iMarine Final Report

FORTH Institute of Computer Science

iMarine Data e-Infrastructure Initiative for Fisheries Management and Conservation of Marine Living Resources

Duration: 01/11/2011 – 30/09/2014 (35 months)

Contact Person: Yannis Tzitzikas, tzitzik@ics.forth.gr

Contents

1 iMarine in a nutshell

Marine (Data e-Infrastructure Initiative for Fisheries Management and Conservation of Marine Living Resources) is co-funded by the European Commission, DG Connect Unit, under Framework Programme 7 and involves thirteen international partners. The ultimate goal of iMarine is to contribute to sustainable environmental management with invaluable direct or indirect benefits to the future of our planet, from climate change mitigation and marine biodiversity loss containment to poverty alleviation and disaster risk reduction. The project was launched in November 2011 and ended in September 2014.

iMarine provides an e-infrastructure that facilitates open access and the sharing of a multitude of data, collaborative analysis, processing and mining processing, as well as the publication and dissemination of newly generated knowledge. This is a complex process because it requires coordination with many actors and initiatives across different scientific and operational domains. It is also important to tackle data heterogeneity while relying on a multitude of resources and technologies, some of which are not yet ripe or powerful enough to meet the given requirements.

iMarine products and services are exposed via the iMarine gateway. Currently, a number of gCube applications offering specialized functionalities for managing, processing, and visualising scientific data and textual content are available. The gateway is being further developed as the infrastructure evolves and iMarine services deployed. The iMarine virtual research environment offer a distributed and dynamically created environment, where the data, services, computational and storage resources are governed by dedicated policies. These services and resources are assigned to users via interfaces for a limited timeframe at little or no cost for the providers of the participatory data infrastructures.

2 FORTH involvement

In the context of iMarine FORTH concentrated on (a) the design and implementation of a model for harmonizing the metadata and semantics of Marine datasets, (b) construction of a semantic warehouse for the marine domain containing information from various datasets and (c) offering a set of services for semantically enriching marine-related information.

More specifically FORTH has been involved in the following tasks

- **T3.3**: Harmonization of Metadata, Semantics and Technologies
- **T10.1**: Data Retrieval Facilities
- **T10.4**: Semantic Data Analysis Facilities (***Task Leader***)
- **T11.2**: Data Management APIs
- **T11.3**: Data Consumption APIs

3 Outcome

3.1 MarineTLO

MarineTLO is a top level ontology, 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 used OWL 2, and to evaluate it we use a set of competency queries, formulating the domain requirements provided by the related communities. The latest version (V4, July 2014) of MarineTLO contains 127 classes and 81 properties and it is organized in two abstraction levels: model (schema) and meta-model (meta-schema).

3.2 MarineTLO-based warehouse

The MarineTLO-based warehouse is a semantic repository containing information about marine data coming from various sources. The conceptual backbone of MarineTLO-based warehouse is MarineTLO (see Section 3.1). The warehouse integrates data coming from FLOD[1], ECOSCOPE[2], WoRMS[3], FishBase[4] and DBpedia[5]. The objective of the warehouse is to provide a coherent set of facts about marine species. Just indicatively the following figure illustrates some information about the species Thunnus Albacares, which are derived from different sources. These pieces of information are complementary and are assembled for enabling advanced browsing, querying and reasoning. The latest version of the warehouse (V4, July 2014) contains more than 5.5 million triples with information about species, their scientific and common names, predators, ecosystems and more. The SPARQL endpoint can be accessed through the iMarine gateway (https://i-marine.d4science.org/group/biodiversitylab/sparql-search).

[1] Fisheries Linked Open Data (http://www.fao.org/figis/flod/endpoint/)
[2] http://ecoscopebc.mpl.ird.fr/joseki/ecoscope
[3] World Register of Marine Species (http://www.marinespecies.org/)
[4] http://www.fishbase.org/
[5] http://dbpedia.org/About

