2009

# A New Perspective on Semantics of Data Provenance

Sudha Ram

Jun Liu

# A New Perspective on Semantics of Data Provenance

Sudha Ram, Jun Liu
430J McClelland Hall, Department of MIS, Eller School of Management,
University of Arizona, Tucson, AZ 85721

**Abstract**: Data Provenance refers to the “origin”, “lineage”, and “source” of data. In this work, we examine provenance from a semantics perspective and present the W7 model, an ontological model of data provenance. In the W7 model, provenance is conceptualized as a combination of seven interconnected elements including “what”, “when”, “where”, “how”, “who”, “which” and “why”. Each of these components may be used to track events that affect data during its lifetime. The W7 model is general and extensible enough to capture provenance semantics for data in different domains. Using the example of the Wikipedia, we illustrate how the W7 model can capture domain or application specific provenance.

## 1. Introduction

Data provenance is an overloaded term that has been defined differently by different people. A recent survey [1] reviews the various definitions of provenance in literature. Some researchers define provenance as the origin or source of data [2]. As an example, Buneman puts forth two forms of data provenance, i.e., “why” provenance and “where” provenance [3]. Both “why” and “where” provenance deal with tracing the source from which the data came. Others view provenance as metadata recording the process of experimental workflows, annotations and notes about scientific experiments [4]. In research such as [5, 6], the data generating processes in the form of workflows are the primary entities for which provenance are collected. Due to the lack of consensus on the semantics or meaning of provenance, current efforts on capturing data provenance have focused on only one or two aspects of provenance while ignoring others. As a result, the provenance is often incomplete and cannot be shared across applications. In response to this challenge, we attempt to formally define the semantics of provenance that can be agreed upon by people from different domains. To our knowledge, our research is the first of its kind to explore the “semantics” of provenance.

In this research, we define the W7 model, an ontology that clarifies the semantics of data provenance. The W7 model represents data provenance as a combination of seven interconnected elements including, “what”, “when”, “where”, “how”, “who”, “which”, and “why”. The W7 model is general and extensible enough to capture provenance semantics for data in different domains. Using examples in Wikipedia, we illustrate how the W7 model can help define, capture, and use data provenance.

## 2. Use cases and competency questions

Following the formal methodology for ontology development proposed in [7], we started by collecting use cases from different domains. Given the set of use cases, a set of competency questions were identified. The competency questions are those that our ontology must be “competent” to answer. Our use cases and their corresponding competency questions describe a set of requirements the ontology must satisfy. They helped us understand the intended informal semantics of the concepts and relations to be included in the ontology. We gathered 188 use cases from users in various domains including biology, businesses (such as the manufacturing, defense, and pharmaceutical organizations), and physical sciences. We present some of the use cases as well as the competency questions.

*Use Case 1:* In a missile manufacturing company, an engineer performs a material test to measure the transverse tension fatigue life of a particular material “S2/8552 glass-epoxy”. She then publishes the results and test procedure online. Another engineer discovers the published results years later. Before reusing the results, he verifies whether the results are valid by repeating the test procedure, in the test environment that was described.

*Competency Questions:* A replication of a material test requires recording provenance that is competent in answering the following questions: 1) *how was the material data created*, and 2) *how was the material test conducted in terms of the test procedure, test environment, sample condition, temperature, etc*.

*Use Case 2*: To organize the huge amounts of bio images being generated, the bio-computing lab stores bio images on different storage devices based on their “value”. For instance, images created by a graduate student doing an internship or images that have not been accessed for 5 five years are deemed less valuable.

*Competency Questions:* Use case 2 demonstrates the desire for recording “*who created the bio images*” and “*when the bio images have been accessed*”.

*Use Case 3:* An engineer obtains the composite material “Cycom 381/S2 Uni-glass” and performs a test to measure the tensile strength of the composite. Another engineer in a different lab later performs a test on the same material, again provided by the same vendor. She compares the two results and notices significant differences. She needs to assess whether the differences are because of different test methods or different instruments used in the test.

*Competency Questions:* To determine the quality or reliability of material test results, it is necessary to provide answers to the following two questions: 1*) how were the results generated*, and 2) *which instrument was used in created the data and what were its parameter settings?*

*Use Case 4:* A genetics researcher records in his lab notebook the reason for using specific data records in an *in silico* experiment, e.g., “I chose this restriction enzyme as it cut only three times within 200 base pairs of the SNP”.

