

Towards dimensions and granularity in a unified workflow and data provenance framework

Tanja Auge¹, Sascha Genehr^{2,3,*}, Meike Klettke¹, Frank Krüger⁴ and Max Schröder³

¹University of Regensburg, Germany

²University of Rostock, Germany

³Rostock University Library, Germany

⁴Wismar University of Applied Sciences, Germany

Abstract

Provenance information are essential for the traceability of scientific studies or experiments and thus crucial for ensuring the credibility and reproducibility of research findings. This paper discusses a comprehensive provenance framework combining the two types 1. *workflow provenance*, and 2. *data provenance* as well as their dimensions and granularity, which enables the answering of **W7+1** provenance questions. We demonstrate the applicability by employing a biomedical research use case, that can be easily transferred into other scientific fields. An integration of these concepts into a unified framework enables credibility and reproducibility of the research findings.

Keywords

data provenance, workflow provenance, dimensions, granularity, W7 questions, wetlab data

1. Introduction

The traceability of scientific studies and experiments is essential for the credibility and reproducibility of their findings. Provenance information provide the essential data for this purpose, e.g. the sequence of activities that resulted in the creation of measurement data, the involved persons or the investigated samples. Thereby, provenance can have a significant impact on the scientific value of the research data. Many approaches and definitions of provenance in different contexts have been proposed that cover some aspects of either the scientific domain or the provenance information itself, e.g. workflow systems [1, 2], jupyter notebooks [3], and lab documentation [4, 5]. In this paper, we discuss a unification of these definitions along typical biomedical research processes, which can be easily adapted to other scientific domains.

Scientific investigations in the biomedical domain can be categorized by their kind into: 1. *in-silico*, 2. *in-vitro*, and 3. *in-vivo experiments*. Often, in-silico studies are performed to simulate real world phenomena based on data obtained from previous in-vitro or in-vivo experiments, i.e. experiments in laboratories respectively living organisms. New findings from the in-silico studies are then again used to validate the effects in in-vitro or in-vivo experiments, leading to

LWDA 2024

*Corresponding author.

✉ tanja.auge@ur.de (T. Auge); sascha.genehr@uni-rostock.de (S. Genehr); meike.klettke@ur.de (M. Klettke); frank.krueger@hs-wismar.de (F. Krüger); max.schroeder@uni-rostock.de (M. Schröder)

ORCID 0009-0006-2150-9713 (T. Auge); 0000-0002-1702-6878 (S. Genehr); 0000-0003-0551-8389 (M. Klettke); 0000-0002-7925-3363 (F. Krüger); 0000-0003-1522-494X (M. Schröder)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

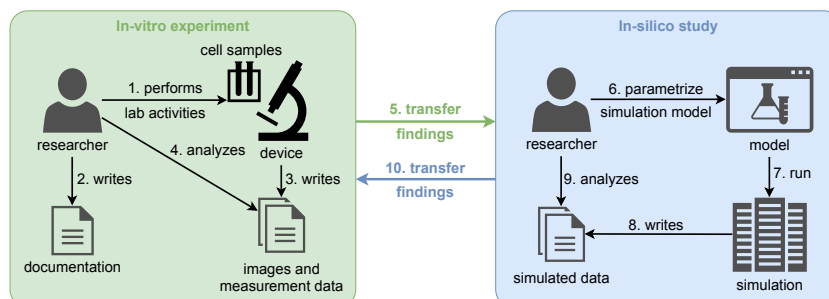


Figure 1: Biomedical use case with two experiment types (in-vitro and in-silico) whose research findings are used to validate and optimize the other. Activities are numbered to illustrate the trajectory.

a closed loop of data exchange in this scientific process with respect to the experimental kind. For the discussion of the particular provenance concepts, we employ the following use case, that we restrict on the two kinds, in-silico and in-vitro, as the challenges and requirements for provenance tracking in in-vitro and in-vivo are quite similar.

Use case. A researcher starts with in-vitro experiments in a wetlab environment measuring the calcium ion mobilization in osteoblasts. In order to measure this real world phenomenon, a series of actions is performed in the laboratory environment. The course of action is standardized in a protocol and documented in a so-called *electronic lab notebook (ELN)*. In particular, this involves the preparation of cell environments as well as the set-up and measuring of the prepared cells using a microscope in conjunction with a stimulation device. Measurement data are collected as microscopy images that are analysed to extract tabular fluorescence intensity data. In order to exclude environmental and other issues, the experiment is repeated multiple times, resulting in a series of measurements as well as documentation following the same experimental protocol.

