EcoLexicon: an environmental TKB

Arianne Reimerink, Pilar León Araúz, Pedro J. Magaña Redondo
University of Granada
Buensuceso, 11 18002 Granada, Spain
E-mail: arianne@ugr.es, pleon@ugr.es, pmagana@decsai.ugr.es

Abstract

EcoLexicon, a multilingual knowledge resource on the environment, provides an internally coherent information system covering a wide range of specialized linguistic and conceptual needs. Data in our terminological knowledge base (TKB) are primarily hosted in a relational database which is now linked to an ontology in order to apply reasoning techniques and enhance user queries. The advantages of ontological reasoning can only be obtained if conceptual description is based on systematic criteria and a wide inventory of non-hierarchical relations, which confer dynamism to knowledge representation. Thus, our research has mainly focused on conceptual modelling and providing a user-friendly multimodal interface. The dynamic interface, which combines conceptual (networks and definitions), linguistic (contexts, concordances) and graphical information offers users the freedom to surf it according to their needs. Furthermore, dynamism is also present at the representational level. Contextual constraints have been applied to reconceptualise versatile concepts that cause a great deal of information overload.

1. Theoretical premises and macrostructure

EcoLexicon[1] is a multilingual knowledge resource on the environment. So far it has 3,042 concepts and 10,597 terms in Spanish, English and German. The user (translators, technical writers, environmental experts, etc) can access it through a friendly visual interface with different modules devoted to conceptual, linguistic and graphical information.

In its construction great care has been taken to develop an internally coherent system. This terminological knowledge base (TKB) is inspired in the theoretical premises of cognitive linguistics (Barsalou, 2003). The Generative Lexicon theory (Pustejovsky, 1995) has guided our conceptual modelling and description procedures.

At a macrostructural level, all knowledge extracted from a specialized domain corpus has been organized in a frame-like structure or prototypical domain event, namely, the Environmental Event (see figure 1; Faber, 2007; León *et al.*, 2008; Reimerink and Faber, 2009). This prototypical domain event or action-environment interface (Barsalou, 2003) provides a template applicable to all levels of information structuring.

The Environmental Event (EE) is conceptualised as a dynamic process that is initiated by an AGENT (either natural or human), affects a specific kind of PATIENT (an environmental entity) and produces a RESULT in a geographical AREA. These macro-categories (AGENT→ PROCESS → PATIENT/RESULT, and LOCATION) are the semantic roles characteristic of this specialized domain, and the EE provides a model to represent their interrelationships at a more specific level.

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

2. Domain ontology

Data in our TKB are primarily hosted in a relational database (RDB). This widespread modeling allowed for a quick deployment of the platform and fed 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, since natural human implicit knowledge cannot be inferred. This is why ontologies arose as a powerful representational model but, 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 legacy system.

Upper-level classes in our ontology correspond to the basic semantic roles described in the EE (AGENT-PROCESS-PATIENT-RESULT-LOCATION). As shown in figure 2, all classes constitute a general knowledge hierarchy derived from each of them. 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:

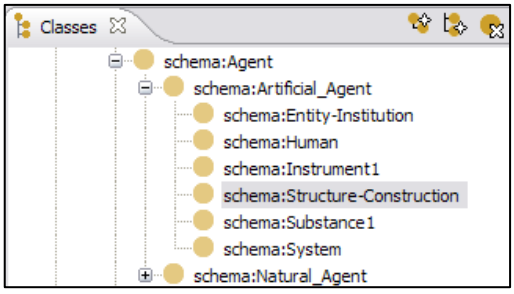


Figure 2. Ontological classes

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 (León Araúz and Magaña Redondo, in press).

3. Conceptual information

Domain ontologies need a set of systematized criteria for conceptual description, because a high degree of generalization jeopardizes accurate inferences. This allows for the application of property characteristics, such as transitivity, symmetry, etc, and restrictions, such as allValuesfrom and someValuesfrom, which enhance user searches (Smith et al., 2004).

Our TKB is designed from a user's perspective, where conceptual information is encoded in fine-grained networks and definitions. The difference between both of them lies in the fact that definitions must be brief statements and only include prototypical information, but the underlying conceptual structure of both is inspired on the same combinatorial criteria, depending on conceptual nature and relational power.

