Managing inner and outer overinformation in EcoLexicon: an environmental ontology

Pilar León Araúz[1], Pedro Javier Magaña Redondo[2] and Pamela Faber[1]

[1] Department of Translation and Interpreting, University of Granada
{pleon, pfaber}@ugr.es
[2] Department of Computer Science and Artificial Intelligence,
University of Granada,
pmagana@decsai.ugr.es

Abstract : EcoLexicon is a Terminological Knowledge Base (TKB) on environment. 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 semantic resources. Nevertheless, before considering the interoperability of other environmental knowledge-based projects, we must first deal with overinformation in our RDB. Such a wide domain as the environment has caused an information overload and contextual constraints seem a plausible way to structure knowledge in a similar way to how things relate in the real world. The global domain is divided into different sub-domains according to multidimensionality. That means that concepts' dimensions are only activated when particular contexts arise. On the other hand, other environmental sources can be used to widen 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.

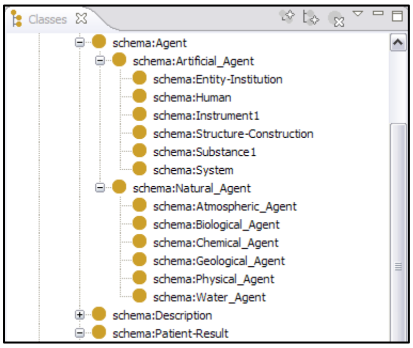
1 Introduction

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. Ontologies enable 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) based on the cognitive linguistics view of frames and semantic roles. Fillmore and Atkins (1992) define frames as a network of concepts related in such a way that

[1] http://manila.ugr.es

one concept evokes the entire system. According to our corpus-based analysis (Faber et al., 2006), the underlying structure of the entire environmental domain can be encoded in various prototypical frames. Consequently, the upper-level classes in our ontology correspond to basic semantic roles like AGENT, PROCESS, PATIENT, RESULT and LOCATION.



```
⊟ schema:Agent
    ⊟ schema:Artificial_Agent
        schema:Entity-Institution
        schema:Human
        schema:Instrument1
        schema:Structure-Construction
        schema:Substance1
        schema:System
    ⊟ schema:Natural_Agent
        schema:Atmospheric_Agent
        schema:Biological_Agent
        schema:Chemical_Agent
        schema:Geological_Agent
        schema:Physical_Agent
        schema:Water_Agent
⊞ schema:Description
⊟ schema:Patient-Result
```

Fig. 1 – Ontological classes

Ontologies provide a suitable schema for sharing and reusing semantic resources making them manageable. According to the Semantic Web initiative, our TKB can benefit from previous works in this field. This could enrich our system with new information, complementing our TKB from a different perspective or even with other contents, such as real-world geographical instances. Nevertheless, information overload not only occurs when interconnecting different systems. 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. However, such a wide domain as the ENVIRONMENT has caused an information overload.

Obviously, users would not acquire any meaningful knowledge if all dimensions of WATER were shown at the same time (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.

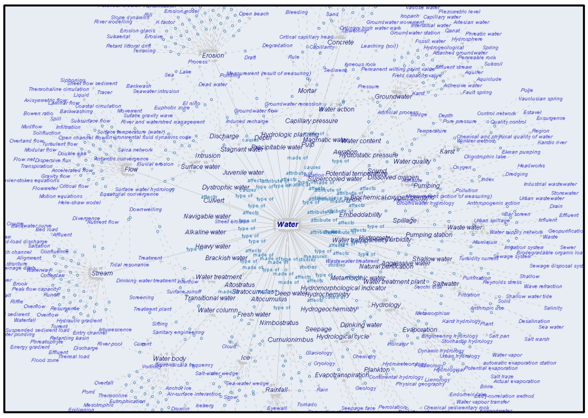


Fig. 2 – Information overload

When it comes to hyponymy, the incompatibility among conceptual facets is even more outstanding. Multidimensionality can usually 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. 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.

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. Frames can thus be applied to sub-hierarchies as well. This is done by dividing the global environmental specialized field in different contextual domains according to corpus-based data: HYDROLOGY, GEOLOGY, METEOROLOGY, BIOLOGY, CHEMISTRY, ENGINEERING, WATER TREATMENT, COASTAL PROCESSES, NAVIGATION. In this way, context domain membership reconceptualises 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. However, in figure 4, the WATER TREATMENT context shows WATER in a new structure with other concepts and relations.

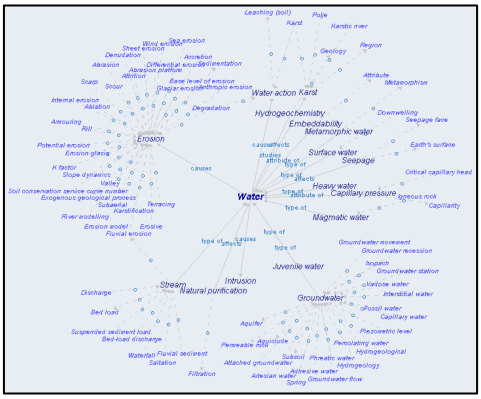


Fig. 3 – WATER in the GEOLOGY context domain

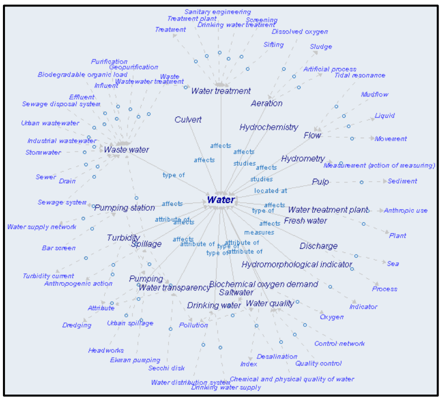


Fig. 4 – WATER in the WATER TREATMENT context domain

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. For example, in the GEOLOGY domain, *type_of* and *causes* stand out from the rest. This implies that in geological contexts WATER is an active *agent* whose multidimensionality is determined by its location or origine (SURFACE WATER, GROUNDWATER, MAGMATIC WATER).

On the contrary, in the WATER TREATMENT domain, *affects* and *attribute_of* are clearly the main relations, which mean that WATER has then a prototypical *patient* role and is generally described in terms of hydro-chemical concepts. Finally, WATER is not always related to the same concept types. In the GEOLOGY context domain, WATER is mainly linked to *natural* entities or processes, while in the WATER TREATMENT context it is primarily related to *artificial* ones.

Reconceptualization does not involve a clear-cut distinction among different context domains, since they can also share certain conceptual propositions. 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. Moreover, they are also dynamic and flexible structures that should evolve over time according to the type and amount of information stored in our TKB. If many other concepts were added to a particular context, new constraints should be developed in accordance with other versatile concepts' special needs. Dynamism would thus help to avoid potential overinformation caused by new data.

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. This is why 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.

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.

Our 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.

Once information can be accessed by using ontological resources, it is easier to connect it with other environmental systems. 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 (Berners-Lee, 2006) is an innovative approach facing this problem. It uses Semantic Web technologies to publish structured data and, at the same time, set links between different data sources, but keeping them as independent resources.

The next step in our development is to connect EcoLexicon with other resources within the same domain. 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 and SWEET ontologies 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 established a sound basis to integrate a legacy system like EcoLexicon 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.

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

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).*

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.