and Biobanking*. **10** (2012) 239–244. doi:10.1089/bio.2011.0047.
[30] S. Miles, P. Groth, S. Munroe, and L. Moreau, PrIMe: A Methodology for Developing Provenance-aware Applications, *ACM Transactions on Software Engineering and Methodology*. **20** (2011) 8:1–8:42. doi:10.1145/2000791.2000792.
[31] T.D. Huynh, M. Ebden, J. Fischer, S. Roberts, and L. Moreau, Provenance Network Analytics, *Data Min Knowl Disc*. (2018) 1–28. doi:10.1007/s10618-017-0549-3.
[32] S. Herbold, A. Trautsch, and J. Grabowski, Global vs. local models for cross-project defect prediction, *Empir Software Eng*. **22** (2017) 1866–1902. doi:10.1007/s10664-016-9468-y.
[33] V. Honsel, D. Honsel, S. Herbold, J. Grabowski, and S. Waack, Mining Software Dependency Networks for Agent-Based Simulation of Software Evolution, in: 2015 30th IEEE/ACM International Conference on Automated Software Engineering Workshop (ASEW), 2015: pp. 102–108. doi:10.1109/ASEW.2015.9.
[34] D. Dominguez-Sal, P. Urbón-Bayes, A. Giménez-Vañó, S. Gómez-Villamor, N. Martínez-Bazán, and J.L. Larriba-Pey, Survey of Graph Database Performance on the HPC Scalable Graph Analysis Benchmark, in: H.T. Shen, J. Pei, M.T. Özsu, L. Zou, J. Lu, T.-W. Ling, G. Yu, Y. Zhuang, and J. Shao (Eds.), Web-Age Information Management, Springer Berlin Heidelberg, 2010: pp. 37–48. doi:10.1007/978-3-642-16720-1_4.
[35] A. Gehani, and D. Tariq, SPADE: support for provenance auditing in distributed environments, in: Middleware 2012, Springer, Berlin, Heidelberg, Montreal, Quebec, Kanada, 2012: pp. 101–120. doi:10.1007/978-3-642-35170-9_6.
[36] V. Torra, G. Navarro-Arribas, D. Sanchez-Charles, and V. Muntés-Mulero, Provenance and Privacy, in: V. Torra, Y. Narukawa, A. Honda, and S. Inoue (Eds.), Modeling Decisions for Artificial Intelligence, Springer International Publishing, 2017: pp. 3–11.
[37] B. Glavic, Big Data Provenance: Challenges and Implications for Benchmarking, in: Specifying Big Data Benchmarks, Springer, Berlin, Heidelberg, 2014: pp. 72–80. doi:10.1007/978-3-642-53974-9_7.
[38] G. Closa, J. Masó, B. Proß, and X. Pons, W3C PROV to describe provenance at the dataset, feature and attribute levels in a distributed environment, *Computers, Environment and Urban Systems*. **64** (2017) 103–117. doi:10.1016/j.compenvurbsys.2017.01.008.
[39] R. Hoekstra, and P. Groth, PROV-O-Viz - Understanding the Role of Activities in Provenance, in: Provenance and Annotation of Data and Processes, Springer-Verlag New York, Inc., Cologne, Germany, 2015: pp. 215–220. doi:10.1007/978-3-319-16462-5_18.
[40] A. Schreiber, and R. Struminski, Visualizing Provenance using Comics, in: 9th USENIX Workshop on the Theory and Practice of Provenance, USENIX Association, Seattle, WA, USA, 2017. https://www.usenix.org/conference/tapp17/workshop-program/presentation/schreiber.
[41] H.B. Lahmar, M. Herschel, M. Blumenschein, and D.A. Keim, Provenance-Based Visual Data Exploration with EVLIN., in: Proceedings of the 21th International Conference on Extending Database Technology, EDBT 2018, Vienna, Austria, March 26-29, 2018., 2018: pp. 686–689. doi:10.5441/002/edbt.2018.85.
[42] M.G. Kahn, J.S. Brown, A.T. Chun, B.N. Davidson, D. Meeker, P.B. Ryan, L.M. Schilling, N.G. Weiskopf, A.E. Williams, and M.N. Zozus, Transparent Reporting of Data Quality in Distributed Data Networks, EGEMs (Generating Evidence & Methods to Improve Patient Outcomes). 3 (2015) 7. doi:10.13063/2327-9214.1052.
[43] N.G. Weiskopf, and C. Weng, Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research, *J Am Med Inform Assoc*. **20** (2013) 144–151. doi:10.1136/amiajnl-2011-000681.
[44] B. Baum, C.R. Bauer, T. Franke, H. Kusch, M. Parciak, T. Rottmann, N. Umbach, and U. Sax, Opinion paper: Data provenance challenges in biomedical research, *It - Information Technology*. **59** (2017) 191–196. doi:10.1515/itit-2016-0031.

Address for correspondence

Corresponding author: marcel.parciak@med.uni-goettingen.de