3.1 Conceptual relations

In order to make our TKB internally coherent, we apply the premises of Generative Lexicon to the conceptual relations encountered in the environmental domain. Generative Lexicon (GL) describes lexical items according to their *qualia* structure. The *qualia* structure is composed of the following roles:

1. Formal role: the basic type distinguishing the meaning of a word;
2. Constitutive role: the relation between an object and its constituent parts;
3. Telic role: the purpose or function of the object, if there is one;
4. Agentive role: the factors involved in the object's origins or "coming into being". (Pustejovsky et al. 2006: 3)

GL and *qualia* structure have been successfully applied to the SIMPLE ontology, where an extended version of the *qualia* structure was developed (Lenci et al., 2000) and in the creation of the Brandeis Semantic Ontology (BSO, Pustejovsky et al., 2006). In the BSO, lexical items consist of three major types: entity, event and property. Each of these is divided into three further hierarchies: natural, artifactual, and complex:

1. Natural types: natural kind concepts consisting of reference only to formal and constitutive *qualia* roles;
2. Artifactual types: concepts making reference to purpose, function, or origin.
3. Complex types: concepts integrating reference to a relation between types. (Pustejovsky et al. 2006: 1)

In the same way as the *extension of the qualia structure*, we have related our concept typology, together with their *qualia* roles, to the inventory of conceptual relations elaborated for our specialized domain. Conceptual relations are associated with a particular *qualia* role, depending on each concept type. As a result, the macrostructure and microstructure of all concepts in the domain are represented in terms of these possible combinations (see figure 3).

The most recurrent concepts of the domain (PHYSICAL OBJECTS and PROCESSES) are the ones that can be linked to others through a greater number of relations. However, there are also certain relations exclusive of a single type, such as *attribute_of*, for properties, and *studies* (for sciences and disciplines). For natural PHYSICAL OBJECT types, apart from the relations traditionally linked to formal and constitutive roles, two non-hierarchical relations have been added. The conceptual relations, *has_location* and *made_of*, are necessary in the description of environmental entities. The *material* that an object is *made of* or its *location* are key properties of subordinate concepts, and can even be the most essential feature. For instance, a GROYNE is not a GROYNE if it is not *located in* the SEA.

Concept nature triggers or restricts the activation of a set of possible relations, but at the same time it determines which other concept types can be linked through each relation. For instance, if a PROCESS activates the relation *part_of*, it can only be related to another PROCESS, since this concept type can never be divided into physical entities. However, relations can also constrain the second concept in a proposition. For example, *made_of* can only activate physical concepts from both sides of the proposition.

Those conceptual relations, specifically conceived for our Environmental TKB, can be enhanced by an additional degree of OWL semantic expressiveness provided by property characteristics. This is one of the main advantages of ontologies, making reasoning and inferences possible. For example, *part_of* relations can benefit from transitivity, as shown in figure 4.

In figure 4, a SPARQL query is made in order to retrieve which concepts are *part_of* Concept 3262, which refers to the concept SEWER. On the right side, DRAINAGE SYSTEM is retrieved as a direct *part-of* relation, whereas SEWAGE COLLECTION AND DISPOSAL SYSTEM and SEWAGE DISPOSAL SYSTEM are implicitly inferred through the Jena reasoner.

