# Creating a Semantic Integration System using Spatial Data

Jennifer Green
Ordnance Survey
Southampton, SO16 4GU, UK
Jenny.Green@ordnancesurvey.co.uk

Catherine Dolbear
Glen Hart
Ordnance Survey
Southampton, SO16 4GU, UK
{Catherine.Dolbear, Glen.Hart}
@ordnancesurvey.co.uk

John Goodwin
Paula Engelbrecht
Ordnance Survey
Southampton, SO16 4GU, UK
{John.Goodwin,
Paula.Englelbrecht}
@ordnancesurvey.co.uk

## ABSTRACT

Data integration is complex often requiring much technical knowledge and expert understanding of the data and its meaning. In this paper we investigate the use of current semantic tools as an aid to data integration, and identify the need to modify these tools to meet the needs of spatial data. Illustrating the benefits of exposing the semantics of integration through creation of a demonstrator.

## Categories and Subject Descriptors

H.3.4 [**Systems and Software**]: Information networks

## General Terms

Design, Experimentation, Theory.

## Keywords

Geospatial, Semantic Web, OWL, RDF, Ontology.

## 1. INTRODUCTION

Increasingly there is a need to provide solutions that integrate data from different datasets originating from diverse data providers. Integrating this data is a non-trivial task; one that is made more difficult by data not only being supplied in different formats, but also being defined using subtly different semantics. Such semantic difference is not always clearly documented and the final system can hide much of the semantics of the integration within a black box model.

The prototype we are developing conducts data integration in a more explicit fashion producing a merged ontology that aims to resolve most, if not all, of the integration issues. However, such an approach has to overcome an additional problem in that there typically exists a semantic gap between the domain described in the ontology, and the related data source.

In order to illustrate that such processes are possible, if not yet efficient, this paper describes a work in progress to build a demonstrator that uses ontologies to describe the domains, mapping those domains to the physical data and finally combining the domains through linking of the ontologies, including linking using spatial relationships.

In doing so we have chosen to model a real world problem, that of predictive modelling of diffuse water pollution. We have simplified the problem to make it manageable but have ensured that it is a reasonable subset of the problem we have chosen.

## 2. BACKGROUND

In the literature there are many documented use cases for semantic integration [9],[8]. Each of these systems developed their own technology to carry out the integration. Much of the work carried out since these early use cases has been concerned with solutions for specific problems associated with semantic integration, such as mapping between ontologies [3] and querying multiple data sources using a single ontology [12]. This has resulted in tools being developed which have started to transfer technologies from academia into the mainstream commercial sector.

None of the current toolsets available allow the spatial attribution of the data sources in this scenario to be queried. Our demonstrator will extend features from currently available tools to make this possible.

Previous use cases create ontologies from existing sources of semantics, such as thesauri [7], or are created by the domain expert involved in the project. The ontologies created in this demonstrator will be created using knowledge elicitation techniques with active involvement from domain experts.

## 3. THE DEMONSTRATOR

The semantic data integration demonstrator enables queries to be passed to the system that are expressed in the vocabulary of the application domain rather than using the terminology of the database schemas. The demonstrator comprises a series of layered ontologies and a translator for converting the data sources into a virtual RDF graph.

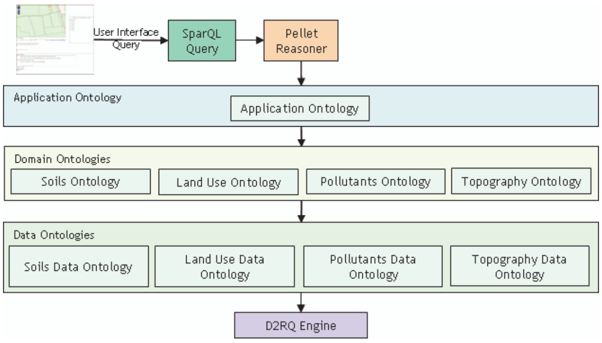


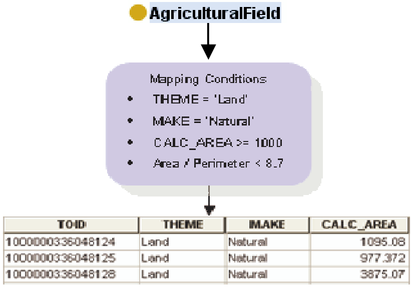
**Figure 1. The demonstrator architecture**

Figure 1 shows the three layers in the ontology stack. The domain ontologies, in the middle of the stack, are written using the involvement of domain experts. Knowledge elicitation techniques such as laddering [5] and semi-structured interviews are carried out on the domain experts. This knowledge is then documented using Rabbit [4], a controlled natural language which is converted in to the OWL DL [10] domain ontologies.

The application ontology links the domain ontologies together, adding in additional application-specific information. Recent research has recognised ontology modularisation techniques [2] that can be used to assist us in the linking process by avoiding the need to import entire domain ontologies. However, much of the application ontology will be created manually, as it requires domain knowledge of the scenario being modelled. For instance in the diffuse pollution scenario, a field in the Topography ontology is linked to a soil type in the Soils ontology based on the soil type that covers the largest proportion of the field. The application ontology is documented in Rabbit and OWL DL with SWRL [6] rules to represent some of the task specific knowledge.

