Journal of Database Management
Volume 28 • Issue 1 • January-March 2017

Improving the Domain Independence of Data Provenance Ontologies:
A Demonstration Using Conceptual Graphs and the W7 Model

Jun Liu, College of Business and Information Systems, Dakota State University, Madison, SD, USA
Sudha Ram, Eller College of Management, University of Arizona, Tucson, AZ, USA

ABSTRACT

Provenance is becoming increasingly important as more and more people are using data that they themselves did not generate. In the last decade, significant efforts have been directed toward developing generic, shared data provenance ontologies that support the interoperability of provenance across systems. An issue that is impeding the use of such provenance ontologies is that a generic provenance ontology, no matter how complete it is, is insufficient for capturing the diverse, complex provenance requirements in different domains. In this paper, the authors propose a novel approach to adapting and extending the W7 model, a well-known generic ontology of data provenance. Relying on various knowledge expansion mechanisms provided by the Conceptual Graph formalism, the authors' approach enables us to develop domain ontologies of provenance in a disciplined yet flexible way.

KEYWORDS

Bunge-Wand-Weber Ontology, Conceptual Graph, Data Provenance, Ontology

INTRODUCTION

Since the start of the new millennium, people have been sharing data in an unprecedented scale and richness. In scientific domains such as biology and chemistry, the trend of "big science" signified by large scale collaborative projects such as the iPlant Collaborative (http://www.iplantcollaborative.org) demands the sharing of data over organizational boundaries and even across disciplines. For businesses, Big Data is a key component in competition, growth and innovation, and much of Big Data originates outside of the company that is absorbing it. With the large-scale proliferation and sharing of data, questions such as "Where did this data come from?", "Who else is using this data?", and "Why is this piece of data here?" are becoming increasingly common (Ram & Liu, 2012). Data provenance, often referred to as "origin", "lineage" "history", or "pedigree" of data, contains the answers to the questions. When data travel beyond the specific setting in which they are generated, it is imperative that the provenance of the data needs to be captured to ensure the trustworthiness of the data.

In the last decade, significant research has been conducted to standardize the semantics of data provenance and develop a shared provenance ontology that allows unambiguous interpretation of provenance, supports interoperability of data provenance between systems, and improves the usability of data provenance by enabling richer queries. One of the earliest efforts in standardizing provenance semantics is the development of the W7 model (Ram & Liu, 2007). The W7 model conceptualizes provenance as consisting seven Ws including *w*hat, *w*hen, *w*here, ho*w*, *w*ho, *w*hich and *w*hy, and it has been adopted in research such as (Lupelli et al., 2015; Narock, Yoon, & March, 2014; Prat &

DOI: 10.4018/JDM.2017010104

Madnick, 2008), etc. Another widely used provenance model is the Open Provenance Model (OPM) (Moreau et al., 2011). The OPM represents the provenance of objects by an annotated causality graph. A causality graph captures the causal dependencies between three types of nodes: artifacts, processes and agents. Other well-known provenance ontologies include Provenance Vocabulary (Hartig & Zhao, 2010) and PROV-DM model (Belhajjame et al., 2012). These generic provenance ontologies are designed to be domain and architecture independent. They support a digital representation of provenance for any "thing" so that provenance can be exchanged between systems by means of a compatibility layer based on a shared provenance model (Moreau et al., 2011).

The generic provenance ontologies such as the W7 model and OPM describe the semantics of data provenance and are independent from a specific task or domain. However, users often have domain-specific and application-specific provenance requirements. Which provenance information is required and at what level of detail significantly vary by discipline, data type, purpose, and project. A software approach that requires the meaning and format of provenance to be standardized is thought by some researchers to be unlikely to meet the needs of various multi-scale research communities (Myers et al., 2003). Indeed, generic ontologies such the W7 and the OPM, even though they are intended to be general and comprehensive enough to cover a broad range of provenance-related vocabularies, are still insufficient for capturing provenance for all types of data in a specific domain without being substantially extended. For instance, the provenance of data on a plant gene may include not only the experimental process by which the data was derived, but also information about what plant part and sample was used in the experiment and how the sample was manipulated. Moreover, a detailed description of the plant, such as its morphology, its Phenotypic information, its ecological environment and development stage is also critical provenance information. Hence, a significant gap exists between diverse, complex, and domain-specific provenance requirements and the generic provenance models such as the W7 model and the OPM. No matter how complete they are, these generic provenance models are far from being sufficient to capture provenance for all types of data in all domains without being adapted and extended.

The goal of our research is to develop a novel approach to bridging the aforementioned gap. Focusing on the W7 model, a well-known generic provenance ontology, we propose a conceptual graph-based approach that enables us to easily adapt and extend the generic W7 model to develop domain ontologies that capture domain-specific provenance requirements. We illustrate the applicability of our approach to different domains by developing domain ontologies of provenance for the iPlant project and for the domain of new product design and development.

THE W7 MODEL

The theoretic underpinning of the W7 Model lies in Bunge's ontology (Bunge, 1977) that includes constructs related to events and history of things. In Bunge's view, the world is made up of things that have properties. Wand and Weber (1990, 1993, 1995) applied Bunge's ontology to modeling objects in information systems and provided a number of examples. For instance, a customer order as a thing have properties including order number, customer number, quality ordered, quality supplied, price, date, and a "processed" flag. Data are also things. A data object (e.g., an image, a tuple in a database, etc.) has a large variety of properties including its content, format, ownership, storage location, access rights, and a number of flags such as *isArchived*, *isPublished*, and *isAnnotated*. Moreover, according to Bunge, "all things are in flux" (Bunge, 1977). A state of a thing comprises a set of property values of the thing at a specific point in space and time, and an event occurs to the thing when the thing changes its state. A sequence of events (or changes of states) that occur to a thing during its lifetime manifests the *history* of the thing. Provenance is often referred to as the pedigree or history of data. Hence, in the W7 model, data provenance is conceptualized as consisting of various events (i.e., state changes) that happens during the lifetime of the data from its creation to destruction (Ram & Liu, 2007). The centerpiece of the W7 model is *"what",* which represents the events. According to Bunge