A Semantic Approach for Learning Objects Repositories with Knowledge Reuse

Isabel Azevedo[1], Rui Seiça[2], Adela Ortiz[1], Eurico Carrapatoso[2], Carlos Vaz de Carvalho[1],

[1] Instituto Superior de Engenharia do Porto, GILT, R. de São Tomé,
4200 Porto, Portugal
{iazevedo, adela, cvc}@dei.isep.ipp.pt

[2] Faculdade de Engenharia da Universidade do Porto, R. Roberto Frias,
Porto, Portugal
ruiseica@gmail.com, emc@fe.up.pt

Abstract. In this paper we discuss how the inclusion of semantic functionalities in a Learning Objects Repository allows a better characterization of the learning materials enclosed and improves their retrieval through the adoption of some query expansion strategies. Thus, we started to regard the use of ontologies to automatically suggest additional concepts when users are filling some metadata fields and add new terms to the ones initially provided when users specify the keywords with interest in a query. Dealing with different domain areas and having considered impractical the development of many different ontologies, we adopted some strategies for reusing ontologies in order to have the knowledge necessary in our institutional repository. In this paper we make a review of the area of knowledge reuse and discuss our approach.

Keywords: Semantic Web, ontology selection, ontology reuse.

1 Introduction

Formal ontologies have been seen as a way to state content specific understandings for many knowledge-sharing tasks, in the form of conceptualizations of the world of interest.

In a study from 2007 [1], the participants were asked about what motivated them to use ontologies. The answers showed that the two main reasons were the need to share a common understanding of the structure of information among people or software agents (69.9 percent), and the requirement to enable the reuse of domain knowledge (56.3 percent).

However, the creation of ontologies leads to some costs, not only time, and it is often not considered a simple process, even if many methodologies have emerged. The creation of ontologies cannot be done automatically and requires, to a considerable degree, human input. Furthermore, unlike what happened in Software Engineering, ontology creation and other Semantic Web methodologies are not fully matured. Actually, the same study already referred before found that 60 percent of the users do not apply any methodology for developing ontologies.

Ontology reuse has been seen as a viable alternative to having ontologies with reduced expenses. But, reusing ontologies is not a simple process. It starts by finding the possible candidates, and then it is necessary to rank them, and select one or more in accordance to those characteristics that were considered important. It could be argued that it is simpler to develop one from scratch but it is not always possible to find an expert and a knowledge engineer that can dedicate the required time to that. Furthermore, it brings new costs in finding, selecting and revision tasks, in which significant automation is not possible, except, to some extent, for the discovery of ontologies.

Nonetheless, due to the incapability of developing ontologies from scratch for a large number of different domains to use in an institutional repository, the reuse of ontologies has emerged as a feasible solution.

The rest of the paper is organized as follow. In section 2 we briefly discuss the foundations of knowledge reuse in general and ontology reuse in particular. In the third section we present the TREE (Teaching Resources for Engineering Education) repository, explain the reasons for considering the reuse of ontologies and the adopted strategies for finding and selecting the appropriate ontologies. Finally, the last section discusses the empirical evaluation carried out, states some conclusions and the planned work for the next months.

2 Knowledge Reuse

Knowledge processes frequently entail the creation or reuse of knowledge, and the methods appropriate for one, may not be convenient for the other [2].

According to Alavi et al., knowledge reuse is the process by which an entity is capable of finding and applying shared knowledge [3]. The reuse of knowledge components probably started to be considered as an important subject when the Knowledge Sharing Effort, sponsored by some American organizations aiming to support the sharing of knowledge among systems, suggested the connection of reusable units to build knowledge-based systems in 1991 [4].

Markus identified three main players in the process of reusing knowledge [5]:

- *Knowledge producer*—the knowledge creator, who registers explicit knowledge or transforms tacit knowledge into explicit,
- *Knowledge intermediary*— the agent that adapts knowledge for reuse, with many roles, including sharing it,
- *Knowledge consumer*—the person or system that recovers the suitable knowledge piece and makes use of it.

The same questions that emerged from knowledge reuse, appeared, to some extent, in the attempts to reuse ontologies. Indeed, reusability is (or should be) an underlying property of ontologies. They are often seen as a way to allow “knowledge sharing and reuse" [4], as the concepts represented in an ontology become explicit, reliable and convenient to share and, then, to reuse.

As ontologies have been more and more used in many domains, several ontologies have been made available on the Web, with different scope and quality. One of the motivations to make ontologies available is that an increase in their use and revision can boost their quality. Also, the applications that use them become (more) interoperable and are provided with a deeper, machine-processable understanding of the underlying domain.

Therefore, ontology reuse started to be considered, following the trend started in knowledge, in general, and even before that, the tendency in the software engineering field.

Pinto recognizes two purposes for reusing ontologies [6]:

- *Ontology merging* - By merging different ontologies, another one is built with the combined concerned parts;
- *Ontology integration* - Different ontologies are used to build a new one, but with modifications and extensions.

