Principles of Provenance
Theme Proposal

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

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

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

Theme topic and brief description Recent research in a variety of settings (databases and data warehouses [10, 7, 20, 8], file systems [17], geographic information systems [6], scientific workflows and grid computation [19, 13], archiving and digital curation [1], and the Semantic Web [14]) has considered the problem of keeping track of metadata about creation and modification history, influences, ownership, and other *provenance* or *lineage* information. Such metadata is essential for making informed judgements about data quality, integrity, and authenticity. In addition, ideas about provenance are now being used in several other areas of computer science such as probabilistic databases [12], adaptive computation [4], security [15, 3] file synchronisation [11], and annotation propagation [5]. Other topics, such as version control and archiving, may also benefit from better understanding of provenance. We believe the time is ripe to develop the foundations of the topic and address questions such as:

- What is and what isn't provenance?
- What underlying problems do real-world uses of provenance address, and how can we formalise correctness for proposed solutions to such problems in computer systems?
- How can we compare models of or approaches to provenance? Is there a "best" or ideal approach to provenance?
- Why is provenance tracking/management considered a "hard problem", despite its seeming obviousness ("just record everything about the history of the data")?
- Where should research efforts be focused in order to best make progress?

There are currently a number of research groups developing systems for managing provenance (see for example surveys [6, 19]), as well as an established biennial forum, the International Provenance and Annotation Workshop [16]. This community includes both applications scientists developing systems for a particular area and computer science researchers developing systems that are meant to be used by applications scientists and other people with data management needs involving provenance.

However, in contrast to problems such as high-performance computing and security that are widely appreciated to be challenging, have a received a great deal of attention, and have a well-developed theory, there is not yet much awareness of or interest in provenance in the broader computer science systems or theory communities. We believe that this is a symptom of a lack of foundational understanding of the problem, despite being listed as a "hard problem" in several recent surveys, such as the British Computing Society's Grand Challenges in Computing Science [2] and (in the US) the InfoSec Council's Hard Problems List [3]. At the same time, we find, both in our own work and in discussion with other researchers, that there is not enough familiarity among researchers with users' motivations and goals for provenance tracking and with the uses to which the information is (or is expected to) be put. Thus, we believe that it is important to bring together scientists and computer science researchers who use or develop of provenance systems with researchers with expertise in other areas we believe are relevant to developing the foundations of the subject in directions that will lead to progress on the actual problems arising in practise.

We believe that applications scientists are already strongly interested in this topic and would see the immediate benefits of participating, but to engage computer science researchers, particularly theoreticians, we need to make a stronger case that there is compelling research to be done in this area that is relevant to mainstream subdisciplines of computer science, not just systems implementation research that has been the focus of most of the work in IPAW. This proposal therefore places greater emphasis theme activities based on mainstream computer science research areas such as concurrency, databases, security, and programming languages. This organisational emphasis on computer science topics rather than on along e-Science disciplinary boundaries (bioinformatics, physics, astronomy, etc.) does not mean that the theme activities will exclude or marginalise applications scientists; this would be self-defeating. Rather, the

goal of each event is to foster interaction and understanding between applications scientists (possibly from several different disciplines) and computer scientists that we currently believe is not taking place.

We have already planned and organised a Workshop on Principles of Provenance that will take place in late November 2007 in Edinburgh.

Proposed leaders with brief description of their areas of specialisation.

- **James Cheney**, Research assistant, University of Edinburgh. Research foci include data provenance, XML processing (compression, database query and update languages, access control), programming language semantics, type systems.
- **Peter Buneman**, Professor, University of Edinburgh; research director, UK Digital Curation Centre (DCC). Research focuses on databases, particularly semistructured data, data provenance, and archiving as applied to scientific databases.
- **Bertram Ludaescher**, Associate Professor, University of California, Davis. Research focuses on scientific workflows, particularly modelling and managing provenance for workflows.

