J Data Semant (2012) 1:11–17
DOI 10.1007/s13740-012-0002-0

EDITORIAL

A Semantic Foundation for Provenance Management

Sudha Ram · Jun Liu

Received: 29 February 2012 / Accepted: 2 March 2012 / Published online: 28 March 2012

Abstract Provenance is a term used to describe the lineage, history, or origin of an object. While provenance originated from the art world, it is now becoming increasingly important in the context of digital objects on the World Wide Web. Large scale scientific collaborations and social media platforms on the web have enabled production and sharing of a variety of digital objects on the web. With the proliferation and sharing of such objects, which include documents, pictures, videos, and more, questions such as "where did this object come from?", "who else is using this object?" and "for what purpose was it generated?" are becoming increasingly common. To ensure that digital objects from different sources can be trusted and used appropriately, it is imperative that the provenance of the digital objects be tracked, recorded, and made available to its users. In this work, we attempt to provide a foundation for understanding provenance, clearly define the semantics of provenance, distinguish provenance from "uses or application" of provenance, suggest a mechanism for managing provenance, and provide important directions for research in provenance management.

Keywords Provenance management · Semantics · Digital data provenance · Data lineage · Semantic Web technologies · Data curation

S. Ram (✉)
University of Arizona, Tucson, USA
e-mail: ram@eller.arizona.edu

J. Liu
Opera Solutions, Inc., San Diego, USA
e-mail: jliu@operasolutions.com

1 Introduction

The term "provenance" is often used in association with a piece of art or literature. Indeed, Merriam-Webster defines provenance as "the history of ownership of a valued object or work of art or literature" [4]. Knowing the provenance of a work of art is of great importance, in that, it can "help to determine the authenticity and to establish the historical importance of a work by suggesting other artists who might have seen and been influenced by it, and to determine the legitimacy of current ownership" [12]. The motivation for understanding provenance of an artwork is also applicable to digital objects such as data generated in various business and scientific domains. In recent years, the development of new technologies such as web services, Semantic Web, and social media platforms, has enabled large-scale, often worldwide, collaborations that involve sharing of data, knowledge and other resources. For instance, in scientific domains such as chemistry and biology, the tendency toward "big science" (i.e., large-scale collaborative science) is increasingly evident. Scientists are performing advanced research tasks in large collaborative projects such as the *iPlant Collaborative* (iPlant, http://www.iplantcollaborative.org), where they use data generated from many different labs and people. With the proliferation and sharing of such data, questions such as "where did this data come from?", "who else is using this data?" and "for what purpose was it generated?" are becoming increasingly common. To ensure that data provided by other sources can be trusted and used appropriately, it is imperative that the provenance of the data be recorded and made available to its users.

In this paper, we attempt to provide a foundation for understanding provenance, clearly define the semantics of provenance, distinguish provenance from "use or application" of provenance, suggest a mechanism for managing provenance,

and provide important directions for research. Our foundation for provenance is informed by our work on provenance management in different domains such as materials management for a large manufacturing organization, scientific data management for the iPlant Collaborative [9], and Wikipedia [2].

2 Uses of Provenance

Provenance may be used for various purposes [7]. For instance, provenance is often used to gauge the quality of a digital object or to decide how trustworthy it is. It may be used for purposes of creating an audit trail to evaluate if any errors were made in processing the object. Sometime provenance acts as the replication recipe for an object. Often it may be used to provide attribution for an object or to resolve disputes over intellectual property or ownership of an object. Lastly, provenance may be used purely for information purposes, so that the user can decide whether the object is appropriate for their needs or not. Note that we use the term "object" in a generic way, and we address the issue of the nature and granularity of the object later. Whatever the purpose of the provenance, it is clear that provenance needs to be recorded so that it can be useful in the future. It is also not possible to anticipate every possible use of provenance, thus it needs to be very comprehensive.

