Semantic Fisheries Data Integration and Analytics

Aileen Kennedy, NUI Galway (Ireland), a.kennedy28@nuigalway.ie
David Currie, Marine Institute (Ireland), david.currie@marine.ie
Enda Howley, NUI Galway (Ireland), enda.howley@nuigalway.ie
Jim Duggan, NUI Galway (Ireland), jim.duggan@nuigalway.ie

Background

The Marine Institute generates and consumes data from a number of different sources. In particular, Fisheries Ecosystem Advisory Services (FEAS) maintains databases containing information on:

- *catch value and volume*, including commercial landings, fishing effort and fleet capacity;
- *biological data* – which includes data on variables such as the number, length, weight, sex and age of fish species in a given location.

Fine-scale spatio-temporal data about vessel position is also stored. The fisheries data typically becomes available in batches rather than being streamed – these batches can range in size from an individual fishing trip, up to all vessel positions in a calendar year. These databases are typically *siloed*, so querying for information that are stored across many databases can be a significant challenge.

Semantic Integration and Analysis

This project aims to semantically integrate and analyse fisheries and marine data - it will use a pipeline approach to data analytics, whereby heterogeneous marine data sources can be combined within an analytics framework (e.g. R), so that key hypotheses can be explored and evaluated by decision makers.

This system will allow queries to be expressed in terms of the items of interest rather than requiring structural knowledge of the underlying databases. This semantic layer will add extra querying power by allowing inferences to be made that are difficult or impossible to process using traditional queries to the underlying relational model. It will also allow the work to be more transferable since it will be based on meaning, rather than a specific database structure.

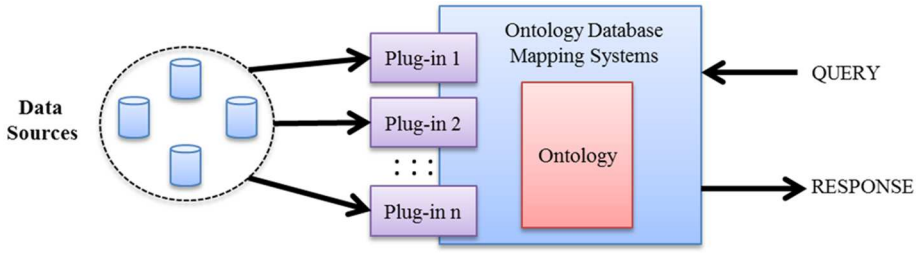

Figure 1: A possible mapping between a semantic model and relational databases

Individual data sources have their own characteristics with respect to noise, accuracy and completeness and data will also be available at different aggregation levels. Techniques for combining data sources must be able to incorporate these features and use them to inform the outputs. The semantic system designed must be extendable, both to facilitate sharing with other researchers and to allow new data sources to be added easily to the system as the project progresses.

The key outputs of the project will be:

1. A novel, open, generic framework for semantic data integration so that heterogeneous marine data sources can be integrated in a meaningful way.
2. An algorithm that will map between semantic queries and different underlying data structures that improves on the performance of the current state-of-the-art whilst retaining expressiveness.
3. A pilot analytics engine to allow the semantic model to be used to answer policy and decision support questions. This engine could use a number of different techniques such as machine learning, simulation, or game theory.

Discussion

This presentation will discuss the progress that has been made on the project and present initial results – these will be targeted at modelling the biological measurements of a fish using a suitable ontology and mapping semantic queries in that domain to the data stored in relational databases.

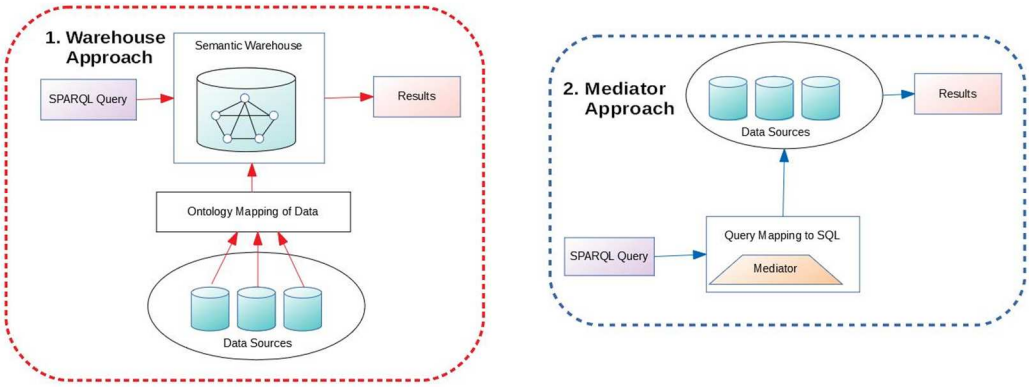

Figure 2: Approaches for data integration

Two different approaches to data integration will be designed and evaluated. In the first approach, the *Warehouse Approach*, the data is extracted from the relational databases, transformed using the ontology and stored in a semantic warehouse. The semantic query language SPARQL is used to query the semantic warehouse directly to obtain the results. In the second approach, the *Mediator Approach*, the SPARQL query is mapped, using an ontology, to SQL - the relational databases are then queried using the SQL query. The two approaches are evaluated with respect to defined performance criteria with the aim of providing a semantic structure to support the design of the analytics engine.

References

Currie D., Howley E., and Duggan J. (2016) A Data Analytics Framework for Ecosystem-Based Fisheries Management. ICES Annual Science Conference 2016

Tzitzikas Y. et al. (2013) Integrating Heterogeneous and Distributed Information about Marine Species through a Top Level Ontology. In: Garoufallou E., Greenberg J. (eds) Metadata and Semantics Research. MTSR 2013. Communications in Computer and Information Science, vol 390. Springer, Cham