Towards Ontological Modelling of Historical Documents

Vanesa Mirzaee*, Lee Iverson*, Babak Hamidzadeh†
*Dept. of ECE, University of British Columbia, Vancouver BC
†Boeing Corporation, Seattle WA
vanesam@ece.ubc.ca, leei@ece.ubc.ca, babak.hamidzadeh@boeing.com

Abstract. In this presentation we describe a methodology we have adopted for coding the semantic structure of a historical document and the resulting semantic model. To do this, we adapted currently available methodologies for ontology engineering to the context of semantic document coding. Using Protégé-2000 we then used this methodology to develop a formal ontological model and finally to encode a historical document covering the evolution of the constitution of modern Iran. The resulting semantic model was then evaluated using Protégé-2000 by direct reference back to the set of competency questions and motivating scenarios used to develop the model. Our implementation was successful in answering these competency questions as well as in providing support for the selected scenarios. The implementation and the evaluation results are presented along with our proposed future work.

1. Introduction

Until recently, it has been assumed that the main advantage of electronic formats over printed matter is the convenience of being able to find the material without having to physically obtain it from a library or other repository. However, once we have this information in a digital format, it is unclear as to how the user might interact with it besides being able to print it and/or read it. We believe that digital documents have the potential to provide us with more functionality than traditional printed matter does.

In particular, we have chosen to use an ontological approach to code documents, thus allowing a community to (1) *share and reuse* their knowledge, (2) *capture the semantics* implicit in the documents, and (3) allow *computational manipulation* of the acquired knowledge. This manipulation of the document's meaning would allow automatic reasoning beyond the simple queries and keyword search provided by current information retrieval methods.

Consider the case of historical document archives. A wealth of historical information is now available in digital form through different resources such as digital libraries. These digital media usually integrate meta-data that provides some information about its content. These collections provide the ability to retrieve the best-matched documents for any search request.

Instead of basing these searches on keywords, it would be ideal for electronic historical archives to provide methods and techniques of posing and resolving historical questions and then providing access to the sources of the claims used to resolve them. For example a historian will want to query relationships between characters, institutions, events, and locations of these events. Similarly, it is vital to capture how these relationships change over time.

To ground our work, we have chosen to examine a particular historical document, "*The History of the Iranian Constitution*," describing the evolution of the modern state of Iran over a 50-year period. We suggest that the methodology adopted will apply equally well to a similar class of documents and that our ontological model will generalize to any document covering similar semantic ground.

Thus, we have presented an approach to representing the knowledge within a historical document that allows such sharing, reuse and automatic reasoning by capturing its semantic content using ontologies. Next we will present the methodology used to build an ontological model to represent the knowledge found in a historical document. We show this with reference to our example ontology developed using Protégé-2000.

2. Methodology and Implementation

Building a well-developed, usable, and sharable ontology represents a significant challenge. There is great diversity in the way ontologies are designed as well as in the way they try to represent the world.

A range of methods and techniques have been reported in the literature regarding ontology building methodologies. However, there is ongoing argument within the ontology community about the best method to build them [10; 6; 2]. Given that the knowledge to be captured usually depends critically on a combination of the domain and the applications being designed to exploit this knowledge [11], it is no surprise that these methodologies are primarily inspired by enterprise modeling or software engineering. For our purposes, it

was important to scale them down and adapted them to facilitate document coding. We divide the ontology building process into the following stages:

1. Identifying the purpose, scope, and users
2. Domain analysis and knowledge acquisition
3. Building a conceptual (informal) ontology model
4. Formalization
5. Evaluation

In our method, we focus on an evolving prototype of the ontology. In this model, at each stage, it is possible to go back to any previous stage of the development process, in order to satisfy emerging requirements. Figure 1 illustrates how these steps are related, and in what order they can be performed to complete the ontology building process. We make every effort to maintain the following criteria for each and every stage of the development process: Clarity; Coherence; Extensibility; Minimal encoding bias; and Minimal ontological commitment [5].

