Data provenance – the foundation of data quality

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

Susan B. Davidson
University of Pennsylvania
Philadelphia, USA
susan@cis.upenn.edu

September 1, 2010

1 What is Provenance?

Provenance for data[1] is often defined by analogy with its use for the history of non-digital artifacts, typically works of art. While this provides a starting point for our understanding of the term, it is not adequate for at least two important reasons. First, an art historian seeing a copy of some artifacts will regard its provenance as very different from that of the original. By contrast, when we make use of a digital artifact, we often use a copy, and and copying in no sense destroys the provenance of that artifact. Second, digital artefacts are seldom "raw" data. They are created by a process of transformation or computation on other data sets; and we regard this process as a part of the provenance. For "non-digital" artefacts, provenance is usually traced back to the point of creation, but no further.

These two issues – the copying and transformation of data – will figure prominently in our discussion of provenance, but rather than attempting a more elaborate definition, we shall start with some examples.

- In September 2008, a news article concerning the 2002 bankruptcy of UAL, United Airlines' parent company, appeared on the list of top news stories from Google News. The (undated) article provoked investor panic and UAL's share price dropped by 75% before it was realized that the article was six years out of date.
- Look up some piece of demographic information on the Web and see if you can get an accurate account of its provenance. A recent Web search for "population Corfu" yielded 14 different values in the first 20 hits ranging from 70,000 to 150,000, several of them given to the nearest person. While some of the values were dated, in only one case was any kind of attribution or citation given.
- Most scientific data sets are not raw observations but have been generated by a complex analysis of the raw data with a mixture of scientific software and human input. Keeping a record of this process – the *workflow* – is all-important to a scientist who wants to use the data and be satisfied of its validity.

  Also, in science and scholarship generally, there has been a huge increase in the use of databases to replace the traditional reference works, encyclopedias, gazetteers etc. These *curated databases* are constructed with a lot of manual effort, usually by copying from other sources. Few of these keep a satisfactory record of the source.

[1] The terms "pedigree" and "lineage" have also been used synonymously with provenance.

- The Department of Defense will be painfully aware of the 1999 bombing of the Chinese Embassy in Belgrade which was variously attributed to an out-of-date map or a misunderstanding of the procedure intended to locate the intended target – in either case a failure to have or to use provenance information.

All these examples clearly indicate that, for data provenance, we need to understand the issues of data creation, transformation, and copying. In the following sections we shall briefly summarize the progress to date in understanding them and then outline some of the practical and major challenges that arise from this understanding. The topics treated here are, not surprisingly, related to the authors' interests. There are a number of surveys and tutorials on provenance that will give a broader coverage [BT07, SPG05, DF08, MFF+08, CCT09, BCTV08, FKSS08]

2 Models of Provenance

As interest in data provenance has developed over the past few years, researchers and software developers have realized that it is not an isolated issue. Apart from being a fundamental issue in data quality, connections have developed with other areas of computer science:

- Query and update languages.
- Probabilistic databases
- Data integration
- Data cleaning
- Debugging schema transformations
- File/data synchronization
- Program debugging (program slicing)
- Security and privacy

In some cases, such as data cleaning, data integration and file synchronization, the connection is obvious: Provenance information may help to determine what to select among conflicting pieces of information. In other cases the connection is less obvious but nevertheless important. For example in program slicing the idea is to trace the flow of data through the execution of a program for debugging purposes.

What has emerged as a result of these studies is the need to develop the right *models* of provenance. Two general models have been developed for somewhat different purposes: *workflow* and *data* provenance – also called coarse- and fine-grain provenance. Both of these concern the provenance of data but in different ways. Incidentally, we have used "data provenance" as a general term and for fine-grain provenance. As we shall see there is no hard boundary between coarse– and fine–grain provenance and the context should indicate the sense in which we are using the term.