*Competency Questions:* The relevant question is *why certain records data were used*.

*Use Case 5:* A scientist, S, is interested in rainfall and water levels in neighboring rivers and lakes for a part of the Sierra Nevada mountain range in California. He is trying to acquire sensor signals captured in Southern California.

*Competency Questions:* Use Case 5 indicates the use of data provenance for data discovery. In this use case, the question the scientist needs to answer is *“where was the data measured”,* so that he can locate the appropriate data.

Table 1: Summary of use cases and their competency questions

| *Competency question* | *Number of use cases* |
|---|---|
| What | 188 |
| How | 156 |
| Who | 145 |
| Which | 91 |
| When | 131 |
| Where | 113 |
| Why | 86 |

Table 1 summarizes the use cases and their competency questions. As an example, the *how* question was necessary to answer in 156 use cases. Our analysis of the use cases and their competency questions indicates that the provenance ontology must contain information regarding *who*, *when*, *where*, *how*, *why* and *which*. Moreover, all of the use cases indicate that the central element of interest is the event that affects each piece of data during its life cycle from birth (creation) to death (deletion or archiving). While many of the use cases point out the need to understand the data creation related provenance, in many cases, other life cycle events are even more useful. For example, *Use case 4* requires us to record the *why* associated the *use* of data. Also, for some domains, the most critical provenance events are *changes in the ownership* of the data and *archiving* of data. As a result, our provenance ontology should be competent to answer the question of "*what"*, i.e., *events* that affect the data. Thus our ontology is anchored around the "what" or the life cycle events affecting the data.

## 3. Conceptualization of provenance based on Bunge's theory

The use case analysis helped us identify the basic components of data provenance including the 7 Ws (*w*hat, ho*w*, *w*hen, *w*here, *w*ho, *w*hich, and *w*hy). We then adopt Bunge's ontology [8] to define these components and identify the relationships between them.

*State, event and history*: The elementary notion of Bunge's ontology is a *thing*. The *state* of a thing is the set of property values of the thing at a given time. Bunge's ontology postulates that everything changes, and every change is a change of state of things, that is the change of properties of things. A change of state is termed an *event*. It follows that an event occurs when a thing acquires or loses a property or changes the value of a property. Based on the constructs of *event* and *state*, Bunge defines the concept of history: History of a thing is a sequence of *event*s that happens to the thing.

*Action, agent, time and space*: These are constructs related to events. An event on a thing occurs when it is *acted* upon by another thing, which is often a human or a software *agent.* An event happens in *time* and *space*.

Data are also "things". Bunge's theory regarding *history* and *events* is a perfect match for defining data provenance and its semantics since data provenance is often referred to as the pedigree or *history* of data. More importantly, our use case analysis indicates that data provenance is really all about various *events* that affect data during its life cycle. Thus, the constructs in Bunge's ontology including *history*, *event*, *action*, etc. lay a theoretical foundation for defining provenance and its components. We define provenance and the 7 Ws and develop connections between them using the constructs in Bunge's ontology.

## 4. An ontological model of data provenance – the W7 model

We conceptualize data provenance as consisting of seven interconnected elements including what, when, where, who, how, which, and why.

*Definition (Provenance).* Provenance of some data *D* is a set of n-tuples: *p(D)* = {< *What, When, Where, How, Who, Which, Why* >}. *What* denotes an event that affected data during its lifetime; *When* refers to the time at which the event occurred ; *Where, is* the location of the event; *How*, is the action leading up to the event; *Who*, is agents involved in the event; *Which*, are the programs or instruments used in the event; and *Why*, the reasons for the events. We therefore name our ontological model for provenance the W7 model. A graphical representation of the W7 model is shown in Figure 1. We represent the W7 model as conceptual graphs (CGs) developed

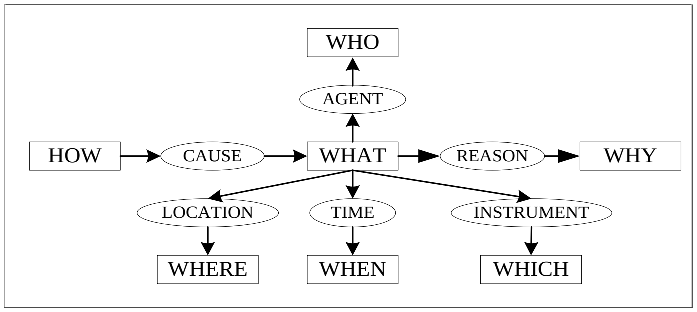

Figure 1. Overview of the W7 model

by Sowa [9], which has been widely as a language for ontology. The boxes in CGs represent concepts and the bubbles are the relationships. As shown in Figure 1, *what*, i.e., events, is the anchor of our model. In essence, data provenance includes events and various information (including who, how, when, where, which and why) associated with and describing the events.

