A Conceptual-Model-Based Computational Alembic for a Web of Knowledge*

David W. Embley[†], Stephen W. Liddle[‡], Deryle Lonsdale[**], George Nagy[††], Yuri Tijerino[‡‡], Robert Clawson[†], Jordan Crabtree[†], Yihong Ding[†], Piyushee Jha[††], Zonghui Lian[†], Stephen Lynn[†], Raghav K. Padmanabhan[††], Jeff Peters[†], Cui Tao[†], Robby Watts[†], Charla Woodbury[†], and Andrew Zitzelberger[†]

[†]Department of Computer Science
[‡]Department of Information Systems
[**]Department of Linguistics and English Language
Brigham Young University, Provo, Utah, 84602

[††]Department of Electrical, Computer, and Systems Engineering
Rensselaer Polytechnic Institute, Troy, New York, 12180

[‡‡]Department of Applied Informatics
Kwansei Gakuin University, Kobe-Sanda, Japan

The current web is a web of linked pages. Frustrated users search for facts by guessing which keywords or keyword phrases might lead them to pages where they can find facts. Can we make it possible for users to search directly for facts embedded in web pages? Instead of a web of human-readable pages containing machine-inaccessible facts, can the web be a web of machine-accessible facts superimposed over a web of human-readable pages? Ultimately, can the web be a WoK (a Web of Knowledge) that can provide direct answers to factual questions and support these answers by referencing and highlighting relevant base facts embedded in source pages?

Answers to these questions call for distilling knowledge from the web's wealth of heterogeneous digital data. But how? Our computational alembic must turn raw symbols contained in web pages into knowledge and make this knowledge accessible via the web. Further, the computational alembic must successfully break down barriers to WoK creation and usage. Currently, several barriers are too high: the barrier of creating machine-readable content (i.e., of creating populated ontologies); the barrier of annotating human-readable, web-page content with respect to ontologies; and the barrier of learning to query machine-readable content. Thus, WoK creation and usage faces three main technical challenges: (1) automatic or sufficiently easy creation of ontologies, (2) automatic or sufficiently easy annotation of web pages with respect to these ontologies, and (3) simple, but accurate, query specification, usable without specialized training. Meeting these basic challenges can simplify WoK content creation and access to the point that the vision of a web of knowledge can become a reality.

* This material is based upon work supported by the National Science Foundation under grant no. 0414644 and grant no. 0414854.

Conceptual modeling plays a foundational role in creating a computational alembic to actualize these ideas. An ontology is a conceptualization of a real-world domain in terms of objects, relationships, generalizations, specializations, and aggregations with constraints over these conceptualizations. Indeed an ontology can be thought of as a conceptual model grounded formally in a logic system. Automatic and semi-automatic ontology generation from data-rich, semi-structured web pages is akin to reverse engineering structured data into conceptual models. Automatic and semi-automatic annotation of web pages can proceed bottom-up—can occur as a by-product of ontology generation via reverse engineering. Or annotation can proceed top-down—can come from extraction ontologies in which instance recognizers attached to conceptual object sets and relationship sets extract data on web pages with respect to conceptual models comprised of these object and relationship sets. For query processing, conceptual models grounded in description logics form a template to which free-form queries can be matched to yield formal queries that can be processed by standard query engines.

Conceptual-modeling research can help actualize the WoK vision by:

- providing an answer to the question about how to turn syntactic symbols into semantic knowledge;
- showing how to establish a workbench with toolkits to convert heterogeneous digital data into knowledge under the auspices of an ontology;
- exploring the synergistic interplay among ontology, epistemology, and logic for the advancement of knowledge to provide new ways to think computationally about what knowledge is and how knowledge is acquired; and
- providing a basis for untrained users to query and reason over fact-filled ontologies.

Our WoK Demo[1] illustrates the foundational presence of conceptual modeling in creating a WoK. Specifically, it shows how to create ontologies and annotate pages with respect to these ontologies, and it shows how to query and display annotated content. Ontology creation and usage in HTML-page annotation can be automatic, semi-automatic, or human specified in a user-friendly mode of interaction. User confirmation and correction is always possible, so that the user has the last say, but in many cases, automatically created ontologies and automatically annotated web pages are immediately usable within the WoK. Query specification can range from free-form conjunctive queries to a formal query language. Applications include scientific data (e.g., genes), geopolitical data (e.g., Canadian demographic statistics), family-history data (e.g., genealogical information), and commodity sales and services (e.g., car sales and apartment rentals).

[1] See www.deg.byu.edu and www.tango.byu.edu.