2.1 Workflow provenance

As indicated in our examples, scientific data sets are seldom "raw". What is published is usually the result of a sophisticated workflow that takes the raw observations (e.g. from a sensor network or a microarray image) and subjects them to a complex set of data transformations that may involve informed input from a scientist. Keeping an accurate record of what was done is crucial if one wants to verify or, more importantly to repeat an experiment.

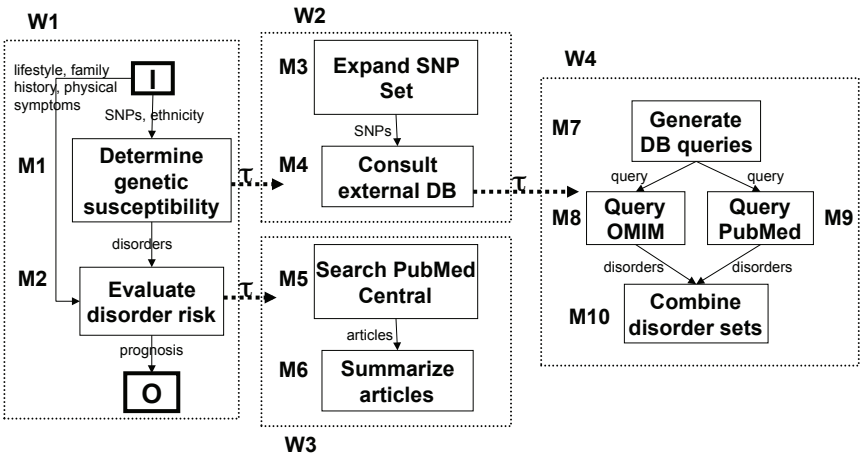

Figure 1: Disease Susceptibility Workflow

For example, the workflow in Figure 1 estimates disease susceptibility based on genome-wide SNP array data. The input to the workflow (indicated by the node or *module* labeled I) is a set of SNPs, ethnicity information, lifestyle, family history, and physical symptoms. The first module within the root of the workflow (the dotted box labeled W1), M_1, determines a set of disorders the patient is genetically susceptible to based on the input SNPs and ethnicity information. The second module, M_2, refines the set of disorders the patient is at risk for based on their lifestyle, family history, and physical symptoms. M1 and M2 are complex processes, as indicated by the τ expansions to the subworkflows labeled W2 and W3, respectively. To understand the prognosis for an individual (the output indicated by the module labeled O), it is important to be able to trace back through this complex workflow and see not only what processing steps were executed, but examine the intermediate data that is generated, in particular the results returned by OMIM and PubMed queries in W4 (modules M8 and M9, respectively).

2.2 Data Provenance

This is concerned with the provenance of relatively small pieces of data. Suppose, for example, on a visit to your doctor, you find that your office telephone number has been incorrectly recorded. It is probable that it was either entered incorrectly or somehow incorrectly copied from some other data set. In either case one would like to understand what happened in case the incorrect number occurs in other medical records. Here we do not want a complete workflow for all medical records – such a thing is probably impossible to construct. What we are looking for is some simple explanation of how that telephone number arrived at where you saw it, e.g. when and where it was entered and who copied it. More generally, as depicted in Figure 2, the idea is to extract an explanation of how a small piece of data evolved, when one is not interested in the whole system.

2.3 Convergence of models

It is obvious that there has to be some convergence of these two models. In the case of a telephone number, the operation of interest is almost certainly copying and a "local" account of the source can be given. By contrast, in our workflow example, since we know nothing about the processing involved in M3 "Expand SNP Set", then each resulting SNP in the output set can only be understood to depend on *all* input SNPs and ethnicity information, rather than one particular SNP and ethnicity information. But most cases lie

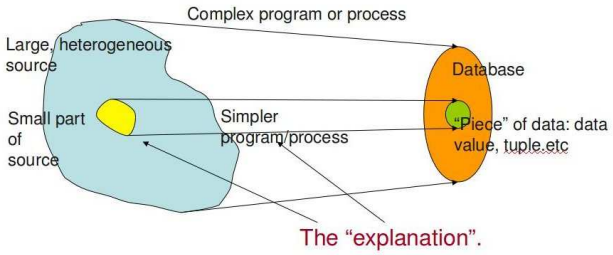

Figure 2: An informal view of data provenance

somewhere between these two extremes.