Tables 2 summarizes the definition of each of the 7 Ws and shows the correspondence between the Ws and Bunge's ontology concepts. For interested readers, please refer to our previous research [10] for a more detailed discussion of each of the 7 Ws.

Table 2: Definition of the 7 Ws

| *Provenance Element* | *Construct in Bunge's ontology* | *Definition* |
|---|---|---|
| What | Event | An event (i.e. change of state) that happens to data during its life time |
| How | Action | An action leading to the events. An event may occur, when it is *acted* upon by another thing, which is often a human or a software agent |
| When | Time | Time or more accurately the duration of an event |
| Where | Space | Locations associated with an event |
| Who | Agent and other things | Agents including persons or organizations involved in an event |
| Which | | Instruments or software programs used in the event |
| Why | - | Reasons that explain why an event occurred |

In [11], Simmhan et al argue that due to the diverse needs across disciplines, it is challenging to develop a standard model for capturing provenance. To address this concern, we developed the W7 model as a generic ontology of provenance that captures the semantics of data provenance and can thus be applied to various domains. However, for our model to be of any practical use, it must be easily adaptable to address domain specific provenance needs. We use the "type definition" mechanism developed by Sowa [9] to provide the domain specific extension of the W7 model. The CG formalism enables to explicitly define the semantics of a concept via a type definition. As an example, in the domain of design and manufacturing, *how* often refers to a material test, using which material data is created. The specification of the test and the material sample used in the test are critical provenance that needs to be captured. We thus formally define "material test", as shown in Figure 2.

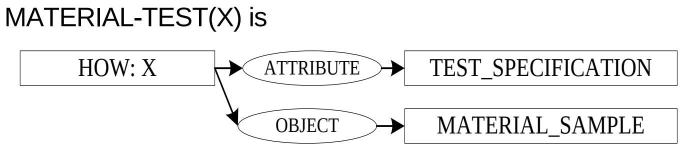

Figure 2. Type definition of the concept "material-test"

The CG in Figure 2 defines MATERIAL-TEST as a subtype of HOW. A material test is carried out upon material samples and it has an attribute "test specification". Type definitions represent the semantics and necessary attributes of a concept that have been agreed upon by people in a domain and therefore can be used to provide domain specific extensions of the W7 model.

## 5. Application of the W7 model – the Wikipedia example

We use Wikipedia as an example to illustrate the application of the W7 model to harvest and structure data provenance. Table 3 summarizes the application of the W7 model in Wikipedia. *What* or events that affect a Wikipedia page are primarily creation, modification and destruction of the page. Other events may include "quality assessment" (e.g., a page may be designated as a featured page) or "change in access rights" (e.g., a page may be locked to prevent editing by anonymous editors). The *"How" construct* for a page modification event may be sentence insertion/update/deletion, link insertion/update/deletion, reference insertion/update/deletion, and reverts (see Table 3). These are actions made by editors that may lead to the modification of a page. *Who* represents the editors of a Wikipedia page. The Wikipedia distinguishes between three types of users: 1) administrators, 2) registered editors, and 3) anonymous editors. *When* refers to the time an event occurs. In the Wikipedia, a timestamp is automatically recorded in the database whenever an event occurs. *Where* in the Wikipedia represents the IP address from which an editor makes a change. *Which* in Wikipedia refers to bots, i.e., software that automatically edits Wikipedia pages. The Wikipedia allows an editor to input *why*, i.e., justifications for a change, in the "comment" field.