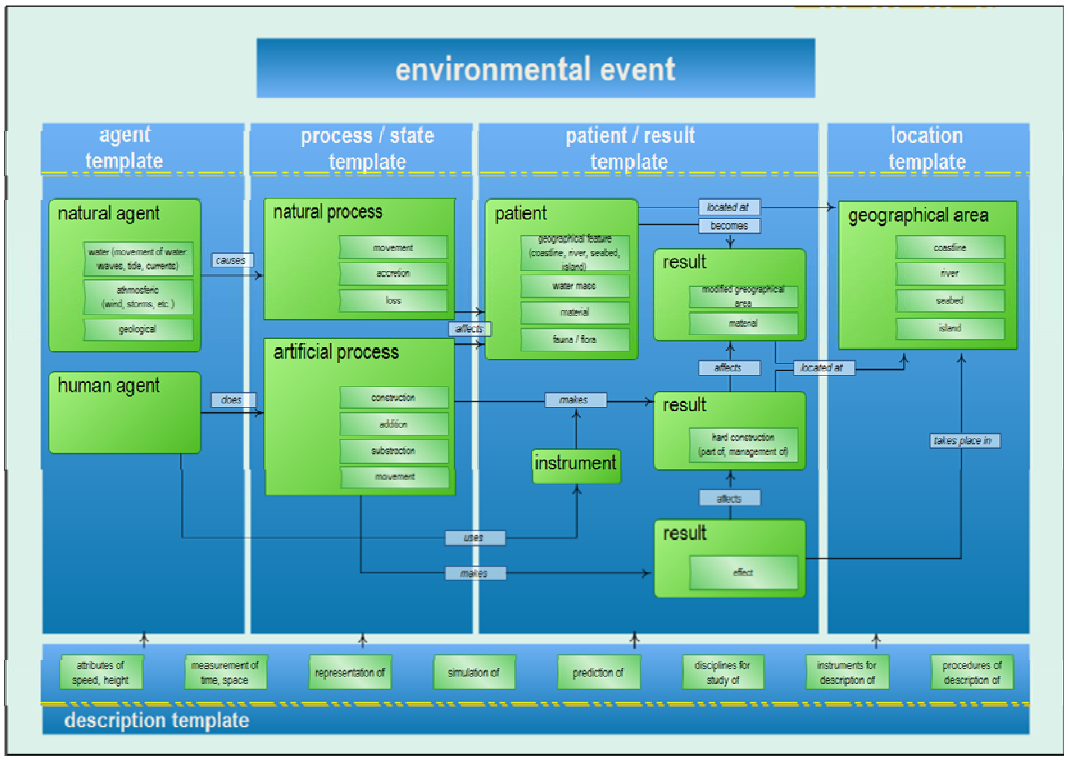


Figure 1. The Environmental Event (EE)

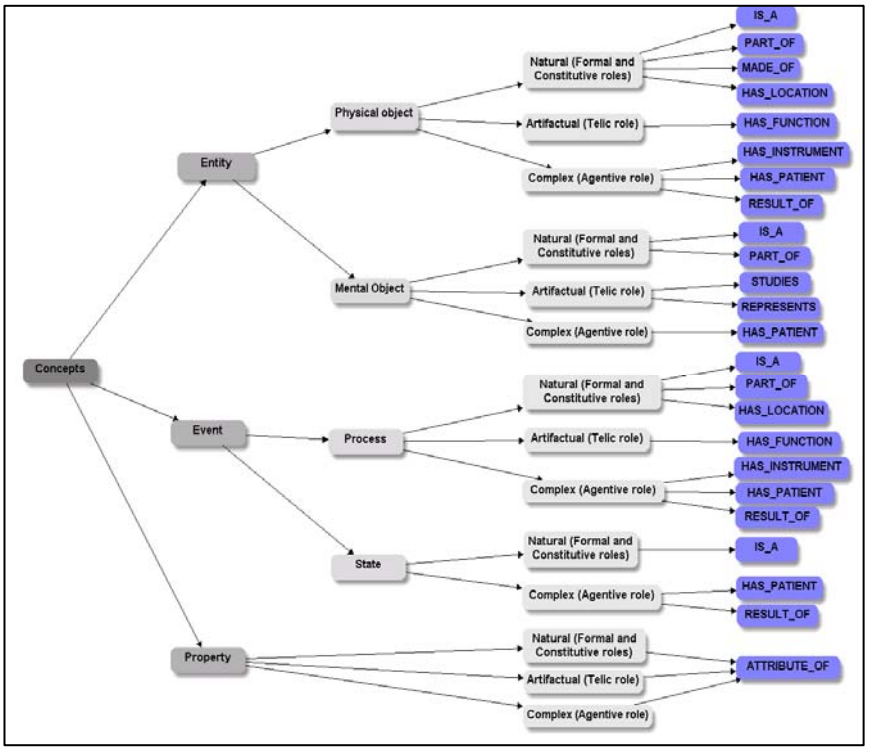


Figure 3. Combination of the concept typology and conceptual relations with Pustejovsky's *qualia* roles.