Please describe the application areas that would benefit from the outcomes of this theme. Data provenance is important in many scientific disciplines that are now taking advantage of databases, Grid computing, and other advanced computer systems. For example:

- In "curated" databases (especially widespread in bioinformatics) that are built up by a combination of manual and automatic data integration techniques, provenance is considered essential for providing context information that a scientist referring to the data can use to determine how trustworthy or believable it is. In most cases, provenance is managed either by hand or by developing customised application code. Developing automatic techniques for managing provenance will relieve curators of these burdens.
- In Grid or workflow computation, provenance is considered important for several purposes. Provenance helps to identify and re-use results that have already been computed, thus avoiding redundant computations. In addition, since a workflow may be nondeterministic, provenance information is needed to record what actually happened when such a workflow was run, ensuring repeatability. Moreover, in scientific workflows, data provenance can often help the scientist interpret "surprising" results, i.e., help debug faulty runs (wrong parameter settings, data or tool bindings, etc), or confirm new findings. Finally, data and computationally-intensive workflows often require runtime monitoring, e.g. to steer the computation, restart "dead branches", etc. A common understanding of such "runtime provenance" v.s. the more common "post mortem" data lineage is also desirable.
- In digital libraries, preservation, and curation, provenance is considered essential for evaluating the context, meaning, and trustworthiness of a digital object. Standards such as the OAIS Reference Model [1] place prominent emphasis on the importance provenance. However, the OAIS model presently considers provenance for atomic digital objects with no internal structure, such as whole documents, and the content of the provenance records in the OAIS model is left to the archivist's discretion. The model does not yet address how to express provenance for parts of large documents or other important classes of digital objects (such as databases) that evolve over time, nor does it specify principles for how a computer system should maintain provenance.

Please describe the technical areas that would be engaged and developed as a result of this theme. The technical areas that are relevant to this theme are primarily within computer science. This includes systems research on provenance (in databases, workflow systems, operating systems, Semantic Web systems, etc.) but also covers more theoretical or speculative research in other areas which we believe are closely related, including database theory, programming language semantics, security, and concurrency theory.

Although theoretical computer science research may seem not to be particularly relevant to provenance in computer systems, we believe that the key obstacle to making progress on provenance is currently the lack of understanding of the ideas underpinning provenance — that is, what makes provenance information correct, complete, or otherwise satisfactory. We believe that theoretical computer science provides the best tools for solving this problem, yet there is currently little work on provenance in this community, or even recognition of the problems of provenance, in part due to lack of interaction between theorists and computer systems and application scientists.

Digital preservation and curation are also closely related to this theme. We expect that members of the DCC, both in Edinburgh and elsewhere, will be interested in (and hopefully participate in) theme activities, and we will explore the possibility of organising a coordinated workshop or other outreach activities for disseminating the results of the theme.

Are there other similar projects to the proposed theme? What would be their relationship to/involvement in this program? We are aware of the following research groups pursuing related problems.

- University of Southampton, PASOA/Provenance projects — Moreau and several other European sites developing provenance infrastructure for service-oriented architecture
- University of Chicago/Computation Institute — Foster and Wilde lead group developing Grid/scientific computing systems including support for provenance
- Harvard University — PASS (Provenance Aware Storage Systems) project led by Margo Seltzer
- University of Pennsylvania — faculty including Davidson, Tannen, Ives, Pierce are working actively on provenance, data synchronisation, and bioinformatics
- University of Washington — Suciu leads research on the foundations of probabilistic databases, an area with close connections to provenance
- Stanford — Widom leads research group developing Trio, a database system providing support for lineage, uncertainty, and incomplete information
- University of California, Santa Cruz — Tan leads research group focusing on provenance, annotation propagation, and applications such as debugging schema mappings
- University of California, Davis — Ludaescher leads research group focusing on scientific workflow processing, with provenance being one area of emphasis

We already have ongoing interactions many of these groups, many of which include collaborators of theme co-leaders. However, each of these projects is focused on provenance within a specific domain, such as databases or workflows, and few of them involve participation from other areas which we believe are crucial for developing the foundations of provenance, such as programming languages and security. We believe all of these areas can benefit from increased interaction with applications scientists facilitated by theme activities.

Identify a focus that will ensure the effort is most likely to be productive, i.e. a specific test application domain/current unsolved research challenge. Provenance has been studied extensively in databases; however, only recently have the theoretical foundations of database provenance begun to emerge. Initial approaches have focused on ad-hoc definitions that proved hard to generalise beyond simple query languages. Moreover, almost all of this research has considered only tracking provenance through database queries, and has ignored the important issue of keeping track of the provenance of a database as it is updated — an essential need in curated databases [7].

We have found (as a result of recent and ongoing work) that appropriate foundations for provenance in databases can be found using the theory of expressiveness for database query and update languages [8] and techniques from programming languages such as dependency analysis and program slicing [9]. However, much remains to be done, both in increasing our understanding beyond simple cases and in putting theory into practise. We hope to be able to attract visitors including both scientific database curators (there are several potential participants located in Edinburgh and elsewhere in the UK), as well as computer scientists such as our colleagues on the papers [7, 8, 9], for visits overlapping for a week or more, who will give talks and further develop foundational understanding of provenance for data that is stored in or moves among databases.

