Understanding the Semantics of Data Provenance to Support Active Conceptual Modeling

Sudha Ram, Jun Liu
430J McClelland Hall, Department of MIS, Eller School of Management,
University of Arizona, Tucson, AZ 85721
Email: {ram, junl}@eller.arizona.edu
URL: http://adrg.eller.arizona.edu/

Abstract: Data Provenance refers to the lineage of data including its origin, key events that occur over the course of its lifecycle, and other details associated with data creation, processing, and archiving. We believe that tracking provenance enables users to share, discover, and reuse the data, thus streamlining collaborative activities, reducing the possibility of repeating dead ends, and facilitating learning. It also provides a mechanism to transition from static to active conceptual modeling. The primary goal of our research is to investigate the semantics or meaning of data provenance. We describe the W7 model that represents different components of provenance and their relationships to each other. We conceptualize provenance 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. A homeland security example illustrates how current conceptual models can be extended to embed provenance.

1 Introduction

Data Provenance refers to the lineage or history of information including its origin, key events that occur over the course of its lifecycle, and other details associated with its creation, processing, and archiving. It is the background knowledge that enables a piece of data to be interpreted correctly and to support learning. We believe that tracking provenance, such as the processing and usage history of data, enables users to share, discover, and reuse the data, thus streamlining collaborative activities and reducing the possibility of repeating dead ends.

Despite its critical importance, current approaches to capturing provenance of data have not been particularly effective. As suggested by [1], data provenance needs to be captured with the hope that it is comprehensive enough to be useful in the future. However, due to the lack of consensus on the semantics or meaning of provenance, the concept has not been well-defined in the literature. For instance, some researchers define provenance as the origin of data and its movement between databases [2], while others view it as the process of transformation of data [3]. Accordingly, current efforts aimed at capturing data provenance typically focus on some aspects of provenance while ignoring others. As an example, [4] identifies two kinds of provenance – “why” and “where”. The former refers to the source data that had some influence on the creation of the data of interest; the latter specifies the location(s) in the databases from which the data was extracted. We believe that provenance includes more than what is captured in [4]. In some application domains, provenance may include the literature reference where data were first reported, the history in terms of how the data was created and transformed, the series of experimental procedures by which it was derived from other data, and the sequence of ideas leading to an experiment. Consequently, to generate a complete record of data provenance, it is desirable to gain a deep understanding of the semantics of provenance and identify the key concepts associated with it. To our knowledge, none of the existing work has explored the “semantics” of provenance.

The primary goal of our research is to investigate the semantics or meaning of data provenance. We have developed a generic model called the W7 model that represents data provenance as a combination of seven interconnected elements including, “what”, “when”, “where”, “how”, “who”, “which”, and “why”. Each of these elements may be used to track provenance and may be applied to different domains such as homeland security. Further, we demonstrate how our W7 model can help in active conceptual modeling.

2 Provenance Semantics – The W7 Model

Conventional conceptual models do not provide a straightforward mechanism to explicitly capture the semantics of data provenance, and it is still unclear how provenance information can be linked with the application data at the conceptual level. In response to this problem, we propose an ontological model called the W7 model to capture the semantics of data provenance.

2.1 Theoretic basis – Bunge’s ontology and other philosophical work

To understand the semantics of provenance and identify concepts associated with it, we use Bunge’s ontology [5, 6] as our starting point. Table 1 summarizes the concepts from Bunge’s ontology that are appropriate for understanding data provenance.

The data stored in an information system is a thing and thus has history that is represented as a sequence of events or state changes. Based on the observation that data provenance describes history of data, we propose to define provenance by recording all events that happen to data during its lifetime. Each of these events happens to a data object when it is created or destructed or when it is acquires/loses one or more of its properties or changes the values of its properties. In database applications, these events center around the lifecycle of data which includes creation, updates, and deletion of data.

