This is the final version of the following article:

León Araúz, Pilar, Pedro Javier Magaña, and Pamela Faber. 2009. Managing inner and outer overinformation in EcoLexicon: an environmental ontology. In *8ème conférence internationale Terminologie et Intelligence Artificielle*. Toulouse.

You can find more articles authored by LexiCon Research Group members at <http://lexicon.ugr.es>.

Managing inner and outer overinformation in EcoLexicon: an environmental ontology

PILAR LEÓN ARAÚZ
PEDRO JAVIER MAGAÑA REDONDO
PAMELA FABER
Universidad de Granada

Abstract

EcoLexicon is a Terminological Knowledge Base (TKB) on environment enhanced by both linguistic and knowledge representation techniques. Our TKB is primarily hosted in a relational database (RDB) but at the same time integrated in an ontological model. Ontologies provide a suitable schema for sharing and reusing semantic resources. Nevertheless, before considering the interoperability of other environmental knowledge-based projects, we must first deal with overinformation in our RDB. The final aim of EcoLexicon is to guide the knowledge acquisition process of end users. However, such a wide domain as the environment has caused an information overload. Contextual constraints seem a plausible way to structure knowledge in a similar way to how things relate in the real world. Thus, the global domain is divided into different sub-domains according to multidimensional concepts. That means that concepts' dimensions are only activated when particular contexts arise. On the other hand, other environmental sources, such as those offered by ENVO and SWEET ontologies, provide us with the possibility of widening knowledge according to the Semantic Web initiative. Linked Data provide a useful and easy mechanism for the interaction of current infrastructures keeping them as independent resources.

Keywords : ontologies, overinformation, contextual constraints, linked data.

1. Introduction

EcoLexicon[1] is a Terminological Knowledge Base (TKB) on environment enhanced by both linguistic and knowledge representation techniques. Our TKB is primarily hosted in a relational database (RDB) but at the same time integrated in an ontological model. TKBs can find in ontologies a powerful representational model, as they add the semantic expressiveness lacking in RDBs. This enables potential queries to be richer, since reasoning techniques can be applied to extract implicit information. In turn, the design of ontologies can also benefit from the theoretical background of linguistics, especially from cognitive approaches.

Our TKB is structured around an Environmental Event (EE) which provides the conceptual underpinnings for the location of conceptual sub-hierarchies (Faber et al. 2006).

