Usage of 'provenance': A Tower of Babel
Towards a concept map

Position paper for the Life Cycle Seminar, Mountain View, July 10, 2006

Luc Moreau

June 29, 2006

At the recent International Provenance and Annotation Workshop [MF06], during a Gong Show of outlandish, "outside-the-box" ideas, I presented an analysis of the use of the word 'provenance' in some of the 28 preprints selected for presentation. Over 1100 occurrences of the word 'provenance' were identified, and out of these, 124 different usages were extracted. As will be shown in this position paper, such extensive use of the word 'provenance' results in unnecessary ambiguities and even sometimes contradictions.

Ambiguity in the use of language and concepts is particularly inappropriate for provenance, because disparate components, systems, developers, etc. have to document what happens, in order for third parties to be able to determine the provenance of data products. Hence, all parties must have a good shared understanding for such systems to work in practice. Moreover, it is a type of information that is often expected to last a long time, well after the data products have been produced (e.g., provenance of aircraft test results, important scientific experiment results); consequently, people interested in the provenance of the data products may have no way to communicate with those involved in their production, again requiring a good shared understanding to be established beforehand.

The paper [CJL+06], inspired by previous work such as [GJM+06, Mil06, GMM06], introduces a new terminology attempting to address some of these concerns, but feedback on the paper indicates some misunderstanding and suggests that clarifications are needed. Hence, in this position paper, we introduce a definition of provenance, discuss confusing usages of the word 'provenance', and present a concept map that summarises key associated concepts and their relationships; importantly, such a concept map helps avoid ambiguities of current usages.

1 Definition

The Oxford English Dictionary defines provenance as: "(i) the fact of coming from some particular source or quarter; origin, derivation; (ii) the history or pedigree of a

work of art, manuscript, rare book, etc.; concretely, a record of the ultimate derivation and passage of an item through its various owners."

Hence, we can regard provenance as the derivation from a particular source to a specific state of an item. The description of such a derivation may take different forms, or may emphasize different properties according to interest. For instance, for a work of art, provenance usually identifies its chain of ownership; alternatively, the actual state of a painting may be understood better by studying the different restorations it underwent.

The above dictionary definition also identifies two distinct understandings of provenance: first, *as a concept*, it denotes the source or derivation of an object; second, *more concretely*, it is used to refer to a record of such a derivation. Against this background, a computer-based *representation* of provenance is crucial for users to reproduce results, to perform analysis and reasoning, and to decide whether they have confidence in electronic data.

We note that the provenance of a data product is historical information about such a data product, and therefore is metadata for the data product.

2 Usages of the Word Provenance

While the following study of the word 'provenance' in the papers presented at IPAW'06 [MF06] is not meant to be exact, it is likely to represent typical usages of the word in current scientific publishing. Essentially, four broad usage categories have been identified.

"*Something provenance*" Two different usages fall into this category. The frequent expression '*data provenance*' means the origin of data. Such an example is not ambiguous, as opposed to an expression such as '*process provenance*', for which only the context can help the reader decide whether it is meant to denote the origin of a process, or a process-oriented view of provenance.

"*Provenance something*" In this very frequently used category, we find expressions such as '*provenance record*' or '*provenance traces*', which refer to the representation of provenance in computer systems. Also frequent are expressions such as '*provenance system*' or '*provenance architecture*', which refer to the software infrastructure capable of building and using a representation of provenance for our data products. Such examples are clear, but the expression '*provenance recorder*', which is often meant to denote a software component that records part of the representation of the provenance of a data product, fails to convey the precise intended meaning by simply adopting these two words.

"*Provenance of something*" This category is generally fine but not very often used, probably because of its heavier nature, and usually preferred to be abbreviated as '*something provenance*', with the ambiguity raised above.

"*To do something with/on/for Provenance*" In this category, we find the frequently used expressions '*to capture provenance*' or '*to track provenance*', which do not make the distinction between provenance as a concept or provenance as a representation, which may be crucial in some specific contexts.

Inevitably, we found a few 'provenance' pearls, which exhibit the ambiguity that we have identified and the associated meanings of the word. Is '*provenance history*' the history of history? When saying '*provenance metadata*', do we mean metadata of metadata? By '*actor provenance*', do we mean the provenance of an actor? Is '*data provenance data*' an element of the representation of the provenance of a data item? Is '*provenance of interactions*' the description of interactions or what caused interactions to occur? Is '*prospective provenance*' the history of the future? Finally, when saying '*we do provenance*', are we making history?

The conclusion of this exercise is that the research community studying the concept of provenance needs both to identify associated key concepts and to use terms to denote them in a consistent manner. In the following section, we present some of these concepts in the form of a concept map.

3 A Concept Map for Provenance

Figure 1 presents a concept map inspired from our previous work on specifying an open provenance architecture [GJM$^+$06]. Since our aim is to identify the provenance of electronic data, we define *the provenance of a data product as the process that led to that data product.* In our concept map, we distinguish the concept of provenance from its representation in a computer system.

A user who cares about the provenance of a data product has an interest in specific kind of information related to the process that led to the data product (cf. restoration vs ownership for a work of art); likewise, their needs can identify how far in the past the information should come from (are we going as far as the big bang?). Hence, we see the representation of provenance as the result of a provenance query, which operates over the documentation of a process. Such documentation has a concrete representation, which we refer to as the p-structure. It consists of a structured set of p-assertions made by the different components of the application, whose execution produced the data product we consider the provenance of. Such p-assertions are asserted by software components and describe the components' involvment in a process.

