Report on the Principles of Provenance Workshop

James Cheney
University of Edinburgh
jcheney@inf.ed.ac.uk

Peter Buneman
University of Edinburgh
opb@inf.ed.ac.uk

Bertram Ludäscher
University of California, Davis
ludaesch@ucdavis.edu

Abstract

Provenance, or records of the origin, context, custody, derivation or other historical information about a (digital) object, has recently become an important research topic in a number of areas, particularly databases. However, there has been little interaction between researchers across subdisciplines of computer science working on related problems. This article reports on a workshop on Principles of Provenance held in Edinburgh, Scotland in November 2007, which facilitated interaction among researchers working on provenance in databases, security, information retrieval, Semantic Web, and software engineering settings, as well as developers and database administrators who are currently working with provenance in practice, or foresee the need to do so in the near future.

1. Introduction

Provenance is, informally speaking, information describing the origin, derivation, history, custody, or context of an object — either a physical object such as the Mona Lisa, or a digital object such as a biological database. Provenance is important for digital artifacts because it is useful for understanding the authenticity, integrity and trustworthiness of online information. Accordingly, it has attracted a great deal of research interest recently in many different areas of computer science. For example,

- In databases, provenance has been studied in the context of data annotation [2] and in data warehouses as a means of helping trace information in a view to relevant "source" data in the underlying databases [6].
- In scientific workflow systems, provenance is maintained in order to ensure repeatability and avoid expensive recomputation [3, 12].
- In bioinformatics and other scientific databases, provenance information recording the change history of a database is considered essential for determining its scientific value [4].
- In security, provenance is now considered a challenging part of the problem of providing integrity for data in networked systems [1].
- In Semantic Web systems, provenance is being studied as a form of "proofs" or "explanations" that need to be provided to users to help them understand the meaning of results of inference-based search [7].

In addition, related ideas and techniques also seem to play a role in other areas such as programming languages (source locations in debugging and error messages) and software engineering (version control, configuration management). However, although there are several communities actively working on provenance in different settings, and other communities with established techniques for studying similar problems, there is little interaction between these communities; moreover, there is a great deal of variation of definitions, goals, and techniques even within communities. Yet there has been, to our knowledge, no single forum at which researchers involved in provenance in all of these areas meet and exchange ideas.

Interest in data provenance continues to grow both in the database and in the (scientific) workflow communities. Recent workshops on provenance such as the International Provenance and Annotation Workshop (IPAW) [11] and Provenance Challenge[1]; however, these events have primarily attracted participation from systems researchers involved in developing provenance tracking systems for workflows, which we believe is only one aspect of provenance (albeit an important one).

In June, two of the authors (Buneman and Cheney) along with Nate Foster and Benjamin Pierce (University of Pennsylvania) organized an informal one-day workshop at the University of Pennsylvania on "Principles of Provenance". Several of the speakers from the workshop were invited to contribute to an issue of the IEEE Data Engineering Bulletin on data provenance. In particular, Wang-Chiew Tan's article in that issue provides a comprehensive overview of provenance in database research [14].

We organized a subsequent public workshop on Principles of Provenance that took place on November 19–20, 2007 in Edinburgh, Scotland in the International Centre for Mathematical Sciences, James Clerk Maxwell House, with public calls for abstracts and participation. Twelve abstracts were submitted, all of which were accepted for presentation,

[1] http://twiki.ipaw.info/bin/view/Challenge

and three additional talks were solicited from invited workshop participants. The presentations included discussions of both recently published work and work in progress.

2. Contributions

The workshop consisted of six sessions over one and one-half days. Each session consisted of talks followed by an open discussion involving the speakers and participants.

2.1 Session 1: Provenance in Practice

The first session comprised two talks. Frank Kauff (Universität Kaiserslauten) presented an overview of the WASABI (Web Accessible Sequence Analysis for Biological Inference) system. WASABI is a biological data management system being developed as part of the AFTOL (Assembling the Fungal Tree of Life) project, in joint work with Cymon Cox (Natural History Museum, London, UK), and Francois Lutzoni (Duke University). At present, provenance is not integrated into this system at an essential level, but this is an important requirement for future system development.