3 Nature and Granularity of Objects for Provenance Tracking

Before we start to discuss the semantics of provenance, it is important to note that provenance can be recorded or tracked for different types of objects. Hence, the notion of an object is fundamental to provenance semantics. For instance on Wikipedia, an object may refer to an article or, at a finer granularity, it may refer to each paragraph of an article or, moving to an even finer level, to each sentence or each word of an article. In the case of a materials management domain for a manufacturing organization, an object may refer to a relational database, or to finer granules, such as each relational table that records the physical properties (e.g. tensile strength) for samples of a composite material, or, each tuple in such a table, or even each attribute value in the same table. In a social media environment such as YouTube (http://www.youtube.com), an object may refer to a video, or in the case of Flickr (http://www.flickr.com) an object may refer to an album of photographs or an individual photograph. Based on this it is apparent that provenance can be tracked for different kinds of objects and it is important to identify or define the types of objects in each environment or system where provenance is going to be tracked. It is also desirable to track provenance at a fine level of granularity of objects because it can then be aggregated to generate provenance for a coarse granule.

4 Object Life Cycle as a Foundation for Provenance Semantics

Typically, provenance for a piece of art (such as a painting) starts from the time the painting is created and gets accumulated during its life as it passes from one owner to the next. Similarly, our fundamental philosophy for defining provenance semantics stems from the life cycle of digital objects. We consider that every object has a life that is defined by events starting with its "birth" till its "death." "Birth" of an object refers to its creation, while "death" of an object refers to its deletion or in many cases archiving or retirement of the object.

The notion of starting with an object and defining the important events in its life is fundamental to defining the semantics of provenance. This has its foundation in philosophy and ontologies. According to Bunge [1], the history or lifecycle of a thing is manifested by a sequence of events that affect the thing during its lifetime. An event occurs when the thing changes one or more of its properties. A digital object has a variety of properties including its content, format, owner, storage location and access rights. An event affects a digital object when it is created, changes one or more of its properties, or when it is ultimately destroyed.

From our examination of various domains mentioned above, it is clear that life cycle events can be divided into two mutually exclusive classes, those that affect the content of the object and those that do not. For example, events such as storage, transfer, and archiving do not affect the content of an object. However, events such as review, modification and annotation are related to the content of an object. Therefore, it is important to identify each event precisely within these two classes in the life of an object and decide which ones are necessary to track. Some events such as "access" may occur many times during the life of an object, while others, such as "creation", only occur once during the life of an object. The reason events are so fundamental to the semantics of provenance is because the rest of the semantics of provenance can be understood with respect to this "anchor" point. We refer to events as the "WHAT" of provenance.

5 Anchoring Provenance Elements Around Object Life Cycle

Once the important events in the life of an object are identified, the rest of provenance simply defines the details of the events. These include the "time (WHEN)" at which the event occurs, the "agent (WHO)" responsible for the event, the "mechanism (HOW)" or actions that cause the event to occur, the "place (WHERE)" where the event occurs, the "instrument (WHICH)" that makes the event happen, and the "reason (WHY)" the event occurs. For example, consider

Table 1 Definition of provenance semantics

Provenance element	Construct in Bunge's ontology	Definition
What	Event	An event that affects a data object during its lifetime
How	Action	One or more actions that lead to the event
Where	Space	Location of the event. If data traveled (or was copied) from one location to another during the event, "where" refers specifically to where the data came from
When	Time	Time of the event
Who	Agent and other things	Individuals or organizations involved in the event
Which		Software or instruments used in the event
Why	N/A	Reasons that explain why the event occurred

a photograph as an object. The first event or "What" in its life is the "creation" of the photograph, "When" refers to the time (e.g. 4 PM February 24, 2012) at which the photograph was taken (i.e. created), "Who" refers to the person (photographer Mr. Powers) who took the photo, "Where" refers to the location (Tucson, AZ) of the photo, "Which" refers to the camera (Sony Cybershot DSC H10) used for taking the photo, "How" refers to the mechanism (recorded by zooming in from 20 ft), and "Why" refers to the reason (for including in a magazine article) for taking the photo. After taking the photo, it is stored on a hard drive. Thus "Transfer" becomes another event associated with the photograph, and the rest of the provenance describes this event, that it was stored at 5 PM, February 24, 2012 (When) by copying (How), from Camera to Hard Drive (Where), by Mr. Powers (Who). In this case, "why" and "which" may not be relevant and therefore are not tracked. Thus, the semantics of provenance may be defined by what we term as the W7 model and we refer readers to our earlier work that formally defines the semantic constructs of such a model [6,8].