We coined new terms to denote the representation of process documentation (p-structure) and its constituents (p-assertions). Intuitively, the former would correspond

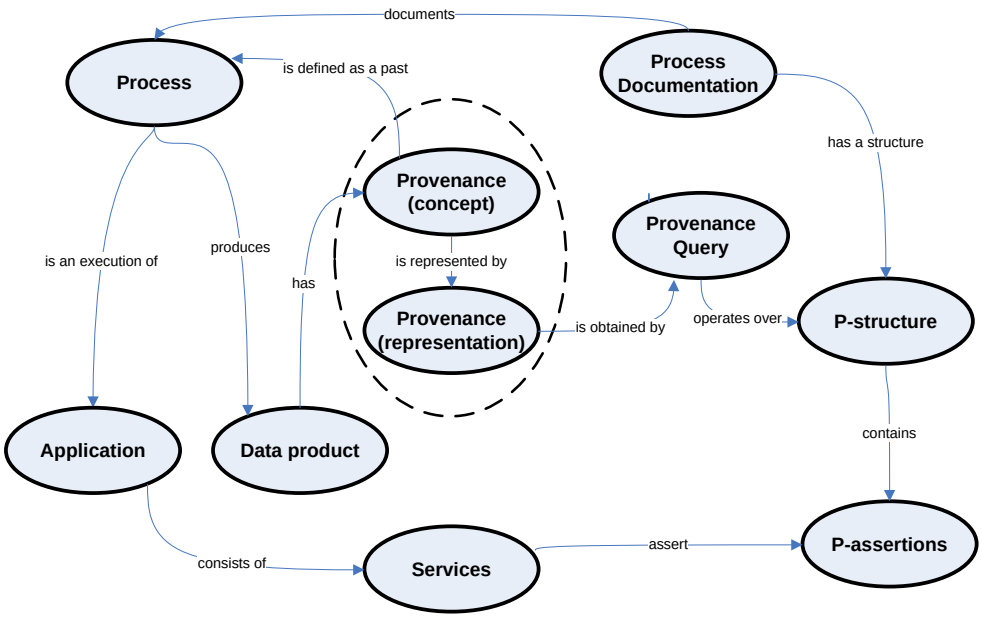

Figure 1: Concept Map for Provenance

to a 'provenance trace', while the latter to a 'provenance record'. We prefer using such coined words to make it explicit that they are *not* provenance.

We note that, while our discussion focused on electronic data and computer-based applications, nothing in this concept map restricts us to this. In fact, the provenance of a physical artefact could also be obtained by querying the documentation of the physical process that led to this artefact.

Scientists in multiple application domains also regard human-entered annotations as crucial to capture the motivation or reasons of experiments, or even analysis of produced results. In our view, such user annotations are also data products, produced by the execution of an application (for instance, an annotation capturing tool), which can be documented like any other process.

If all software components produce a description of their execution, process documentation would be very complete, and from it, many useful information items (such as in the dublin core) could be extracted automatically by provenance queries. Exposing the outcome of such provenance queries, user annotations, part of or entirety of process documentation is crucial for data preservation. The act of selecting and exposing such information is the responsibility of the curator, and the aggregated information is referred to as ancillary information in [CJL+06].

4 Conclusion

It is important for the community to identify concepts, their relationships, and associate them with names to avoid ambiguities and contradictions. We have presented here a map of concepts related to provenance, and it would be very beneficial to extend this map to encompass data preservation related concepts.

Once concepts have been clearly defined and have broad support from the community, it becomes easier to recommend actions. For instance, standardisation [MI06] has been advocated as a means to promote inter-operability of systems. We can now precisely identify candidates for such standardisation.

References

[CJL+06] Simon Cox, Rachel Jones, Bryan Lawrence, Natasa Milic-Frayling, and Luc Moreau. Interoperability issues in scientific data management (version 1.0). Technical report, The Technical Computing Initiative, Microsoft Corporation, March 2006.

[GJM+06] Paul Groth, Sheng Jiang, Simon Miles, Steve Munroe, Victor Tan, Sofia Tsasakou, and Luc Moreau. D3.1.1: An architecture for provenance systems. Technical report, University of Southampton, February 2006. Available from http://eprints.ecs.soton.ac.uk/12023/.

[GMM06] Paul Groth, Simon Miles, and Steve Munroe. Principles of high quality documentation for provenance: A philosophical discussion. In Luc Moreau and Ian Foster, editors, *International Provenance and Annotation Workshop (IPAW'06),*, volume 4145 of *Lecture Notes in Computer Science*. Springer, May 2006. To Appear.

[MF06] Luc Moreau and Ian Foster, editors. *Proceedings of the International Provenance and Annotation Workshop (IPAW'06)*, volume 4145 of *Lecture Notes in Computer Science*, Chicago, IL, May 2006. Springer. To Appear.

[MI06] Luc Moreau and John Ibbotson. Standardisation of provenance systems in service oriented architectures — white paper. Technical report, University of Southampton, 2006. Available from http://eprints.ecs.soton.ac.uk/12198/.

[Mil06] Simon Miles. Electronically querying for the provenance of entities. In *Proceedings of the International Provenance and Annotation Workshop 2006 (IPAW 2006)*, Lecture Notes in Computer Science. Springer, 2006. To appear.