Curt Tilmes of the NASA Goddard Space Flight Center discussed provenance tracking in climate science data processing systems. Such systems deal with large volumes of data obtained from satellites and then subjected to a large number of processing steps in order to produce data that are useful for scientists and other interested parties. Provenance tracking for such data is challenging because of its volume and because the preferred algorithms used for processing the data also tend to change over time, both as a direct result of improvements to the software and as a result of changes to the hardware, operating system, and library environment in which the analyses run.

Both talks provided an excellent start to the workshop by focusing attention on the importance and difficulty of provenance tracking in practice — including not only the system-development challenges of tracking the information efficiently but also the theoretical challenge of determining what exactly should be tracked in order to accomplish a particular aim. There is also a significant *organizational* challenge because many organizations do not place a high priority on retaining provenance information since it is expensive but may not provide short-term benefits. Therefore it is crucial that provenance techniques be inexpensive and provide a clear benefit or they will not be used.

2.2 Session 2: Security

Uri Braun, representing the Provenance-Aware Storage Systems (PASS) Team at Harvard University, presented "Why provenance needs its own security model". The talk presented several examples of security problems involving provenance, such as an employee's performance review, where an employee should have access to some data but *not* its provenance, or a National Intelligence Estimate, where the data's provenance is (partly) public knowledge but the data itself should remain secret. The talk then discussed potential shortcomings of existing security models and argued that a new, provenance-aware security model is needed to deal with such problems. In particular, provenance has significant implications on privacy, anonymity, and other areas of security that are of current interest.

Brian Corcoran of the University of Maryland gave a talk entitled "Combining Provenance and Security Policies in a Web-based Document Management System", covering joint work with Nikhil Swamy and Michael Hicks. The authors have developed a Wiki-like system that provides secure *mandatory access control*, ensuring that secret data cannot be leaked to unauthorized users, and which also tracks provenance describing how the Wiki pages have been modified over time. Security policies can take provenance into account, and the policies are expressed in a high-level programming language that provides provable guarantees.

The final talk in this session was by Corin Pitcher (DePaul University), on "Programming Trustworthy Provenance", joint work with Andy Cirillo, Radha Jagadeesan, and James Riely. This work addressed the problem of developing secure and trustworthy decentralized systems, in which provenance is often an important part of security policies. For example, one agent may be trusted to check and re-certify data received from certain other agents. The talk then discussed techniques for certifying that Java-like programs satisfy the security policy, via program analysis techniques based on a form of authorization logic.

2.3 Session 3: Information Retrieval and the (Semantic) Web

The third session consisted of talks about provenance in information retrieval and on the (Semantic) Web. The first talk, on "Provenance in Semantic Web Applications", was given by Sergej Sizov (University of Koblenz-Landau), describing joint work with Bernhard Schueler and Steffen Staab. The authors argue that provenance is an important part of the "proof layer" in the Semantic Web, since provenance is part of the explanation that a system should provide. They introduce a model for provenance for SPARQL queries over RDF data, in which each RDF triple carries an annotation and provenance is computed by combining the annotations.

Andreas Harth presented "Towards a Social Notion of Provenance", describing joint with with Axel Pollres and Stefan Decker (National University of Ireland, Galway). The talk focused on making use of provenance information already (implicitly) present in Semantic Web data sources, including URLs, HTTP metadata, domain name registries, and so on.

Finally, Erin Fitzhenry (Oregon State University) gave a talk entitled "The Use of Provenance in Information Retrieval". This work, joint with Simone Stumpf and Thomas Dietrich, addresses the problem of "desktop search", or information retrieval in personal computer operating systems such as Windows. There is some evidence that people re-

member relationships among documents better than specific details such as the title, creation date or document keywords. Thus, provenance information recording the relationships among documents may be useful for improving desktop search. TaskTracer, currently under development by the authors, is a system that records high-level events such as opening or closing a document, copying and pasting data, sending or receiving email and attachments, etc. This information is stored in a repository and can be browsed or searched. Work on evaluating the system's usefulness for real users is under way, but evaluation is a challenging problem.

2.4 Session 4: Software Engineering and Dependability

The fourth session included two speakers from the University of Edinburgh, Perdita Stevens and Conrad Hughes. Both are involved in research on software engineering and dependability.

