An enabling semantic pipeline for fisheries observation data to improve sustainable management of fisheries resources

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

Abstract

Fisheries scientists typically collect and analyse data such as commercial landings, fishing effort and fleet capacity - there is also a requirement for biological sampling data and spatio-temporal data in relation to vessel positioning. These data sources provide large quantities of heterogeneous data. As commercial fisheries data is often private it has tended not to follow interoperable standards and, as such, is more difficult to integrate. This work aims to use semantic web techniques to integrate and analyse heterogeneous marine and fisheries data sources. The Observation and Measurement (O&M) Ontology is used to provide a generic, non-domain specific ontology that can be used to allow interoperability between the different data sources to provide a standardised framework. The aim of the work is to create a semantic data management infrastructure that integrates fisheries observational data at a national level to aid decision support systems. Once proven at a national level, the generic pipeline can be scaled and migrated to a regional level. The integrated data could provide a data platform for machine learning prediction and forecasting techniques to aid in the sustainable management of fisheries resources.

Context

O&M is a domain neutral international standard information model. The scope of O&M includes *in-situ* observations, remote sensing, *ex-situ* observations, numerical models and simulations, and forecasts. It can include any action whose result is an estimate of the property value. The INSPIRE European standards include a Species Distribution theme in which the O&M standards have been identified as being relevant however its purpose is not to directly record observations rather aggregations of such [2]. Our approach is to implement O&M to allow the integration and interoperability of fisheries data at a detailed level [3], [4] - we also use appropriate controlled vocabularies from sources such as ICES and NERC Vocabulary Server for this reason

System Design and Implementation

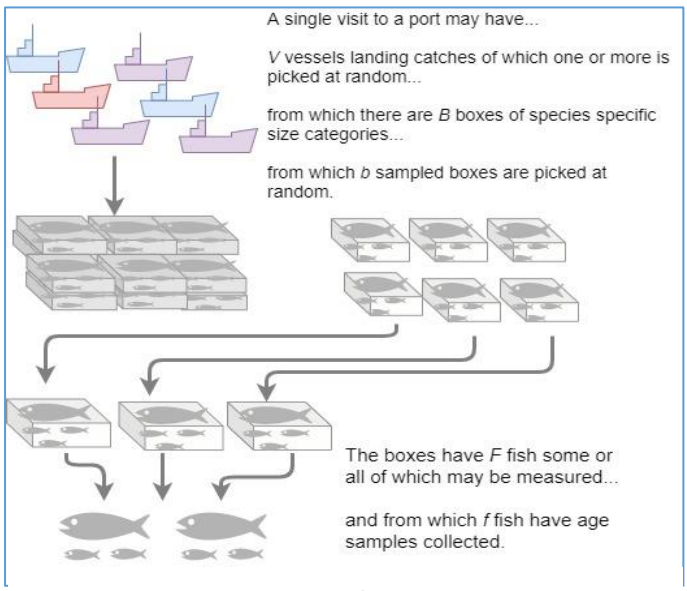

Figure 1: Sampling Process

The existing data set, which is gathered according to multi-stage hierarchical sampling schemes (an example of which is shown in Figure 1), is stored on a number of heterogeneous relational databases and is composed of both fine grain spatial data and coarser grained biological sampling data. The data is typically siloed and querying across many different databases can be a significant challenge. Our system extracts the data from the appropriate databases using SQL. The extracted data is transformed to RDF triple format, with each triple composed of a *subject, predicate* and *object.* The structure of the triples is determined by the structure of the ontology.

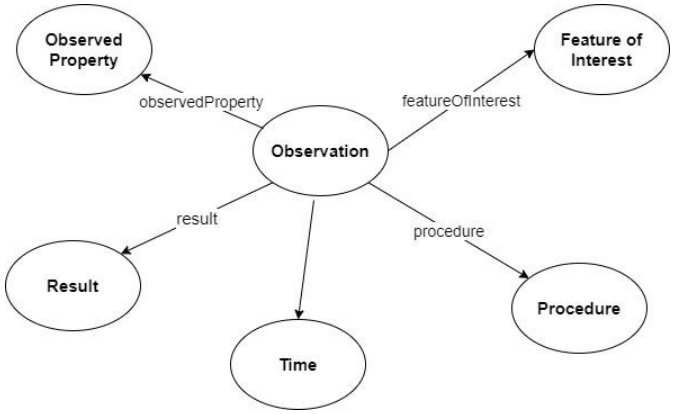

Figure 2: Simple Observation in O&M Ontology

The ontology is implemented in OWL (Web Ontology Language) using classes and properties from the "oml-lite" and "samfl" ontologies to ensure that it is generic and interoperable [1] (Figure 2). Each physical sample can have a number of different measurements e.g. weight, length, age etc. Each measurement has its own class which stores the result of the measurement and other data such as observed property, time and procedure.

The technical architecture includes a number of different elements: the ontology is written in OWL, using Protégé; the data, stored in relational databases, are extracted and transformed to RDF using Python; the resulting serialised RDF data is stored in a triple store and queried using SPARQL, an RDF query language.

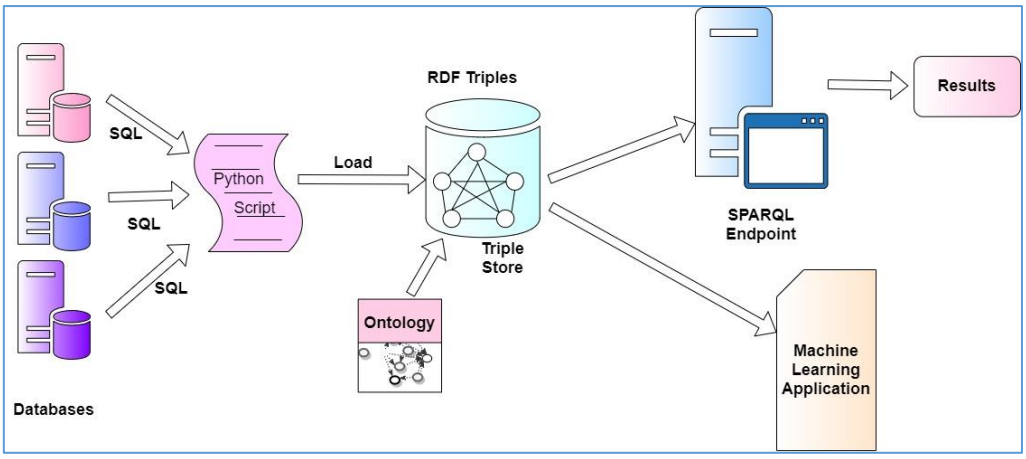

Figure 3: System Architecture

Results and Future work

We have applied the techniques discussed here in the paper to commercial fisheries sampling data and will present the outputs. The use of SPARQL to query the integrated data could enable users to write more intuitive queries rather than the complex SQL. Insight can also be gained through the use of inference on the transformed data, which would not be possible with the underlying relational databases. The pipeline integrating the data into a more standardised, interoperable format opens up opportunities for further analytics and application of machine learning techniques.

References

[1] Cox, S.J., 2017. Ontology for Observations and Sampling Features, with Alignments to Existing Models. Semantic Web, 8(3), pp.453-470

[2] INSPIRE Thematic Working Group Species Distribution, 2013, D2.8.III.19 INSPIRE Data Specification on Species Distribution – Technical Guidelines

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

[4] Kennedy A., Currie D., Howley E., and Duggan J. (2018) Semantic Fisheries Data Integration and Analytics, IMDIS Conference 2018