The final set of ontologies in the stack are the data ontologies. These map each data source to the concepts in the related domain ontologies. This helps to close the semantic gap between the data source and the domain concepts, for instance within the topographic domain ontology there is a concept of a field and yet the topographic data source does not explicitly define such a concept. The data ontology also allows spatial relations present in the domain ontologies to be mapped to functions on the data source. This is illustrated by the mapping shown in Figure 2 showing an Agricultural Field shown which requires spatial operations (e.g. perimeter calculations) in order to be instantiated from the database.

The data ontology is written in the D2RQ mapping language [1]. The translator that mediates between the data ontology and the data sources is a modified D2RQ Engine which has been enhanced to allow the use of spatial operators provided by the Oracle Spatial database.



| TOID | THEME | MAKE | CALC_AREA |
|---|---|---|---|
| 1000000336048124 | Land | Natural | 1095.08 |
| 1000000336048125 | Land | Natural | 977.372 |
| 1000000336048128 | Land | Natural | 3875.07 |

**Figure 2. Mapping the domain concepts to the source data.**

Using the data ontology and the translator the data sources are then exposed as a virtual RDF which can be queried using SparQL [11].

As our ontologies are written in OWL DL we preserve the OWL DL semantics in the query by using the A Box query engine in the Pellet reasoner [13]. The addition of the reasoner to the demonstrator is an important step as if we use only SparQL we lose all of the OWL expressivity, for example within the scenario we define the concept of a polluted area. The instances of a polluted area are inferred by the reasoner using the axioms within the ontology e.g. if a field has recently been fertilised and was then heavily rained upon then that field will produce pollutants. However if only using a SparQL query then the user would need to create a query of all fields which were recently fertilised and suffered heavy rain. This illustrates the importance of being able to use the domain knowledge captured within the ontology.

Although we want to expose the knowledge and definitions used to model the scenario we do not want users to need to use semantic technologies, such as SparQL, to build their queries. To this end we have designed a user interface that will hide the semantic queries from the user, providing a mapping interface to visualise the spatial attribution. This will be used to input details to the query as well as displaying the results.

# 4. CONCLUSIONS

As a data provider we know that data integration is of paramount importance to our customers. It is hoped that this demonstrator will prove the benefits and highlight some possible problems of semantic integration that arise when introducing these technologies. Thus far, the major issues appear to be: immaturity of technology to deal with spatial data; scaling issues when using virtual RDF and the difficulties in modelling complex tasks. The benefits are seen in the exposure of integration assumptions, and the ability to link using domain concepts.

# 5. REFERENCES

[1] Bizer, C., Seaborne, A. 2004. D2RQ - Treating Non-RDF Databases as Virtual RDF Graphs, *In Proc of the 3rd International Semantic Web Conference*.

[2] Ernesto, J.R., Grau, B.C., Sattler, U., Schneider, T., Berlanga, R. 2008. Safe and Economic Re-Use of Ontologies: A Logic-Based Methodology and Tool Support, In *Proc. of 5th European Semantic Web Conference*.

[3] Euzenat, J., Shvaiko, P. 2007. *Ontology Matching*, Springer.

[4] Hart, G., Johnson, M., Dolbear, C. 2008. Rabbit: Developing a Control Natural Language for Authoring Ontologies, *In Proc. of the 5th European Semantic Web Conference.*

[5] Hinkle, D. 1965. The change of personal constructs from the viewpoint of a theory of construct implications, Ohio State University.

[6] Horrocks, I., Patel-Schneider, P.F., Boley, H., Tabet, S., Grosof, B., Dean, M. 2004. SWRL: A Semantic Web Rule Language Combining OWL and RuleML, W3C.

[7] Hyvonen, E., Vilijanen, K., Tuominen, J., Seppala, K. 2008. Building a National Semantic Web Ontology and Ontology Service Infrastructure - The FinnONTO Approach, In *Proc. of 5th European Semantic Web Conference.*

[8] King, R.L., Durbha, S., Younan, N.H. 2005. Interoperability in Costal Zone Monitoring Systems, In *Proc. of 31st International Symposium on Remote Sensing of Environment.*

[9] Lin, K., Ludascher, B. 2003. A System for Semantic Integration of Geologic Maps via Ontologies, In *Proc. of Semantic Web Technologies for Searching and Retrieving Scientific Data.*

[10] McGuinness, D.L., van Harmelen, F. 2004. OWL Web Ontology Language Overview, W3C.

[11] Prud'hommeaux, E., Seaborne, A. 2008. SPARQL Query Language for RDF, W3C.

[12] Wu, Z., Chen, H., Wang, H., Wang, Y., Mao, Y., Tang, J., Zhou, C. 2005 Dartgrid: a Semantic Web Toolkit for Integrating Heterogeneous Relational Databases, in *Proc. of 4th International Semantic Web Conference.*

[13] Sirin, E., Parsia, B., Grau, B.C., Kalyanpur, A., Katz, Y. 2007. Pellet: A Practical OWL-DL Reasoner, *Journal of Web Semantics*