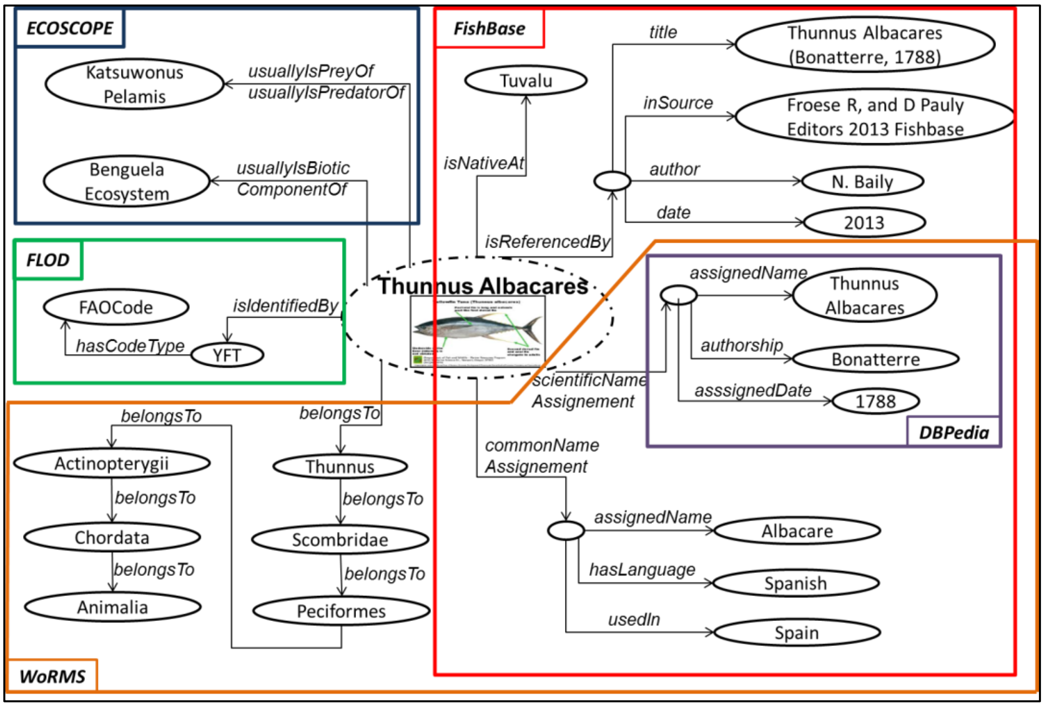


Figure 1. Integrated information about Thunnus Albacares from different sources

3.3 MatWare

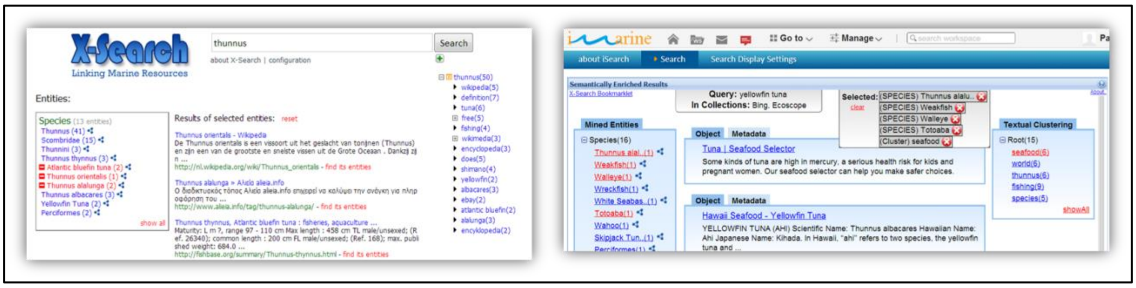
MatWare is a framework that automates the process of constructing semantic warehouses. MatWare automatically fetches contents from the underlying sources using several access methods (including SPARQL endpoints, HTTP accessible files, JDBC, gCube services). The fetched data are then stored (in their original form or after transforming them) in a RDF triplestore. It also supports several levels of modeling the provenance (e.g. provenance of URIs and literals, of triples etc.). One of the distinctive features of MatWare is that it allows evaluating the connectivity of the resulted warehouse. Connectivity refers to the degree up to which the contents of the semantic warehouse form a connected graph that can serve, ideally in a correct and complete way, the query requirements of the semantic warehouse, while making evident how each source contributes to that degree. To this end MatWare supports several metrics. These metrics allow someone to get an overview of the contribution (to the warehouse) of each source (enabling the discrimination of the important from the less-important sources) and to quantify the benefit of such a warehouse. These metrics include matrixes of common URIs/literals between different sources, percentages of the unique contribution of triples of each source, average degree of the entities of each source, complementarity factors (e.g. the number of sources that provided unique triples for the entities of interest), etc.

