Linking Domain-Specific Knowledge to Encyclopedic Knowledge: an Initial Approach to Linked Data

Pilar León Araúz[1], Pamela Faber[1], Pedro J. Magaña Redondo[2]

[1] Department of Translation and Interpreting, University of Granada
Buensuceso, 11 18002 Granada, Spain
[2] Andalusian Centre for the Environment (CEAMA), University of Granada
Avda. Mediterráneo s/n, 18071 Granada, Spain
{pleon, pfaber, pmagana}@ugr.es

Abstract. Linked Data creates a shared information space by publishing and connecting resources in the Semantic Web. However, the specification of semantic relationships between data sources is still a stumbling block. One solution is to enrich ontologies with multilingual and concept-oriented information. Usefully linking entities in the Semantic Web is thus facilitated by a semantic-oriented cross-lingual ontology mapping framework in which knowledge representations are not restricted to a particular natural language. Accordingly, this paper describes a preliminary approach for integrating general encyclopedic knowledge in DBpedia with EcoLexicon, a multilingual terminological knowledge base on the environment.

Keywords: terminology, knowledge representation, linked data, multilinguality

1 Introduction

Knowledge bases play an increasingly important role in enhancing the intelligence of Web as well as in supporting information integration [1]. In this respect, the Semantic Web is an extension of the current Web in which information is given a well-defined meaning, better enabling computers and people to work in cooperation [2, 3]. This refers to people all over the world, who speak different languages. As Cimiano [4] states, the Semantic Web has the potential for dealing with cross-linguistic mappings since its content is structured much like a database and thus is language-independent.

The awareness of linguistic complexity has intensified over the last ten years as the number of Internet webpages in other languages has soared. This is a challenge for usefully linking entities in the Semantic Web because this process requires some sort of semantic-oriented cross-lingual ontology mapping framework in which knowledge representations are not restricted to the use of a particular natural language [5]. However, without a coherent description of concepts and terminological variants that take into account the categorization of real world entities by other language communities, the Semantic Web will never be truly multilingual. We thus propose a model for integrating general encyclopedic knowledge in DBpedia with our domain-specific resource, EcoLexicon (http://ecolexicon.ugr.es), a multilingual terminological knowledge base (TKB) on the environment.

2 EcoLexicon

EcoLexicon was initially implemented in Spanish, English, and German though more languages are currently being added (Modern Greek, Russian, Dutch and French). So far it has a total of 3,271 concepts and 14,646 terms. One of its main assets is its multilinguality, but as an added value, its user interface also includes semantic networks, graphical resources, definitions and contextual information that enhance the representation of conceptual and terminological knowledge [6]. Nevertheless, the focus here is on terminological and semantic information since these data sets are the ones used for linking EcoLexicon to DBpedia. For every environmental concept, multilingual choices are made available. The users can then click on any of them and obtain terminological information, such as whether a linguistic designation is a synonym, acronym, register, or stylistic variant. For instance, the concept EBB CURRENT has a total of 16 different designations in Spanish, English, German, and Greek since all registers and linguistic varieties are accounted for. In this entry, *ebb tide* appears as a non-technical variant for the concept EBB CURRENT. Even in large term bases, multilingual variety is rarely represented in an exhaustive way. However, this type of information is invaluable because not only does it provide users with multiple options for text comprehension and production, but it is also useful for conceptual disambiguation (section 3). Concepts are also displayed in dynamic semantic networks linked to other concepts. Nevertheless, problems can arise when it is a question of browsing networks of very general concepts, which carry an excessive load of information (Fig. 1).