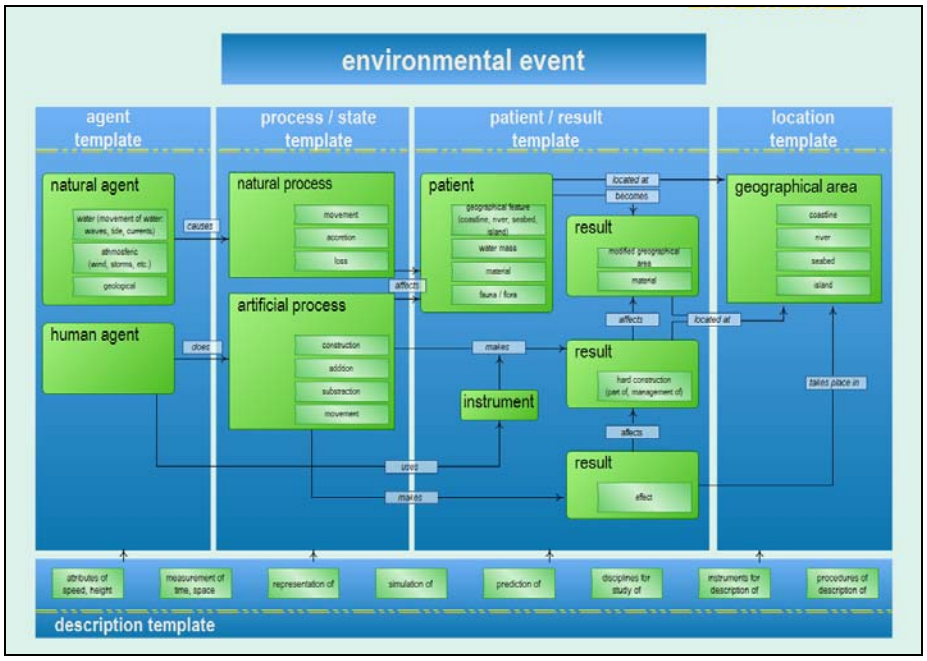

Fig. 1 – Environmental Event

The EE is based on the cognitive linguistics view of frames and semantic roles. According to Fillmore and Atkins (1992), frames are defined as a network of concepts related in such a way that one concept evokes the entire system. Consequently, the upper-level classes in our ontology correspond to the basic semantic roles described in the EE (AGENT-PROCESS-PATIENT-RESULT-LOCATION), all derived from a general knowledge hierarchy. This structure enables users to gain a better understanding of the complexity of environmental events, since they give a process-oriented general overview of the domain.

On the other hand, ontologies provide a suitable schema for sharing and reusing semantic resources. According to the Semantic Web initiative, our TKB can benefit from previous works in this field. Furthermore, reasoning techniques can be applied to discover and extract new information, thus increasing the richness of potential queries. Nevertheless, many projects have not been conceived as such from the beginning. As a result, these legacy systems need to find some mechanism to get integrated in ontological models.

Most such information comes from relational databases (RDB), as it is the case of EcoLexicon. In our approach, we emphasize the importance of storing semantic information in the ontology, while leaving the rest in the relational database. In this way, we can continue using the new ontological system, while at the same time feeding the database. This entails linking RDB stored information with an ontological system.

Once information can be accessed by using ontological resources, it is easier to connect it with other environmental resources. Reusability is often based on data merging, but that would lead to a heterogeneous blending of diverse data founded on very different aims. Linked data is an innovative approach facing this problem. It uses Semantic Web

technologies to publish structured data and, at the same time, set links between data from one data source to data within other data sources (Heath et al. 2008), but keeping them as independent resources.

Nevertheless, before considering the interoperability of other environmental knowledge-based projects, we must first deal with overinformation in our own TKB.

2. EcoLexicon: a context-based resource

The final aim of EcoLexicon is to guide the knowledge acquisition process of end users, both for communicative and cognitive purposes. This involves the design of a user-friendly interface where concepts are related in a meaningful way. Based on the EE, conceptual networks in EcoLexicon are structured around a set of different vertical and horizontal relations, some of which are domain-specific. However, such a wide domain as the ENVIRONMENT has caused an information overload:

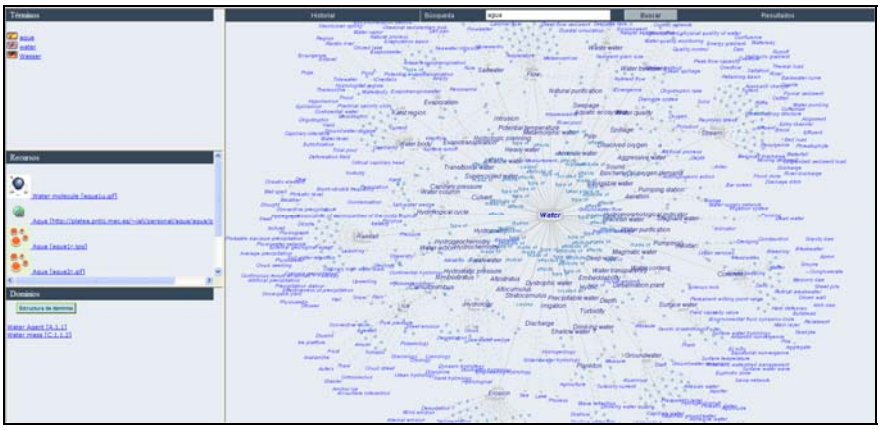

Fig. 2 – Information overload

Obviously, users would not acquire any meaningful knowledge if all dimensions of WATER were shown at the same time, as in figure 2. Overinformation results from a high degree of multidimensionality, which is especially prevalent in what we call *versatile concepts*. Versatile concepts, as WATER, are usually general concepts involved in a myriad of events. For instance, in figure 2, WATER is linked to the same extent to diverse natural and artificial processes, such as EROSION or DESALINATION. However, WATER will never activate those relations at the same time, as they evoke completely different situations, where WATER is an *agent* in the first one and a *patient* in the second one.

