Semantic framework for complex knowledge domains

Marta González
Robotiker-Tecnalia
Parque Tecnológico Edif 202
Zamudio, 48170, Spain
+34 946002266
marta@robotiker.es

Stefano Bianchi
Softeco Sismat S.p.A.
Via de Marini 1, WTC Tower
Genoa, 16149, Italy
+39 010 6026 1
stefano.bianchi@softeco.it

Gianni Vercelli
DIST- Università di Genova
Via All'Opera Pia 13
Genoa, 16145, Italy
+39 010 3532814
gianni.vercelli@unige.it

ABSTRACT
Large amounts of scientific digital contents, potentially available for public sharing and reuse, are nowadays held by scientific and cultural institutions which institutionally collect, produce and store information valuable for dissemination, work, study and research. Semantic technology offers to these stakeholders the possibility to integrate dispersed heterogeneous yet related resources and to build value-added sharing services (overcoming barriers such as e.g. knowledge domain complexity, different classification, language, data format, localization) by exploiting semantic annotation and building virtual content aggregation schemas on top of distributed collections. Applications in real cases are anyway often hampered by difficulties related to the proper formalization of complex scientific knowledge (ontology engineering) and the classification of contents (semantic annotation). This paper illustrates the lessons learnt in applying the Semantic Web specifications to support content management and sharing in complex knowledge domains and provides practical example of application in an EC-funded project.

Categories and Subject Descriptors
Semantic web in use, science, complex knowledge domain

Keywords
Mixed annotation, hierarchical free tags, ontology learning, ontology merging, semantic services.

1. INTRODUCTION
Internet and digitization facilities make large amounts of contents available for public sharing and reuse: this is particularly valid for scientific organizations which institutionally collect, produce and store information valuable for work, study and research in many different contexts. Nevertheless, complexity of the knowledge domain, lack of unified classification and vocabulary (differences in terminology, syntax and semantics), language, heterogeneous data formats and distributed physical location often hamper access to and use of contents.
This poster describes a semantic framework designed to ease content management and sharing functionalities in complex application domains by means of explicit knowledge formalization. The framework is the result of European project AquaRing - "Accessible and Qualified Use of Available Digital Resources about Aquatic World In National Gatherings" – (partially funded by the EC eContentplus programme) which is aiming to set up a European cross-border digital collection space in the aquatic environment by integrating distributed digital collections provided by several European science centres, aquaria and natural history museums [1].

2. FORMALIZING A COMPLEX ENVIRONMENT
At the beginning of the project one of the first tasks launched was the analysis of the aggregated collection of digital resources to be available on the AquaRing web site. From this analysis it was inferred that the AquaRing knowledge domain covers the following subjects: Aquatic and Marine Activities and Technology, Marine Biology / Aquatic Sciences / Environment, Marine Culture and Leisure, Law, Education and Awareness.
The main obstacle appeared when trying to formalize a complex knowledge environment (vast knowledge domain + a certain level of dynamicity) such as the aquatic domain. The topics covered represent 75% of the earth (water) and the interactions with the remaining 25% that includes the biological species (e.g. humans who use that 75% of the earth as a working place (e.g. fishermen), for leisure, shipping and industry using marine products for food processing, for substance extractions, etc); also considering complementary and in some cases essential aspects such as the geographical location, the habitats, the environmental matter (e.g. pollution, climate change,...), educational issues, law.

The dynamicity of this knowledge domain is clearly demonstrated by Fishbase, the online resource containing the list of fish species, which held 25,000 species in October 2006 and which at the moment of writing this poster classifies 30,300 species. Or by the recent discovery, in May 2008, of a 380-million-year-old fossil fish that shows an unborn embryo and umbilical cord.

Once the search for existing ontologies in the domain was started, it came out that the Aquatic Domain in general is poorly covered by real ontologies, while a great number of thesauri exist (e.g. ASFA[2], GEMET[3], AGROVOC[4]) provided by reliable

Figure 1. Oldest Live-Birth Fossil Found

sources such as FAO (Food and Agriculture Organization of the United Nations) and the European Environment Agency.

Whenever possible, only largely accepted ontologies developed by authoritative institutions in the domains addressed were considered (acting as selection factors their correctness, multilingualism, use extent and scientific acceptance) so to apply only state-of-the-art knowledge models. Seven ontologies were adopted for semantic annotation of AquaRing resources.

AquaRing had to wait for the results of another European project, NeOn[5], in September 2007 to obtain a first set of ontologies provided by FAO's use case: Biological Species, Fishing Areas, Land Areas and Vessels.