In a second step, an in-silico study is set up to further elaborate the stimulation parameter settings based on the real world data from the in-vitro experiments. The measurements are used to specify and parameterize a simulation model for this phenomenon. Specifically, the simulation model in combination with the computing environment is run under different simulation settings. The simulation is repeated in order to reduce random influences. New findings are then interpreted from all simulations, which provide insights for the optimization of stimulations. Lastly, these findings are validated in in-vitro experiments again (cf. Figure 1).

In this entire sequence of experiments, provenance information is needed to support the credibility and reproducibility of the research findings. As such, activities, involved persons and entities (e.g. data, samples, devices and models) are encoded in the provenance information.

Contribution. 1. We discuss a unification of workflow and data provenance including the dimension and granularity based on the current state of the art, and 2. we extend the **W7** provenance question concept [6] to **W7+1** provenance questions and discuss dimensions and granularity for this unification.

2. Provenance Concepts

Provenance is a very broad term with many meanings and definitions. In general, provenance refers to the origin of an object or captures information about the creation or evolution process of an object – in our example these objects are simulation models and measurement data. With respect to the (*scientific*) *workflow*, provenance information encode the tracing of the particular research objects resulting in the scientific finding. In the context of a (*scientific*) *database* [7], provenance includes information about the origin of a data element and details about its scientific processing (parameters, software versions, etc.). Thus, provenance information enable the verification of (scientific) processes as well as database queries [8, 9, 10].

2.1. Provenance Types

Workflow provenance. This type refers to the process of a dataset’s derivation in the form of a scientific workflow. As described in [8] a (*scientific*) *workflow* is a directed graph where nodes represent arbitrary functions or modules in general with some input, output, and parameters. Edges model a predefined data or control flow between these modules. It includes information about the workflow’s procedure, deviations from it, and the execution in general.

Workflow provenance is often expressed in terms of description logics and derived concepts like ontologies, e.g. the PROV-O ontology [11]. The *PROV-O ontology* distinguishes three core concepts: *entities*, *activities*, and *agents*. Entities are data or objects and can be derived from other entities. Activities can generate or use entities. Agents can perform or control activities or produce entities. An example for our biomedical use case in PROV terms is shown in Figure 2.

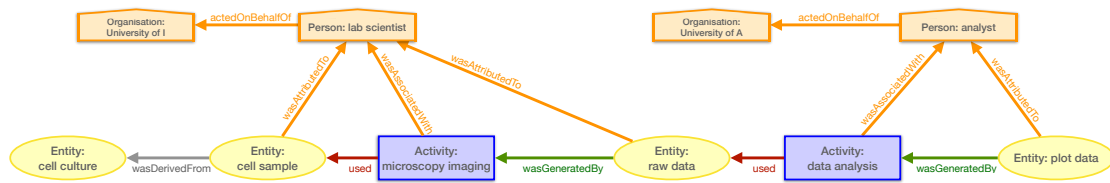


Figure 2: Biomedical in-vitro experiment with organizations and persons (agents), cell cultures and samples as well as data sets (entities), and research activities (activity) including their relationships in the PROV standard. Note that PROV relationship direction is typically from the result to the origin.

Data provenance. Data provenance plays a role in data wrangling, which can often be expressed in terms of relational algebra. It describes the derivation of a piece of data from data sets, typically the result of a database query Q . Considering our use case, the following query encodes the comparison of the fluorescence intensity from two stimulation experiments with respect to their voltage: `SELECT voltage_2 FROM R NATURAL JOIN S WHERE intensity_1 < intensity_2`, where two tables R and S are joined on a unique join attribute `sample_id`, against a source database D (consisting of R and S). The query result is stored in a table T :

sample_id	intensity_1	voltage_1
1	40.027	0.9
2	41.038	1.4
2	41.033	1.4

r_1
 r_{2,t_1}
 r_{2,t_2}

sample_id	intensity_2	voltage_2
1	40.375	1.0
1	39.998	1.3
1	42.001	1.0

s_1
 s_2
 s_3

voltage_2
1.0

$r_1 \cdot s_1 + r_1 \cdot s_3$
or $\{(r_1, s_1), (r_1, s_3)\}$

The rows (*tuples*) are provided with *provenance IDs* r_1, s_1, \dots . In the case of evolving databases (cf. Section 3.2), additional time stamps are required, processed by $r_{2,t_1}, r_{2,t_2}, \dots$