Table 1. Selected Ontological concepts from Bunge (adapted from [5]and [7])

Ontological Concept	Explanation
Thing	A thing, the elementary unit in the ontology, is a substantial individual endowed with its properties. Two or more things (composite or simple) can be associated into a composite thing
Property Intrinsic Mutual	Things possess properties. A property that is inherently a property of a thing is called an intrinsic property. A property that is meaningful in the context of two or more things is called a mutual or relational property
State	The property value vector of a thing at a given time
Event Qualitative Quantitative	An event is a change of state of a thing. An event of a thing is qualitative when it acquires or loses one or more properties. An event is quantitative when the thing changes its property values
History	A sequence of events or state changes of a thing
Action	The history of a thing changes when it is under actions of other things
Space and time	Everything exists in space and time. The history of any thing has a nonempty projection onto the spacetime, a unification of space and time

However, simply recording *what* events occur is not sufficient to meaningfully represent the provenance of data. To provide insightful provenance knowledge, it is necessary to identify and explicitly describe various details describing the events. Bunge's ontology includes some constructs related to events such as space, time and actions. According to [5], space and time provide the basic framework of events. An event happens and history of a thing changes when it is under actions of an agent or other things. Hence, a causal relationship exists between an action and an event [5]. A further investigation of the relationship between events and actions leads us to other philosophical work such as [8] and [9] that help identify other important concepts related to events. In [8], Davis defines an action as a "doing" in which an event occurs when an agent wants or has other reasons for the event to happen. An action provides the causal explanation, i.e., "how"-explanation, while the agent's intentions, beliefs and other mental events provide the reason explanation, i.e., "why"-explanation for the event. Chisholm, on the hand, stresses the importance of the agent for an event by proposing the notion of "agency causation": An agent brings about an event by undertaking something [9]. Based on these philosophical works pertinent to events and actions, we identify concepts that provide insights to events including actions (how), reasons (why) and agents (who and which) in addition to space (where) and time (when).

2.2 Overview of the W7 model

Based on our theoretical analysis of data provenance, we conceptualize provenance as consisting of seven interconnected dimensions including what, when, where, who, how, which, and why.

Definition 1. Provenance is defined as a n-tuple *P = (WHAT, WHEN, WHERE, HOW, WHO, WHICH, WHY, OCCURS_AT, HAPPENS_IN, LEADS_TO, BRINGS_ABOUT, IS_USED_IN, IS_BECAUSE_OF)*, where

- *WHAT* denotes the sequence of events that affect the data object; *WHEN*, the set of event time; *WHERE*, the set of all locations; *HOW*, the set of all actions leading up to the events; *WHO*, the set of all agents involved in the events; *WHICH*, the set of all devices; *WHY*, the set of reasons for the events. The formal definition of each of these 7 Ws is given later.

- *OCCURS_AT* is a collection of pairs of the form *(e, t)*, where *e* ? *WHAT* and *t* ? *WHEN*. *HAPPENS_IN* is a collection of pairs of the form *(e, l)*, where *l* denotes a location. *LEADS_TO* represents pairs of the form *(e, h)*, where *h* ? *HOW* and represents an action leading up to the event *e*. *BRINGS_ABOUT* is a collection of pairs $(e, \{a_1, a_2, \ldots, a_k\})$, where $a_1, a_2, \ldots, a_k$? *WHO*, indicating that more than one agent can cooperate to bring about an event, and *IS_USED_IN* is also a collection of pairs $(e, \{d_1, d_2, \ldots, d_k\})$, where $d_1, d_2, \ldots, d_k$? *WHICH*. Finally, *IS_BECAUSE_OF* is a collection of pairs of the form $(e, \{y_1, y_2, \ldots, y_k\})$, where $y_1, y_2, \ldots, y_k$? *WHY*.