Other two ontologies were programmatically developed using as source an XML version of ASFA (Marine Biology) thesaurus from FAO and manually modified afterwards. A database view of EUNIS[6] Habitats classification provided by the European Environment Agency was the source for the Habitats ontology development. Finally, as no appropriate ontology was available to cover the educational part of the AquaRing knowledge domain, the EDUcation ontology was developed merging different reference models such as LOM, DC-Ed AP and LRE among others.

The scientists approved the ontologies adopted, but knowledge gaps were still detected. The solution then adopted a formal/informal annotation schema using the different specialized ontologies plus an additional hierarchical free tagging approach, generating a unique ontology by means of an ontology learning approach. The hierarchical free tagging approach exploits the relevant background knowledge of scientific partners involved in resources annotation to complement and extend the domain semantic description by means of free tags (free keywords) referred to the knowledge area covered by each ontology an using reliable thesauri as source, whenever possible.

An ontology-learning technique automatically merges ontologies and hierarchical free tags using content annotation as information source and exploiting the relations that are implicitly established when ontologies instances are used to annotate contents. An ontology editor tool was developed in order content providers access to the generated ontology (see Figure 2) and modifies it (terms translations, relationship pruning, free tags editing).

In the case of free tags related to biological species a connection to uBio (Universal Biological Indexer and Organiser) is provided allowing the creation of a new biological specie instance and associating it to the corresponding family and vernacular names.

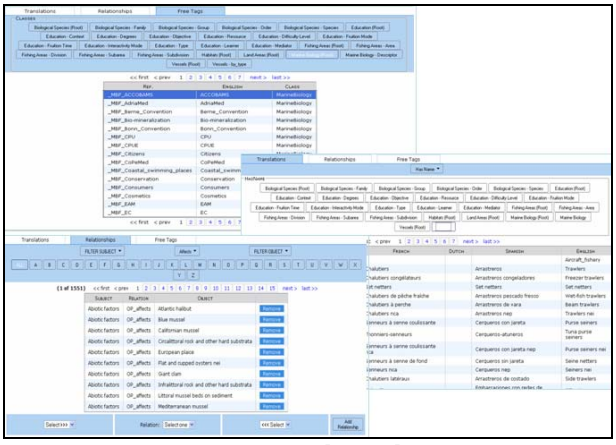

Figure 2: Ontology Editor

Multilingual semantic services, addressed to the AquaRing site visitors and customized depending on visitor profile, are being developed over the resulting ontology, as the semantic search engine, semantic content browser, semantic knowledge map and semantic virtual exhibition.

3. CONCLUSIONS

The semantic framework presented in this poster is flexible, extendable and applicable to complex knowledge domains,

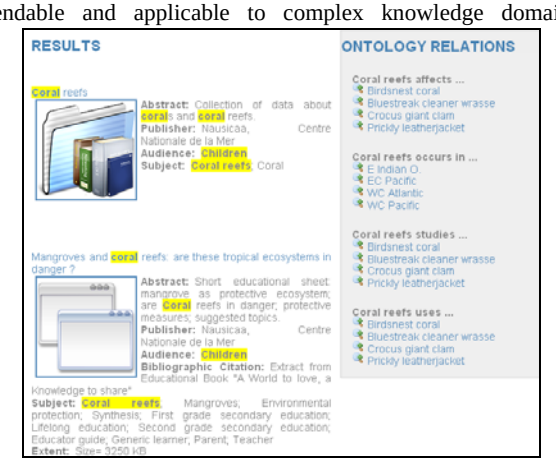

Figure 3: Semantic search engine

allowing seamless integration of distributed resources into a single virtual collection and its exploitation through multilingual semantic value-added services (multilingual semantic search, semantic content browser, etc.).

The most valuable result is the interoperability of resources according to an assessed knowledge model deriving from state-of-the-art ontologies plus informal annotation. In all cases, value-added services have been designed and extensively tested, providing positive feedback on how semantics can improve the way information is generated, navigated and searched. Future work will consider the improvement of the degree of automation of semantic annotation, thus facilitating the integration of new content providers, one of the AquaRing project final purposes as well as the interoperability with the European Data Library.

4. ACKNOWLEDGMENTS

The framework described in this paper has been partially funded by the European Commission within the eContentplus programme. The authors would like to thank all partners involved in the AquaRing project for their precious support.

5. REFERENCES

[1] AquaRing Project. http://www.aquaringweb.eu

[2] ASFA Aquatic Sciences and Fisheries Abstracts. (http://www.fao.org/fishery/asfa)

[3] GEMET General Multilingual Environmental Thesaurus (http://glossary.eea.europa.eu/EEAGlossary/G/General_Multilingual_Environmental_Thesaurus

[4] AGROVOC Multilingual Agricultural Thesaurus (http://www.fao.org/aims/ag_intro.htm)

[5] NEON Project: Lifecycle Support for Networked Ontologies (http://www.neon-project.org/web-content/)

[6] uBio. (http://www.ubio.org