3.4 XSearch

XSearch is a meta-search engine that reads the description of an underlying search source (OpenSearch compliant), queries that source, analyzes the returned results in various ways and also exploits the availability of semantic repositories. The key features of XSearch are the following:

- provision of textual clustering of the results. Clustering is performed on the textual snippets of the returned results, but clustering of the entire contents is also supported.
- provision of Named Entity Recognition (NER) on the results. NER can be performed either over the textual snippets or over the entire contents. Various methods for ranking the identified entities are supported.
- faceted search-like exploration of the results. The results of clustering and entity mining are visualized and exploited in a faceted and session-based interaction scheme that allows the user to restrict his/her focus or information need gradually, and exploits the results of the previous steps.
- on-click semantic exploration of a Knowledge Base. XSearch provides the necessary linkage between the mined entities and semantic information (Linked Data). In particular, by exploiting a Semantic Knowledge Base (accessible through a SPARQL endpoint), the user can retrieve more information about an entity by querying and browsing – at real time – this Knowledge Base.
- entity discovery and exploration during plain Web browsing. XSearch also offers entity discovery and exploration while user is browsing on the Web. Specifically, the user is able to inspect the entities of a particular Web page by simply clicking a bookmarklet (a special bookmark) and then to semantically explore the properties of the identified entities. Namely, the user can at real-time exploit the aforementioned functionality while browsing.

X-Search is fully configurable in terms of the supported categories of entities, the underlying Knowledge Bases and the way the system queries the Knowledge Bases. XSearch has been developed as a web application, as a portlet (withing iMarine gateway), as a bookmarklet and as an API (see Section 3.5). The default configuration of XSearch exploits the contents of the MarineTLO-based warehouse (see Section 3.2) for linking search results with semantic information.



Thunnus orientalis - Wikipedia
De Thunnus orientalis is een vissoort uit het geslacht van tonijnen (Thunnus) en zijn een van de grootste en snelste vissen uit de Grote Oceaan . Dankzij zij n ...
http://nl.wikipedia.org/wiki/Thunnus_orientalis - find its entities

Thunnus alalunga » Αλαλιο aleia.info
Ο διαδικτυακός τόπος Αλιεία aleia.info επιχειρεί να καλύψει την ανάγκη για πληρ οφόρηση του ...
http://www.aleia.info/tag/thunnus-alalunga/ - find its entities

Thunnus thynnus, Atlantic bluefin tuna : fisheries, aquaculture ...
Maturity: L m ?, range 97 - 110 cm Max length : 458 cm TL male/unsexed; (R ef. 26340); common length : 200 cm FL male/unsexed; (Ref. 168); max. publi shed weight: 684.0 ...
http://fishbase.org/summary/Thunnus-thynnus.html - find its entities

Query: yellowfin tuna
In Collections: Bing, Ecoscope

Tuna | Seafood Selector
Some kinds of tuna are high in mercury, a serious health risk for kids and pregnant women. Our seafood selector can help you make safer choices.

Hawaii Seafood - Yellowfin Tuna
YELLOWFIN TUNA (AHI) Scientific Name: Thunnus albacares Hawaiian Name: Ahi Japanese Name: Kihada. In Hawaii, "ahi" refers to two species, the yellowfin tuna and ...

Figure 2. XSearch as a web application (left), XSearch as a portlet (right)

3.5 XSearch API

An API exposing the XSearch functionality has been developed. The API offers the required methods for supporting the semantic post-processing of search results. The API offers the following services:

- Post-processing of search results service which is responsible for performing textual clustering, entity mining or both in the top results as they are returned from the specified search system.

- Matching identified entity service which is responsible for linking the name of an entity with a resource in a knowledge base.
- Enriching identified entity service which is responsible for enriching an entity with semantic information.
- Identifying entities in a web-accessible document service which retrieves the contents of a document which is web accessible and performs entity mining over these contents.

