Contributors:
Name: Yannis Tzitzikas
Short CV:
Yannis Tzitzikas is Assistant Professor in the Computer Science Dep. at University of Crete (Greece) and Associate Researcher in Information Systems Lab at FORTH-ICS (Greece). Before joining UofCrete and FORTH-ICS he was postdoctoral fellow at the University of Namur (Belgium) and ERCIM postdoctoral fellow at ISTI-CNR (Pisa, Italy) and at VTT Technical Research Centre of Finland. His current research focuses on flexible and dynamic methods for integrating information (structured and unstructured), and on exploratory search. The results of his research have been published in more than 70 papers in refereed international conferences and journals, he has received two best paper awards (at CIA'2003 and ISWC'07), and currently he actively participates to two FP7 Research Infrastructure Projects (iMarine and SCIDIP-ES), one Network of Excellence (APARSEN) on Digital Preservation, and one COST Action (MUMIA).
Home page: www.ics.forth.gr/~tzitzik

Type: **Research contribution**
Title of Presentation: **Constructing, Evaluating and Exploiting Domain Specific Semantic Warehouses based on Linked Data**

Summary of the presentation (100 words):
In many applications one has to fetch and assemble pieces of information coming from more than one sources and/or SPARQL endpoints. We will describe the corresponding requirements and challenges, and then we will present a process for constructing such a semantic warehouse. We will focus on the aspect of *quality* and *value* and for this reason we will introduce various metrics for quantifying the *connectivity* of the warehouse. We will demonstrate the behavior of these metrics in the context of a real and operational semantic warehouse, and how the warehouse is exploited in an exploratory search process.

Extended abstract of the presentation:
An increasing number of datasets are already available as Linked Data. For exploiting this wealth of data, and building domain specific applications, in many cases there is the need for fetching and assembling pieces of information coming from more than one SPARQL endpoints (and other kinds of sources). These pieces are then used for constructing a warehouse for offering more complete browsing and query services (in comparison to those offered by the underlying sources).

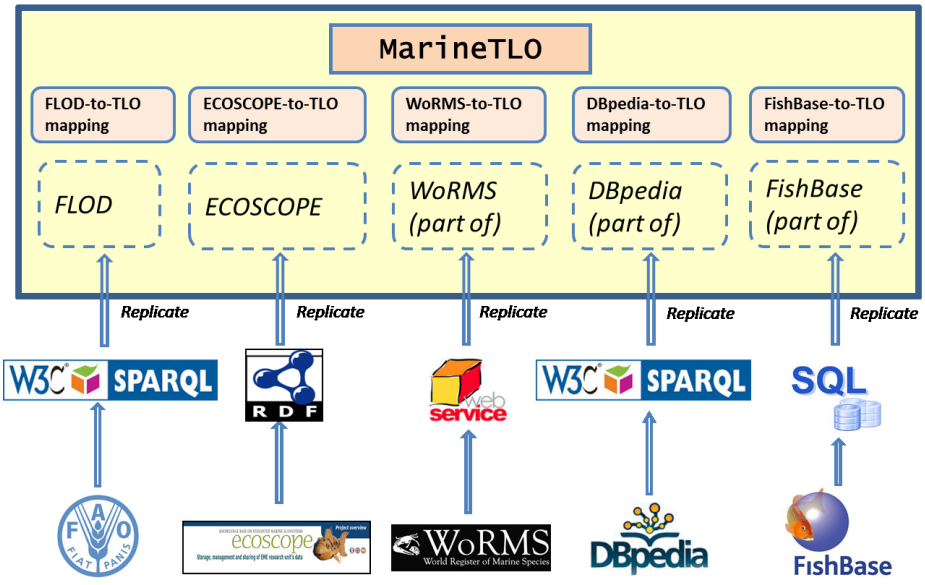
We can distinguish domain independent warehouses, like the Sindice RDF search engine, or the Semantic Web Search Engine (SWSE), but also *domain specific*. In this presentation we will focus on the requirements for building domain specific warehouses. Such warehouses aim to serve particular needs, particular communities of users, consequently their "quality" requirements are higher. It is therefore worth elaborating on the process that can be used for building such warehouses, and on the related difficulties and challenges. In brief, for building such a warehouse one has to tackle various challenges and questions, e.g. how to define the objectives and the scope of such a warehouse, how to connect the fetched pieces of information (common URIs or literals are not always there), how to tackle the various issues of provenance that arise, how to keep the warehouse fresh, i.e. how to automate its construction or refreshing, and how to measure the quality and value of such warehouse.

The context of this work is the ongoing iMarine project (FP7, Research Infrastructures, http://www.i-marine.eu/) that offers an operational distributed infrastructure that serves hundreds of scientists from the marine domain. As regards semantically structured information, the objective is to integrate information from various marine sources, specifically from WoRMS (http://www.marinespecies.org/), Ecoscope, Fishbase (http://www.fishbase.org/), FLOD (http://www.fao.org/figis/flod/) and DBpedia.

One of the challenges in the **iMarine** project is how users could experience a coherent source of facts about marine entities, rather than a bag of contributed contents. Queries like *"Given the scientific name of a species, find its predators with the related taxon-rank classification and with the different codes that the organizations use to refer to them"*, could not be formulated (and consequently nor answered) by any individual source. To formulate such queries we need an expressive conceptual model, while for answering them we also have to assemble pieces of information stored in different sources.