As stated in our earlier work, we adopt Bunge's ontology [1] in defining provenance. Bunge defines the *history* of a thing as a sequence of events that occurred to a thing during its lifetime. An event happens to a thing as it is acted upon by an agent or other things. Also, an event occurs in space and time. Hence, we define provenance as consisting of various events that happen during the lifetime of a data object from its creation to destruction, and then include *how*, *who*, *when*, *where*, *which* and *why* associated with each event. Table 1 shows these 7 Ws and explains the mappings between them and the constructs in Bunge's work.

Figure 1 shows the W7 model and its conceptual representation [11]. The overall structure of provenance (shown in Fig. 1a) includes the 7 Ws (represented as concepts in boxes) and the relationships (represented in ovals) between them. *What* is the anchor of our W7 model. Each of the 7 Ws has its subtypes, as shown in Fig. 1b. For instance, *What* has subtypes including *creation*, *modification*, *publication*, and *ownership. How,* representing an *action* leading to an event, can be classified into *single action* and *complex action*. As shown in Fig. 1b, we identify actions associated with different types of event. For instance, an ownership event affects an object when it is *purchased* or *donated*. Data creation often results from the *measurement* or *observation* of a physical phenomenon.

6 Provenance Semantics for Different Domains

While a model such as W7 defines the general semantics of provenance, it can be tailored to suit the needs of different domains.

Figure 2 below depicts the semantics of provenance for a subset of objects in the iPlant Collaborative [9]. The specific objects are phylogenetic tree datasets. The events relevant to these objects are creation, modification, reconciliation, and sharing (see Fig. 2a). Typically, plant biologists (users of the iPlant Collaborative) "create" a phylogenetic tree dataset by importing it (how), then they may "modify" it by deleting a species, and finally "share" it for use by others. Figure 2b represents an example of an event graph. A domain model of provenance can include a number of event graphs, each of which represents the semantics associated with an event. The event graph shown in Fig. 2b indicates that plant biologists often create a phylogentic tree by importing (how) it from some external database (where). Our previous research [9] proposed a graph-based approach to deriving event graphs from the W7 model shown in Fig. 1.

Similarly, Fig. 3 shows semantics of provenance for a materials management domain. As shown in Fig. 3a, the provenance of a data object such as, the tensile strength values of samples of a composite material used in manufacturing, includes details associated with its creation, modification and approval. Figure 3b indicates that some test