3.6 XLink

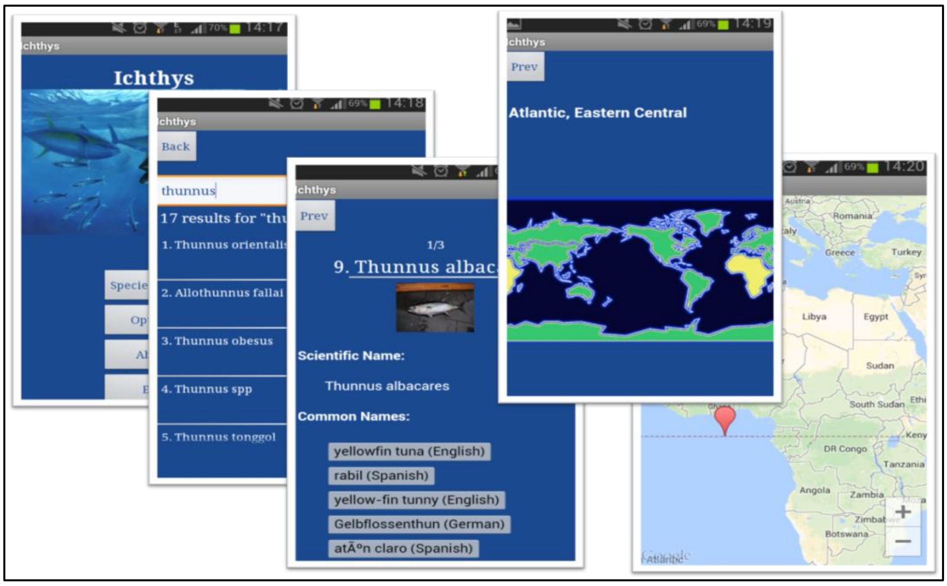
XLink is a fully configurable (Linked Data-based) Named Entity Extraction (NEE) tool which allows the user/developer to easily define the categories of entities that are interesting for the application at hand by exploiting one or more (online) Semantic Knowledge Bases (Linked Data). The user is also able to update a category and specify how to semantically link and enrich the identified entities. This enhanced configurability allows XLink to be configured for different contexts, for building domain-specific applications (e.g. for identifying drugs in a medical search system, for annotating and exploring fish species in a marine-related web page, etc.).

X-Link is based on Gate ANNIE[6] and supports both gazetteers (lists of names) and natural language processing functions. Gate ANNIE is a ready-made information extraction system which contains several components (e.g. Tokeniser, Gazetteer, Sentence Splitter, Orthographic Coreference, etc.). We have extended Gate ANNIE in order to be able to create a new supported category and update an existing one (using gazetteers) by exploiting the Linked Data. Currently, X-Link exports the results in XML and CSV.

[6] https://gate.ac.uk/ie/annie.html

3.7 Ichthys

A related activity that was not funded by iMarine, but was done in the time frame of iMarine was the design and implementation of a prototype Android application, called Icthys. It aims at providing information about marine species by querying MarineTLO-based warehouse. In brief, the user enters some basic information about what he is looking for (e.g. the common name of a species in Greek, or the scientific name of a species, etc.) and the application is responsible for constructing the appropriate SPARQL queries, submit them to the warehouse, retrieve the results, and integrate them and present them to the user. The following figure shows some indicative screenshots of the Icthys application.



Scientific Name:

Thunnus albacares

Common Names:

yellowfin tuna (English)
rabil (Spanish)
yellow-fin tunny (English)
Gelbflossenthun (German)
atÃºn claro (Spanish)

Figure 3. Some indicative screenshots of Ichthys Android application

4 Publications

[1]. P. Fafalios, I. Kitsos, Y. Marketakis, C. Baldassarre, M. Salampasis, Y. Tzitzikas. ***Web Searching with Entity Mining at Query Time. Multidisciplinary Information Retrieval***, 5th International Retrieval Facility Conference, IRFC 2012, Vienna, Austria, July 2-3, 2012, Proceedings, pp 73-88, DOI 10.1007/978-3-642-31274-8_6.

[2]. P. Fafalios, I. Kitsos, Y. Tzitzikas. ***Scalable, flexible and generic instant overview search***. In Proceedings of the 21st international conference companion on World Wide Web (WWW '12 Companion). ACM, New York, NY, USA, 333-336. DOI 10.1145/2187980.2188042.

[3]. P. Fafalios, M. Salampasis and Y. Tzitzikas. ***Exploratory Patent Search with Faceted Search and Configurable Entity Mining***. Proceedings of the 1st International Workshop on Integrating IR technologies for Professional Search, in conjunction with the 35th European Conference on Information Retrieval (ECIR'13), Moscow, Russia. Paper in Volume 968 of CEUR Workshop Proceedings.