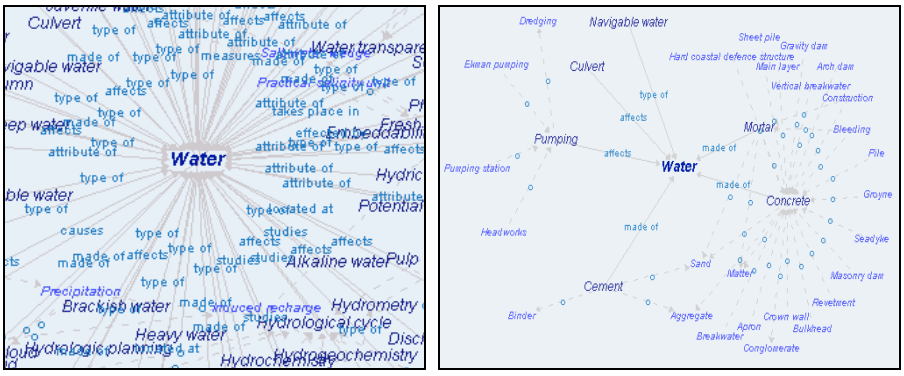

Fig. 1. and **2.** Context-free and context-based semantic network of WATER

Overloaded concepts, such as WATER, share multiple relations with many other concepts, but they rarely, if ever, activate all those relations at the same time since this would evoke completely different and incompatible scenarios. 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 [7]. The area of environmental knowledge was thus divided into a set of contextual domains (e.g. HYDROLOGY, GEOGRAPHY, OCEANOGRAPHY, CIVIL ENGINEERING, ENVIRONMENTAL ENGINEERING, etc.) and the relational power of concepts was

constrained accordingly. Thus, when constraints are applied, the network of WATER within the CIVIL ENGINEERING domain is recontextualized and becomes more meaningful (Fig. 2).

EcoLexicon is primarily hosted in a relational database (RDB), but at the same time it is integrated in an ontological model. Semantic information is stored in the ontology, while leaving the rest in the relational database [7]. This is important because the linked data process not only involves the transformation of data to RDF format, but also includes the use of terminologies, controlled vocabularies, and ontologies to describe triples attributes in a systematic way and as reference conceptual models to support an integrated view of data and semantic interoperability between datasets [8]. As seen in Fig. 3, contextual domains have inspired the design of our ontology classes. The ontology is automatically retrieved from the data stored in our RDB, according to the following assumption: if a concept *c* is part of one or more propositions allocated to a contextual domain *C*, *c* will be an instance of the class *C*. EcoLexicon keeps multilingual terminological information and ontological information separate. Each terminological entry has different word forms linked to the same natural language definition, constrained by the knowledge represented in the ontology concept [9].

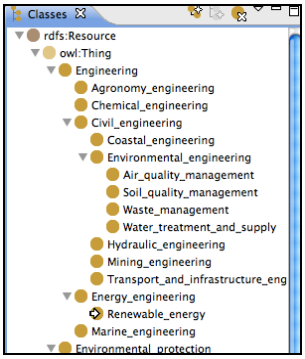

Fig. 3. Ontological classes

3 Linking EcoLexicon to DBpedia

Linked Data is an important initiative for creating a shared information space by publishing and connecting structured resources in the Semantic Web [10]. However, the specification of semantic relationships between data sources is still a stumbling block. Our initial proposal is to integrate EcoLexicon with DBpedia through the *sameAs* property, because: (1) DBpedia is at the core of the Linked Data initiative; (2) users can complete their knowledge acquisition process through a guided access to encyclopedic knowledge.

Linking data sources from DBpedia can be quite straightforward since different tools, such as the ontology editor TopBraid Composer, can automatically suggest the links. However, because of lexical variation and the lack of univocity in both general and specialized knowledge, automatic mappings are not always viable. Furthermore, although establishing an identity relation initially may appear to be a simple task,

matching two entities, both at the syntactic and the semantic levels, is often far from easy [11]. Problems with text searching and entity matching highlight the fact that a word is more than a mere string of characters. The following are basically the same problems that have plagued linguists over the years: polysemy, homonymy, synonymy, and different levels of specificity [12]. There are also other aspects of lexical meaning that lead to confusion, such as the fact that: (i) the meaning of a term can vary, depending on the context; (ii) meanings can change in time and space; (iii) different languages reflect different mappings of reality, which may coincide totally, partially, or not at all.