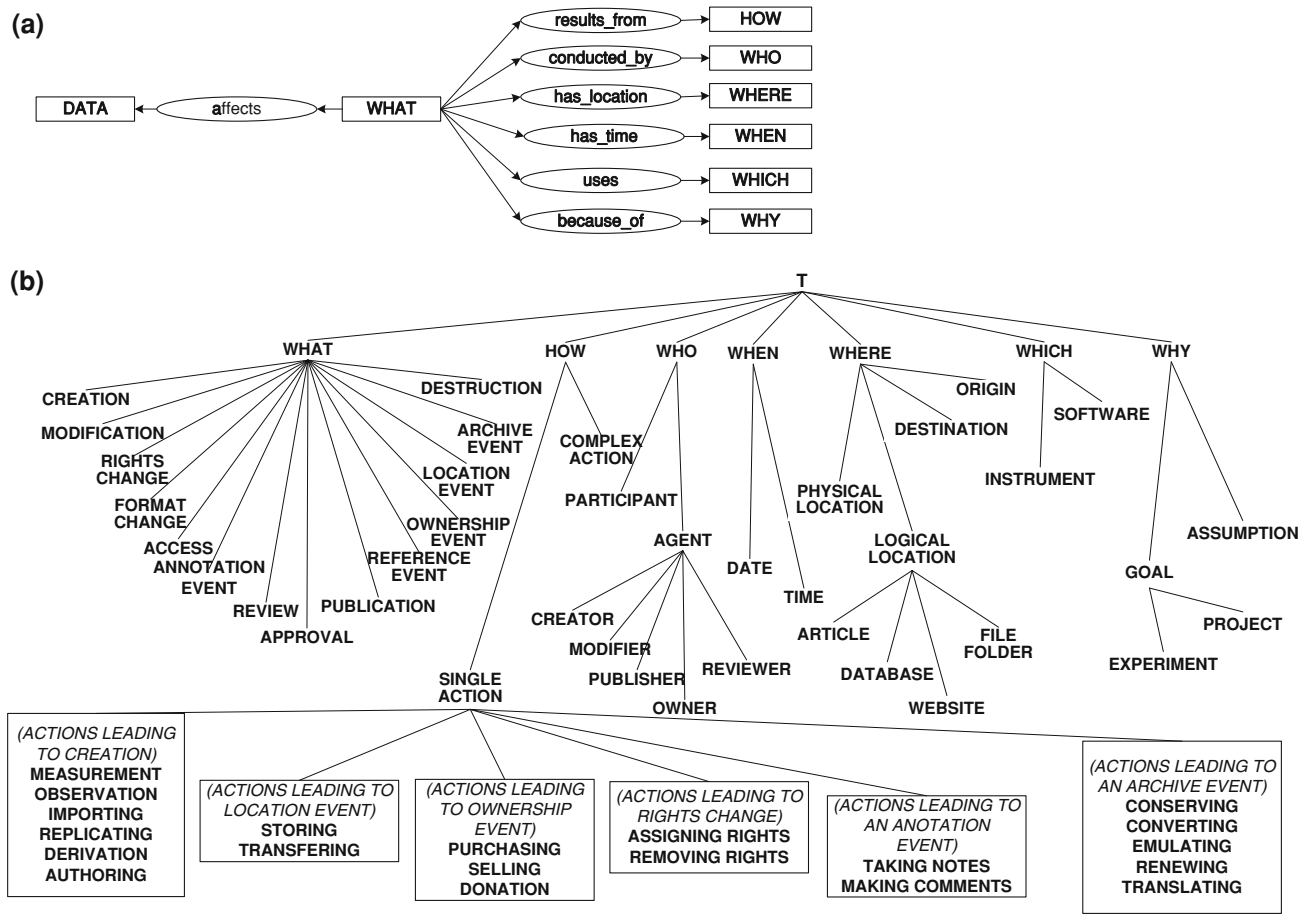

Fig. 1 Provenance semantics using the W7 model

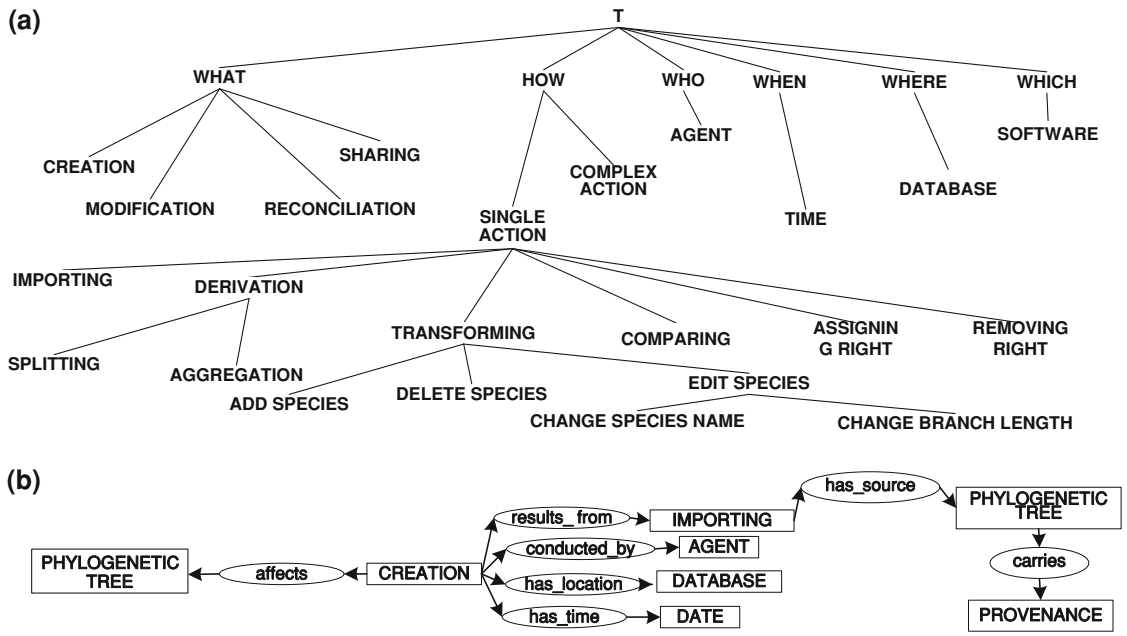

Fig. 2 Provenance semantics for a phylogenetic tree dataset [9]

