Provenance: On and Behind the Screens

Melanie Herschel
Institute for Parallel and Distributed Systems
University of Stuttgart
Stuttgart, Germany
melanie.herschel@ipvs.uni-stuttgart.de

Marcel Hlawatsch
Visualization Research Center
University of Stuttgart
Stuttgart, Germany
marcel.hlawatsch@visus.uni-stuttgart.de

ABSTRACT

Collecting and processing *provenance*, i.e., information describing the production process of some end product, is important in various applications, e.g., to assess quality, to ensure reproducibility, or to reinforce trust in the end product. In the past, different types of provenance meta-data have been proposed, each with a different scope. The first part of the proposed tutorial provides an overview and comparison of these different types of provenance.

To put provenance to good use, it is essential to be able to interact with and present provenance data in a user-friendly way. Often, users interested in provenance are not necessarily experts in databases or query languages, as they are typically domain experts of the product and production process for which provenance is collected (biologists, journalists, etc.). Furthermore, in some scenarios, it is difficult to use solely queries for analyzing and exploring provenance data. The second part of this tutorial therefore focuses on enabling users to leverage provenance through adapted visualizations. To this end, we will present some fundamental concepts of visualization before we discuss possible visualizations for provenance.

1. PROVENANCE TYPES AND VISUALIZATIONS

Provenance generally refers to any information that describes the production process of an end product, which can be anything from a piece of data to a physical object (food, chemical compound, business report, etc.). Thus, in general, provenance information includes meta-data about entities, processes, activities, and persons involved in the production process. As discussed below, provenance information exists under various forms, i.e., different *provenance types* exist. Essentially, these different types have been motivated by varying application requirements and usage scenarios.

Visualization deals with the visual representation of data. Since the visual channel is the largest information channel of the human, visualization enables the analysis and exploration of large amounts of data, especially when combined with suitable interaction methods. Considering the variety of provenance types, it is clear that there is no single visualization concept that can represent all types of provenance in the best possible way. To be most efficient, the visualization must be developed and adapted to each type of provenance and each type of application. This tutorial therefore presents general visualization concepts and approaches applicable to provenance data. The goal is to provide basic knowledge for designing, developing, and applying provenance visualization.

Structure. In the remainder of this section, we provide some background on both provenance and visualization in the context of provenance. Section 2 summarizes the learning objectives and the tutorial outline. Tutorial format and audience are covered in Section 3 before the tutorial's instructors are briefly introduced in Section 4.

1.1 Provenance Types

As mentioned above, different types of provenance have been proposed, guided by varying application needs. We classify the prevalent types of provenance into four main types of provenance, namely *data provenance*, *workflow provenance*, *information systems provenance*, and *provenance meta-data*. Roughly speaking, these four types form a type hierarchy, as illustrated in Figure 1. Essentially, as we move up from one level to a more specific level of the provenance type hierarchy, the more specialized the provenance type becomes, i.e., its domain reduces, meaning that either the set of possible processes or the set of possible provenance data models reduces, as illustrated in Table 1.

We now briefly introduce the different types in more detail, starting with the most general one.

SIGMOD'16, June 26-July 01, 2016, San Francisco, CA, USA
© 2016 ACM. ISBN 978-1-4503-3531-7/16/06. . . $15.00
DOI: http://dx.doi.org/10.1145/2882903.2912568

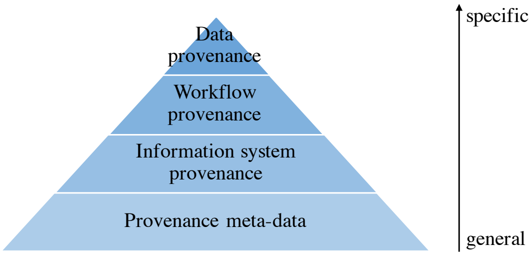

Figure 1: Provenance type hierarchy.

Prov. type	Process type	Prov. data model
data provenance	structured query language	based on specific data model and query language of corresponding process type
workflow prov.	workflows	standards-based
information system provenance	processes supported by information systems	standards-based
prov. meta-data	anything	anything

Table 1: Restrictions on process type and provenance data model on provenance (prov.) types.