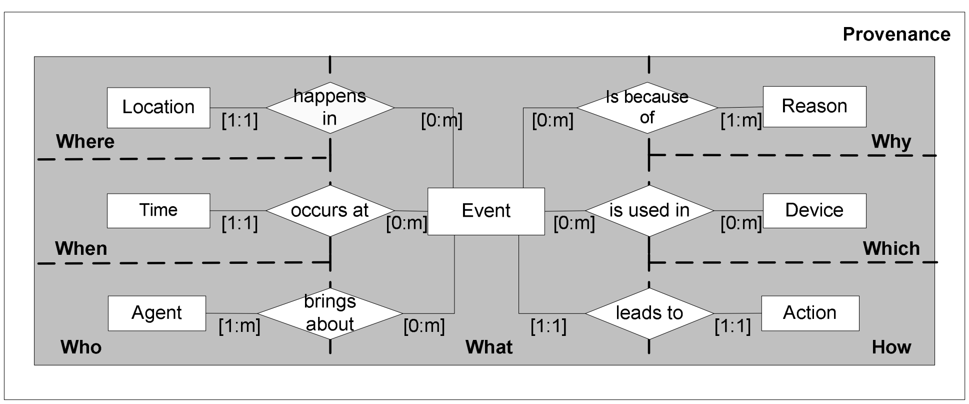

Fig. 1. Overview of W7 Model

Fig.1 provides an overview of the W7 model. We allow data provenance to be specified for data objects at different granularity levels. For example, for some applications, provenance may be specified on instances of an entity class, and their provenance be aggregated to form the provenance of the entity class or a subset of it.

Alternatively, data objects that are subset of a parent dataset may share their provenance with the parent and yet be different as a whole. “What”, i.e., a sequence of events, is the anchor of our W7 model. The other provenance components are semantically related to “what” and describe various details about the events. An event “occurs at” a time and “happens in” a location. A “leads_to” relationship exists between “action” and “event”, indicating the causal relationship between actions and events. An action is taken by some *agents* using some *devices* for some *reasons*, which is reflected by the various relationships existing between “what” and the elements “who”, “which” and “why”. We discuss the semantics of each element and provide a graphical representation using the Unifying Semantic Model (USM) [10], an extended version of the Entity-Relationship (ER) model [11]. The USM is used in provenance modeling for two purposes. First, the USM extends the ER model with constructs such as super/subclasses (represented by “—⟨S⟩► ”) and groups/aggregates (represented by “—⟨Group⟩► ”), thus providing a formal and precise expression of provenance semantics. Second, it enables us to directly apply the W7 model to conceptual modeling with easy adaptations.

2.3 What

The fundamental building block of the W7 model is the element “what”. Its semantics are defined as follows.

Definition 2. *WHAT* is a sequence of events $<e_1, e_2, \ldots, e_n>$ that affect a data object during its life time.

We categorize events by drawing upon Bunge’s theory. A data object such as an entity instance in a relational database or a document in a digital library is a thing often composed of other things (e.g. an entity instance is a composite thing made up of attribute values). A data object as a thing has intrinsic and mutual or relational properties. The intrinsic properties of a data object normally include its content and composition, while its mutual properties of a data object provide its contextual information such as its ownership, custody, rights, etc. An event of thing is a change of state. In addition to creation or “coming into being” event [6] and destruction, an event happens to a data object when it acquires or loses one or more of its intrinsic or mutual properties (i.e., qualitative events) or when it changes its property values (i.e., quantitative events). Accordingly, we classify events into creation, transformations, destruction, and contextual events. Transformations of a data object represent changes that happened to its intrinsic properties such as its content. Contextual events are manifested by changes made to mutual properties. They can be further classified into ownership changes, changes of storage location, etc. Fig. 2 presents a graphic representation of “what”.

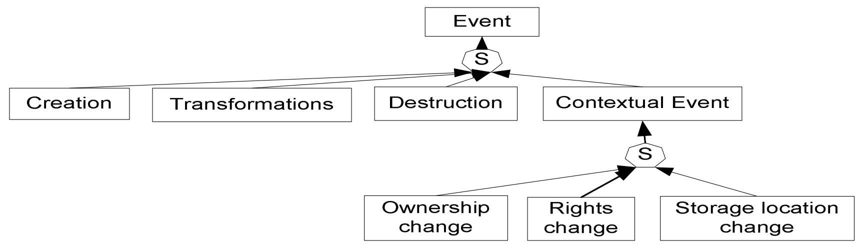

Fig. 2. Semantics of "What"

2.4 When

The semantics of "when" are shown in Fig. 3. Different from existing temporal data models such as [12, 13] that capture valid time, i.e., a time period during which a fact is true in the real world, and transaction time, i.e., a time period during which a fact is stored in the database, our model focuses on recording time of various events that affect data during its lifetime. As an example, given a script of a correspondence between two terrorist suspects, we capture the duration or time period during which the script is recorded as its creation time. When the script is stored in a database, we record the data storage time. When the script is accessed/used, we capture the time period during which it is used. Associating a timestamp with each event provides a detailed timeline of the events and enables us to reconstruct the history of the data.