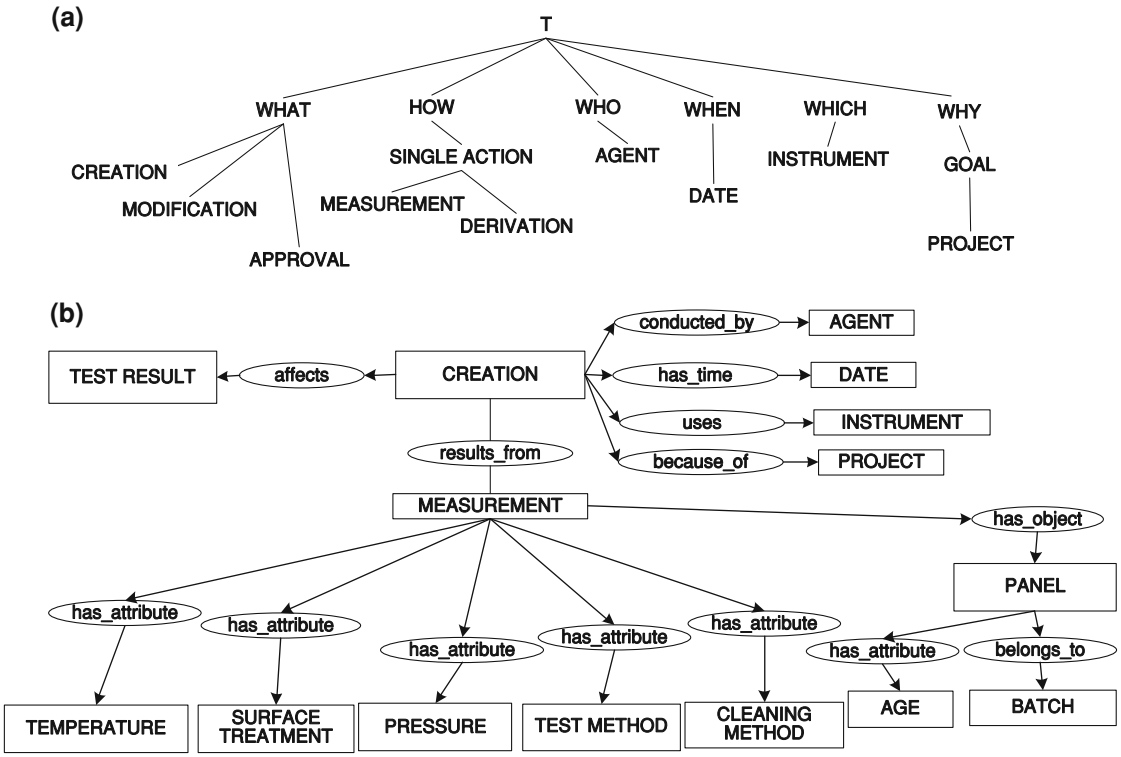

Fig. 3 Provenance semantics for a materials properties dataset

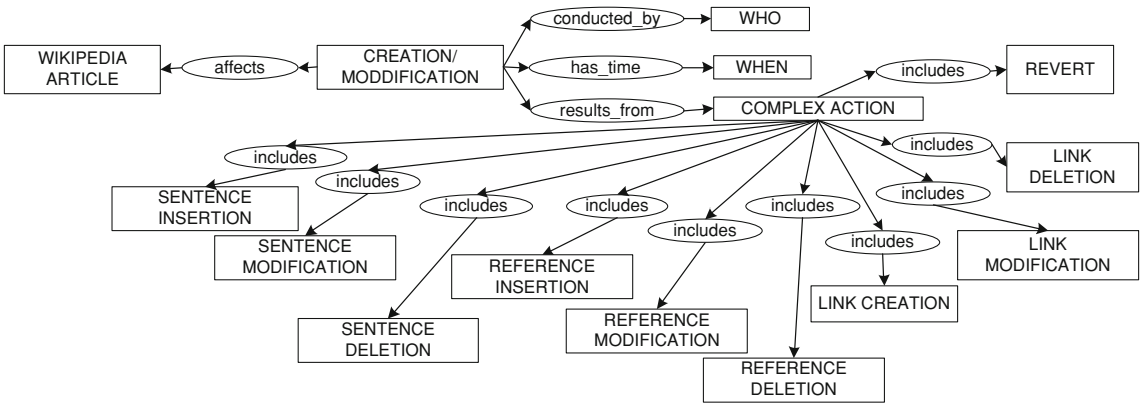

Fig. 4 Provenance semantics for a Wikipedia article

result (e.g., tensile strength value) can be created (what) by an agent (who) on a specific date (when) through measurement (how) by an instrument (which) for a specific project (why). Figure 3b also indicates that it is necessary to record details about the measurement such as which cleaning and test method have been used in the measurement. We also create event graphs for capturing the semantics of other events such as modification and approval.

Finally, an event graph for a Wikipedia article is shown in Fig. 4. It shows that creation or modification of a Wikipedia article results from a complex action that includes a number of single actions, such as sentence insertion/modification/deletion, reference insertion/modification/deletion.

From these three domains, it is evident that while a model such as W7 defines the general semantic elements of provenance, not all of these elements are relevant for every domain. It is also up to each domain or system to decide the level of detail required within each element of provenance. Thus the iPlant Collaborative may decide that they do not need to track every "access" of a phylogenetic tree, but they may want to track every "share" event for a phylogenetic tree. The semantics of provenance as defined in the earlier sections can be

physically represented in a variety of formats such as a relational database, XML, RDF or a noSQL database.

7 Research Directions for Provenance Semantics and Management

Ideally, provenance should be tracked at source, i.e. at each point in the life of an object. Post hoc tracking of provenance is a very difficult problem and is almost impossible to achieve based on our experience. A general architecture for a provenance tracking system called PROMS [10] is shown in Fig. 5.

We recommend that provenance be tracked at a fine granular level of objects because we do not know when and how it will be used in the future. It is important to note that provenance will accumulate over time as objects go through their life cycle. Thus, the volume of provenance can easily be several orders of magnitude larger than the object itself. Indeed, provenance can be considered to be a form of "BIG DATA" [3]. Based on our experience with the various domains of science, manufacturing, and social media websites, we recommend a "Google Analytics (http://www.google.com/analytics)" or "webserver log" like approach for tracking and managing provenance based on the lifecycle of objects. While many of the provenance elements can be automatically tracked, the only element that is likely to need human input is "why" or the reason for each event. It is therefore advisable to devise a mechanism to automatically harvest provenance every time an object is touched in any way in the digital world.