Table 3: Application of the W7 model in Wikipedia

| *Provenance Element* | *Application to a Wikipedia article* |
|---|---|
| What | Creation, modification, destruction, quality assessment, access rights change |
| How | Sentence insertion/update/deletion, link insertion /update/deletion, reference insertion/update/ deletion, revert (reverting the article to a previous version) |
| Who | Administrators, registered editors, and anonymous editors |
| When | Timestamps of the events |
| Where | IP address of the editor |
| Which | Software used in editing the page |
| Why | User comments |

Harvesting data provenance in the Wikipedia requires little human effort. The Mediawiki software used by the Wikipedia is set to automatically capture the *what*, *who*, *when*, *where*, and *which.* The *how* provenance can be derived by comparing two versions of a page using the *diff* function. Only the *why* provenance demands manual input. Applying the W7 model to the Wikipedia enables us to harvest provenance of the Wikipedia pages in a structured and comprehensive way. Data provenance in the Wikipedia has widely been used to automatically assess the quality of Wikipedia pages. As an example, [12] suggests metrics such as "rigor" (total number of changes made for the article) and "diversity" (total number of unique editors for the article) as measures of quality. In our recent study [13], we track every action by an editor that affects the life of a Wikipedia article from its creation to the present time. We classify roles by mining the provenance, i.e., various actions carried out by a contributor on an article. We then further identify collaboration patterns based on provenance in terms of *who* does *what.* The collaboration patterns derived from data provenance have been proved to be correlated with data quality of Wikipedia pages.

## 6. Conclusion and Future Research

In conclusion, the focus of our research is on investigating the semantics of provenance. We have developed a generic provenance model, i.e., the W7 model, to represent these semantics. We identify various elements of provenance such as "what", "where", "when", "who", "how", "which" and "why" and present the semantics of each of these elements. Our W7 model is inspired by theoretical work such as Bunge's ontology as well as our empirical analysis of provenance use in many application domains. It is a generic model of data provenance and is intended to be easily adaptable to represent domain specific provenance requirements. Using the Wikipedia as an example application, we illustrate the use of the W7 model to harvest and track data provenance. We are continuing to use this model to harvest and track provenance in a variety of other application domains.

## References

[1] S. Ram and J. Liu, "A Semiotics Framework for Analyzing Data Provenance Research," *Journal of computing Science and Engineering*, vol. 2, pp. 221-248, 2008.

[2] P. Buneman, S. Khanna, and W. C. Tan, "Data Provenance: Some Basic Issues," Proceedings of FSTTCS, New Delhi, India, 2000.

[3] P. Buneman, S. Khanna, and C. T. Wang, "Why and Where: A Characterization of Data Provenance," in *Lecture Notes in Computer Science*, vol. 1973, pp 316-330, Springer, 2001.

[4] J. Frew and R. Bose, "Earth System Science Workbench: A Data Management Infrastructure for Earth Science Products," Proceedings of 13th International Conference on Scientific and Statistical Database Management, Fairfax, VA, 2001.

[5] R. Bose, "A conceptual framework for composing and managing scientific data lineage," Proceedings of 14th International Conference on Scientific and Statistical Database Management, 2002.

[6] M. Greenwood, C. Goble, R. Stevens, J. Zhao, M. Addis, D. Marvin, L. Moreau, and T. Oinn, "Provenance of e-Science Experiments - experience from Bioinformatics," Proceedings of UK e-Science All Hands Meeting, Nottingham, UK, 2003.

[7] M. Grueninger and M. Fox, "Methodology for the Design and Evaluation of Ontologies," Proceedings of Workshop on Basic Ontological Issues in Knowledge Sharing, Montreal, Quebec, Canada, 1995.

[8] M. Bunge, *Treatise on Basic Philosophy: Vol. 3: Ontology I: The Furniture of the World*. Boston, MA: Reidel, 1977.

[9] J. Sowa, *Conceptual structures: Information processing in Mind and Machine*. Reading, MA: Addison-Wesley, 1984.

[10] S. Ram and J. Liu, "Understanding the Semantics of Data Provenance to Support Active Conceptual Modeling," in *Lecture Notes in Computer Science*, vol. 4521, pp 17-29, Springer-Verlag, 2007.

[11] Y. Simmhan, B. Plale, and D. Gannon, "A Survey of Data Provenance Techniques," Indiana University, Technical Report IUB-CS-TR618, 2005.

[12] A. Lih, "Wikipedia as Participatory Journalism: Reliable Sources? Metrics for Evaluating Collaborative Media as a News Resource," Proceedings of 5th International Symposium on Online Journalism, 2004.

[13] J. Liu and S. Ram, "Who Does What: Collaboration Patterns in the Wikipedia and Their Impact on Data Quality," Proceedings of nineteenth Annual Workshop on Information Technologies and Systems(WITS 2009), Phoenix, Arizona, USA, December, 2009.