Definition 3. *WHEN* represents a set of timestamps $\{t_1, t_2, \ldots, t_n\}$ associated with various provenance events.

While some events may be instantaneous, others may occur over an interval of time. Accordingly, we specify two disjoint subsets of *WHEN* including *INTANT* and *DURATION. INSTANT* is a set of instants. Each instance is a point on the time line. *DURATION* represents a set of durations. A duration refers to the time between two instants with a start and an end. Hence, let *INSTANT_VALUE* denote a set of values an instant can take, and we define two functions: *Start: DURATION → INSTANT_VALUE* and *End: DURATION → INSTANT_VALUE.* Moreover, *DURATION* should be well-formed, which entails a constraint $\forall t$? *DURATION, Start(t)<End(t).*

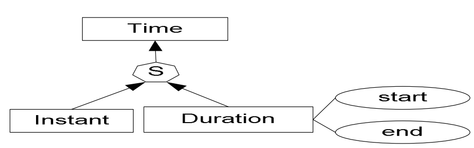

Fig. 3. Semantics of "When"

2.5 Where

The element "where" in the W7 model captures event locations. We provide a graphic representation of "where" in Fig. 4.

Definition 4. *WHERE* denotes a set of locations $\{l_1, l_2, \ldots, l_n\}$, where various events happen.

The most common forms of representing locations are physical and geographical. Physical locations specify the position of places or points based on a global coordinate system, while geographical locations signify an area or boundary governed by a common law and are normally organized hierarchically. Correspondingly, we capture these two concepts as two subsets of *WHERE* including *PHYSICAL_LOCATION* and *GEOGRAPHICAL_LOCATION*. In addition to physical and geographical location, we introduce the concept of *logical location*, which links a data object to its location in a server or database. This concept is important since data may travel between information sources due to location change events such as storage and transfer. The logical location can often be represented by a URI.

A location can be typed into *source* and *destination*. Let *LOCATION* represent a set of locations, and we specify a function that maps a transaction location to its type as *Type*: $LOCATION \rightarrow T$, where $T = \{Source, Destination\}$.

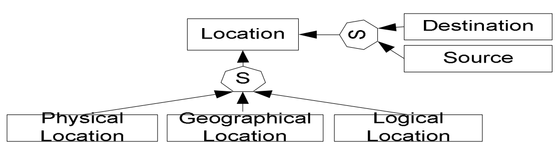

Fig. 4. Semantics of "Where"

2.6 How

"How" documents *actions* that lead to the occurrence of an event. An action is seen as a system of "doings", where agents work on certain objects in order to obtain a desired outcome. Actions are causes of event, and events are brought into being as results of actions performed by agents. Consider the case of creation of a data object. A data object can be created as a result of an action such as an observation or a measurement. It can also be generated by deriving it from existing data or acquiring it from other people.

Information regarding actions normally includes:

- *Preconditions* that refer to conditions that must hold prior to the enactment of an action.

- *Methods* that provide detailed descriptions about what has been done and capture various action parameters.

- *Inputs* and *outputs* that refer to data objects that are manipulated by the enactment of an action. An action can thus be seen primarily as a process of transforming a set of inputs into outputs.

- *Sources* refer to people or media we acquire data from. As an example, a CIA agent is a source of a terrorist report.

Definition 5. HOW is defined as a tuple *(ACTION, PRECONDITIONS, METHODS, INPUTS, OUTPUTS, SOURCES,* $\mathcal{P}, \mathcal{M}, \mathcal{I}, \mathcal{O}, \mathcal{S}$*)*, where

- *ACTION* = $\{h_1, h_2, \ldots, h_n\}$ is a set of actions, and the previously mentioned concepts such as *Preconditions, Methods, Inputs,* and *Resources* are also defined as sets.

- $\mathcal{P}$: *ACTION* → *PRECONDITIONS* is a function that maps an action to its preconditions; $\mathcal{M}$: *ACTION* → *METHODS* maps an action to its methods; $\mathcal{I}$: *ACTION* → *INPUTS* maps an action to its inputs; $\mathcal{O}$: *ACTION* → *OUTPUTS* maps an action to its outputs; and $\mathcal{S}$: *ACTION* → *SOURCES* associates an action with the sources.

