Combining Data Integration with Natural Language Technology for the Semantic Web

Dean Williams, Alexandra Poulovassilis
{dean,ap}@dcs.bbk.ac.uk

School of Computer Science and Information Systems, Birkbeck College, University of London

1 Introduction

The Semantic Web requires us to be able to integrate information from a variety of sources, including unstructured text from web pages, semi-structured XML data, structured databases, and metadata sources such as ontologies. Integration of heterogeneous data sources is a problem that has been addressed by several recent data integration systems, one of which is the AutoMed system being developed at Birkbeck and Imperial Colleges (`http://www.doc.ic.ac.uk/automed`). In data integration systems, several data sources, each with an associated local schema, are integrated to form a single virtual database with an associated global schema. If the data sources conform to different data models, then these need to be transformed into a common data model as part of the integration process. The AutoMed system uses a low-level graph-based data model, the HDM, as its common data model, and bi-directional schema transformation pathways to transform and integrate heterogeneous schemas [3].

There is clearly potential for using this approach for information integration in the Semantic Web, but a number of extensions are required. In particular, while data in a wide range of structured and semi-structured formats has been dealt with by previous data integration systems, natural language sources and ontologies have not. In this paper we present a method of extracting data and metadata from natural language sources and integrating it with other structured and semi-structured data sources. We describe how existing metadata can be used to assist in this extraction process.

Our approach combines Information Extraction (IE) technology with the AutoMed data integration system. The resulting system, called **ESTEST (Experimental Software for Extracting Structure from Text)**, makes use of existing metadata such as database schemas, natural language ontologies, and domain-specific ontologies, to assist the IE process from text. The Resource Description Framework (RDF) is an emerging standard for representing and sharing ontolological data, and in [6] we have shown how RDF and RDFS can be represented in the HDM, so that RDF/S description bases can be treated as AutoMed data sources. In ESTEST, once new data and new metadata have been extracted from the text, they are integrated with the existing data and metadata. This extraction and integration process can be reiterated as required.

While the new data and new metadata discovered by ESTEST may be expressed in a variety of data models, these can all be mapped into the HDM, and hence for ESTEST we have developed a native HDM repository to store all the new data and metadata, described in [4].

2 ESTEST

ESTEST makes use of AutoMed for its data integration aspects and of an IE system to extract structured information from text. Figure 1 illustrates the ESTEST system architecture, and below we briefly dscribe each of its main components. We refer the reader to the full paper [5] for more details and for an extended example illustrating the use of ESTEST in the area of Road Traffic Accident analysis.

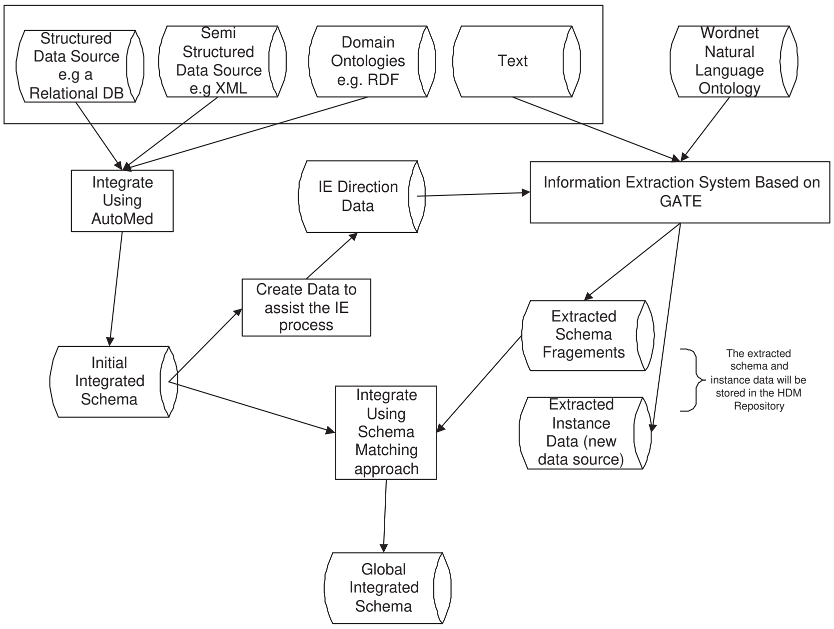

Fig. 1. Overview of the ESTEST System

Initial Integration. The available data sources other than the text (e.g. structured, semi-structured, and domain ontologies) are first integrated into a single virtual global schema, using AutoMed schema transformation pathways. This global schema can then be queried by submitting queries expressed in AutoMed's IQL query language to its global query processor [1].

