An Ontology-Driven Approach for Harmonizing and Integrating Environmental Information

Yu-Chi Chu[1], Su-Mei Huang[2], Chin-Cheng Lien[2], Chen-Chau Yang[3]

Abstract

The paper describes an ontology-driven methodology to be used for the integration of heterogeneous environmental information that may be distributed in government agencies. The methodology adopts a series of systematic processes to construct a localized domain ontology which may be used as the driven force or guidelines for environmental information integration. As an example application, the Taiwan EPA's Environmental Data Repository Project (EDR) is used to demonstrate the feasibility and applicability of the proposed methodology.

1. Introduction

For making sensible, justifiable, and legally correct decisions to protect our earth, both government agencies and private sectors need detailed information regarding the current state of the environment and ongoing developments. Currently it is very difficult to share environmental data since the information typically resides on geographically disparate and heterogeneous databases (systems). These systems often do not facilitate access by secondary users and frustrate attempts to draw data together to form a more comprehensive understanding of environmental conditions and actions. Therefore, there is a major demand for appropriate systems and adequate tools to provide integrated information for managing the issues of environmental protection.

The term ontology was originally used in philosophy to describe a theory of "being and existence." In the field of artificial intelligence it was adopted to describe knowledge models that provide definitions of vocabulary used to describe a certain domain. Hence, the use of ontologies for the explication of implicit and hidden knowledge is a possible approach to overcome the problem of semantic heterogeneity. During the past years, ontologies have gained importance in many areas of applications, like interoperability, design of intelligent systems, and lately for knowledge management and electronic commerce. In addition, ontologies are being used for a more precise specification of the semantic information content of the underlying data as well as of the user's information needs. Some domain specific ontologies are also being viewed as vehicles of capturing semantic information content independent of the underlying syntactic and structural representation of the data.

The following section describes the proposed methodology with an example application. The systematic processes for ontology extraction, alignment, as well as integration and enrichment will be illustrated in detail. Section 3 offers some concluding remarks.

2. Methodology

For the purpose of integrating environmental information, we want ontologies to be practical and useful artefacts. This means that the effort required to construct new ontologies must be minimized and overall effort required to construct an ontology must be amortized over multiple uses and users. We take several

[1] Department of Environmental Monitoring and Information Management,
Environmental Protection Administration, Taiwan, R.O.C.
[2] Department of Computer Science, SooChow University, Taiwan, R.O.C.
[3] College of Electrical and Computer Engineering, St. John's University, Taiwan, R.O.C.

steps towards attaining that goal. First, we provide an environment that assists users to build prototype ontologies by extracting knowledge from the existing information sources such as database schemas or DTD for XML documents. Secondly, the terms of concepts in the built prototype ontologies will align with existing common ontologies and some of the domain specific thesauri such as GEMET. The alignment may provide a semantic consistency among different ontologies. For example, the names of concepts facility-names, plant-names, site-names, and contaminant-sites will be aligned to a term in GEMET called "*site*." After this process, the ontologies will be called aligned ontologies. Thirdly, the aligned ontologies will be merged and combined to form an integrated ontology.

2.1. Systematic processes

The core component of our approach is the integrated ontology, which is dedicated to the domain of environmental protection. Figure 2.1 depicts the process of the proposed method. We explain the main steps of the method we propose for building an integrated ontology from existing knowledge sources such as the environmental databases owned by local governments and EPA at the central government level.

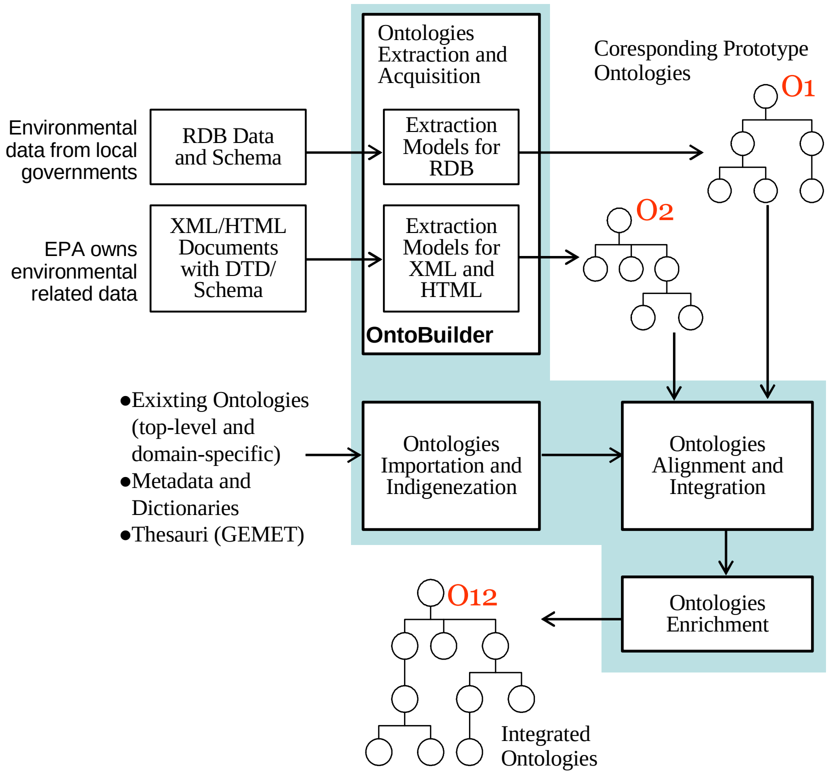


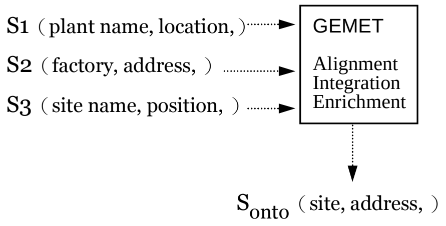
Figure 2.1: The process of the integration of environmental domain ontologies