Following [14], we classify actions into *primitive* and *complex* (see Fig. 5). Accordingly, we specify that *ACTION* consists of two subsets *PRIMITIVE_ACTION* and *COMPLEX_ACTION*. An action is considered to be primitive if no decomposition will reveal any further information which is of interest. Complex actions, on the other hand, may be arbitrarily complex activities and can be decomposed into primitive actions that happen sequentially or simultaneously. Moreover, previous research such as [3] has been focused on capturing the procedures used for processing the data, by describing the workflow of an experiment. Accordingly, we define a "depends_on" relationship that captures the control flow of primitive actions within a complex action such as concurrency, sequence, etc.

Definition 6. A complex action $c = (P, DEPENDS\_ON)$, where

- $P = \{p_1, p_2, \ldots, p_k\}$ is a set of primitive actions that constitute the complex action *c*.

- $DEPENDS\_ON = \{d_1, d_2, \ldots, d_k\}$ is a set of relationships. Each relationship *d* is an ordered pair (p_i, p_j), where p_i, p_j ? *P*.

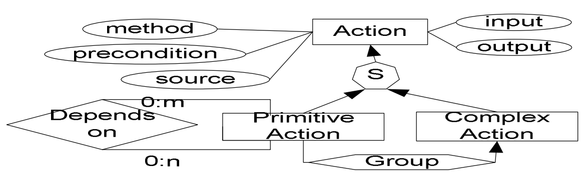

Fig. 5. Semantics of "How"

2.7 Who

"Who" refers to *agents* who bring about the events. The USM diagram of "who" is shown in Fig. 6. An agent is "an intentional entity", that is it has some idea of purpose that guides its actions [14]. We use the term “agent” refer to human agents including *persons* and *organizations*. Artificial agents such as software applications are captured in “which”. An agent assumes a *role* in an event, and a role is defined as “a coherent set of activities to be assigned to an agent as a functional responsibility”[15]. Each agent plays a certain role to make some contributions to the action leading up to an event. For instance, a federal agent may play the role of supervisor in creating the script of a suspicious correspondence.

Definition 7. *WHO* is a triple *(AGENT, ROLE,* $\mathcal{RL}$ *)*, where

- $AGENT = \{a_1, a_2, \ldots, a_n\}$ is a set of agents that are involved in various events.
- $ROLE = \{r_1, r_2, \ldots, r_n\}$ is a set of roles agents are allowed to assume.
- $\mathcal{RL}: AGENT \rightarrow ROLE$ is a function that associates an agent with the role she played in a particular event.

WHO includes two subsets, i.e., a set of persons *PERSON* and a set of organizations *ORGANIZATION*. We often need to capture the *position* and *affiliation* of a person. When a person participates in her affiliation, she is no longer entirely free to choose her goals and actions. Instead, she accomplishes some activities according to her position. A *position*, which is called organizational role in [14], represents a set of responsibilities of an individual in her affiliation. As a result, we specify a function $\mathcal{PA}: PERSON \rightarrow POSITION$? *AFFILIATION* that maps a person to her position and affiliation.

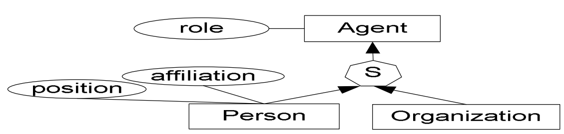

Fig. 6. Semantics of “Who”

2.8 Which

The element “which” describes which *devices* are used in data creation, analysis, and transformation. Devices can be distinguished into *instruments* (e.g. equipments and hardware) and *applications*. When an event involves a device, some level of detail about the device in which it is hosted should be captured. Moreover, some actions are specifically supported or offered by certain devices, whereby the characteristics and capability of the devices may play an integral role in describing the behavior of the action.

As shown in Fig. 7, the information related to a device is logically divided into three classes depending on the type of information they provide, namely device *description*, *function* and *settings*. Device description contains basic information related to a device such as its name, vendor, version, etc. A device’s function can be specified in terms of the variables of the device itself, e.g., a battery’s function is often specified as providing an electric voltage measured in volts. More frequently, a device is composed of parts or components, and its function is expressed in terms of the variables of its components. As an example, a computer may have a CPU of 2.0 GHz and memory of 256 MB. Different from the functional properties that rarely change throughout a device's lifetime, its *settings* contain volatile information pertaining to the device such as current level of CPU usage and remaining power level of a computer. The settings of a device often vary among applications, and it specifies the performance of the components of a device during an event.