Please list any people who have agreed to actively collaborate. The following people have agreed to speak at or attend the Workshop on Principles of Provenance we are organising in November.

- Umut Acar, Amal Ahmed (Toyota Technological Institute, Chicago)
- Khalid Belhajjame (University of Manchester)
- Brian Corcoran (University of Maryland)
- Erin Fitzhenry, Simone Stumpf (Oregon State University)
- Andreas Harth (National University of Ireland, Galway)
- Frank Kauff (University of Kaiserslauten)
- Natalia Kwasnikowska (University of Hasselt and Transnational University of Limburg)
- Luc Moreau (University of Southampton)
- Corin Pitcher (DePaul University)

- Karen Renaud (University of Glasgow)
- Timos Sellis (National Technical University of Athens)
- Sergey Sizov (University of Koblenz-Landau)
- Val Tannen (Pennsylvania)
- Curt Tilmes (NASA Goddard Space Flight Center)
- Dirk van Gucht (Indiana University)
- Jan Van den Bussche (University of Hasselt and Transnational University of Limburg)
- Stijn Vansummeren (University of Hasselt and Transnational University of Limburg)

We expect that many of them and their colleagues would be interested in visiting during the theme.

Sketch of who is probably working in the area, and/or might be interested. We have in mind the following applications scientists whom we might invite for research visits or to give talks:

- Teresa Attwood (Manchester)
- Wendy Bickmore (MRC, Western General Hospital, Edinburgh)
- Bob Mann (Royal Observatory, Edinburgh)
- Ian Foster, Mike Wilde (Chicago)
- Jim Frew (University of California, Santa Barbara)

We have in mind the following computer scientists whom we would like to involve:

- Martin Abadi (University of California, Santa Cruz/Microsoft Research)
- Susan Davidson, Zack Ives, Benjamin Pierce, Steve Zdancewic (Pennsylvania)
- Michael Hicks (Maryland)
- Gerome Miklau (Massachusetts)
- Christopher Re, Dan Suciu (Washington)
- Wang-Chiew Tan (University of California, Santa Cruz)

Identify the current key research challenges in the area. The main research challenge in this area is to define and develop a broad, rigorous understanding of provenance: what it is, what it can be used for (or what users expect it to provide), how we can compare and evaluate proposed techniques, and how we can maintain provenance when data moves between systems or across different data models.

What are the plausible outcomes (deliverables) from the theme? Will they be entirely theoretical, or will there be some experiments and/or software produced? The goal of the theme is to bring together applications scientists and researchers who are working directly on provenance with researchers in other areas that we believe are directly related. We believe that this will result in a significant amount of collaboration that will help establish the foundations of provenance as a broad research area within computer science (and perhaps touching upon other disciplines). We plan to encourage visitors to draft joint position papers which we expect will develop into high-quality research papers suitable for publication in high-profile conferences and journals.

Many of the visitors we envision inviting are actively related in developing systems for the use of application scientists. Thus, we hope that ideas developed in the research fostered by the theme will be transferred directly into these systems, but this is not the immediate goal of the theme.

Sketch of the kind of events (focus/scope) proposed, where they would be held and who would participate.

- **Focused joint symposia** We have in mind several intensive one to two week visitor programs ("symposia"), each involving 3-4 visitors (roughly evenly split among application scientists and computer scientists) as well as local participants from both communities. Each symposium will consist of talks, tutorials and discussions; however, beyond asking each visitor to give a talk at the National e-Science Centre (open to the public), interact with the other participants, and for the participants as a group to produce a short written report, we plan to provide the participants with the flexibility to organise their time as they see fit. Each symposium will focus on one of the following areas:

- **Provenance in workflows and concurrent computation** Workflows and other tools for high-level Grid programming (as well as related Semantic Web/Web Service technologies) provide simple models of very complex computational systems. Provenance is needed in this setting to avoid repeating redundant computations, to ensure "repeatability" of experiments, and to make the underlying process realising a computation "transparent" and "trustworthy". Although it is fairly clear what provenance information suffices to improve efficiency by avoiding repeated computations, it is not clear yet what exactly we mean by "repeatability", "transparency" or "trustworthiness", especially in the complex distributed systems which enact workflows. This symposium would aim to provide rigorous answers for these questions that accurately reflect the needs of application scientists.
- **Provenance in and among databases** So far, most work on provenance in the context of databases has considered tracking provenance of data through views or data warehouses. However, in real e-Science situations, data warehouses are just one of several scenarios where provenance is essential. Provenance is also important for "curated" databases in biology that are updated frequently, and in which data is frequently shared or copied among databases. Although recent research has begun to address this question, so far this work has not been incorporated into systems that e-Scientists regularly use. This symposium will focus on raising the awareness of the real provenance-related problems of database curators among database researchers and setting an agenda for future research that will address their needs.
- **Programming language techniques for understanding provenance** Techniques from the study of programming languages and software engineering provide many interesting parallels with provenance in e-Science and other settings. For example, version control and configuration management systems have long been used in software engineering to manage and keep consistent large programming projects being edited by many users concurrently, but lack the ability to fully explain the history of the data. The concept of *traceability* in software engineering seems closely related to provenance and is currently an active research area. As another example, a great deal of research in program analysis has focused on understanding how one part of a program (or its input) "depends" on another part. These ideas have recently borne fruit in the study of language-based techniques for efficient adaptive computation (including kinetic algorithms and other incremental scientific computation models [4]) and for ensuring that computer systems do not "leak" sensitive information [18], and Cheney, Ahmed and Acar [9] have recently shown how the same ideas can provide a rigorous foundation for provenance in databases. The same ideas may also be applicable to understanding how to assess the integrity guarantees provided by provenance in databases, workflows, or the many other models of computation used in e-Science. This symposium will focus on exploring this connection.
- **Provenance and accountability in secure software systems** Provenance is not just relevant to e-Science and databases but also to recent regulations concerning data transparency and accountability in industry, business, and government, such as the Sarbanes-Oxley Act, enacted in reaction to scandals such as the collapse of Enron Inc. in the United States. These regulations establish obligations concerning retaining the history of the data used by a company to make decisions that organisations must meet or face legal liability; yet currently, computer systems provide little built-in support for doing so, and the organisations must therefore develop custom applications. Such systems must be both "secure" in the classical sense (that is, data or privileges must not leak to unauthorised parties) and "transparent" (that is, it should not be possible for trusted individuals to abuse the system by forging or rewriting history). These concerns are also relevant to proprietary research data such as chemical and pharmaceutical industries (regulated by the FDA Title 21 CFR 11 in the US and REACH in the EU). This symposium aims to bring together computer security researchers and application scientists or industrial partners that are interested in this area.

Although each of the above topics is anchored in a subdiscipline of computer science, we plan to invite both computer science researchers and applications scientists (or industrial partners in the last case) with interests in the areas. Each visit will begin with talks or tutorials by application scientists (or, in one case, industry participants) describing their current needs and practices and identifying currently unmet challenges or unknown issues; that is, to provide answers to questions such as

- How do scientists want to use provenance information?
- How do they do so now?
- What are the limitations of what is already done?

which could help define the context for the rest of the programme.

For example, for the database-oriented symposium, we plan to invite several local scientists involved in database curation, such as Dr. Bob Mann (Royal Observatory of Edinburgh, Dr. Wendy Bickmore (MRC, Edinburgh) to describe their current needs and problems, as well as researchers working on adding provenance support to database systems. Although it might also be sensible to organise an initial agenda-setting workshop focusing exclusively on the provenance practices and needs of application scientists, we believe that this structure would not be as beneficial because it is unlikely that all of the computer science researchers whom we believe would benefit from such a workshop would attend, partly due to scheduling concerns and partly because such a workshop may not be perceived as worthwhile because of its lack of clear relevance to an existing area of study in computer science proper. Our proposed "joint symposium" structure aims to make attendance worthwhile for both application scientists and computer scientists, while making best use of the time available for interaction and exchange of ideas, and taking advantage of the deep pool of applications scientists and computer scientists present in Edinburgh.

- **Workshops** At present, we have already organised a workshop on Principles of Provenance to be held in November, following a more informal meeting organised in June with Benjamin Pierce and Nate Foster at the University of Pennsylvania. This workshop could be used to kick off the theme, by taking advantage of the opportunity to advertise and advise participants of the workshop of the possibility of visiting for an extended period to participate in the theme. We may organise a second workshop to be held in late 2008 to present the results of the work undertaken during the theme, along with other outreach or dissemination activities (possibly jointly with the DCC). We may also organise an external, public workshop with invited and peer-reviewed contributed papers, perhaps collocated with a major relevant conference such as VLDB or the USENIX Technical Conference.

What difference will be generated by running the theme for 6 or 12 months? We believe that running the theme for 6 months would be feasible, but due to the many constraints on the schedules of organisers and potential participants, and the long lead times needed for organising workshops and extended visits, we believe that running for 12 months would be preferable in order to maximise the benefit of the theme.

