

PROMS: A System for Harvesting and Managing Data Provenance

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

Data Provenance refers to the lineage of data in terms of various key events that occur over the course of its lifecycle and other related information associated with its creation, processing, and archiving. Provenance enables users to share, discover, and reuse data, thus streamlining collaborative activities, reducing the possibility of repeating dead ends, and facilitating learning. To capture and represent the semantics of data provenance, we have developed an ontological model of provenance called the W7 model that conceptualizes data provenance as a combination of seven interconnected elements including “what”, “where”, “when”, “how”, “who”, “which”, and “why”. The element “what” describes *events* that affect data, including creation, use, storage, transformation, and archiving of data. The other elements are connected to “what” and describe different aspects of the events. The element “when” records the event time, and “where” captures the event location. “How” documents actions leading up to the events. “Who” refers to people or organizations in various relations to the events. “Which” describes the instruments or software applications used in the events. Finally, “why” is defined as the decision rationale behind the creation, transformation, or other events that affect data.

Using new product design and development as the application domain, we have designed and developed a **PRO**venance **M**anagement **S**ystem (PROMS) for harvesting and using data provenance. The architecture of our system is shown in Fig. 1. The Provenance Capture Module records user-provided provenance using templates based on the W7 model. Utilizing natural language processing functions provided by GATE, it also supports semi-automatic harvesting of provenance by extracting information stored in electronic notebooks with little manual intervention. The harvested provenance is recorded in the data provenance knowledge base implemented using a relational database. PROMS retrieves data and its provenance upon request from a user via a web-based graphical user interface. The Provenance Navigation Module allows the user to view and navigate provenance in a convenient way. It enables the user to query data via the provenance. For instance, it allows the user to formulate queries such as “retrieve all the electro-mechanical property values (for a specific alloy) created by Josh Cohn in July, 2006”.

PROMS has been used to harvest and store data provenance in the context of WIKI’s. The Wikipedia contains more than 1,400,000 articles and most of these articles undergo frequent changes. PROMS helps monitor the provenance of Wikipedia pages to track changes made to a page. It also tracks who made the changes, how and why the changes were made and at what time. It uses this provenance to generate warnings about potential vandalism threats to Wikipedia pages.

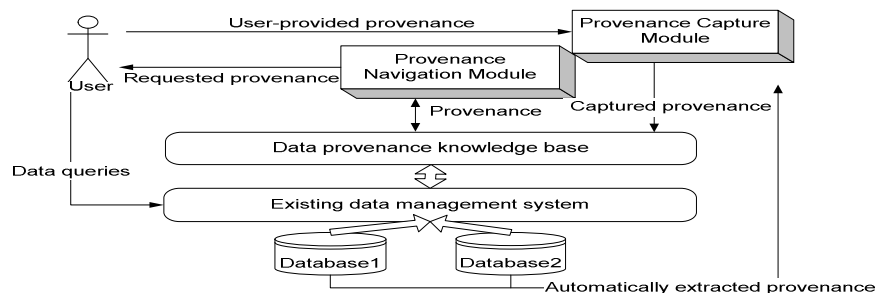


Fig. 1. Architecture of PROMS