First, even if we stay within the confines of relational query languages, we may be interested in how a record was formed. For example, the incorrect telephone number may have found its way into your medical record because that record was formed by the incorrect "join" of two other records. In this case you would not only be interested in *where* the telephone number was copied from, but also in *how* that record was constructed. To this end [GKT07] describes the provenance of a record by a simple algebraic term. This term can be thought of as a characterization of the "mini-workflow" that constructed the record.

At the other end of the spectrum, workflow specification languages such as Taverna and Kepler do not simply treat their component programs and data sets as "black boxes". They are capable of executing instructions such as applying a procedure to each member of a set or selecting from a set each member with a certain property (see, for example, [TMG+07]). These are similar to operations of the relational algebra, so within these workflow specifications it is often possible to extract some finer grain provenance information.

There is certainly hope [ABC+10] that further unification of models is possible, but whether there will ever be a single model that can be used for all kinds of provenance is debatable. There is already substantial variation in the models that have been developed for data provenance, and although there is a standardization effort for workflow provenance, it is not yet clear that there is one satisfactory provenance model for all varieties of workflow specification.

3 Capturing provenance

Creating the right models of provenance assumes that provenance is being captured to begin with. However, the current state of practice is largely manual population, i.e. humans must enter the information. We summarize some of the advances being made in this area:

- Schema development: Several groups are reaching agreement on what provenance information should be recorded, whether the capture be manual or automatic. For example, within scientific workflow systems the Open Provenance Model (http://openprovenance.org/), see also [MFF+08]) has been developed to enable exchange between systems. Various government organizations are also defining the metadata that should be captured, e.g. PREMIS (Library of Congress), Upside (Navy Joint Air Missile Defense), and the Defense Discovery Metadata Specification (DDMS).
- Automated capture: Several scientific workflow systems (e.g. Kepler, Taverna, and VisTrails) automatically capture processing steps as well as their input and output data. This "log data" is typically

stored as an XML file, or put in a relational database, to enable users to view provenance information. Other projects focus on capturing provenance at the operating system level, e.g. the Provenance Aware Storage System (PASS) project at Harvard.

- Interception: Provenance information can also be "intercepted" at observation points in a system. For example, copy-paste operations could be intercepted so that where a piece of data was copied from is recorded. An experimental system of this kind has been proposed in [BCC06]. Another strategy is to capture calls at multi-system "coordination points" (e.g. enterprise service buses).

3.1 Why capture is difficult and the evil of Ctrl-c Ctrl-v

One of the simplest forms of provenance to describe is that of manual copying of data. Much data, especially in curated databases is manually copied from one database to another. While describing this process is relatively simple and while it is a particularly important form of provenance to record, progress in realizing this has been slow. There are several reasons.

First, the database into which a data element is being copied may have no field in which to record provenance information; and providing, for each data value another value that describes provenance could double (or worse) the size of the database unless special techniques such as those suggested in [BCC06] are used.

Second, the source database may have no accepted method of describing from where (from what location in the database) the data element was copied. This is related to the data citation problem.

Third, if we are to assume that the provenance information is to be manually transferred or entered, we may be requiring too much of the database curators. Typically – and understandably – these people are more interested in extending the scope of their database and keeping it current than they are with the "scholarly record".

This last point illustrates that one of the major challenges is to change the mindset of people who manipulate and publish data and of the people who design the systems they use. For example, data items are often copied by a copy-paste operation, the unassuming Ctrl-c Ctrl-v keystrokes that are part of nearly every desktop environment. This is where much provenance gets lost. Altering the effect of this operation and the attitude of the people (all of us) who use it is probably the most important step to be made towards automatic recording of provenance.