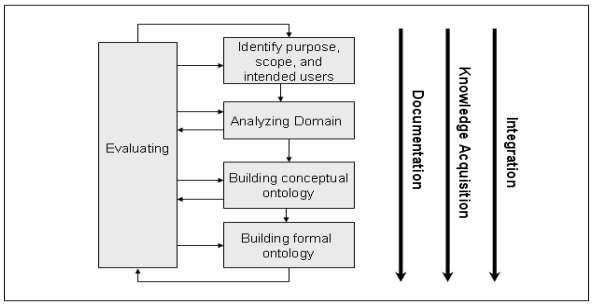

Figure 1 Our ontology development process. Integration, Knowledge Acquisition, and Documentation are carried out throughout the entire development process.

2.1. Identify purpose, scope, and intended users

The main purpose for building this ontology is to capture the semantics of a historical document, especially the temporal and dynamic aspects of the concepts and their interrelations. To promote sharing, reuse and enable better integration with existing knowledge sources we relied heavily on the consensual terminology available in general ontologies. The selected audience included both the general public, historians and biographers who might directly access the semantic models.

The requirements gathered were formulated as a set of competency questions and motivating scenarios that our model must answer and provide support for. A few of these competency questions are presented in Table 1.

Table 1 Some Competency Questions

1. Who was *Person* P?
2. In What *Events* were *Person* P involved?
3. What *Positions* did *Person* P held?
4. When did *Person* P held these *Positions*?
5. Who was taking over *Person* P's *Position* Po?
6. What was the governmental position hierarchy at the time *Person* P holds *Position* PO?

2.2. Domain analysis and knowledge acquisition

Using the competency questions and scenarios we then produce a set of concepts and terms covering the full range of information that the ontology must characterize to satisfy the requirements identified previously. In this phase, we use knowledge acquisition techniques such as brainstorming, in conjunction with informal analysis of the text to gather all potential relevant terms into a glossary [3].

This glossary includes the terms, their definition or description, and may include additional information, such as examples that help understanding these definitions. In order to provide definitions for the terms, we consulted dictionaries such as the Merriam Webster Dictionary and the Oxford Dictionary as well as general purpose ontologies such as SUMO [15], Ontolingua [12], and WordNet [16].

2.3. Building an informal ontological model

Once we have a relatively complete glossary of terms, we identify concepts, relations within the concepts, and their attributes. We use the guideline provided in [11] to do so. The results are stored in document tables called the Concept Dictionaries [3]. At this stage, the concepts are structured into naturally occurring groups using a combination of the approaches introduced in [11] and [7]. We categorized our concepts into five concept dictionaries relating to *people*, *places*, *events*, *documents*, and *time*. Each of these categories holds the concepts that are most related

For the next step, we use the previously generated concept dictionaries, along with the motivating scenarios and a middle-out approach to develop our graphical conceptual ontology model. Our conceptual model not only represents the concept taxonomy but also the other (non-taxonomic) relations that hold amongst the concepts within our domain.

Throughout the ontology building stages, we queried existing ontology libraries, such as Ontolingua [12], DAML [4], and SUMO [15] to search for similar or related terms and relations that might be useful. This was done in order to speed up the development process as well as to gain a better insight of how to build a particular area or set of concepts within our ontology. Thus we were able to build our ontology on a well-grounded structure. In particular, the time concepts were derived from general time ontologies [12; 15; 17] and the temporal relations in TELOS [10]. Events were based on Sowa's thematic roles or case relations [14]. Places were defined using standard ontologies for geographic information representation and categorization [1; 8].

Figure 2 shows the top-level concept hierarchy in our domain. We identified five central concepts within our ontology: AGENT, PLACE, EVENT, DOCUMENT, and

TIME. Every other concept in this domain is defined around these primitive concepts.

An important characteristic of the proposed ontological model is its capability to represent temporally dynamic concepts. This is of particular importance for historical data since the concepts and the relations between them change and evolve through time. This is accomplished by associating a time interval with each relation, as was done in Telos [9]. Additionally, this model not only captures the relationships between the concepts but also demonstrates the interrelated hierarchal structure within them. An example of such hierarchical structures found within our document is the governmental position hierarchy. In this hierarchy, not only do the people that hold positions change but the structure itself evolves throughout time.