Is the topic of the theme so specific that it can really all be "tied up" in 6 or 12 months time, or should there be some follow-on? If so how might the follow-on be funded? The topic is very broad. The purpose of the theme is certainly not to solve the problem of provenance, indeed it is not clear what this would mean (though we hope to have a clear understanding of this by the end of the theme). We expect the theme to inspire subsequent research papers and successful grant proposals.

Are there opportunities for co-funding from other sources? 50% of Cheney's salary will continue to be paid by existing funding. Some travel and visitor expenses related to Cheney and Buneman's ongoing related work on provenance may also be covered by this source.

Cheney has applied for EPSRC and Royal Society fellowships; if one of them is awarded, it will come into effect in the last 3–4 months of 2008.

Project Plan The following is an outline of the activities of the theme. Note that the specific assignment of topics to visitor blocks is highly dependent on the (as yet unknown) schedules of likely participants, so is subject to change.

- January – March 2008
  - (January) Opening lecture: "Why does provenance need principles?", James Cheney
  - (February) Visitor block 1:"Foundations of workflow provenance"
    Possible visitors: Freire, Davidson, Plale, application scientists that make use of Grid or workflow technologies
- April – June 2008
  - (May/June) "Provenance in and among databases"
    Possible visitors: Ives, Tan, Vansummeren, database curators
- July – September 2008
  - (July) Visitor block 3: "Provenance in programming languages and software engineering"
    Possible visitors: Ahmed, Acar, application scientists involved in
  - (August/September) External workshop (possibly affiliated with VLDB)
- October – December 2008
  - (October) Visitor block 4: "Provenance and accountability in secure software systems"
    Possible visitors: Zdancewic, Hicks, Seltzer

- (November) Workshop
- Closing lecture presenting summary of results, James Cheney

References

[1] Reference model for an open archival information system (oais). Technical Report CCSDS 650.0-B-1, Consultative Committee on Space Data Systems, 2003. ISO standard 14721:2003, available at `http://nost.gsfc.nasa.gov/-wwwclassic/documents/pdf/CCSDS-650.0-B-1.pdf`.

[2] Grand challenges in computing — research. Technical report, British Computing Society, 2004.

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

[4] Umut A. Acar, Guy E. Blelloch, and Robert Harper. Selective memoization. In *POPL*, pages 14–25, 2003.

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

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

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

[8] Peter Buneman, James Cheney, and Stijn Vansummeren. On the expressiveness of implicit provenance in query and update languages. In *ICDT 2007*, number 4353 in LNCS, pages 209–223. Springer, 2007.

[9] James Cheney, Amal Ahmed, and Umut A. Acar. Provenance as dependency analysis. In M. Arenas and M. I. Schwartzbach, editors, *Proceedings of the 11th International Symposium on Database Programming Languages (DBPL 2007)*, number 4797 in LNCS, pages 139–153. Springer-Verlag, 2007.

[10] 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.

[11] J. Nathan Foster, Michael B. Greenwald, Jonathan T. Moore, Benjamin C. Pierce, and Alan Schmitt. Combinators for bi-directional tree transformations: A linguistic approach to the view update problem. In *POPL*, 2005.

[12] Todd J. Green, Grigoris Karvounarakis, and Val Tannen. Provenance semirings. In *PODS*, 2007. To appear.

[13] Paul Groth, Michael Luck, and Luc Moreau. A protocol for recording provenance in service-oriented grids. In *OPODIS '04*, 2004.

[14] Jingwei Huang and Mark S. Fox. Uncertainty in knowledge provenance. In Christoph Bussler, John Davies, Dieter Fensel, and Rudi Studer, editors, *ESWS*, volume 3053 of *LNCS*, pages 372–387. Springer, 2004.

[15] Peng Li, Yun Mao, and Steve Zdancewic. Information integrity policies. In *Workshop on Formal Aspects in Security & Trust (FAST 2003)*, 2003.

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

[17] Kiran-Kumar Muniswamy-Reddy, David A. Holland, Uri Braun, and Margo Seltzer. Provenance-aware storage systems. In *USENIX*, pages 43–56. USENIX, June 2006.

[18] Andrei Sabelfeld and Andrew Myers. Language-based information-flow security. *IEEE Journal on Selected Areas in Communications*, 21(1):5–19, 2003.

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

[20] Nicholas E. Taylor and Zachary G. Ives. Reconciling while tolerating disagreement in collaborative data sharing. In *SIGMOD*, pages 13–24. ACM Press, 2006.