3.2 Definitions

The notion of *qualia* is also applied to the definitional templates of EcoLexicon. For example, even though a PROCESS can activate all the relations shown in figure 3, the prototypical definitional structure is constrained. A NATURAL PROCESS only activates the formal role, since this is the minimum information needed for description (see table 1). In contrast, an ARTIFICIAL PROCESS activates both the formal *quale* (the action itself) and the constitutive *quale*, since artificial processes are generally composed of several steps or actions (see table 2). Furthermore, an ARTIFICIAL PROCESS always has a purpose (telic *quale*) and in certain engineering operations, an instrument may be used, which would also add the agentive rol.

NATURAL PROCESS: A succession of actions that happen or take place	
	▪ FORMAL ROLE

Table 1. Definitional template of NATURAL PROCESS.

ARTIFICIAL PROCESS: A succession of actions and steps carried out for a specific purpose	
	▪ FORMAL ROLE ▪ CONSTITUTIVE ROLE ▪ TELIC ROLE ▪ (AGENTIVE ROLE)

Table 2. Definitional template of ARTIFICIAL PROCESS.

In EcoLexicon, the linguistic expression of the definitional template results in a definition such as the one shown in Table 3. A PROCESS like DREDGING shows all of the four roles with different specialized fillers and the same category template is applied to other processes in the same paradigm, such as PIPING, PUMPING, etc.

DREDGING
Removal of beach material from underwater [FORMAL ROLE] by pumping, extracting and piping it [CONSTITUTIVE ROLE] by means of a dredger [AGENTIVE ROLE] in order to maintain water depths in rivers, canals or harbours and to obtain material for construction or beach nourishment [TELIC ROLE].

Table 3. Definition of DREDGING

4. Linguistic and graphical information

Context-based information is not only included in the representation of conceptual relations, but also expressed linguistically. The TKB provides the user with the following additional information: linguistic contexts, concordances and images.

Linguistic contexts help the user achieve a level of understanding of a specialized domain. The linguistic contexts included in the TKB go beyond the relations expressed in the definition. In Table 4, for example, GROYNE is not only defined as a COASTAL DEFENSE STRUCTURE. Other relevant information is included as well: they are cost-effective and many coastal communities prefer other solutions.

Groynes are extremely cost-effective coastal defense measures, requiring little maintenance, and are one of the most common coastal defense structures. However, groynes are increasingly viewed as detrimental to the aesthetics of the coastline, and face strong opposition in many coastal communities.

Table 4. Linguistic context of GROYNE.

Three types of concordances are included in each entry of EcoLexicon: conceptual, phraseological and verbal. These concordances allow the users to widen their knowledge from different perspectives. Conceptual concordances show the activation of conceptual relations in the real use of terms. Phraseological concordances help the user in acquiring specialized discourse. Thirdly, verbal concordances highlight the most frequent verbal collocations, which offer, again, both linguistic and conceptual information.

Figure 5 shows the conceptual concordances in the entry of GROYNE. Linguistic markers such as *designed to* and *provide* explicitly relate the concept to its function, *shore protection* and *trap and retain sand*.

Finally, the third type of contextual information added to the entry are images. These images are selected according to their most salient functions (Anglin et al., 2004; Faber et al., 2007) or in terms of their relationship with the real-world entity that they represent to illustrate the relations a concept can express. Table 5 shows an example of how several images are explicitly related to the conceptual relations expressed in the definition of GROYNE.

5. Overinformation: reconceptualization in contextual domains

In knowledge representation, concepts are very often classified according to different facets or dimensions. This phenomenon is widely known as *multidimensionality* (Kageura, 1997). The representation of multidimensionality enhances knowledge acquisition providing different points of view in the same conceptual system. However, not all dimensions can always be represented at the same time, since their activation is context-dependent. This is the case of certain versatile concepts involved in a myriad of events, such as SEDIMENT. In EcoLexicon this has led to a great deal of information overload (see figure 6), which jeopardizes knowledge acquisition.