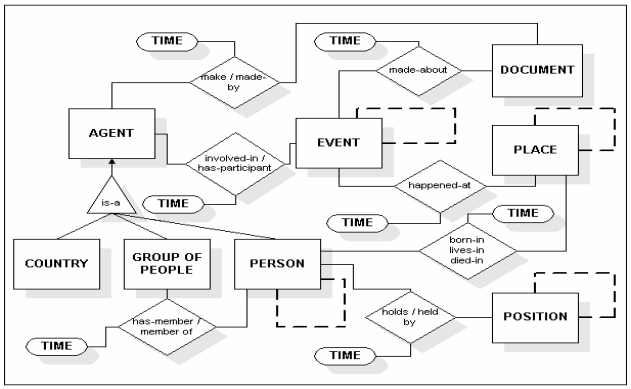

Figure 2 Overview of main concepts and relations in our history ontology.

2.4. Building a formal ontological model

The next step in our approach was to build a formal ontology based on the conceptual model. After a review of available ontology development environments, we selected Protégé-2000 [13] to formalize and instantiate our ontology. Our selection was based on the tool's expressiveness, flexibility, customizability, scalability, extensibility, and usability. Significantly, it also provided us with the facilities to test and evaluate our model.

Additionally, Protégé-2000 provides facilities to impose constraints to concepts and relations. While creating the ontology, it is necessary to make general assertions about fundamental concepts, and be able to later test and ensure these assertions hold across the entire knowledge-base. For example, in our ontology it was useful to assert common-sense constraints such as:

- All instances of Person have exactly one birth-date.
- A Person's birth-date must precede the death-date.
- Every Event in which a Person is involved, must take place between his or her birth-date and death-date.
- For any given time interval there can only be one person holding the position "king".

Figure 3 illustrates part of the governmental hierarchy that holds for a given time interval.

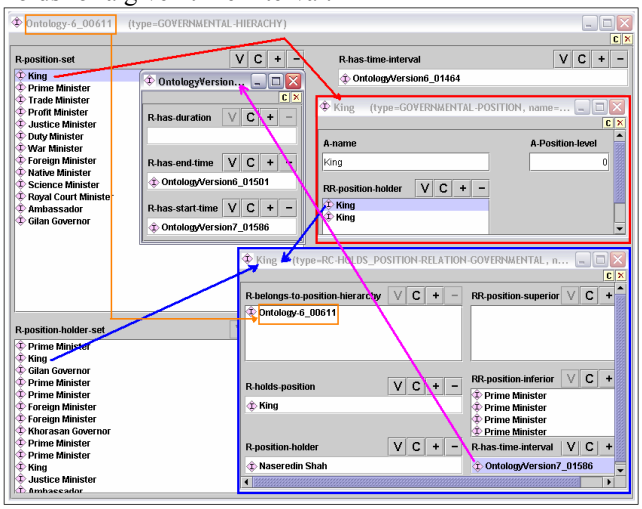

Figure 3 This figure illustrates an instance of the position hierarchy that holds for a specific time interval in our domain. In this example, the position "king" (top right window) was held by a person called "Mozafaredin Shah" (bottom right window) at that time interval. We can also see the inferiors and superiors of that person in that specific time period (within the bottom right window).

2.5. Evaluation

After designing, building, and formalizing our ontology using Protégé and enforcing constraints on attributes and relations, we used the knowledge acquisition forms provided in Protégé to instantiate our history ontology. Over seven hundred and fifty (750) instances were extracted from the history book and included in our ontology. Amongst these instances we find people, places, documents, and events.

In order to evaluate the correctness and completeness of the created ontology, we use the query and visualization facilities provided by Protégé-2000. We use the built-in query engine for the simple query searches and use additional query plug-in provided to create more sophisticated searches. We also use protégé-visualization plug-ins to browse the ontology and ensure its consistency. Visualization aids were particularly helpful when trying to understand hierarchical relations.