Perdita Stevens' talk on "Model transformations, traceability and provenance" discussed some aspects of software engineering research, such as Model-Driven Design/Architecture (MDD/MDA) and traceability, which seem analogous to provenance in other settings. Traceability has long been considered an important part of verification and validation in software engineering; however, automating the capture and maintenance of traceability remains challenging, similar to provenance. Moreover, a key aspect of MDD/MDA is understanding how changes made to one "model" of a software system affects, or propagates, to other models. Stevens discussed recent work in this area including her paper [13] which was recognized as Best Paper of MODELS 2007.

Conrad Hughes presented his work on "Synchronising Diversely Implemented Databases to Support Administration of Clinical Research" in collaboration with Stuart Anderson and Mark Hartswood of the University of Edinburgh. Hughes has been working with the National Health Service in the UK to develop a system by which research grant administrators and researchers can share data such as grant proposals. The system is built on top of the Harmony data-synchronization system [8]. In this system, provenance was not considered a crucial need and is not automatically maintained; instead the users of the system are expected to communicate with each other to resolve conflicts. In fact, it was found that keeping explicit track of changes could be counterproductive, because of the possibility of instigating a "blame" culture.

2.5 Session 5: The Open Provenance Model

The fifth session consisted of one talk, given by Luc Moreau of the University of Southampton, on "The Open Provenance Model", which is being developed by Moreau together with Juliana Freire (University of Utah), Jim Myers, Joe Futrelle (NCSA), and Patrick Paulson (PNNL). The Open Provenance Model[2] is an outgrowth of the two Provenance Challenges which were organized after the first IPAW workshop in 2006. The first Provenance Challenge was proposed as a benchmark for comparing different approaches to recording, storing, and querying provenance for workflows. This revealed that different solutions made several different reasonable-seeming design choices. The second Challenge also required participants to consider how to make their provenance records interoperable with those of other participants. Subsequently, Moreau, Freire, Myers, Futrelle, and Paulson have developed a data model that distills the lessons learned from the two Provenance Challenges.

This talk sparked a lively discussion, including questions such as: What validity/integrity constraints should OPM-style provenance satisfy? What principles (besides syntactic well-formedness) should guide developers in applying OPM to record provenance in their systems? What does the structure of an OPM record tell us about the "real" behavior or semantics of the process it is supposed to capture?

2.6 Session 6: Databases

The final session consisted of talks on provenance in databases and data warehouses. Stijn Vansummeren (University of Hasselt/Transnational University of Limburg) presented joint work with Buneman and Cheney on the *expressiveness* of techniques for recording provenance [5]. In particular, the talk introduced a model of provenance for queries and updates in the nested relational data model, a generalization of SQL. In this setting, it is important to determine what provenance behaviors can be expressed by a provenance-propagating query or update, just as it is important to understand the expressiveness of ordinary query languages. Interestingly, this approach shows that updates are more expressive than queries when provenance is taken into account, because in-place updates cannot be simulated by queries.

Panos Vassiliadis (University of Ioannina) presented joint work on "Data Provenance in ETL Scenarios", joint work with Timos Sellis, Dimitris Skoutas (National Technical University of Athens), Alkis Simitsis (IBM Almaden). The talk initially presented an overview of the authors' work on designing high-level workflow languages for programming ETL (Extract-Transform-Load) processes. In this setting, it is a considerable challenge to show where records in the result "come from" in the input or ETL process. Moreover, although updates to the source data are currently supported efficiently, updates to the results, schemas, or workflows are not. Thus, there may be challenging open problems to be addressed involving provenance and ETL workflows.

Natalia Kwasnikowska (Hasselt University and Transnational University of Limburg) presented "A formal model for dataflows, runs of dataflows, and provenance within runs",

[2] See `http://twiki.ipaw.info/bin/view/Challenge/OPM` and `http://eprints.ecs.soton.ac.uk/14979/1/opm.pdf` for more about the Open Provenance Model

joint work with Jan van den Bussche that extends previous work on modeling dataflow repositories [10]. This work addresses the problem of recording "runs" of scientific computations, specified using the nested relational calculus. The goal of this work is to provide a clear, formal model explaining what information is stored to record a run and to provide the ability to query this detailed provenance information.