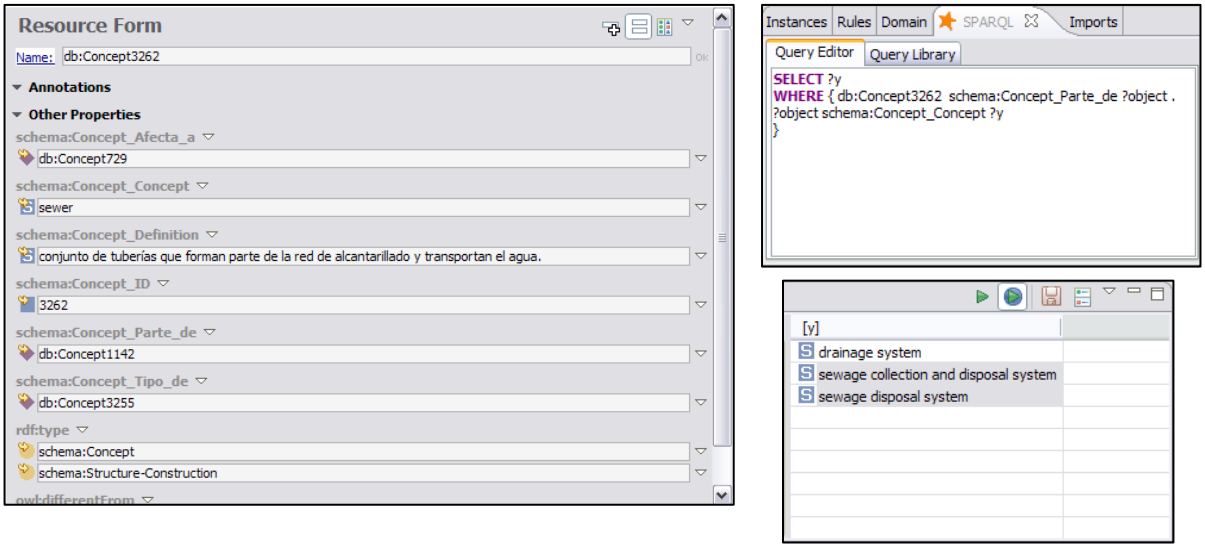


Figure 4. Concept SEWER in the ontology and inferred transitivity

TYPE

od only qualitatively. 5-3. Groins. a. General. (1) Groynes are barrier-type structures that extend from the
ks at the coast of this area existed of hard measures (groynes and seawall). The groins could not be kept-up
of inlet stabilization works (e.g., jetties, terminal groynes, offshore breakwaters)\par on the shorelines ad
wearing away of land by the action of natural forces. Groyne -- a shore protection structure, usually built pe
beach nourishment. Some coastal structures, including groynes, breakwaters and sills, can have a positive effe
ocial implications Coastal defence structures such as groynes and detached breakwaters generally increase the

MATERIAL

vant to most rubble mound structures such as seawalls, groynes, and breakwaters. It should be noted that in thi
sand, nourishing the beach compartments between them. Groynes may be made of wooden or rocky materials. They
ct will also include adding notching to existing stone groynes and extending outfall pipes. Unger said a sche

LOCATION

olution includes the following three elements: a rock groyne extending seaward from the existing shoreline, a
are intended consequences for people and wildlife. Groynes are structures built out from the shoreline, typ
line from receding. Along almost the entire coastline groynes are present, only at Warnemunde mitte there are

FUNCTION

the alignment of the updrift shoreline shifts as well. Groyne fields are designed to trap and retain sand, nour
ents, and thus a greater erosion rate. Breakwaters and groynes are effective in retaining sand and reducing ero
nearer the\par barrier may also be considered, using groynes to impede updrift movement of material at the\pa
ions to anchor the fill material. In either instance, groynes provide shore protection by modifying longshore
acceptable erosion of the downdrift shore. At first, a groyne field interrupts the longshore movement of sand i
ngth of beach to be protected. The basic purposes of a groyne are to modify the longshore movement of sand and

Figure 5. Conceptual concordances in the entry of GROYNE

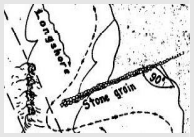
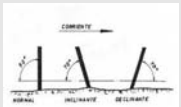
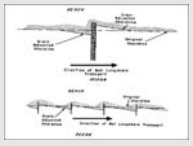
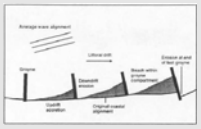
GROYNE		
Formal role	▪ hard coastal defence structure [**IS_A**],	
Constitutive role	▪ *default value* (concrete, wood, steel, and/or rock) [**MADE_OF**]	
Formal role	▪ perpendicular to shoreline [**HAS_LOCATION**]	 
Telic role	▪ protect a shore area, retard littoral drift, reduce longshore transport and prevent beach erosion [**HAS_FUNCTION**]	 

Table 5. The convergence of linguistic and graphic descriptions of GROYNE.