3. Conclusions and Future Work

In this work we confronted the limitations of traditional electronic documents. In particular we were interested in capturing the semantics of a historical document to allow for richer retrieval, reuse and manipulation of its embedded knowledge than is capable with standard text manipulation tools.

After adapting existing methodologies to the problem of text coding, we developed an ontology motivated by historical and biographical needs and the contents of the book "*The History of the Iranian Constitution*." Our implementation allowed us to get an overview of the general concepts in this book, relationships amongst these concepts and provided us with different methods for visualizing dynamic hierarchical structures of both governmental positions and geopolitical interdependencies. Additionally, this model captures the changes that these relations undergo through time (dynamicity). The temporal aspects of the knowledge we captured proved to be useful in making our representation more accurate and realistic.

Protégé-2000 provides us with facilities to define our concepts and relations with relative ease. However, in order to assign time attributes to our relations, we are required to either use the "RELATION" plug-in or reify our concepts and relations, which could prove to be difficult to naïve users. For future work, we intend to develop an interface using Protégé as the core modelling system which facilitates defining time-based relations by naïve users or experts in other domains.

References

[1] Alani, H., Jones, C. and Tudhope, D. (2000). "Ontology-Driven Geographical Information Retrieval." GIScience 2000..

[2] Beck, H. and Pinto, H. S. (2003). "Overview of Approach, Methodologies, Standards, and Tools for Ontologies." The Agricultural Ontology Service (UN FAO).

[3] Blazquez, M., Lopez, M. F., Perez, A. G. and Juristo, N. (1998). "Building Ontologies at the Knowledge Level Using the Ontology Design Environment." In Proceedings of KAW'98, Banff, Canada.

[4] DAML (DARPA Agent Markup Langauage) Ontology Library http://www.daml.org/ontologies/.

[5] Gruber, T. R. (1995). "Toward principles for the design of ontologies used for knowledge sharing." International Journal of Human-Computer Studies, 43(5-6), 907-928.

[6] Lopez, M. F. and Perez, A. G. (2002). "Overview and Analysis of Methodologies for Building Ontologies." Knowledge Engineering Review, 17(2), 129-156.

[7] Lopez, M. F., Perez, A. G., Sierra, J. P. and Sierra, A. P. (1999). "Building a Chemical Ontology Using Methontology and the Ontology Design Environment." Ieee Intelligent Systems & Their Applications, 14(1), 37-46.

[8] Mark, D. M., Skupin, A. and Smith, B. (2001). "Features, Objects, and other Things: Ontological Distinctions in the Geographic Domain." Conference On Spatial Information Theory (COSIT), Morro Bay, CA, USA.

[9] Mylopoulos, J., Borgido, A., Jarrke, M. and Koubarakis, M. (1990). "Telos: Representing Knowldege About Information Systems." ACM TOIS. 325-362.

[10] Noy, N. F. and Hafner, C. D. (1997). "The state of the art in ontology design - A survey and comparative review." AI Magazine, 18(3), 53-74.

[11] Noy, N. F. and McGuinness, D. L. (2001). "Ontology Development 101: a Guide to Creating Your First Ontology." Stanford, CA, Stanford University.

[12] Ontolingua. www.ksl.stanford.edu/software/ontolingua/ Knowledge System Labratory, Stanford University.

[13] Protege-2000 http://protege.stanford.edu/index.html.

[14] Sowa, J. F. (2000). "Knowledge Representation: Logical, Philosophical, and Computational Foundation." Pacific Grove, CA. Brooks Cole Publishing Co.

[15] SUMO (Suggested Upper Merged Ontology) http://ontology.teknowledge.com/.

[16] WordNet http://www.cogsci.princeton.edu/~wn/.

[17] Zhou, Q. and Fikes, R. (2002). "A Reusable Time Ontology." Proceeding of the Ontologies for the Semantic Web Workshop, AAAI National Conference