Provenance meta-data. Provenance meta-data can be any type of data describing the source, the process of acquisition, the context or environment of acquisition etc. of some data. This most general type of provenance provides users with the widest degree of freedom to model, store, or access the provenance of any type of end product or production process. We distinguish provenance meta-data from other meta-data based on its intended purpose. Indeed, whereas general meta-data aims at assigning meaning to data, provenance information is descriptive of the data derivation process [13]. Note that these meta-data types may overlap. For instance, the resolution of a digital photograph can be seen as meta-data translating the quality of the photograph. At the same time, it is a key parameter in the process of taking and post-processing the photograph. Without further restrictions, provenance meta-data can be modeled, stored, or queried in any conceivable way, justifying the generality of this first type of provenance. This type of provenance, for instance, applies on proprietary solutions for provenance management (e.g., Informatica Big Data Management, IBM InfoSphere Information Governance Catalog, SAS Lineage).

Information system provenance. The problems of proprietary solutions for provenance management, in particular the interoperability between different information systems, have initiated standardization efforts on both the syntax (e.g., XML [5] or RDF [29]) and semantics of provenance meta-data (e.g., [11, 21]). One significant result of these efforts is a W3C standard, W3C PROV [20], that includes a clearly defined data model, ontology, and mechanisms for accessing and querying provenance. In summary, information system provenance covers provenance for various different types of production processes supported by information systems that complies with some standard representation of provenance. As we further move up the provenance hierarchy, the processes for which provenance are collected become more specific.

Workflow provenance. Due to requirements such as documentation, repeatability, accountability, or parameter evaluation, provenance has become an integral part of workflow management tools (as surveyed for instance in [12, 14]). To represent provenance, these typically rely on a standard representation of provenance for better exchange and interoperability, which are already at the heart of information systems provenance. What distinguishes workflow provenance from information system provenance is the restriction to a very specific family of processes, i.e., workflows. This specialization entails a specialization of the semantics of provenance meta-data. In particular, when considering workflow provenance, we distinguish between retrospective and prospective workflow provenance or different provenance granularities.

Data provenance. Data provenance (surveys include [10]) allows to track the processing of individual data items (e.g., tuples) at the “highest resolution”, i.e., the provenance itself is at the level of individual data items (and the operations they undergo). Collecting data provenance typically applies on structured data models and declarative query languages or dataflow languages with clearly defined semantics of individual operators or functions (SQL [16], Spark [19]). This is necessary to either recover data provenance of individual data items after running a data transformation, or to extend data items or operators to pass on their provenance annotation as the data is processed. Clearly, the data model of the generated provenance meta-data highly depends on the data model of the processed data and the semantics of operators of the data manipulation language.

As we will discuss in more detail in Section 2, this tutorial’s first learning objective is to give the audience an introduction to the four introduced types of provenance. Thus, this tutorial provides an in-depth survey on provenance. Additionally, we will discuss details of selected provenance methods to allow for a more informed decision when faced with a problem or application that may benefit from provenance.

1.2 Visualization

The primary goal of visualization is to find a suitable visual representation of data that provides the viewer with the information she requires or is interested in. This requires usually emphasizing or hiding parts or aspects of the data. To be effective, the visualization must be designed and adapted for the specific application, its data, and its context. It is therefore not possible to find a single visualization method that represents all the different types of provenance in an effective and useful way. Thus, this tutorial will introduce common visualization concepts and principles with the goal to provide the audience with basic knowledge to select or develop provenance visualization for their application.

Perceptual aspects. Human perception is an important aspect when developing a visualization because it influences how the visual information is processed and interpreted by the viewer [28]. Three Important topics in this context are (i) the usage and the effect of colors [6], (ii) issues with 3D perception and reasons why information visualization typically uses only 2D representations [26], and (iii) the usefulness of animation [27].

Visual metaphors. There are established and common visual metaphors for different classes or aspects of data. For instance, relational data or graphs can be represented with node-link diagrams, adjacency matrices, or adjacency lists [18]. There are different ways to represent the time dimension of temporal data, e.g., by using a spatial dimension or animation [1]. Glyphs are a powerful visual metaphor that can encode many different data dimensions [4]. Furthermore, the visualization can make use of different layouts for the data dimensions, e.g., Cartesian or radial layouts [8].

Presentation and interaction concepts. There are some basic principles for the presentation of the data and how the user can interact with the visualization. Multiple coordinated views can be used to provide different visual representations for the same data to combine the advantages of different visualization methods [22]. The different views must be synchronized, i.e., if something is changed in one view, the other views must be updated accordingly. Strongly related to this is the concept of "brushing and linking" [7]: selecting data elements in one view also highlights them in the other views. Another concept is "overview first, details on demand" [24]. Intuitively, the idea is to initially provide an overview of the data and then allow the user to interactively explore and analyze parts of it in detail.