Definition 8. WHICH is a tuple *(DEVICE, SETTINGS, DESCRIPTION, FUNCTION, $\mathcal{S}$, $\mathcal{D}$, $\mathcal{F}$),* where

- *DEVICE* = $\{d_1, d_2, \ldots, d_n\}$ is a set of devices used in various events. It consists of two disjoint subsets *INSTRUMENT* and *APPLICATION.*

- *SETTINGS* denotes a set of settings a device can take, *FUNCTION* represents a set of device functions, and *DESCRIPTION* denotes a value set of descriptions a device can take.

- $\mathcal{S}$: *DEVICE* $\rightarrow$ *SETTINGS,* $\mathcal{D}$: *DEVICE* $\rightarrow$ *DESCRIPTION,* and $\mathcal{F}$: *DEVICE* $\rightarrow$ *FUNCTION* represent mappings from a device to its settings, description, and function.

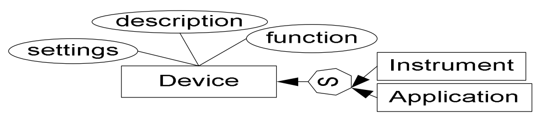

Fig. 7. Semantics of "Which"

2.9 Why

In this subsection, we define the semantics of "why" and provide a USM representation of the semantics in Fig. 8.

Definition 9. WHY represents a set of reasons $\{y_1, y_2, \ldots, y_n\}$ for various provenance events.

Our scheme for representing "why" is based largely on the Belief-Desire-Intention Model [16], which identifies *beliefs*, *desires* and *intentions* as significant factors that affects decision making. Beliefs represent knowledge of the world, desires are goals assigned to the agent, and intentions are commitments by an agent to achieve particular goals. Here, we collapse desires and intentions into *goals*. As a result, we specify two subsets of *WHY*, i.e., *BELIEF* and *GOAL.* The former represents a set of beliefs and the latter a set of goals.

A natural way to answer "why" questions is by tracing them to goals. For example, why a milestone is established in an anti-terrorist action can be related to the goal that the action be completed on time. Explicit representation of goals is important because it allows us to study a specific event from an intentional point of view. We also define an "is_reduced_to" relationship to capture the goal-subgoal structure (See Fig. 8). This relationship corresponds to the classical reduction operator in the problem reduction approach to problem solving. A goal can have several parent goals as it can occur in several reductions. We define *IS_REDUCED_TO* = $\{s_1, s_2, \ldots, s_n\}$ as a set of relationships representing goal-subgoal structures. Each relationship *s* is an ordered pair (g_i, g_j), where g_i, g_j ? *GOAL*. Furthermore, each relationship *s* ? *IS_REDUCED_TO* should not be symmetric. Thus, we specify a constraint on *s* as *s* ? *IS_REDUCED_TO* and s^{-1} ? *IS_REDUCED_TO* ? $s = s^{-1}$.

The other important concept associated with "why" is the concept of *belief.* Agents have a subjective view of the world, where they form their beliefs. Different from the goals an agent intends to fulfill through an action, beliefs refer to what an agent believes prior to the action, and they form the background upon which an agent can choose to act in a particular way [17]. We further classify beliefs into *assumptions* and *hypotheses* (see Fig. 8).

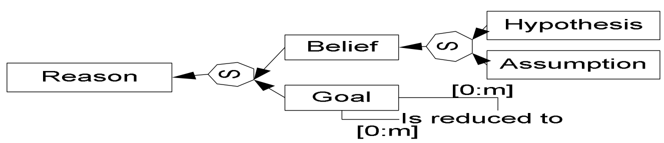

Fig. 8. Semantics of "Why"

3 Active Conceptual Modeling with Provenance Annotations

As discussed in [18], a serious problem in today's data modeling practices is that database design approaches have viewed data models as representing only a snapshot of the world and recommend ignoring variations of information as well as the causes and other details of those variations during data modeling. In response to this problem, Chen et al. propose active conceptual modeling [19] that describes all aspects of the world, its activities, and its changes under different perspectives, thus providing a multilevel and multi-perspective view of reality. Active conceptual modeling requires us to capture provenance knowledge in terms of *what* event/change may happen to the data, at the stage of conceptual modeling. Moreover, we need to identify provenance components such as "where", "when", "how", "who", and "why" behind the "what" to provide insights about the changes. Our W7 model captures the semantics of the various provenance components, thus providing a foundation for explicitly capturing data provenance in active conceptual modeling.