When it comes to hyponymy, the incompatibility among conceptual facets is even more outstanding. Multidimensionality can occur at an intracategorial level, based on the internal structure of concepts. This means that a concept may be classified according to different perspectives but still in the same context, causing the well-known phenomenon of multiple inheritance. For example, different dimensions, like *salinity* or *location* can be hydrological parameters to classify diverse WATER subtypes, such as FRESH WATER, BRACKISH WATER and SALT WATER or SURFACE WATER and GROUNDWATER. These dimensions are compatible enough to share the same conceptual network, because they all describe physical properties of WATER. Moreover, both dimensions are also related, since FRESH, BRACKISH or SALT WATERS can be at the same time either SURFACE or GROUNDWATER.

Nevertheless, hyponymic dimensions show a different nature depending on the external situations where a concept may appear. In that sense, even though WATER subtypes like PRECIPITABLE WATER, DRINKING WATER and NAVIGABLE WATER represent the same dimension *function*, they are not strict coordinate concepts. They only share the same hyperonym, but they will never evoke a common scene. In this line, Barsalou (2005) states that a given concept produces many different situated conceptualizations, each tailored to different instances in different settings. Thus, context can be said to be a dynamic construct that triggers or restricts knowledge.

Our claim is that any specialized domain contains sub-domains in which conceptual dimensions become more or less salient depending on the activation of specific contexts. This means that concepts tend to be intermingled in different situations, and they are not always related to the same concepts or through the same relations. It seems pretty clear that certain dimensions are only activated when particular contexts arise and, as a result, a more believable representational system should account for re-conceptualization according to the situated nature of concepts.

Frames can thus be applied to sub-hierarchies as well. This is done by dividing the global environmental specialized field in different contextual domains: HYDROLOGY, GEOLOGY, METEOROLOGY, BIOLOGY, CHEMISTRY, ENGINEERING, WATER TREATMENT, COASTAL PROCESSES, NAVIGATION. In this way, context domain membership reconceptualise versatile concepts restricting their relational behaviour.

Contextual constraints are neither applied to individual concepts, since one concept can be activated in different contexts, nor to individual relations, because concepts can make use of the same relations although with different values. Constraints are instead applied to each conceptual proposition. For instance, CONCRETE is linked to WATER through a *part_of* relation, but this proposition is not relevant if users only want to know how WATER naturally interacts with landscape. Consequently, the proposition WATER *part_of* CONCRETE will only appear in an ENGINEERING context.

As a result, when constraints are applied, WATER only shows relevant dimensions for each context domain. In figure 3 WATER is just linked to propositions belonging to the context of GEOLOGY.

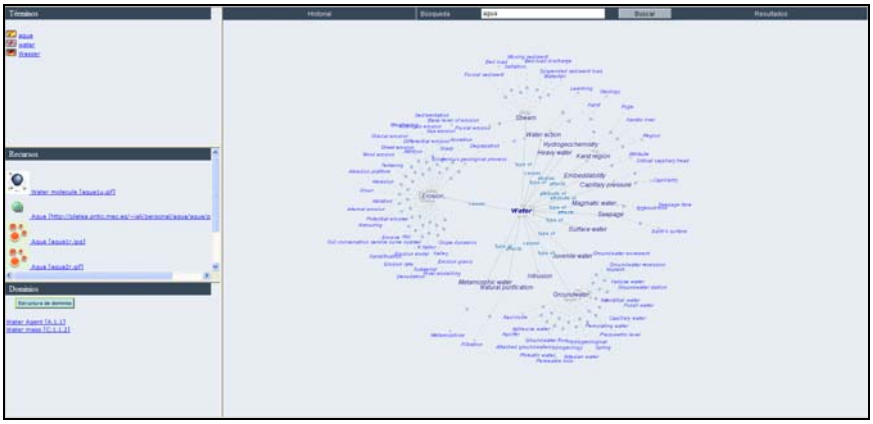

Fig. 3 – WATER in the GEOLOGY context domain

However, in figure 4, the WATER TREATMENT context shows WATER in a new structure with other concepts and relations:

Fig. 4 – WATER in the WATER TREATMENT context domain

Finally, HYDROLOGY is the most prototypical context domain of WATER. This is why the concept still appears embroiled in a complex network, although it is more illustrative than the primitive one:

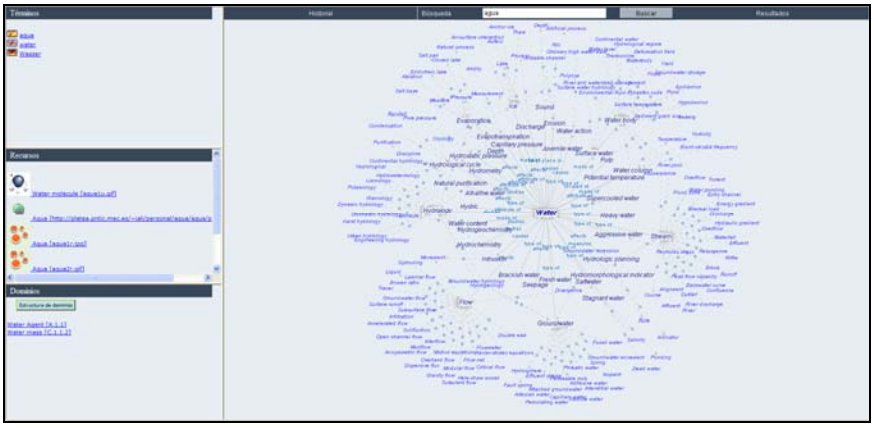

Fig. 5 – WATER in the HIDROLOGY context domain

This means that contexts are dynamic and flexible structures that should evolve over time according to the type and amount of information stored in our TKB. In this way, if many other concepts were added to the HYDROLOGY context, other constraints should be developed. For example, it could be divided into two different domains like GROUNDWATER and SURFACE WATER HYDROLOGY. Dynamism can thus help to avoid potential overinformation caused by new data.

Comparing the context-free WATER network with its context-based representation we can see that reconceptualization affects the relational behaviour of concepts in several ways. First of all, the number of conceptual relations changes from one context to another, as WATER is not equally relevant in all context domains. Furthermore, relation types are also different in each context, which also informs about the changing nature of WATER'S internal structure in each case. For example, in the HYDROLOGY context domain, most relations are

type_of and *attribute_of*, whereas in the GEOLOGY domain, *causes* stands out from the rest. This implies that in geological contexts WATER is a much more active *agent* than in the HYDROLOGY context domain, where the concept is rather more subject to general description. On the contrary, in the WATER TREATMENT domain, *affect* is clearly the main relation, which means that WATER has then a prototypical *patient* role. Finally, WATER is not always related to the same concept types. In HYDROLOGY and GEOLOGY context domains, WATER is mainly linked to *natural* entities or processes, while in the WATER TREATMENT context it is primarily related to *artificial* ones.

However, different context domains can also share certain conceptual propositions. For instance, the HYDROLOGY context has much in common with the GEOLOGY domain. This is due to the fact that multidisciplinarity gives rise to fuzzy category boundaries and, as a result, contextual domains can form their own hierarchical structure. GEOLOGY and HYDROLOGY belong to the same paradigm, as they even constitute the common discipline of hydrogeology. WATER TREATMENT is only partially related to HYDROLOGY through concepts making reference to water analysis, like HYDROMORPHOLOGICAL INDICATOR, HYDROCHEMISTRY or HYDROMETRY. However, it does not share a single concept with GEOLOGY.

On the other hand, overinformation can affect conceptual networks at different stages: (1) when the first concept is versatile, as shown above (2) and when specific concepts are linked to one of the versatile concepts in the first hierarchical level, since the second level will spread the whole first-level network of the versatile concept.

This is why contextual constraints have been applied at all levels, reconceptualising any concept somehow linked to versatile ones, whether they show an information overload or not. In this way, the context-free network of EROSION in figure 6 becomes restricted by the HIDROLOGY context domain shown in figure 7.

