W3C® > Semantic Web Use Cases and Case Studies

Case Study: Online Resource for Information on Aquatic Sciences

Marta González Rodríguez, Spain

July 2009

tecnalia
Corporación Tecnológica

Introduction

Aquaring is a portal on European resources of the Aquatic world, offering a multilingual semantic search engine, a semantic tag cloud adapted to user context, map, exhibitions, etc. The initial aim was to offer a multilingual and unique access point to a heterogeneous and distributed collection of digital resources. This collection was provided by European expert organisations, keeping in mind the possible future integration into Europeana: the European Digital Library.

A careful analysis of the digital collection reveals that the Aquatic world is a complex knowledge domain, due to its heterogeneity and dynamicity. The main features of these digital resources are their multilingualism, different formats and terminology, distributed locations, differences in their targeted audience, and a continuously growing knowledge domain.

The knowledge domain covered by the digital resources ranges from marine species to marine biology, and it includes environment, location, regulations, impact, and educational issues. Also, all this knowledge is dynamic due to results of scientific research. For instance, FishBase (http://www.fishbase.org/), an online resource with practically all fish species known to science, contained 25,000 species in October 2006 and, two years and a half later, this number was 31,200; and, in aquatic environments, fishes are not the unique living beings.

Once the search for existing ontologies in the field was initiated, it turned out that the Aquatic Domain, in general, is poorly covered by real ontologies, while a great number of well-known and reliable thesauri exist—ASFA, GEMET, AGROVOC,…— provided by reliable sources such as the FAO (Food and Agriculture Organisation of the United Nations) or the European Environment Agency. Unsuccessful initiatives tried to integrated these thesauri into a single ontology, resulting in an enormous ontology, hard to manage and to maintain.

Whenever possible, only generally accepted ontologies, developed by authoritative institutions in the relevant domains, were considered (taking into account their correctness, multilingualism, extend of usage, and scientific acceptance) in order to apply only state-of-the-art knowledge models. Following this analysis, four existing orthogonal multilingual ontologies were selected and other three were created; one from a thesaurus, other from a database view, and the last one from scratch exploiting the partners' scientific knowledge. These seven ontologies are used for the semantic annotation of the digital resources, each of them covering a sub-domain of the Aquatic domain: Biological species, Marine Biology, Vessels, Fishing Areas, Land Areas, Habitats, and Education.

However, even this was not enough; indeed, important gaps were detected in the knowledge domain covered by the adopted ontologies. For instance, *the biological species* ontology provided by FAO only covers species for trade, leaving out the rest of species.

The Solution

An ad-hoc **metadata editor** was developed that allowed the use of several ontologies and free tags during annotation time. The supported metadata schema is DCMI (Dublin Core Metadata Initiative) Metadata Terms, thanks to its capacity to host ontology instances and free tags, documents translations management, expression in RDF, and standardisation of the Dublin Core Metadata Element Set (DCMES).

The use of free tags as a means to exploit the annotators' relevant scientific background knowledge is justified by the need of the management of the continuous growth of the knowledge domain. The use of free tags should comply with a soft constraint: each free tag should be linked to the sub-domain it adds to, yielding an approach that has been referred to as *"hierarchical free tags"*. Annotators were advised to look up already existing terms in reliable and well known thesauri before creating a new free tag.

In the specific case of biological species, an annotator dealing with a resource about a concrete species that is not considered by the *biological species* ontology can use the scientific name of the species as a free tag to annotate the resource. Following this approach the digital resources are annotated in a mixed fashion, using seven orthogonal ontologies, and creating a collection of free tags classified according to the knowledge sub-domain covered by each of the seven orthogonal ontologies.

In order to provide services over the annotated digital resources, an ontology learning process appeared as a requirement; learning from the annotations it creates a unified ontology that relates the instances according to predefined relationships. Several services are offered over the **unified ontology** and the annotated digital resources, such as a customised web portal, multilingual semantic search engine, and semantic tag cloud adapted to user context, map, exhibitions, etc.