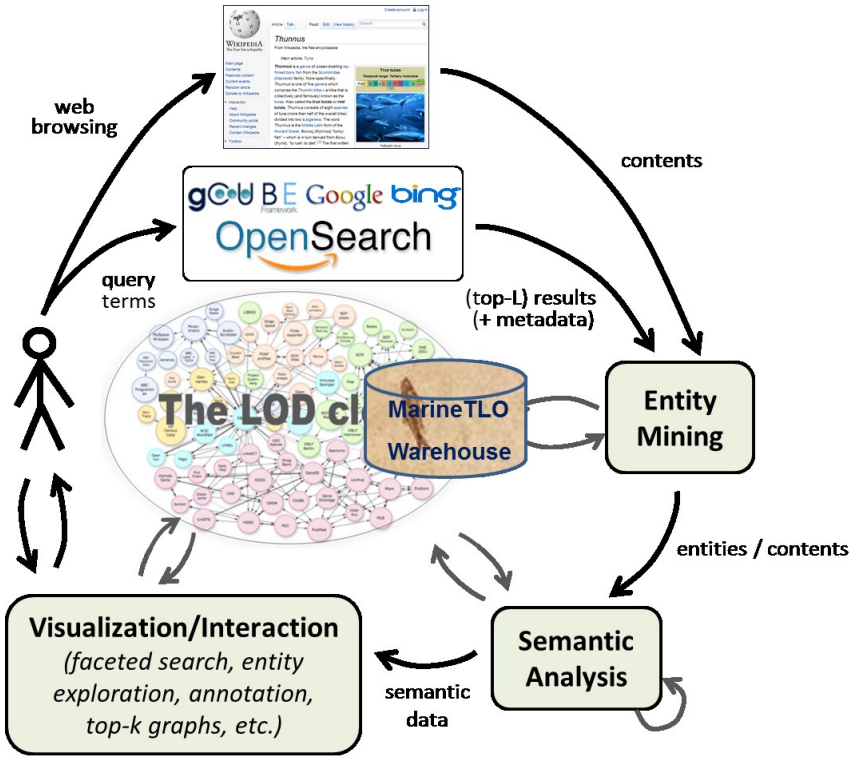
For this reason we have designed and implemented a top level ontology, called **MarineTLO**, which is generic enough to provide consistent abstractions or specifications of concepts included in all data models or ontologies of marine data sources and provide the necessary properties to make this distributed knowledge base a coherent source of facts relating observational data with the respective spatiotemporal context and categorical (systematic) domain knowledge. It can be used as the core schema for publishing Linked Data, as well as for setting up integration systems for the marine domain. It can be extended to any level of detail on demand, while preserving monotonicity. For its development and evolution we have adopted an iterative and incremental methodology where a new version is released every two months. For the implementation we use OWL 2, while for the needs of its evaluation we use a set of query requirements provided by the related communities.

For answering complex queries, we have to assemble pieces of information stored in different sources. For this reason, we have established a process (supported by a tool that we have developed for this purpose) for creating **MarineTLO-based warehouses** that integrate information coming from various sources. To fetch the data we have to use a plethora of access methods (SPARQL endpoints, HTTP accessible files, JDBC), while for connecting the fetched data we have to define schema mappings, transformations, as well as rules for instance matching. The current version of the warehouse integrates information coming from WoRMS, ECOSCOPE, FLOD, FishBase and DBPedia, contains around 3 million triples, and provides harmonized and integrated information for about 37,000 distinct marine species.



Another big challenge nowadays is how to integrate structured data with unstructured data (documents and text). The availability of harmonized structured knowledge about the marine domain can be exploited for a semantic post-processing of the search results (over dedicated or general purpose search systems). Specifically the work done in the context of iMarine so far, proposed a method to enrich the classical (mainly keyword based) searching with entity mining that is performed at query time. In particular, the

results of entity mining (entities grouped in categories) complement the query answers with information which can be further exploited by the user in a faceted and session-based interaction scheme.



This means that instead of annotating and building indexes for the documents (or web pages), the annotation can be done at *query time* and using the desired entities of interest. These works show that the application of entity mining over the *snippets* of the top hits of the answers can be performed at real-time, and indicated how semantic repositories can be exploited for specifying the entities of interest and for providing further information about the identified entities.

The integrated warehouse is now operational and it is exploited in various applications, including generators of fact sheets or for enabling exploratory search services that offers semantic post-processing of search results).

The presentation will attempt to present as coherent whole, the research results presented at:

- Y. Tzitzikas, C. Alloca, C. Bekiari, Y. Marketakis, P. Fafalios, M. Doerr, N. Minadakis, T. Patkos and L. Candela, Integrating Heterogeneous and Distributed Information about Marine Species through a Top Level Ontology, Proceedings of the 7th Metadata and Semantic Research Conference, MTSR'13, Thessaloniki, Greece, November 2013.
- I. Kitsos, K. Magoutis, Y. Tzitzikas, Scalable entity-based summarization of web search results using MapReduce, Journal on Distributed and Parallel Databases (accepted for publication in the special issue "Scalable Data Summarization on Big Data" scheduled for 2013)
- P. Fafalios and Y. Tzitzikas, X-ENS: Semantic Enrichment of Web Search Results at Real-Time, Demo paper at SIGIR 2013
- P. Fafalios, M. Salampasis and Y. Tzitzikas, Exploratory Patent Search with Faceted Search and Configurable Entity Mining, 1st International Workshop on Integrating IR technologies for Professional Search, in conjunction with the 35th European Conference on Information Retrieval (ECIR'13)
- P. Fafalios, I. Kitsos, Y. Marketakis, C. Baldassarre, M. Salampasis and Y. Tzitzikas, Web Searching with Entity Mining at Query Time, IRFC 2012, Vienna, Austria, July 2012.