In relational databases, data provenance is often specified in the form of *provenance polynomials* [12] or *witness basis* [13]. While the former specifies a concrete calculation rule in the form of a polynomial defined by a commutative semi-ring $(\mathbb{N}[X], +, \cdot, 0, 1)$ with $+$ for duplicates (resulting from projection or union) and \cdot for (natural) joins [14], the witness basis describes the set of all relevant witnesses. A witness itself contains all the tuple IDs needed to reconstruct a tuple. Then, the tuple in the table T can be explained by the provenance polynomial $r_1 \cdot s_1 + r_1 \cdot s_3$ or the witness basis $\{\{r_1, s_1\}, \{r_1, s_3\}\}$. Both representations contain information about the natural join ($r_1 \cdot s_1$ or $\{r_1, s_1\}$) and the duplicate ($t + t'$ or $\{\{t\}, \{t'\}\}$) for the witnesses $t = \{r_1, s_1\}$ and $t' = \{r_1, s_3\}$, which are generated by answering the query Q .

2.2. Provenance Dimensions and Granularity

Dimensions. The literature [8, 10] distinguishes between three dimensions of workflow provenance: *retrospective provenance*, *prospective provenance* and *evolution provenance*. A provenance solution may support only one, two, or all three of them. While *retrospective provenance* provides information about past workflow executions and data derivations, *prospective provenance* captures the structure and static context of a workflow [8]. The latter is independent of any workflow execution or input data and can be understood as a recipe for future workflow executions. Retrospective provenance, however, preserves information on the resources that are accessed or generated during execution. *Evolution provenance* reflects changes made between two iterations of a workflow.

While the three provenance dimensions do not rely on each other's presence [8], it may be beneficial to capture multiple dimensions in order to provide detailed information about the executed processes, the underlying procedures, and their relation to other research processes. For data provenance, to the best of our knowledge, the three dimensions have not yet been discussed. In Section 3.2, we discuss potential applications of dimensions in data provenance.

Granularity. The granularity of a provenance model depends on the level of details with which a workflow is described [8]. This level of detail is defined by seven *provenance questions* (cf. Section 3.1). Depending on how many of the provenance questions [6] are answered, the granularity of a model can be considered *coarse-* or *fine-grained*. While the former yields a broad description of the process, the latter allows for a more detailed description of a process.

3. Combining the provenance concepts

Combining data provenance with provenance of workflows provides a more granular insight into research findings by providing insights into the data origin. This combination is particularly important in biomedical research or data science applications, as the data processing steps are largely carried out outside the (relational) databases; this concerns algorithms for data pre-processing, data cleaning and data transformation. In order for their effects to be assessable, the workflow provenance information must be added to the data provenance information.

While workflow provenance typically is encoded in the form of knowledge graphs using W3C PROV [11], the data provenance can be encoded within specialized data files (entities) that

are also integrated into the workflow model. Alternatively, the data provenance concept can be extended from tuples in databases to the file level so that data provenance and workflow provenance are specified and processed in parallel but uncoupled (cf. Section 3.2).

3.1. Provenance Questions

Traceability and reproducibility raise many questions about where things come from and how they were processed: *Which* datasets are affected by an error or bug? *How* are datasets affected by modifying a parameter? *Why* is an expected value *not* included (in the result)? The concept of workflow provenance typically answers seven question types including their combinations (**W7**, [6]), whereas data provenance addresses three of these question types: *how*, *why*, and *where*. However, we propose that also the additional *why not* question, as it is known from *provenance games* [15] in the context of data provenance, can be specified on workflow provenance as well as *what*, one of the coarse-grained **W7** questions, on data provenance. Though under a closed world assumption, the answering of the remaining negated questions such as *where not* is conceptually valid, but in real world settings infeasible.

Regarding data provenance, answers to *why not* provide an explanation, why an expected result is not part of the query result. For example, the query Q from above would, for a value range of $[1.0, 1.5]$ for `voltage_2`, never yield the result 1.3, leading to the insight that the query would to be modified. *Why not* regarding workflow provenance has, to the best of our knowledge, not been stated in the literature. This question, however, contains valuable information to workflow optimization and analysis, as it gives potential reasons about certain choices in experiment design, such as why a particular procedure has been followed instead of another. We suggest as answers, for instance, notes, design comments, or warnings that are collected during experiment planning and execution.

Schema changes can result in dirty data, including rounding errors or omitting characters (spaces or leading zeros). To solve this problem, we defined *what* also for data provenance [16]. It stores the data type of all attributes. In workflow provenance, however, *what* could be answered by the order of processes that were performed [17]. *When*, *who*, and *which* are only defined for workflow provenance as they are not relevant outside the workflow scope.

These extensions of the **W7** questions, we propose to call **W7+1**. The **W7+1** question types as well as their classification according to the provenance types workflow provenance (highlighted in blue) and data provenance (highlighted in orange) are shown in Figure 3.