The foregoing illustrates that while, in many cases, our provenance models tell us what should be captured, our systems may require fundamental revision in order to do this.

4 Related topics

In this section we briefly describe a number of topics concerning the use and quality of data that are closely related to provenance.

4.1 Privacy

Although capturing complete provenance information is desirable, making it available to all users may raise *privacy concerns*. For example, intermediate data within a workflow execution may contain sensitive information, such as the social security number, a medical record, or financial information about an individual; in our running example, the set of potential disorders may be confidential information. Although certain users performing an analysis with a workflow may be allowed to see such confidential data, making it available through a workflow repository, even for scientific purposes, is an unacceptable breach of privacy. Beyond

data privacy, a module itself may be proprietary, meaning that users should not be able to infer its behavior. While some systems may attempt to hide the meaning or behavior of a module by hiding its name and/or the source code, this does not work when provenance is revealed: Allowing the user to see the inputs and outputs of the module over a large number of executions reveals its behavior and violates module privacy (see [DKPR10, DKRCB10] for more details).. Finally, details of how certain modules in the workflow are connected may be proprietary, and therefore showing how data is passed between modules may reveal too much of the structure of the workflow; this can be thought of as *structural privacy. There is therefore a tradeoff between the amount of provenance information that can be revealed and the privacy guarantees of the components involved.*

4.2 Archiving, citation, and data currency

Here is a completely unscientific experiment that anyone can do. One of the most widely used sources of demographic data is the World Factbook[2]. Until recently it was published annually, but is now published on-line and updated more frequently. The following table contains for each of the last ten years, the Factbook estimate of the population of China (a 10-digit number) and the number of hits reported by Google on that number at the time of writing (August 2010)

Year	Population	Google hits
2010	1,338,612,968	15,100
2009	1,338,612,968	15,100
2008	1,330,044,544	6,800
2007	1,321,851,888	15,700
2006	1,313,973,713	5,600
2005	1,306,313,812	20,700
2004	1,298,847,624	4,900
2003	1,286,975,468	3,280
2002	1,284,303,705	1,710
2001	1,273,111,290	32,700
2000	1,261,832,482	777

The number of hits that a random 10-digit number gets is usually less than 100, so we have some reason to think that most of the pages contain an intended copy of a Factbook figure for the population of China. The fluctuation is difficult to explain: it may be to do with release dates of the Factbook; it may also be the result of highly propagated "shock" stories of population explosion.

Does this mean that most Web references to the population of China are stale? Not at all. An article about the population of China written in 2002 should refer to the population of China in that year, not the population at the time someone is reading the article. However, a desultory examination of some of the pages that refer to old figures reveals that a substantial number of them are intended as "current" reference and should have been updated.

What this does show up is the need for *archiving* evolving data sets. Not only should provenance for these figures be provided (it seldom is) but it should also be possible to verify that the provenance is correct. It is not at all clear who, if anyone, is responsible for archiving, and publishing archives of, an important resource like the Factbook. The problem is certainly aggravated by the fact that the Factbook is now continually updated rather than being published in annual releases.

The story is the same for many curated data sets, even though tools have been developed for space-efficient archiving [BKTT04], few publishers of on-line data do a good job of keeping archives. What this means is that, even if people keep provenance information, the provenance trail can go dead.

[2]The Central Intelligence Agency World Factbook `https://www.cia.gov/library/publications/the-world-factbook/`

Along with archiving is the need to develop standards for *data citation*. There are well-developed standards – several of them – for traditional citations. The important observation is that citations carry more than simple provenance information about *where* the relevant information originates – it also carries useful information such as authorship, a brief description – a title – and other useful context information. It is often useful to store this along with provenance information or to treat it as part of provenance information. It is especially important to do this when, as we have just seen, the source data may no longer exist.