A solution to some of these problems can be found when ontologies are enriched with multilingual and concept-oriented information, as reflected in the field of environmental knowledge, but manual work is still necessary to a certain extent. Nevertheless, instead of mapping one-to-one manual correspondences, we can take advantage of the semantics contained in each resource. In our approach, the term strings of EcoLexicon are compared with those from DBpedia, enhanced by those data sets that include multilingual choices and variants as well as category membership. To illustrate our data linking proposal, we have chosen four concepts: GROIN, BANK, ACCRETION, WASTEWATER TREATMENT PLANT and the pseudocode of the general matching algorithm is shown in the following table:

```
for each w:word in ecolexicon
for each cp:concept in dbpedia
    w' = stem(w); cp' = stem(cp)
    if str_compare(v', cp') > word_threshold
      multi_e = multilingual_variants(v)
      multi_g = multilingual_variants(cp)
      if multingual_compare(multi_e, multi_g) > multilingual_threshold
        result.add(pair(v, cp))
        related_instances = instances_of(context(v))
        for each i:instance in related_instances
          if look_for_text(comment_properties(cp), i) > text_threshold
            result.add(pair(v, i))
```

The concept GROIN in DBpedia is not designated by its most frequent form but by a geographical variant (*groyne*). The fact that EcoLexicon stores all lexical variations of each concept allows us to identify the same entity in both resources by comparing the string of all our English monolingual variants with the entries in DBpedia. However, if the search was only performed for the string *groin*, DBpedia would redirect to a disambiguation page, since GROIN can also refer to a part of the human body. In this case, with the help of the English variant *groyne* and the French equivalent *épi*, the concept can be easily disambiguated.

The case of BANK is similar to that of GROIN. Nevertheless, it is necessary to add other parameters to the linking rule since *bank* is polysemic at a cross-linguistic level. For example, as in English, the Spanish term *banco* can refer to a geographic landform or a financial institution, and there are not many other common multilingual equivalents in DBpedia for disambiguation. In DBpedia, this domain-specific entry is named, and differentiated from others, as BANK (GEOGRAPHY). In order to match this entry and not any of the others, it is necessary to add a context-based rule. Therefore, this match will occur in the following situations: (1) when the word in brackets matches the string of any of our contextual classes or their linguistic variants; (2)

when any term, in any language, associated with any concept belonging to the same contextual class as the search concept appears in one or more of the values of the following properties: *dbpedia-owl:abstract*, *dcterms:subject*, *rdfs:comment,* or *dbpedia-owl:wikiPageRedirects*. In this case BANK in EcoLexicon belongs to the classes, GEOGRAPHY, GEOLOGY and OCEANOGRAPHY, as do many other concepts, such as SHORELINE, ESTUARY, RESERVOIR, SLOPE, RIVER, MARSH, etc, all of which are contained in the properties *dbpedia-owl:abstract*, *dcterms:subject* and *rdfs:comment*. Furthermore, since the disambiguating word in brackets coincides with the EcoLexicon class GEOGRAPHY, the second step is not even required in this case.

Nevertheless, there is a similar but even more complex example in the concept ACCRETION. ACCRETION is polysemic in different languages as well as within the environmental domain. This time disambiguation is not only performed in order to differentiate other domains from the environmental one. On the contrary, three different senses (concepts) in EcoLexicon, designated by the same terms in all languages and with no variants, have to be matched with three out of the five entries in DBpedia. In DBpedia, the term *accretion* may be related to the fields of FINANCE, ASTROPHYSICS, ATMOSPHERE, GEOLOGY, or COASTAL MANAGEMENT, of which only the last three are included in EcoLexicon. In EcoLexicon, the concepts belong to the classes of ATMOSPHERIC SCIENCES, GEOLOGY, and OCEANOGRAPHY, respectively. The concepts related to FINANCE and ASTROPHYSICS are ruled out through the same context-based rule as in BANK. However, this rule must be further specified in order to disambiguate the DBpedia entries of ACCRETION (ATMOSPHERE), ACCRETION (GEOLOGY) and ACCRETION (COASTAL MANAGEMENT). In this case, matching the concepts in common with those included in the property values and those that belong to the same contextual class as each of the concepts designated by *accretion* is insufficient, since all three concepts are closely interrelated. For instance, the key terms *ice* or *droplet*, only present in the property values of ACCRETION (ATMOSPHERE), could seem enough to disambiguate the concept. However, the concepts designated by these terms belong to both our ATMOSPHERIC SCIENCES and GEOLOGY classes. Apart from their obvious relation to the atmosphere, they are also related to geological concepts, such as AVALANCHE or EROSION. Therefore, at this point, disambiguation is still necessary between ACCRETION (GEOLOGY) and ACCRETION (ATMOSPHERE). As for the property values of ACCRETION (COASTAL MANAGEMENT), there are certain terms in the property values, such as *erosion*, *sediment, beach,* and *weather* that can point to all of the three classes (i.e. *weather* to ATMOSPHERIC SCIENCES, *erosion* and *sediment* to GEOLOGY, and *beach* to both GEOLOGY and OCEANOGRAPHY). Consequently, for these cases, one more variable is added to the matching algorithm: from all the contextual classes to which key concepts may belong, only the most frequent one will be used for disambiguation. This means that if most concepts included in the property values of ACCRETION (COASTAL MANAGEMENT) are mostly activated in propositions framed within the OCEANOGRAPHY class, then both concepts are equivalent.