Val Tannen concluded the workshop by giving a short talk on recent work on *provenance semirings*, an approach he has been developing with Green and Karvounarakis [9]. In this work, it was shown that semiring-valued relations generalize a number of existing variations on the relational model, including probabilistic databases, incomplete databases. Moreover, semiring expressions are closely related to some existing models of provenance such as lineage and why-provenance.

3. Conclusions

Provenance is a growing research topic in several areas, yet it is currently not very well-understood and there is not much understanding of provenance across subdisciplines. The Principles of Provenance workshops have, we believe, helped bring researchers already working on different aspects of provenance into contact with one another, helped encourage foundational research on provenance, and helped to interest and inform newcomers to the area.

To follow up this workshop, we have applied for and been awarded funding by the UK eScience Institute for a *theme*, or year-long program of workshops, lectures, and research visitors. The Principles of Provenance Theme[3] will run from April 2008 through March 2009 and we currently plan to structure it around four or five week-long workshops/visitor programs based on relevant areas of computer science, such as scientific workflows, databases, programming languages and software engineering, and security.

Acknowledgments The Principles of Provenance workshops have been partly supported by funding from the United Kingdom Engineering and Physical Sciences Research Council.

References

[1] INFOSEC hard problem list. Technical report, INFOSEC Research Council, 2005. `http://www.infosec-research.org/-docs_public/20051130-IRC-HPL-FINAL.pdf`.

[2] Deepavali Bhagwat, Laura Chiticariu, Wang-Chiew Tan, and Gaurav Vijayvargiya. An annotation management system for relational databases. *VLDB Journal*, 14(4):373–396, 2005.

[3] Rajendra Bose and James Frew. Lineage retrieval for scientific data processing: a survey. *ACM Comput. Surv.*, 37(1):1–28, 2005.

[4] Peter Buneman, Adriane Chapman, and James Cheney. Provenance management in curated databases. In *SIGMOD 2006*, pages 539–550, 2006.

[5] Peter Buneman, James Cheney, and Stijn Vansummeren. On the expressiveness of implicit provenance in query and update languages. In Thomas Schwentick and Dan Suciu, editors, *ICDT*, volume 4353 of *Lecture Notes in Computer Science*, pages 209–223. Springer, 2007.

[6] Yingwei Cui, Jennifer Widom, and Janet L. Wiener. Tracing the lineage of view data in a warehousing environment. *ACM Trans. Database Syst.*, 25(2):179–227, 2000.

[7] Paulo Pinheiro da Silva, Deborah L. McGuinness, and Rob McCool. Knowledge provenance infrastructure. *IEEE Data Eng. Bull.*, 26(4):26–32, 2003.

[8] J. Nathan Foster, Michael B. Greenwald, Jonathan T. Moore, Benjamin C. Pierce, and Alan Schmitt. Combinators for bidirectional tree transformations: A linguistic approach to the view-update problem. *ACM Trans. Program. Lang. Syst.*, 29(3):17, 2007.

[9] Todd J. Green, Grigoris Karvounarakis, and Val Tannen. Provenance semirings. In *PODS '07: Proceedings of the twenty-sixth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 31–40, New York, NY, USA, 2007. ACM.

[10] Jan Hidders, Natalia Kwasnikowska, Jacek Sroka, Jerzy Tyszkiewicz, and Jan Van den Bussche. A formal model of dataflow repositories. In Sarah Cohen Boulakia and Val Tannen, editors, *DILS*, volume 4544 of *Lecture Notes in Computer Science*, pages 105–121. Springer, 2007.

[11] Luc Moreau and Ian T. Foster, editors. *Proc. International Provenance and Annotation Workshop*, volume 4145 of *LNCS*. Springer, 2006.

[12] Yogesh Simmhan, Beth Plale, and Dennis Gannon. A survey of data provenance in e-science. *SIGMOD Record*, 34(3):31–36, 2005.

[13] Perdita Stevens. Bidirectional model transformations in qvt: Semantic issues and open questions. In Gregor Engels, Bill Opdyke, Douglas C. Schmidt, and Frank Weil, editors, *MoDELS*, volume 4735 of *Lecture Notes in Computer Science*, pages 1–15. Springer, 2007.

[14] Wang-Chiew Tan. Provenance in databases: Past, current, and future. *IEEE Data Eng. Bull.*, 30(4):3–12, 2007.

[3] `http://wiki.esi.ac.uk/Principles_of_Provenance`