The semantics of provenance as described here form a foundation for developing a standard model of provenance. Thus, efforts such as the Open Provenance Model [5] will benefit from the philosophy, semantics, and structure described for the W7 model [6,8,9]. Using this approach the OPM effort can consider defining event types for various domains and the actions (how) associated with each type of event as well as the other details. While the OPM has some elements of provenance defined in W7, we believe the underlying foundation and semantics we have prescribed will help make it extensible and adaptable to different domains.

It would be worth exploring the issue of provenance inheritance and aggregation as well as the levels of granularity of objects for which provenance should and could be tracked. Inheritance of provenance can only occur from a finer granule or object to a coarser object which is different from the traditional concept of inheritance. Moreover, objects maybe of different kinds—in addition to datasets, or relational databases, objects may be software modules, workflows, documents, images and others. Thus, provenance semantics should be defined and refined for each of these types of objects and their various granularities. These, in turn, should lead to the development of query languages for provenance that focus on the semantics of provenance.

The issue of how much provenance to expose to different users is an open issue. Intelligence agencies require detailed provenance tracking for their objects. The provenance is then accessed by users with various levels of security clearance. These security levels will drive access to not only data but also provenance. In some cases, data access may only be allowed at an aggregate level, which, in turn, will need mechanisms to control the exposure of provenance for objects at different levels of granularity.

It is inevitable that objects will get created and modified in a system that tracks provenance and then it may exit the current system via email or perhaps some other off-line mechanism. Such "export" events should be recorded and tracked, although what happens to the object outside of the current system will not be possible to track. Finally, if such an object returns to the current system, mechanisms to recognize it and seamlessly link its new "life" to its old "life" based on an analysis of its provenance need to be devised.