Figure 1: Semantic Tag Cloud (a larger version of the image is also available)

An **ontology editor** for non experts in ontology engineering was developed offering free tags, translation and relationships editing. In the specific case of the free tags attached to the biological species domain, the ontology editor offers a connection to the uBIO (Universal Biological Indexer and Organizer) taxonomical web service, that returns the set of common (or vernacular) names of the species for every language in response to the scientific name. The data retrieved through uBIO is incorporated into the unified ontology. The editor is then able to place the species under the corresponding family.

Multilingual Approach

As stated by our initial goal, ie, to "offer a multilingual and unique access point to a heterogeneous and distributed collection of digital resources about the Aquatic world", the digital contents to be annotated, as well as the services to be provided from the web site, should support multilingual description and retrieval.

To achieve this challenge a semantic multilingual framework was designed and developed:

- All the **ontologies** used during annotation were multilingual, supporting the translation of each instance in the six allowed languages, and open to the incorporation of new languages. The unified ontology, result of the ontology learning process, also follows the same multilingual structure of the source ontologies.
- The **metadata editor** developed in the project applied the metadata record and instances approach of Dublin Core: the main record describes a resource in its original language and format, whereas the related instances describe the different translations of the original resource, even when the translations are not in the same format as the original resource. The general instance creation rules are:
  - It is mandatory to create at least one instance attached to each metadata record.
  - It is mandatory that an annotation of the resource in English exists. It could be the metadata record or an instance.
- The **ontology editor** allows translating the unified ontology instances when no previous translation was present in the corresponding original ontology or when the instances are derived from free tags.

The services, accessible at the web site, exploit this framework. For instance, the semantic search engine is provided with an independent language search machine, it uses concepts (=meaning), and not its representation in a language, to look for the most relevant resources for a search term.

Following this approach, using, e.g., Spanish as the search language the system can retrieve a document written in Lithuanian in case no Spanish translation exists. But, in this case, the system will provide the English description of the Lithuanian document (thanks to the instance creation rules explained above). Why? Because if the resource and/or the authors are relevant then a translation can be asked for or at least the user knows that the resource exists somewhere. Also, at any time the user can tell the system to retrieve only resources written in a specific language.
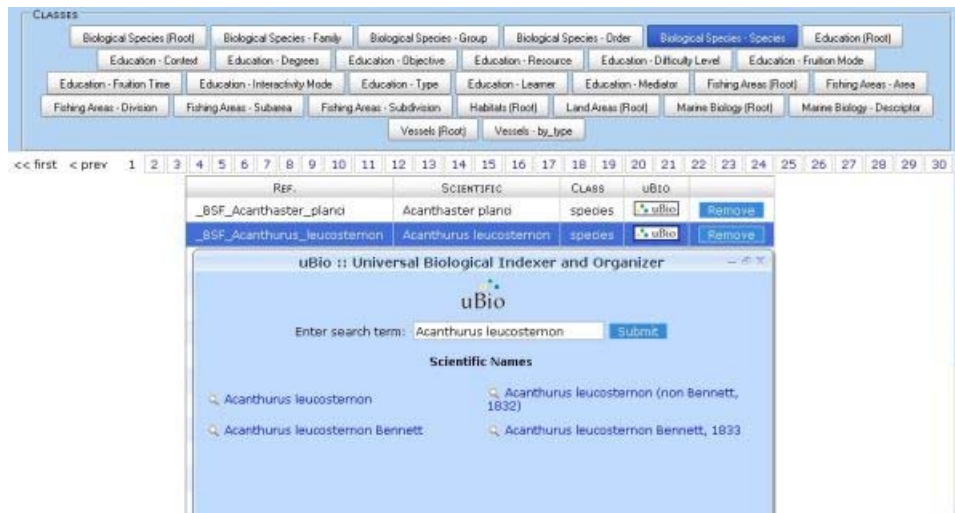
Figure 2: Ontology editor (a larger version of the image is also available)

Key Benefits of Using Semantic Web Technology

- To create a unique access point to aggregate heterogeneous and distributed resources
- To reuse and align existing ontologies
- It offers a way to exploit scientific background knowledge not previously expressed in existing ontologies
- To easily manage multilingual and heterogeneous information
- To share data between web portals
- To offer added value multilingual search and browsing tools over the aggregated data collection
- To customise the information, i.e., the portal according to the visitor context: language, audience type, preferences.