By providing an overview on these important visualization aspects, this tutorial will support the attendees in selecting an appropriate visualization method for their type of provenance (Section 1.1) and application. Furthermore, it will help them to assess the suitability of existing provenance visualization methods for their applications.

Provenance visualization. Examples for such visualization methods include our work [17] on visualizing workflow provenance, VisTrails [9], or the work by Shrinivasan and Wijk [25] on supporting analytical reasoning. A recent survey [23] provides further examples.

2. TUTORIAL OUTLINE

The goal of the tutorial is twofold. First, it aims at giving an introduction and general understanding of provenance types and techniques to obtain provenance information. Second, it provides an introduction to visualization in general and its application to provenance. With such a broad survey, this tutorial will equip non-experts with the essential knowledge necessary to begin research in the field of provenance in general, with an emphasis on provenance exploration, analysis, and processing based on visual user interaction. More specifically, after attending the tutorial, we expect the audience to:

- Obtain a general understanding of types of provenance and typical applications leveraging these different types. Thus, when confronted with a problem requiring or benefiting from provenance, adequate solutions may be developed more rapidly, accurately, and systematically.
- Get a deeper understanding of basic methods and algorithms employed to represent and obtain provenance information as well as an overview of the state of the art of more advanced methods.
- Understand the importance of visualization to represent and interact with data and know which visualization concepts and techniques are best amenable to provenance-based applications.

To achieve these learning objectives, we propose the following tutorial outline.

1. Motivation and provenance overview. We begin the tutorial with a motivation for provenance metadata, based on multiple real-life applications from various domains. Possible domains include business analytics, source code auditing, scientific data processing, query debugging and fixing, or visualization. Based on these use-cases, we introduce the provenance type hierarchy depicted in Figure 1. We show examples of systems operating at each of these levels (including those in our introduction) and clearly highlight the differences distinguishing the different provenance types, along the lines briefly covered in Section 1.1.

2. Provenance types. The remainder of the first part of the tutorial delves into details of individual provenance types, primarily focusing on workflow provenance and data provenance. The most general provenance type that essentially represents a "placeholder" for further provenance types as well as information system provenance will be briefly covered through a high-level discussion and illustration. For the remaining provenance type, we begin with the summary of an in-depth survey of the state of the art, based on novel classifications. For instance, data provenance research may be classified as illustrated in Figure 2, significantly extending the classification previously proposed in [10]. We then discuss selected solutions representative of each provenance type in a bit more detail. Whereas workflow provenance will be detailed as proposed in [9, 2], our discussion of data provenance will introduce provenance semi-rings (how-provenance) [15] and their missing-data provenance counterpart [3].

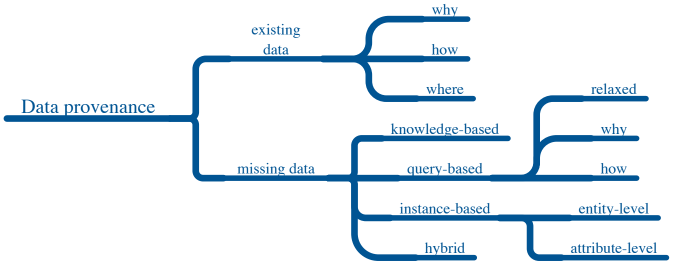

Figure 2: Data provenance survey classification.

3. Visualization basics The second part of the tutorial is related to visualization. We first discuss basic information visualization concepts applicable to provenance data as introduced in Section 1.2. These include perceptual aspects related to color, 2D and 3D representations, and animation. We discuss different visual metaphors, like glyphs or graph representations, that can be and have been used to represent provenance data. We exemplify possible mappings of the temporal dimension and different visualization layout strategies. Furthermore, we introduce common presentation and interaction concepts like multiple coordinated views. Some of these concepts are then demonstrated in the following step-by-step example of provenance visualization.

4. Visualization for provenance. On the example of our work [17], we will demonstrate how to design and develop a suitable visualization for workflow provenance (Figure 3). We will discuss different design choices and what visualization methods are suitable for the considered application scenario.

5. Open research issues. After surveying the state of the art, we will discuss selected open research issues at the intersection of provenance and visualization. Some of these are ongoing research within a national collaborative research center[1]. In particular, we will provide details and food for thought on (i) provenance visualization quality quantification and its potential uses, (ii) visualization-guided pay-as-you-go provenance computation and collection, and (iii) semi-automatic provenance visualization optimization.