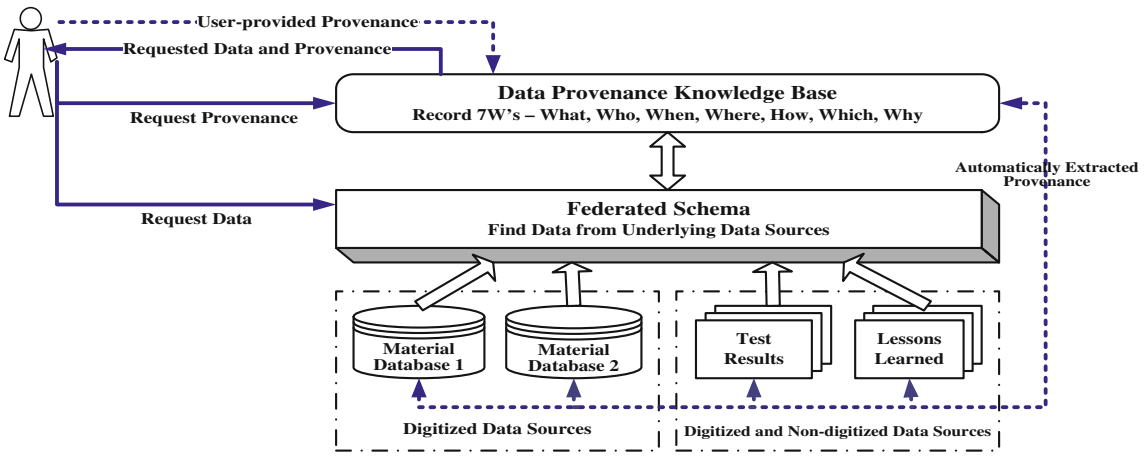

Fig. 5 Architecture of PROMS

Provenance is supposed to be immutable and any attempts to change it should themselves be tracked and recorded. Mechanisms similar to "watermarking" may also be needed to authenticate provenance and to make sure that provenance is never separated from the object itself. Finally, it is not clear what happens to provenance when an object is "deleted". Destruction of an object may either completely obliterate its provenance or the provenance may remain without being attached to any object, which destroys the integrity of the provenance record.

We hope this analysis of the foundations of provenance and its semantics will provide food for thought and foster several new research directions for the readers.

Acknowledgments This research was supported in part by grants from the National Science Foundation #DBI-0735191, and IIS#0455993, and the Science Foundation of Arizona.

References

1. Bunge M (1977) Treatise on basic philosophy: ontology I. The furniture of the world, vol 3. Reidel, Boston
2. Liu J, Ram S (2011) Who does what: collaboration patterns in the Wikipedia and their impact on article quality. ACM Trans Manag Inf Syst 2(2):23 (article 11)
3. Lynch C (2008) Big data: how do your data grow? Nature 455:28–29
4. Merriam-Webster Online. http://www.m-w.com/home.htm
5. Moreaua L, Clifford B et al (2011) The open provenance model core specification (v1.1). Future Gener Comput Syst 27(6):743–756
6. Ram S, Liu J (2007) Understanding the semantics of data provenance to support active conceptual modeling. Lecture notes in computer science, vol 4512. Springer, Berlin, pp 17–29
7. Ram S, Liu J (2008) A semiotics framework for analyzing data provenance research J Comput Sci Eng 2(3):221–248
8. Ram S, Liu J (2009) A new perspective on semantics of data provenance. In: Proceedings of the first international workshop on the role of semantic web in provenance management, Washington, DC
9. Ram S, Liu J (2010) Provenance management in biosciences. Lecture notes in computer science, vol 6413. Springer, Berlin, pp 54–64
10. Ram S, Liu J et al (2006) PROMS: a system for harvesting and managing data provenance. In: Proceedings of the 16th annual workshop on information technologies and systems, Milwaukee
11. Sowa J (1999) Conceptual graphs: draft proposed American National Standard. Lecture notes in artificial intelligence, vol 1640. Springer, Berlin, pp 1–65
12. Tan W (2004) Research problems in data provenance IEEE Data Eng Bull 27:45–52