1. *Selection of underlying information sources, standards, laws and regulations, classifications, etc.* In this step, we select the sources of information that we consider relevant to the target domain, for example, environmental protection, or e-Business for a certain industry. They usually provide taxonomy of concepts and terminologies used in the domain.
2. *Ontology extraction and acquisition.* This step performs the process of knowledge acquisition from the sources of information previously selected and adapts them to form a prototype ontology for each knowledge source. This activity can be performed using tools such as Ontolingula and WebODE. We have been implementing a simple tool set for this activity called OntoBuilder, which focuses on the relational database schema and XML documents. In addition, we also employ the procedure of "protocol analysis" with domain experts. This procedure consists of asking users to describe various types of domain applications, the data used in such applications, and the terms used in their field. Several such sessions of protocol analysis may result in a standard set of terms and inter-relationships among these terms.
3. *Ontology importation and indigenization.* In this step, we import ontologies that are existing ontologies in target domains but might usually be used in other countries or areas. For example, in the domain of environmental protection, the OECD[4] countries commonly adopt a classification, which can be seen as an ontology for identification of environmental industry. Most countries in European area use GEMET, which is a thesaurus for unifying the terminology in environmental use. Those ontologies are helpful to construct new ontologies. However, when importing those ontologies, we need to tailor them to fit the feature in the area or country where ontologies will be applied. For example, if we want to adopt the OECD classification for use in Taiwan, we have to delete some of the items in the classification such as forest management, reforestation because they are not under the jurisdiction of the Environmental Protection Administration in Taiwan.
4. *Ontology alignment and integration.* This activity consists of two processes. First, align the prototype ontologies with imported ontologies or upper level ontologies. This process may include adjusting the name or terminology in the prototype ontologies and making them consistent with each other. Secondly, we combine and merge the prototype ontologies to form an integrated ontology.
5. *Ontology enrichment.* Traditionally, most of the ontologies merely represent taxonomy of concepts, where others may just include some attributes for them. In this activity, we will enrich the integrated ontologies with extra information where available.

2.2. Example application

We use the Environmental Data Repository Project (EDR)[5] of the Taiwan Environmental Protection Administration (TEPA) as a pragmatic example to illustrate the process of implementation. EDR is an integrated data warehouse system that provides a single point of access to data extracted from several major TEPA databases, including the Air Pollution Control System, the Water Permit Database, the Hazardous Waste Control System, and the Toxic Release Database. We construct the integrated ontology by extracting the domain knowledge from some of the information sources and aligning the concepts in the ontology with the laws and regulations of Taiwan EPA. The ontology becomes a major component to drive the integration process of the EDR project. Based on the implementation of a prototype system of EDR project, we may justify the applicability of our approach.

[4] Organization for Economic Co-operation and Development
[5] http://edb.epa.gov.tw



(a) An example of term integration

(b) EDR homepage

Figure 2.2: Example application

As shown in Figure 2.2(a), each information source might maintain the data regarding the potential pollution sites. However, they do not use the same database schema and cause a conflict between the terms that is used to identify the sites. Ontology can assist us in overcoming this conflict, and develop a consistent view through which information can be integrated. Figure 2.2(b) shows the EDR homepage, which we have implemented based on the proposed methodology.

3. Conclusions

This paper presents a methodology based on ontology-driven related technologies to integrate environmental information. It is shown, to some degree, that ontology can provide assistance in solving the heterogeneous problems among diverse information sources. Our approach may serve as an infrastructure component for integrating environmental data with known, but differing, collections of data. As for future work, recent advancements including Web services, Ajax and knowledge management might be integrated with the proposed approach to design and implement a more sophisticated and practical system.

Bibliography

Chou, K.W. (2007): An ontology-based knowledge management system for flow and water quality modeling, *Advances in Engineering Software*, 38: 172-181.

Chu, Y.C. (2001): *Integrating Heterogeneous Information Sources through Ontology-Driven Model and Data Quality Analysis*, Ph.D. Dissertation, National Taiwan University of Science and Technology.

Chu, Y.C. and S.M. Huang, C.C. Yang (2005): Ontology-based government information integration, *Proceedings of National Computer Symposium* (NCS 2005), Tainan, Taiwan (in Chinese)

Europen Environment Agency, GEMET web pages, http://www.eionet.europa.eu/gemet

Purvis, M. et al. (2003): A multi-agent system for the integration of distributed environmental information, *Environmental Modeling & Software*, 18: 565-572.