3. EXPECTED AUDIENCE

The target audience for this ninety minute tutorial consists of both students and researchers with a general interest in the research fields of provenance and visualization, both being listed as SIGMOD 2016 topics of interest. Additionally, provenance is a relevant field for further areas of interest to SIGMOD, including data warehousing, database monitoring and tuning, database usability, information extraction, information retrieval, knowledge discovery, data integration and cleaning, scientific databases, and uncertain databases.

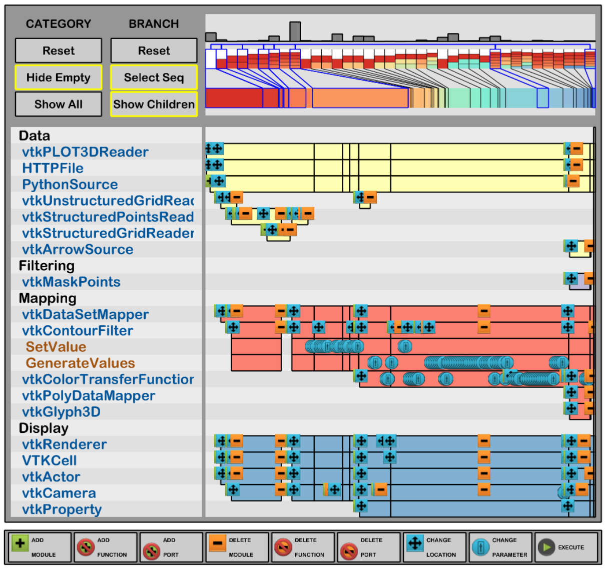

Figure 3: Provenance visualization example. The visualization shows the changes in a visualization workflow made by the user when using VisTrails [9].

As prior knowledge, we expect basic knowledge on databases (which any SIGMOD attendee typically has). Building on this knowledge, we will provide a brief self-sufficient motivation and introduction to essential preliminaries of provenance to enable non-specialists to follow the subsequent detailed discussion. Given the scope of SIGMOD, we do not expect any prior knowledge on visualization and will devote more time to the introduction of fundamental concepts.

4. BIOGRAPHIES

Melanie Herschel is a professor at the University of Stuttgart. Her primary research interests include provenance, data integration, entity resolution, and data engineering. Currently, she is a primary investigator within the Collaborative Research Center "Quantitative Methods for Visual Computing" (SFB-TRR 161), where she explores synergies between visualization and provenance. She is also the leading scientist of the Nautilus project, leveraging provenance for query debugging and semi-automatic fixing. Recently, she has been the area co-chair at ICDE 2015 and the publicity chair of SIGMOD 2016.

Marcel Hlawatsch has a PhD in computer science and works at the Visualization Research Center of the University of Stuttgart (VISUS), Germany, as a research associate. His research topics cover, amongst others, the visualization of provenance data, (dynamic) graphs, and eye tracking data. He is also the manager of the Collaborative Research Center "Quantitative Methods for

[1] Collaborative Research Center "Quantitative Methods for Visual Computing" (SFB-TRR 161): http://www.sfbtrr161.de

Visual Computing” (SFB-TRR 161) at the University of Stuttgart funded by the German research foundation DFG.

Acknowledgements. The authors thank the German Research Foundation (DFG) for financial support within projects B01 and D03 of SFB/Transregio 161.

5. REFERENCES

[1] W. Aigner, S. Miksch, H. Schumann, and C. Tominski. *Visualization of Time-Oriented Data*. Human–Computer Interaction Series. Springer London, 2011.

[2] Y. Amsterdamer, S. B. Davidson, D. Deutch, T. Milo, J. Stoyanovich, and V. Tannen. Putting lipstick on pig: Enabling database-style workflow provenance. *Proceedings of the VLDB Endowment (PVLDB)*, 5(4), 2011.

[3] N. Bidoit, M. Herschel, and A. Tzompanaki. Efficient computation of polynomial explanations of why-not questions. In *International Conference on Information and Knowledge Management, (CIKM)*, 2015.

[4] R. Borgo, J. Kehrer, D. H. Chung, E. Maguire, R. S. Laramee, H. Hauser, M. Ward, and M. Chen. Glyph-based visualization: Foundations, design guidelines, techniques and applications. *Eurographics State of the Art Reports*, 2013.

[5] R. Bose and J. Frew. Composing lineage metadata with XML for custom satellite-derived data products. In *International Conference on Scientific and Statistical Database Management (SSDBM)*, 2004.

[6] C. A. Brewer. Color use guidelines for data representation. *Section on Statistical Graphics, American Statistical Association*, 1999.

[7] A. Buja, J. A. McDonald, J. Michalak, and W. Stuetzle. Interactive data visualization using focusing and linking. In *International Conference on Visualization (VIS)*, 1991.