To avoid this information overload, we have divided the environmental field into different contextual domains according to corpus information and expert collaboration: HYDROLOGY, GEOLOGY, METEOROLOGY, BIOLOGY, CHEMISTRY, ENGINEERING, WATER TREATMENT, COASTAL PROCESSES, NAVIGATION. The environment is a multidisciplinary domain where different scientific disciplines converge. Nevertheless, they deal with the same subject in different terms. For example, WATER is found in the SEA or in a WATER TREATMENT PLANT, it may be related to CLOUDS or DESALINATION, but WATER is still WATER. Though its basic definition will not vary across different contexts, reconceptualization must somehow be represented.

We have manually established these different contexts in accordance with the information extracted from concordances. For instance, when *water* is found near terms like *irrigation* or *plants*, the conceptual relations conveyed by concordances are associated with the AGRICULTURE context. However, when *water* is followed by *supply*, *population* or *treatment* it activates different relations conveyed by certain lexico-syntactic patterns such as *used for*, *available for*, etc. It then acts as a patient in the WATER TREATMENT domain. Thus, domain membership reconceptualizes versatile concepts restricting their relational behaviour.

Contextual constraints are neither applied to individual concepts nor to individual relations (León et al., 2009; León Araúz and Magaña Redondo, in press). Constraints are instead applied to conceptual propositions. For instance, SLUDGE is linked to sediment through a *type_of* relation, but this proposition is irrelevant if users only want to know what kinds of SEDIMENT are found in nature as a result of geological processes. Consequently, the proposition SLUDGE *type_of* SEDIMENT will only appear in a WATER TREATMENT context. As a result, when constraints are applied, SEDIMENT only shows relevant dimensions for each context domain. In figure 7 SEDIMENT is just linked to propositions belonging to the context of WATER TREATMENT.

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. On the other hand, conceptual dimensions are noticeably reduced with a coherent and consistent method based on a cognitive approach.

6. The user interface

All the information contained in previous sections converges in the user interface, shown in figure 8.

Each entry includes:

- Access to the ontological structure, under the tag 'Domains', where the exact position of the concept in the domain hierarchy is shown. GROYNE, for example, *is_a* CONSTRUCTION (bottom-left corner of the window);
- Access to conceptual relations, displayed in a dynamic network of related concepts (right-hand side of the window);
- Access to the concept definition (shown when the cursor is placed on the concept). The definition for GROYNE is the linguistic expression of the template for DEFENSE STRUCTURE and therefore includes relations such as *is_a*, *made_of*, and *has_function*;
- Access to the terminological units, under the tag 'Terms', designating the concept in English and Spanish: '*groyne*' and its variant '*groin*', and '*espigón*', respectively (top left-hand corner);