[4]. P. Fafalios and Y. Tzitzikas. ***X-ENS: Semantic Enrichment of Web Search Results at Real-Time***, Demo Paper, Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval (demo paper), SIGIR 2013, Dublin, Ireland.

[5]. Y. Tzitzikas, C. Allocca, 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***, 7th Metadata and Semantics Research Conference, MTSR 2013, Thessaloniki, Greece.

[6]. Y. Tzitzikas, C. Allocca, C. Bekiari, Y. Marketakis, P. Fafalios and N. Minadakis. ***Ontology-based Integration of Heterogeneous and Distributed Information of the Marine Domain***, ERCIM News 2014 (96), Special theme: Linked Open Data, January 2014.

[7]. Y. Tzitzikas, N. Minadakis, Y. Marketakis, P. Fafalios, C. Alloca, and M. Mountantonakis. ***Quantifying the Connectivity of a Semantic Warehouse***. In procs of the 4th International Workshop on Linked Web Data Management (LWDM 2014) 2014.

[8]. Y. Tzitzikas, N. Minadakis, Y. Marketakis, P. Fafalios, C. Alloca, M. Mountantonakis, I. Zidianaki. ***MatWare: Constructing and Exploiting Domain Specific Warehouses by Aggregating Semantic Data***. In 11th Extended Semantic Web Conference (ESWC'14), Anissaras, Crete, Greece, May 2014.

[9]. M. Mountantonakis, C. Allocca, P. Fafalios, N. Minadakis, Y. Marketakis, C. Lantzaki, Y. Tzitzikas. ***Extending VoID for Expressing the Connectivity Metrics of a Semantic Warehouse***, In 1st International Workshop on Dataset PROFILing & fEderated Search for Linked Data (PROFILES'14), co-located with ESWC'14, Anissaras, Crete, Greece, May 2014.

[10]. P. Fafalios, M. Baritakis, and Y. Tzitzikas. ***Configuring Named Entity Extraction through Real-Time Exploitation of Linked Data***. 4th International Conference on Web Intelligence, Mining and Semantics (WIMS'14). Thessaloniki Greece, June 2014.

[11]. P. Fafalios and Y. Tzitzikas. ***Post-Analysis of Keyword-based Search Results using Entity Mining, Linked Data and Link Analysis at Query Time***. 8th IEEE International Conference on Semantic Computing (ICSC 2014). Newport Beach, California, USA. June 2014.

[12]. P. Fafalios and Y. Tzitzikas. ***Exploratory Professional Search through Semantic Post-Analysis of Search Results***, In “Professional Search in the Modern World”, Lecture Notes in Computer Science (LNCS), Springer, 2014 (accepted for publication as a State-of-the-Art volume in LNCS).

[13]. P. Fafalios and P. Papadakos. ***Theophrastus: On Demand and Real-Time Automatic Annotation and Exploration of (Web) Documents using Open Linked Data***. Journal of Web Semantics: Science, Services and Agents on the World Wide Web, Elsevier (ISSN: 1570-8268), 2014.

5 Links

- iMarine
    - http://www.i-marine.eu/
- MarineTLO
    - http://www.ics.forth.gr/isl/MarineTLO/
    - http://www.ics.forth.gr/isl/ontology/MarineTLO

    - http://wiki.i-marine.eu/index.php/Top_Level_Ontology
- MarineTLO-based warehouse
    - http://wiki.i-marine.eu/index.php/MarineTLO-based_warehouse
    - https://i-marine.d4science.org/group/biodiversitylab/sparql-search
- MatWare
    - http://www.ics.forth.gr/isl/MatWare/
- XSearch
    - http://www.ics.forth.gr/isl/X-Search/
    - http://wiki.i-marine.eu/index.php/XSearch
    - http://139.91.183.72/x-search/
    - http://139.91.183.72/x-search-metadata-groupings/
    - https://i-marine.d4science.org/group/marinesearch/advanced
- XSearch-API
    - https://gcube.wiki.gcube-system.org/gcube/index.php/XSearch-Service-API
- XLink
    - http://www.ics.forth.gr/isl/X-Link/
    - http://wiki.i-marine.eu/index.php/X-Link