We propose to capture data provenance requirements of the users using provenance annotations. Rather than creating new constructs in a conceptual model, we augment a conventional conceptual model such as an Entity-Relationship model with annotations that represent the provenance information associated with data captured in the core data model without changing the core data model. Using an example in the domain of homeland security, we illustrate our annotation-based approach to active conceptual modeling with provenance annotations.

Provenance knowledge is indispensable in various applications. In particular, it is critical in the domain of homeland security, where given some intelligence information, provenance regarding the information such as how and when it was collected by whom is required to evaluate the quality of the information and avoid false intelligence. Consider the homeland security application described in the core data model in Fig. 9. Nowadays, organizations and ordinary citizens are called upon to report suspicious activities that might indicate terrorist threats. As a successful example, the Pan American Flight School reported that Zacarias Moussaoui seemed "extremely interested in the operation of the plane's doors and control panel", which leads to Moussaoui's subsequent arrest prior to 9/11 [20]. However, intelligence information such as terror reports may be false, out-of-date, and from unreliable sources, which calls for a data model with provenance semantics. Hence, we record events such as the creation and transformations of a terror report at the conceptual level using provenance annotations and by doing so, make the conceptual schema "active".

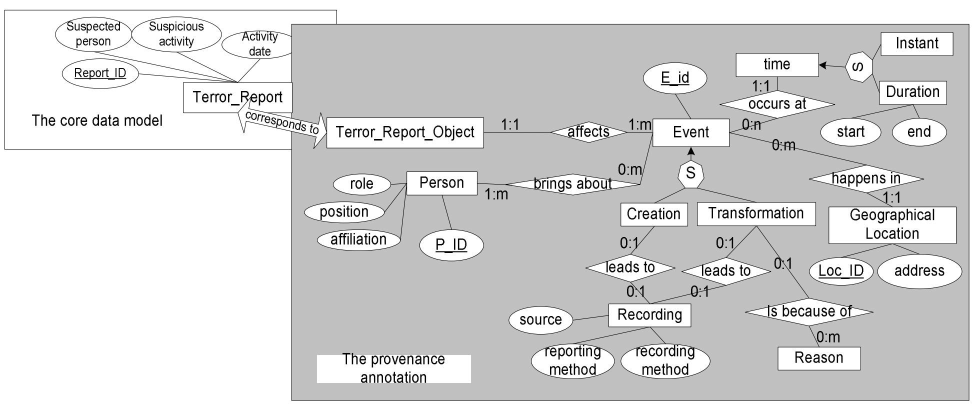

Fig. 9. A Sample Conceptual Model with Provenance Annotations

The annotation shown in Fig. 9 is developed based on our W7 model and is represented in the USM. The entity "Terror_Report_Object" in the annotation corresponds to the entity "Terror_Report" defined in the core data model and represents instances of "Terror_report" (see Fig.9). "Who", in this case, describes persons who play different roles in creating the data such as the agent who receives the report as well as the agent who records it. The actions leading up to the creation or transformations are normally recordings. We capture the source (i.e. the informant) of the report and instantiate the attribute "method" in the W7 model into "reporting method" and "recording method". When a transformation/update event happens to the data, we capture reasons "why" the information is updated in addition to "how", "who", "where" and "when". Via annotations, we enable a supplementary layer of abstraction that describes the data provenance semantics and naturally extends the semantics of a conventional data model.

Recording data provenance is critical in the domain of homeland security. It supports the following activities:

- *Information reliability*: To enforce national security, the right people must collect the right information from the right sources to identify real security threats. In our example, capturing who reported the threat via what reporting method assists in evaluating data reliability. Provenance regarding how the report was recorded or updated by who also helps ensure that the information can be trusted.