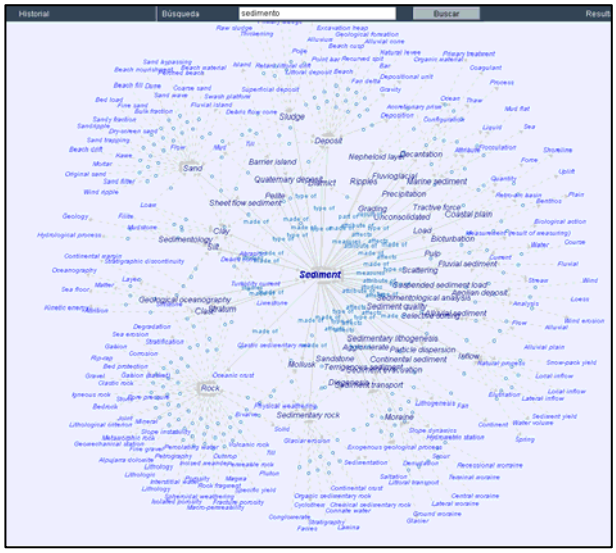


Figure 6. Information overload

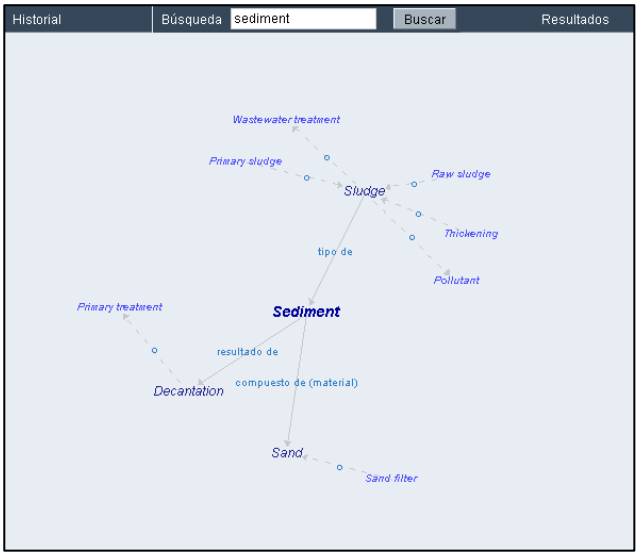


Figure 7. SEDIMENT in the contextual domain of WATER TREATMENT

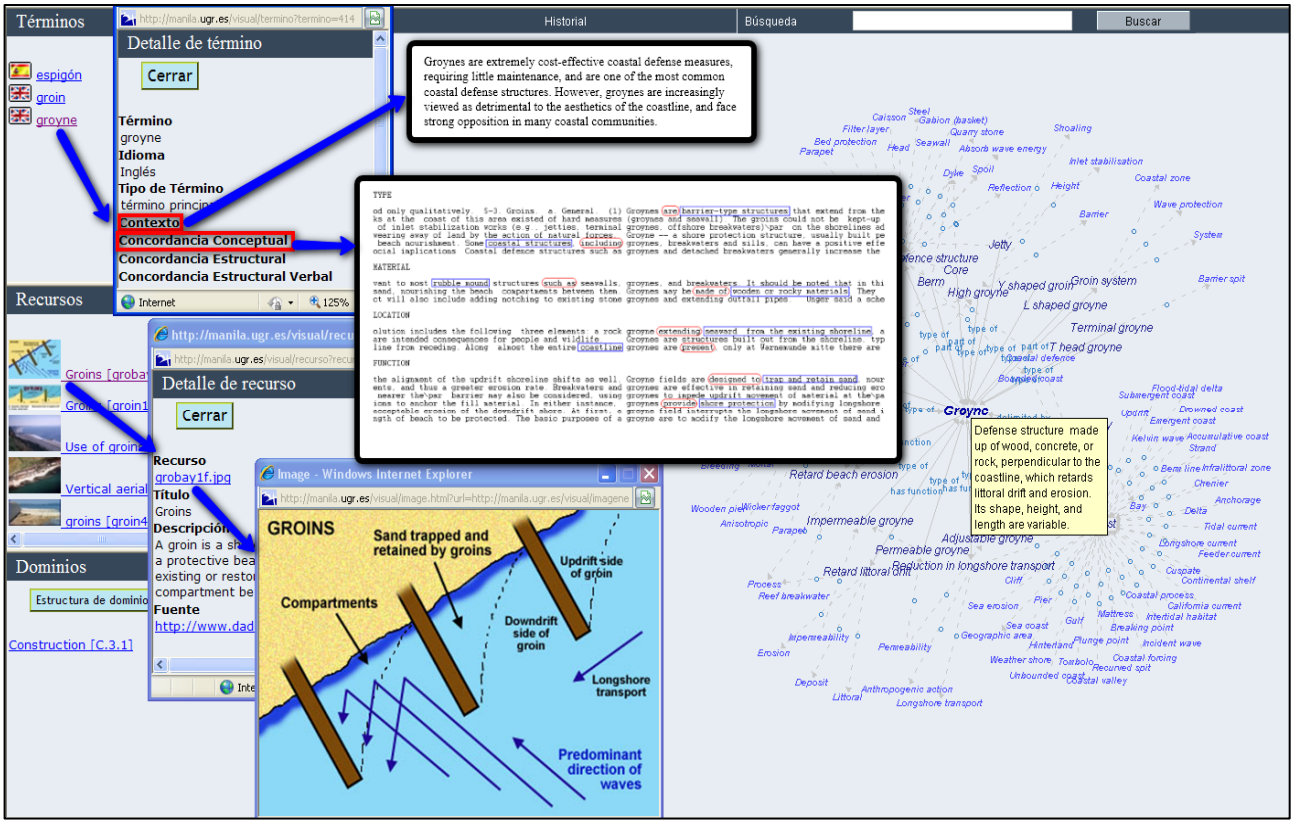


Figure 8. EcoLexicon user's interface in the entry of GROYNE

- Access to contexts (top window with black contour) and concordances (bottom window with black contour) when clicking on the terms.
- Access to graphical resources clicking on the links in the box 'Resources' (in the left-hand margin towards the middle).