Fig. 6 – EROSION in the context-free network

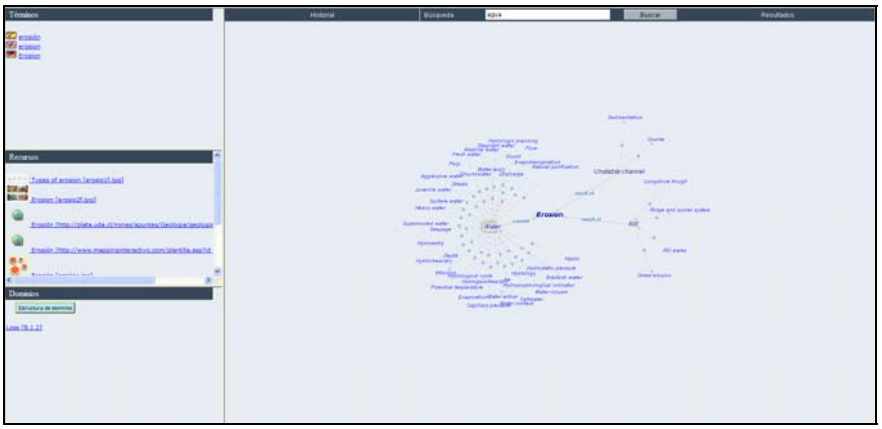

Fig. 7 – EROSION in the HYDROLOGY context domain

3. Linked data: connecting environmental data across the web

As mentioned above, our domain knowledge is represented by using a relational database. This widespread modeling let us do a quick deployment of the platform and feed the system from very early stages. Nevertheless, relational modeling has some limitations. One of the biggest ones is its limited capability to represent real-world entities. Ontologies arose as an excellent alternative, but keeping all the development carried out so far was our priority.

In our approach, we emphasize the importance of storing semantic information in the ontology, while leaving the rest in the relational database. In this way, we can continue using the new ontological system, while at the same time feeding the database. This entails linking RDB stored information with an ontological system.

Nevertheless, this is not an easy task, since both representational models have remarkable differences. In contrast to relational databases, ontologies are highly expressive relational structures where concepts are described in very similar terms to those used by humans. Thus, relational models are suited to organize data structure and integrity, whereas ontologies try to specify the meaning of their underlying conceptualization (Barrasa, 2007).

Consequently, the upper-level classes in our ontology correspond to the basic semantic roles described in the Environment Event (Fig. 8).

These ontological classes are fed through the extraction of stored information in the database. This is done by using the D2RQ tool, which provides a usage scenario where relational databases are maintained as non-legacy applications (Bizer and Seaborne, 2004). D2RQ is a declarative language to describe mappings between both systems. Moreover, these mappings can be conditional, which allows for feeding every class just with its corresponding instances.

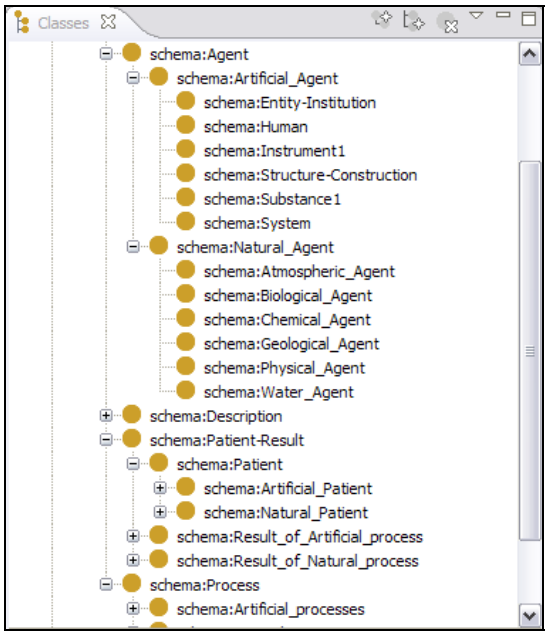

Fig. 8 – Ontological classes

The next step in our development is to connect this environmental resource with other resources within the same domain. Several techniques have been proposed so far. Former attempts deal with automatic mediation algorithms in order to map and merge between ontology schemas (de Brujin et al., 2006). However, a remarkable drawback is that the schemas do not always remain public. On the other hand, many systems provide interfaces to interact with their structured data, the well-known APIs (Application Programming Interfaces). This fact has enabled many developers to combine information from different data sources creating new services known as mashups (Zang et al., 2008). Nevertheless, APIs have several disadvantages. Most of the interfaces are proprietary and it is not possible to set links between data objects.