- *Information currency*: Some types of intelligence information may have a very short shelf-life. As an example, after Saddam Hussein fled Baghdad, information about him being spotted at a specific location changed six to eight times a day [20]. Capturing provenance such as: when the report of his being spotted was created and updated could be used to avoid being misled by old or out-of-date information.

- *Pattern recognition*: Provenance could help discover certain out-of-the-norm behavior patterns, which would helpful for predicting and preventing potential terrorist threats. As an example, a sudden increase in the number of threat reports from people in the same region within a short time period may indicate a terrorist plot. Also, the "who" part of our provenance could help us identify key reliable sources and forestall unreliable sources from feeding false intelligence.

4 Conclusion and Future Research

In conclusion, our research focus 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 aspects in detail. Our W7 model is inspired by our investigations of theoretical works such as Bunge's ontology as well as our observations of provenance issues in application domains including biology, new product design and development, digital archiving and homeland security. It is a generic model of data provenance and is intended to be easily adaptable to represent domain or application specific provenance requirements in conceptual modeling. Using homeland security as an example application, we apply our W7 model to support active conceptual modeling with provenance annotations. We are investigating how provenance might be automatically identified and recorded. In the future, we will investigate the effectiveness of our approach by applying it to different application domains.

References

[1] D. Pearson, "The Grid: Requirements for Establishing the Provenance of Derived Data," presented at Workshop on Data Derivation and Provenance, Chicago, Illinois, 2002.

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

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

[4] P. Buneman, S. Khanna, and C. T. Wang, "Why and Where: A Characterization of Data Provenance," in *Lecture Notes in Computer Science*, vol. 1973, Jan Van den Bussche, Ed.: Springer, 2001.

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

[6] M. Bunge, *Treatise on Basic Philosophy: Vol. 4: Ontology II: A World of Systems*. Boston, MA: Reidel, 1979.

[7] Y. Wand and R. Weber, "On the deep structure of information systems," *Information Systems Journal*, vol. 5, 1995.

[8] L. Davis, *Theory of action*. Englewood Cliffs, NJ: Prentice-Hall, 1979.

[9] R. Chisholm, "The Agent as Cause," in *Action Theory*, M. Brand and D. Walton, Eds. Dordrecht, Holland/Boston, USA: D. Reidel, 1975.

[10] S. Ram, "Intelligent Database Design Using the Unifying Semantic Model," *information and Management*, vol. 29, pp. 191-206, 1995.

[11] P. P. Chen, "The entity-relationship model - toward a unified view of data," *ACM Trans. Database Syst.*, vol. 1, pp. 9-36, 1976.

[12] R. T. Snodgrass and I. Ahn, " Temporal Databases," *Computer*, vol. 19, pp. 35-42, 1986.

[13] V. Khatri, S. Ram, and R. Snodgrass, "Augmenting a Conceptual Model with Geospatio-temporal Annotations," *IEEE Trans. Knowledge and Data Eng.*, vol. 16, pp. 1324-1338, 2004.

[14] M. Koubarakis and D. Plexousakis, "A formal framework for business process modeling and design," *Information Systems*, vol. 27, pp. 299-319, 2002.

[15] B. Curtis, M. Kellner, and J. Over, "Process modeling," *Communication of ACM*, vol. 35, pp. 75-90, 1992.

[16] M. Georgeff, B. Pell, M. Pollack, M. Tambe, and M. Wooldridge, "The Belief-Desire-Intention Model of Agency," presented at the 5th International Workshop on Intelligent Agent: Agent Theories, Architectures, and Languages, Paris, France, 1999.

[17] K. Konolige and M. E. Pollack, "A Representationalist Theory of Intention," presented at the Thirteenth International Joint Conference on Artificial Intelligence (IJCAI-93), 1993.

[18] G. Allen and S. March, "Modeling Temporal Dynamics for Business Systems," *Journal of Database Management*, vol. 14, pp. 21-36, 2003.

[19] P. P. Chen, B. Thalheim, and L. Wong, "Future direction of conceptual modeling," in *Conceptual Modeling: Current Issues and Future Directions, Lecture Notes in Computer Science, No. 1565*, P. P. Chen, Ed. Berlin: Springer-Verlag, 1998, pp. 294-308.

[20] L. English, "Information Quality: Critical Ingredient for National Security," *Journal of Database Management*, vol. 16, pp. 18-32, 2005.