Users do not have to see all this information at the same time, but can browse through the different windows and resources according to their needs.

7. Conclusions

In this paper we have presented EcoLexicon from several points of view. We have briefly explained its theoretical underpinnings, the methodology we apply for knowledge extraction and representation, and we have shown how all this information is presented to the end user. The internal coherence at all levels of a dynamic knowledge representation shows that even complex domains can be represented in a user-friendly way.

EcoLexicon combines the advantages of a relational database, allowing for a quick deployment and feeding of the platform, and an ontology, enhancing user queries. Our next step is the development of an environmental community through the Linked Data technology. We plan to link EcoLexicon to other environmental ontologies such as EnvO and Sweet. However, the success of this approach will largely depend on the proliferation of other shared initiatives.

8. References

Anglin, G. Vaez, H. and Cunningham, K.L. 2004. "Visual representations and learning: the role of static and animated graphic". Visualization and Learning, vol. 33, 865-917.

Barsalou, L. 2003. "Situated simulation in the human conceptual system." *Language and Cognitive Processes* 18(5/6), 513–562.

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

Faber, P., S. Montero Martínez,, M.C. Castro Prieto, J. Senso Ruiz, J.A. Prieto Velasco, P. León Araúz, C.F. Márquez Linares, and M. Vega Expósito. 2006 "Process-oriented terminology management in the domain of Coastal Engineering". *Terminology* 12: 2, 189-213.

Faber, P., León Araúz, P., Prieto Velasco, J.A., Reimerink, A. 2007. "Linking images and word: the description of specialized concepts". *International Journal of Lexicography*, 20:1.

Kageura, K. 1997. "Multifaceted/Multidimensional concept systems". In Wright, S.E. y Budin, G. (eds.), *Handbook of Terminology Management*: *Basic Aspects of Terminology Management*. Ámsterdam/Philadelphia: John Benjamins. 119-32.

Lenci, A., Bel, N., Busa, F., Calzolari, N., Gola, E., Monachini, M., Ogonowski, A., Peters, I., Peters, W., Ruimy, N., Villegas, M. & Zampollo, A. (2000). "SIMPLE: A General Framework for the Development of

Multilingual Lexicons". *International Journal of Lexicography*, 13, 4, 249-263.
León Araúz, P. and Magaña Redondo, P.J. (in press). "EcoLexicon: contextualizing an environmental ontology". *Terminology and Knowledge Engineering*, Dublin.
León Araúz, P., Magaña, P.J. and Faber, P. 2009. "Building the SISE: an environmental ontology". In *Proceedings of Towards e-Environment*. Prague.
León Araúz, P., Reimerink, A. & Faber. P. 2008. "PuertoTerm & MarcoCosta: a Frame-Based Knowledge Base of Coastal Engineering". *Proceedings of the XVIII FIT World Congress*. CD-ROM, Shanghai: FIT.
Pustejovsky, J. 2001. "Type construction and the logic of concepts". In Bouillon, P. and Busa, F. (eds.), *The syntax of word meaning*, Cambridge University Press, Cambridge.
Pustejovsky J. 1995. *The generative lexicon*. Cambridge, MA: MIT Press.
Pustejovsky, J., Havasi, C., Saur, R., Hanks, P. y Rumshisky, A. 2006. "Towards a generative lexical resource: The Brandeis Semantic Ontology". In *LREC*, Geneva.
Reimerink, A. and Faber, P. (2009). "EcoLex: a frame-based knowledge base for the environment". *Proceedings of Towards e-Environment*. Prague.
Smith, M., Welty, C. y McGuinness, D. (eds.). 2004. *OWL Web Ontology Language Guide*. *W3C Recommendation*s.