Create Data to Assist the IE Process. The global virtual resource can be used to provide data which assists the IE process. For example, lists of entities can be created by submitting queries to the global schema. These lists can then be used by the 'named entity recogniser' components of the IE system (see below). From the global schema, we can also extract information to create templates for grammars. This is described in more detail in the next section, and is based on extracting text information from the global schema e.g. from table and column names in relational schemas.

Information Extraction System Based on GATE. The IE component of ESTEST is based on the GATE system (`http://www.gate.ac.uk`) which allows for a sequence of language processing components to be assembled and marks up annotations on the input text. GATE's language processing components include standard components such as sentence splitters and named entity recognisers. Bespoke components can also be constructed and integrated with the existing standard ones.

We are developing new IE components which will generate templates for the extraction, based on the assumption that the entity types in the existing schema and domain ontologies will be at least a significant subset of the entity types for which we wish to extract information from the text. We are also developing a WordNet component which will make use of its synonym and hyponym structures to allow for alternative lists for words to be found in cases where the textual descriptions of schema elements is restricted, for example to a word in a column name.

The result of the IE process is a set of named annotations over sections of the text. These annotations can be thought of as discovered fragments of schema. These fragments and the text to which they refer are stored in the Extracted Schema Fragments and Extracted Instance Data store, respectively, both of which are implemented using our HDM repository.

Integrate New and Existing Metadata Using Schema Matching. A schema matching algorithm takes each new extracted schema fragment and finds its best match with respect to the global schema, or allows for it to be appended to the global schema. Unlike many other schema matching applications, there will not be much structural information available to assist the matching algorithm and we will rely primarily on element names. However, we will also experiment with using the new instance values extracted, looking to see if these are already present in the extents of candidate schema entities and using any presence to provide evidence of a semantic match.

The Extracted Schema Fragments are integrated with the virtual global schema by means of AutoMed schema transformation pathways which are automatically generated by the above schema matching process. The data in the Extracted Data store can thus be treated as a new AutoMed data source, and queries posed against the virtual global schema will automatically make use of this new data.

3 Conclusion

In this paper we have described the ESTEST system, which extends traditional data integration systems by combining the AutoMed data integration approach with IE technology in order to allow information from ontologies and natural language sources to be integrated with other, semantically related, structured or semi-structured data. ESTEST uses related schema and ontology information to assist the IE process from text. The information extracted is integrated as a new data source with respect to a virtual global schema.

We believe that AutoMed is well-suited to data integration on the Semantic Web:

- The low-level, graph-based nature of the HDM lends itself naturally to modelling both structured and semi-structured information, and for discovering structure in text where both the schema and the instance data may be extended as part of the discovery process.
- AutoMed's bidirectional schema transformation pathways result in easy support of schema evolution, both of local data sources and of integrated virtual schemas (see [2,3]); this is very likely to be needed in the dynamic environment of Web applications.
- AutoMed's fine-grained schema transformations make transformation pathways amenable to automatic or semi-automatic generation.

We are currently finishing a first implementation of the ESTEST system and will evaluate its effectiveness in a number of application areas, including Road Traffic Accident Data and Operational Intelligence Police Reports. There are a number of research directions for further work, including the use of metadata to drive IE, and schema matching where only text and metadata is available.

References

1. E. Jasper, A. Poulovassilis, and L. Zamboulis. Processing IQL Queries and Migrating Data in the AutoMed toolkit. Technical report, AutoMed Project, 2003.
2. P.J. McBrien and A. Poulovassilis. Schema evolution in heterogeneous database architectures, a schema transformation approach. In *Proc. CAiSE'02, LNCS 2348*, pages 484–499, 2002.
3. P.J. McBrien and A. Poulovassilis. Data integration by bi-directional schema transformation rules. In *Proc. ICDE'03*, 2003.
4. D. Williams. The Automed HDM data store. Technical report, Automed Project, 2003.
5. D. Williams and A.Poulovassilis. Combining data integration with natural language technology for the semantic web. Technical report, Automed Project, 2003.
6. D. Williams and A.Poulovassilis. Representing RDF and RDF Schema in the HDM. Technical report, Automed Project, 2003.