

# PROVISIA: Visualization of Data Provenance

Sudha Ram, Jun Liu, Arjhun Thiagarajan

Department of Management Information Systems, Eller College of Management,  
The University of Arizona, Tucson, AZ 85721, USA

Data provenance refers to the source and processing history of data. We have clearly defined the semantics of provenance using the W7 model in our previous research (Ram and Liu 2007). We have designed and developed a **PRO**venance **VI**sualization **S**ystem for the **W**ikiped**IA** (PROVISIA) to harvest and visualize the provenance of Wikipedia articles. We use a subset of the W7 model by tracking the *what*, *how*, *who*, *when* and *why* of Wikipedia articles. In the Wikipedia context, *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 may be sentence insertion/update/deletion, link insertion/update/deletion, reference insertion/update/deletion, reverts, etc. These are actions made by editors that may lead to the modification of a page. *Who* represents the editors of a Wikipedia page. *When* refers to the time an event occurs. *Why*, i.e., justification for a change, is recorded in the “comment” field in Wikipedia.

The architecture of PROVISIA is shown in Fig. 1. The Provenance Capture Module automatically harvests provenance from the Wikipedia. The harvested provenance is recorded in the provenance knowledge base which is implemented using a relational database. The Provenance Navigation Module consists of two major components: a provenance browser and a trend analyzer. The provenance browser visualizes the provenance of a Wikipedia page by tracking the various events that affected the page as well as details about these events. It allows the user to navigate to other Wikipedia pages based on data provenance. For instance, the visualization of the page “genome” indicates that an editor named *AdamRetchless* inserted 8 sentences on Nov.30<sup>th</sup>, 2002. The provenance browser then allows a user to navigate to other pages edited by this person or even other pages created around the same time period. The trend analyzer shows how Wikipedia pages evolve over time. It allows visualization of the trail of changes for a specific page, in terms of the number of editors or the number of inserted sentences during its lifetime. It can also be used to compare the evolution paths of multiple Wikipedia pages based on the number of editors, the number of sentence insertions, and other provenance-related information.

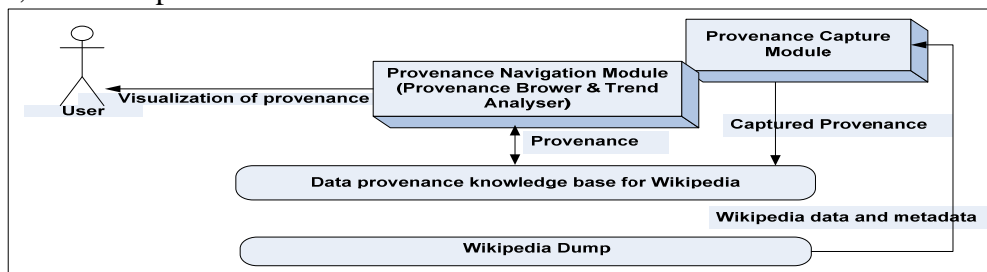


Figure 1: Architecture of PROVISIA

## Reference:

- S. Ram and J. Liu (2007), "Understanding the Semantics of Data Provenance to Support Active Conceptual Modeling," in *Lecture Notes in Computer Science*, vol. 4521, pp 17-29, Springer-Verlag.
- J. Liu and S. Ram (2009), "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.