3.2. Provenance Dimensions and Granularity

Extending the provenance dimensions and granularities towards our unified workflow and data provenance framework, we define them as follows:

Dimensions. The three dimensions of provenance with respect to the documentation within wetlabs are discussed in [18]. Summarizing their arguments, prospective provenance corresponds to the protocol, i.e., the *standard operating procedure (SOP)* in our use case, and retrospective workflow provenance encodes the particular experimental procedure. Evolution provenance for the workflow comes in, when the simulation results reveal new findings so that the underlying SOP is adjusted for future experiments.

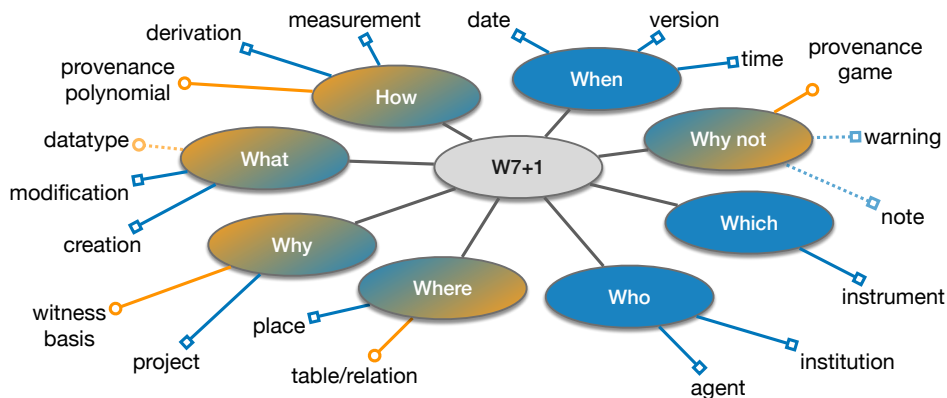


Figure 3: W7+1 provenance questions for workflow provenance (—□) and data provenance (—○) including typical (solid) and new defined (dotted lines) answer options.

Due to the specification of data provenance reasoning over a specific selection of data, it typically encodes the retrospective dimension. However, in the case of *evolving databases* such as from measurement data of wetlab experiments including schema changes and data updates over time, the additional dimension of evolution provenance becomes necessary. It reflects changes made between two states of a database instances by time stamps t_1, t_2, \dots . As updating data means adding a new modified data record, the clue lies in the choice of the provenance IDs which is suffixed by a time stamp [16]. The modified tuples receive the same original number but with a different time stamp, such as r_{2,t_1} and r_{2,t_2} in database D of Section 2.1.

Granularity. Depending on the literature, the terms fine- and coarse-grained provenance have different meanings. While [10] used the former as an alias for data provenance and the latter for workflow provenance, for [8] the granularity definition only exists in the context of workflow provenance, where the coarseness means the level of detail of a workflow description.

In [17], the terms fine- and coarse-grained provenance are specified with respect to different levels of granularity for the answers of provenance questions. Depending on the needs of the user, the question *how* measurement data was created might reflect the entire experiment as a single activity (coarse-grained) or the course of atomic activities in the wetlab (fine-grained). Finding a balance in the granularity of provenance modelling between fine-grained modelling, which clearly impacts storage and computing resources, and coarse-grained modelling which restricts the expressiveness has previously been identified as one of the core challenges [19].

In the context of data provenance, we propose the terms to be specified as follows: While with fine-grained provenance a data element corresponds to the tuples in a database, i.e. a particular measurement data point, with coarse-grained provenance a data element corresponds to the entire measurement file, each provided by a unique ID. By collecting these provenance IDs in a so-called *ID database* – in addition to the provenance IDs, the ID database also contains the file name, file paths and further provenance information – the same evaluations can be performed on the file level as before at the data level. As such, data provenance and workflow provenance can be combined from a unified viewpoint.

4. Conclusion

In this paper, we discussed the combination of workflow provenance and data provenance and their effects on the dimensions and granularity employing a biomedical use case. This discussion represents an initial step towards a unified provenance framework encoding a comprehensive view on research processes by revealing homogeneous concepts within the current literature and necessary extensions. Thus, this work serves as a motivation for the developing of a formal specification of the unified framework in future.

Acknowledgments

SG and **FK** are funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - SFB 1270/2 – 299150580.

Authors' contributions

Author contributions according to CRediT: **TA** Conceptualization, Visualization, Writing - Original Draft. **SG** Conceptualization, Visualization, Writing - Original Draft. **MK** Supervision, Writing - Review & Editing. **FK** Conceptualization, Funding acquisition, Supervision, Writing - Review & Editing. **MS** Conceptualization, Visualization, Writing - Original Draft. All authors read and approved the final manuscript.