Finally WASTEWATER TREATMENT PLANT does not show ambiguity problems because it is a very specialized concept. Nevertheless, this is a good example of how linking data does not always ensure knowledge acquisition since conceptual modeling does not necessarily follow a concrete pattern in all resources. There is thus no assurance that the content is well structured. The definition of *wastewater treatment plant* in DBpedia does not describe the concept at all. In fact, it is incorrectly assigned

to a disambiguation category, and it redirects users to different types of wastewater treatment. In fact, it does not even offer a proper definition of the plant itself. The Spanish version of Wikipedia has a good entry for its equivalent (*estación depuradora de aguas residuales*), but there is no link between them. In this sense, EcoLexicon could serve as a bridge between the multilingual environmental entries in DBpedia that are not correctly linked.

4 Conclusions

This paper has discussed the importance of multilinguality for the Semantic Web and the problems that can arise when knowledge representations in other languages are not taken into account in the linked data process. More specifically, we have compared the term strings of EcoLexicon's concepts GROIN, BANK, ACCRETION, and WASTEWATER TREATMENT PLANT with those from DBpedia, enhanced by multilingual choices and variants as well as category membership. The results show how valid correspondences can be obtained by taking advantage of the semantics contained in each resource.

References

1. Meij, E., Bron, M., Hollink, L., Huurnink, B., De Rijke, M. Mapping Queries to the Linking Open Data cloud: A Case Study Using DBpedia. Web Semantics: Science, Services and Agents on the World Wide Web (2011)
2. Berners-Lee, T., Hendler, J. and Lassila, O. The Semantic Web. Scientific American (2001)
3. Janev, V. and Vranes, S. Applicability Assessment of Semantic Web Technologies Information Processing and Management vol. 47 pp. 507–517 (2011)
4. Cimiano, P. Towards the Multilingual Semantic Web. Lecture given at the University of Granada, February18, 2011. (2011)
5. Fu, B., Brennan, R. and O'Sullivan, D. Cross-Lingual Ontology Mapping and Its Use on the Multilingual Semantic Web. 1st Workshop on the Multilingual Semantic Web (2010)
6. Faber, P. The Dynamics of Specialized Knowledge Representation: Simulational Reconstruction of the Perception–Action Interface. Terminology, vol. 17, pp. 9-29 (2011)
7. León Araúz, P., Magaña Redondo, P. and Faber, P. Managing Inner and Outer Overinformation in Ecolexicon: an Environmental Ontology. 8th International Conference on Terminology and Artificial Intelligence (2009)
8. Cordeiro, K., Marino., T., Campos, M.L., Borges, M.R.S. Use of Linked Data in the Design of Information Infrastructure for Collaborative Emergency Management System. Computer Supported Cooperative Work in Design (CSCWD) pp. 746-711 (2011).
9. Montiel-Ponsoda, E., Aguado de Cea, G., Gómez Pérez, A. and Peters, W. Enriching Ontologies with Multilingual Information. Natural Language Engineering, pp. 1-27 (2010)
10. Bizer, C., Heath, T. and Berners-Lee, T. Linked Data: Principles and State of the Art (2008)
11. Hassanzadeh, O., Kementsietsidis, A., Lim, L., Miller, R.J. and Wang, M. A Framework for Semantic Link Discovery over Relational Data. CIKM '09: Proceeding of the 18th ACM conference on Information and knowledge management. (2009)
12. Dostal, M. and Jezek, K. Automatic Tagging Based on Linked Data: Unsupervised Methods for the Extraction of Hidden Information. Service-Oriented Computing and Applications (SOCA), pp. 1-4 (2010)