The Linked Data approach (Berners-Lee, 2006) provides an efficient mechanism to publish structured information on the web while object data from different data sources can be linked at the same time. This is why we think this methodology can be applied with success in EcoLexicon and other data sources in order to create an environmental community within the Linked Data framework. EnvO (Morrison, 2009) and SWEET (Raskin, 2003) data sources are especially interesting to us. SWEET provides a common semantic framework for various Earth science initiatives whereas EnvO aims at developing a common annotation system for any record in the web community that has an environmental component.

This way, we should be able to have statements like the following in the near future:

```
<http://manila.ugr.es/resource/water>
owl:sameAs
http://purl.org/obo/owl/ENVO#ENVO_00002006.
```

This means that water in EcoLexicon (http://manila.ugr.es/resource/water) would be related to the same concept (expressed as ENVO 00002006) in EnvO (http://purl.org/obo/owl/ENVO#ENVO_00002006), enriching our conceptualization with

any other new data included in these resources. In this way, other resources can equally enhance their systems with our information, which would help to build a real community of shared data.

4. Conclusions

Contextual constraints enrich the system from both a qualitative and quantitative standpoint. On the one hand, they structure knowledge in a similar way to how things relate in the real world, as well as in the human conceptual system. On the other hand, conceptual dimensions are noticeably reduced with a coherent and consistent method based on a cognitive approach. As a result, the situated representation of versatile concepts is a viable solution for managing overinformation and at the same time enhancing knowledge acquisition processes.

We have proven that legacy systems can be integrated in the semantic web. Thanks to this achievement, TKBs can also be linked to other resources through new semantic web technologies like linked data. This step is not concluded yet. In the near future we plan to link EcoLexicon to EnvO and Sweet ontologies extensively. However, the success of this approach will largely depend on the proliferation of other shared initiatives.

Notes

1. http://ecolexicon.ugr.es

Acknowledgements

This research has been partially supported by project FFI2008-06080-C03-01/FILO, from the Spanish Ministry of Science and Innovation and project P06-HUM-01489, from the Andalusian Regional Government. Pedro Magaña holds a FPU scholarship from the Spanish Ministry of Science and Innovation.

References

BARRASA, J. (2007). Modelo para la Definición Automática de Correspondencias Semánticas entre Ontologías y Modelos Relacionales (PhD dissertation, Universidad Politécnica de Madrid).

BARSALOU, L.W. (2005). Situated conceptualization. In H. Cohen. & C. Lefebvre. Eds. *Handbook of Categorization in Cognitive Science* p. 619-650. St. Louis.

BERNERS-LEE, T. (2006). Linked Data. W3C Design Issues.

BIZER, C. & SEABORNE, A. (2004). D2RQ-Treating Non-RDF Databases as Virtual RDF Graphs, *Proceedings of the 3rd International Semantic Web Conference (ISWC2004)*.

DE BRUIJN, J., EHRIG, M., FEIER, C., MARTÍN-RECUERDA, F., SCHARFFE, F., AND WEITEN, M. (2006). Ontology Mediation, Merging, and Aligning.

FABER, P., MONTERO MARTÍNEZ, S., CASTRO PRIETO, M.R., SENSO RUIZ, J., PRIETO VELASCO, J.A., LEÓN ARAÚZ, P., MÁRQUEZ LINARES, C., VEGA EXPÓSITO, M. (2006). Process-oriented terminology management in the domain of Coastal Engineering, *Terminology* 12: 2, p.189-213.

FILLMORE, C.J., ATKINS, B.T.S. (1992). Towards a frame-based lexicon: the semantics of risk and its neighbours. In A. LEHRER & E. F. KITTAY. Eds. *Frames, Fields and Contrasts*, Hillsdale, New Jersey: Lawrence Erlbaum Associates. p. 75-102.

MORRISON, N. (2009). EnvO - Development of an Environmental Ontology. *Proceedings of Towards eEnvironment. Opportunities of SEIS and SISE: Integrating Environmental Knowledge in Europe*.

RASKIN, R. & PAN, M. (2003). Semantic Web for Earth and Environmental Terminology (SWEET). *Proceedings of the Workshop on Semantic Web Technologies for Searching and Retrieving Scientific Data.*

ZANG, N., ROSSON, M. B., AND NASSER, V. (2008). Mashups: who? what? why? In *CHI '08: CHI '08 extended abstracts on Human factors in computing systems*, p. 3171-3176.