References

- [1] K. Belhajjame, K. Wolstencroft, O. Corcho, T. Oinn, F. Tanoh, A. William, C. Goble, Metadata Management in the Taverna Workflow System, in: CCGRID, IEEE Computer Society, 2008, pp. 651–656. doi:10.1109/ccgrid.2008.17.
- [2] I. Altintas, O. Barney, E. Jaeger-Frank, Provenance Collection Support in the Kepler Scientific Workflow System, in: Provenance and Annotation of Data, Springer Berlin Heidelberg, Berlin, Heidelberg, 2006, pp. 118–132. doi:10.1007/11890850_14.
- [3] S. Samuel, B. König-Ries, ProvBook: Provenance-based Semantic Enrichment of Interactive Notebooks for Reproducibility, in: ISWC (P&D/Industry/BlueSky), 2018.
- [4] L. N. Soldatova, D. Nadis, R. D. King, P. S. Basu, E. Haddi, V. Baumlé, N. J. Saunders, W. Marwan, B. B. Rudkin, EXACT2: the semantics of biomedical protocols, BMC Bioinformatics 15 (2014). doi:10.1186/1471-2105-15-s14-s5.
- [5] O. Giraldo, A. García, O. Corcho, SMART Protocols: SeMAntic RepresenTation for Experimental Protocols (2014). doi:10.13140/2.1.2778.5288.
- [6] S. Ram, J. Liu, Understanding the Semantics of Data Provenance to Support Active Conceptual Modeling, in: LNCS, Springer Berlin Heidelberg, 2007, pp. 17–29. doi:10.1007/978-3-540-77503-4_3.
- [7] S. C. Boulakia, W. Tan, Provenance in scientific databases, in: Encyclopedia of Database Systems (2nd ed.), Springer, 2018.

- [8] M. Herschel, R. Diestelkämper, H. Ben Lahmar, A survey on provenance: What for? What form? What from?, *VLDB J.* 26 (2017) 881–906.
- [9] J. Freire, D. Koop, E. Santos, C. T. Silva, Provenance for Computational Tasks: A Survey, *Comput. Sci. Eng.* 10 (2008) 11–21.
- [10] B. Pérez, J. Rubio, C. Sáenz-Adán, A systematic review of provenance systems, *Knowl. Inf. Syst.* 57 (2018) 495–543.
- [11] K. Belhajjame, H. Deus, D. Garijo, G. Klyne, P. Missier, S. Soiland-Reyes, S. Zednik, *PROV Model Primer*, 2013. URL: <https://www.w3.org/TR/prov-primer/>.
- [12] T. J. Green, V. Tannen, The semiring framework for database provenance, in: *PODS*, ACM, 2017, pp. 93–99.
- [13] P. Buneman, S. Khanna, W. C. Tan, Why and where: A characterization of data provenance, in: *ICDT*, volume 1973 of *LNCS*, Springer, 2001, pp. 316–330.
- [14] T. J. Green, G. Karvounarakis, V. Tannen, Provenance semirings, in: *PODS*, ACM, 2007, pp. 31–40.
- [15] S. Köhler, B. Ludäscher, D. Zinn, First-Order Provenance Games, in: *In Search of Elegance in the Theory and Practice of Computation*, volume 8000 of *LNCS*, Springer, 2013, pp. 382–399.
- [16] T. Auge, A. Heuer, Tracing the History of the Baltic Sea Oxygen Level, in: *BTW*, volume P-311 of *LNI*, GI, 2021, pp. 337–348.
- [17] M. Schröder, S. Staehlke, P. Groth, J. B. Nebe, S. Spors, F. Krüger, Structure-based knowledge acquisition from electronic lab notebooks for researchdata provenance documentation, *Journal of Biomedical Semantics* (2022). doi:10.1186/s13326-021-00257-x.
- [18] S. Genehr, M. Bielfeldt, M. Schröder, S. Stählke, B. Nebe, S. Spors, F. Krüger, Modelling three dimensions of provenance for wet-lab experiments: prospective, retrospective, and evolution, in: *Workshop on Metadata and Research (objects) Management for Linked Open Science (3. : 2023 : Online)*, PUBLISSO, 2023. doi:10.4126/FRL01-006444990.
- [19] K. Gierend, F. Krüger, S. Genehr, F. Hartmann, F. Siegel, D. Waltemath, T. Ganslandt, A. A. Zeleke, Provenance information for biomedical data and workflows: A scoping review, *Journal of Medical Internet Research* (2024). doi:10.2196/51297, forthcoming.