5 Conclusion

Provenance is fundamental to understanding data quality. While models for copy-paste provenance [BCC06], database-style provenance [CCT09, BCTV08] and workflow provenance [DF08, FKSS08] are starting to emerge, there are a number of problems that require further understanding, including: operating across heterogeneous models of provenance; capturing provenance; compressing provenance; securing and verifying provenance; efficiently searching and querying provenance; reducing provenance overload; respecting privacy while revealing provenance; and provenance for evolving data sets.

The study of provenance is also causing us to rethink established systems for information storage. In databases, provenance has shed new light on the semantics of updates; in ontologies it is having the even more profound effect of calling into question whether the three-column organization of RDF is adequate.

We have briefly discussed models of provenance in this paper. While there is much further research needed in this area, it is already clear that the major challenges to capture of provenance are in engineering the next generation of programming environments and user interfaces and in changing the mind-set of the publishers of data to recognize the importance of provenance.

References

[ABC+10] Umut Acar, Peter Buneman, James Cheney, Jan Van den Bussche, Natalia Kwasnikowska, and Stijn Vansummeren. A graph model of data and workflow provenance. In *Theory and Practice of Provenance*, 2010.

[BCC06] Peter Buneman, Adriane Chapman, and James Cheney. Provenance management in curated databases. In *Proceedings of ACM SIGMOD International Conference on Management of Data*, pages 539–550, 2006.

[BCTV08] Peter Buneman, James Cheney, Wang-Chiew Tan, and Stijn Vansummeren. Curated databases. In *PODS '08: Proceedings of the twenty-seventh ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 1–12, New York, NY, USA, 2008. ACM.

[BKTT04] Peter Buneman, Sanjeev Khanna, Keishi Tajima, and Wang-Chiew Tan. Archiving Scientific Data. *ACM Transactions on Database Systems*, 27(1):2–42, 2004.

[BT07] Peter Buneman and Wang Chiew Tan. Provenance in databases. In *SIGMOD Conference*, pages 1171–1173, 2007.

[CCT09] James Cheney, Laura Chiticariu, and Wang-Chiew Tan. Provenance in databases: Why, how, and where. *Foundations and Trends in Databases*, 1(9):379–474, 2009.

[DF08] Susan B. Davidson and Juliana Freire. Provenance and scientific workflows: challenges and opportunities. In *SIGMOD Conference*, pages 1345–1350, 2008.

[DKPR10] Susan B. Davidson, Sanjeev Khanna, Debmalya Panigrahi, and Sudeepa Roy. Preserving module privacy in workflow provenance. *Manuscript available at http://arxiv.org/abs/1005.5543.*, 2010.

[DKRCB10] Susan B. Davidson, Sanjeev Khanna, Sudeepa Roy, and Sarah Cohen-Boulakia. Privacy issues in scientific workflow provenance. In *Proceedings of the 1st International Workshop on Workflow Approaches for New Data-Centric Science*, June 2010.

[FKSS08] Juliana Freire, David Koop, Emanuele Santos, and Cláudio T. Silva. Provenance for computational tasks: A survey. *Computing in Science and Engineering*, 10(3):11–21, 2008.

[GKT07] Todd J. Green, Gregory Karvounarakis, and Val Tannen. Provenance semirings. In *PODS*, pages 31–40, 2007.

[MFF+08] Luc Moreau, Juliana Freire, Joe Futrelle, Robert E. McGrath, Jim Myers, and Patrick Paulson. The open provenance model: An overview. In *IPAW*, pages 323–326, 2008.

[SPG05] Yogesh Simmhan, Beth Plale, and Dennis Gannon. A survey of data provenance in e-science. *SIGMOD Record*, 34(3):31–36, 2005.

[TMG+07] Daniele Turi, Paolo Missier, Carole A. Goble, David De Roure, and Tom Oinn. Taverna workflows: Syntax and semantics. In *eScience*, pages 441–448, 2007.