[8] M. Burch and D. Weiskopf. On the benefits and drawbacks of radial diagrams. In W. Huang, editor, *Handbook of Human Centric Visualization*. Springer New York, 2014.

[9] S. P. Callahan, J. Freire, E. Santos, C. E. Scheidegger, C. T. Silva, and H. T. Vo. Vistrails: Visualization meets data management. In *Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data*, 2006.

[10] J. Cheney, L. Chiticariu, and W. C. Tan. Provenance in databases: Why, how, and where. *Foundations and Trends in Databases*, 1(4), 2009.

[11] P. P. da Silva, D. L. McGuinness, and R. McCool. Knowledge provenance infrastructure. *IEEE Data Engineering Bulletin*, 26(4), 2003.

[12] S. B. Davidson and J. Freire. Provenance and scientific workflows: challenges and opportunities. In *International Conference on Management of Data, (SIGMOD)*, 2008.

[13] E. Deelman, G. B. Berriman, A. L. Chervenak, Ó. Corcho, P. T. Groth, and L. Moreau. Metadata and provenance management. In A. Shoshani and D. Rotem, editors, *Scientific Data Management: Challenges, Technology, and Deployment*. Chapman & Hall/CRC, 1st edition, 2009.

[14] J. Freire, D. Koop, E. Santos, and C. T. Silva. Provenance for computational tasks: A survey. *Computing in Science and Engineering*, 10(3), 2008.

[15] T. J. Green, G. Karvounarakis, and V. Tannen. Provenance semirings. In *Principles of Database Systems (PODS)*, 2007.

[16] M. Herschel. A hybrid approach to answering why-not questions on relational query results. *ACM Journal on Data and Information Quality (JDIQ)*, 5(3), 2015.

[17] M. Hlawatsch, M. Burch, F. Beck, J. Freire, C. Silva, and D. Weiskopf. Visualizing the evolution of module workflows. In *International Conference on Information Visualisation (IV)*, 2015.

[18] M. Hlawatsch, M. Burch, and D. Weiskopf. Visual adjacency lists for dynamic graphs. *IEEE Trans. on Visualization and Computer Graphics*, 20(11), 2014.

[19] M. Interlandi, K. Shah, S. D. Tetali, M. Gulzar, S. Yoo, M. Kim, T. D. Millstein, and T. Condie. Titian: Data provenance support in spark. *Proceedings of the VLDB Endowment (PVLDB)*, 9(3), 2015.

[20] P. Missier, K. Belhajjame, and J. Cheney. The W3C PROV family of specifications for modelling provenance metadata. In *Conference on Extending Database Technology (EDBT)*, 2013.

[21] L. Moreau, B. Clifford, J. Freire, J. Futrelle, Y. Gil, P. Groth, N. Kwasnikowska, S. Miles, P. Missier, J. Myers, B. Plale, Y. Simmhan, E. Stephan, and J. V. den Bussche. The open provenance model core specification (v1.1). *Future Generation Computer Systems*, 27(6), 2011.

[22] C. North and B. Shneiderman. Snap-together visualization: can users construct and operate coordinated visualizations? *International Journal of Human-Computer Studies*, 53(5), 2000.

[23] E. Ragan, A. Endert, J. Sanyal, and J. Chen. Characterizing provenance in visualization and data analysis: An organizational framework of provenance types and purposes. *IEEE Trans. on Visualization and Computer Graphics*, 22(1), 2016.

[24] B. Shneiderman. The eyes have it: a task by data type taxonomy for information visualizations. In *IEEE Symposium on Visual Languages 1996*, 1996.

[25] Y. B. Shrinivasan and J. J. van Wijk. Supporting the analytical reasoning process in information visualization. In *SIGCHI Conference on Human Factors in Computing Systems*, 2008.

[26] M. Tory, A. E. Kirkpatrick, M. S. Atkins, and T. Moller. Visualization task performance with 2d, 3d, and combination displays. *IEEE Trans. on Visualization and Computer Graphics*, 12(1), 2006.

[27] B. Tversky, J. B. Morrison, and M. Betrancourt. Animation: can it facilitate? *International Journal of Human-Computer Studies*, 57(4), 2002.

[28] C. Ware. *Information Visualization: Perception for Design*. Information Visualization: Perception for Design. Morgan Kaufmann, 2013.

[29] J. Zhao, C. Wroe, C. A. Goble, R. Stevens, D. Quan, and R. M. Greenwood. Using semantic web technologies for representing e-science provenance. In *International Semantic Web Conference (ISWC)*, 2004.