The more the knowledge is available, the greater the likelihood of being reused. Thus, there are some tools that make easier finding ontologies suitable for some purpose. They do play a role in ontology reuse situations, acting as knowledge intermediaries. These tools are:

- *Ontology registry* – It is an application used to register ontologies, maintaining some description fields, statistics about their contents, and a link to each registered ontology. A registry does not provide a central storage for ontologies, but a searchable list of them.
- *Ontology repository* – It maintains a local copy of ontologies, and their different versions, if they exist. This kind of tools usually only provide browse functionalities.
- *Ontology search engine* – It does not require an active action from ontology developers. This kind of tools automatically searches for and indexes the ontologies they discover. Some examples are Swoogle [7], Watson [8], Sindice [9] and Falcons (http://iws.seu.edu.cn/services/falcons/objectsearch/index.jsp). They vary in the metadata provided for each ontology, but there is not any standard for ontology metadata and exchange.

These kinds of tools facilitate the findability of suitable ontologies, but they differ in the way they describe ontologies and in the provided metadata, usually without any information about domain of interest, creation date or authorship, for instance.

Currently, there are more than 10.000 online ontologies (http://swoogle.umbc.edu/), which do not mean that finding the suitable ontologies have become much easier. The problem has become to select the best ontology from many available. Related to this topic is the ranking of ontologies. Search engines usually rank the found ontologies using one or more of the following approaches:

- *Popularity* – Ontologies are ranked in accordance with the number of times they are referred to in other ontologies. That method is followed by Swoogle (http://swoogle.umbc.edu), with some drawbacks, although consistent with the idea that the most used in other ontologies should had been considered suitable by many people;

- *Ontological structure* – By applying metrics that estimate how elaborate is the knowledge structure of an ontology structure, the “goodness” of an ontology can be estimated. Some of the approaches use the relation between the number of classes and properties [10], the centrality of a class in the whole hierarchy (Centrality measure [11]) or estimates how richness is a concept defined (Density measure [11]);
- *Concept coverage* –It is related to how well a concept is covered in an ontology. The matches between the query term and the labels in an ontology are regarded, and weighted in accordance to how perfect in the matching, usually just carried out at lexical level.

3 TREE repository

As part of the project CASPOE (PTDC/EIA/65387/2006 – Semantic and Pragmatic Characterisation of Learning Objects), a three-year project funded by the Portuguese Foundation for Science and Technology, we have been developing a prototype for an institutional learning object repository, named TREE (Teaching Resources for Engineering Education), which uses ontologies and extracted keywords to represent the semantics of learning resources [4, 5], digital objects that could be used to achieve an educational purpose. It has been populated with some materials from courses taught at ISEP, a higher education institution.

Although TREE is a repository for resources that might be used in engineering courses, it is a heterogeneous document repository as it covers different knowledge areas, such as Law, Linguistics, Environmental Sciences and Mathematical and Computer Sciences. At this moment it is on an Intranet only accessible by students, teachers and staff people.

We use Fedora (Flexible Extensible Digital Object Repository Architecture - http://www.fedora-commons.org/) repository system with some add-ons.

An important factor for the adoption of Fedora was the possibilities offered for the construction of new metadata profiles different from the base version provided. We have adopted metadata standards associated with the practice of teaching and learning, such as IEEE Learning Object Metadata [12], which consider not just metadata fields related to authorship, identification and brief descriptions of content.

Fez (http://fez.library.uq.edu.au/wiki/Main_Page) is used as a Web interface to Fedora; it is an open-source software for creating and maintaining a highly flexible web interface to the Fedora software.

The documents included in the repository are divided into communities related to different scientific areas. Each community has one or more collections assigned to diverse knowledge sub-areas, related to different courses in our institution.

Users can find resources browsing through the communities and collections hierarchical structure, searching through the tags assigned to resources or metadata fields. Under the latter hypothesis, users can search documents through the specification of keywords.

Ontologies are used for the purpose of having the core concepts of a domain well-related in order to improve the performance of information retrieval in the TREE repository. The information in the ontologies should answer what concepts are related to another given one. Thus, the reasons for the adoption of ontologies in the TREE repository are twofold:

- Allow a detailed description of the resources;
- Improve the results by applying a query expansion method.

When a document is submitted to the TREE repository, some metadata fields are automatically or semi-automatically filled. One of those is the keyword metadata (a subelement of the General element in the IEEE LOM standard), which can have multiple values.

The ontologies are used to find additional concepts related to those extracted. Then, the user can accept the desired ones. It is worthwhile to note that this process was found necessary when realized that users are really very concise when filling forms, with detrimental results in subsequent searches.

We developed a module to expand the query terms provided by the TREE users. Each time a query is submitted, every term in it is expanded to related ones. The expansion uses ontologies, considering subsumption or supersumption relations (up to two levels), equivalence and other relations, and instance data, but allowing the users to agree or not with the use of the additional terms.

3.1 Ontology reuse in the TREE repository

Figure 1 generically delineates the architecture of the semi-automated module for ontology reuse. It takes two inputs: a list of domain concepts and the online ontologies found through semantic search engines using the given concepts.

Among many tools that facilitate the discovery of ontologies we selected Swoogle for its easy integration in the repository and the summary supplied with the results. Swoogle applies the algorithm OntoRank, which is quite analogous to PageRank, which is used by Google search engine.

At our institution each course has a description in English, which is mandatory, with some predefined fields to be supplied; one of them is related to the course contents. It is simple to extract the relevant concepts for each course considering that specific field. Nevertheless, those responsible for courses can specify others and disregard or correct some of the extracted ones. That enumeration of the important terms corresponds to one of the recommended steps to follow when developing an ontology.

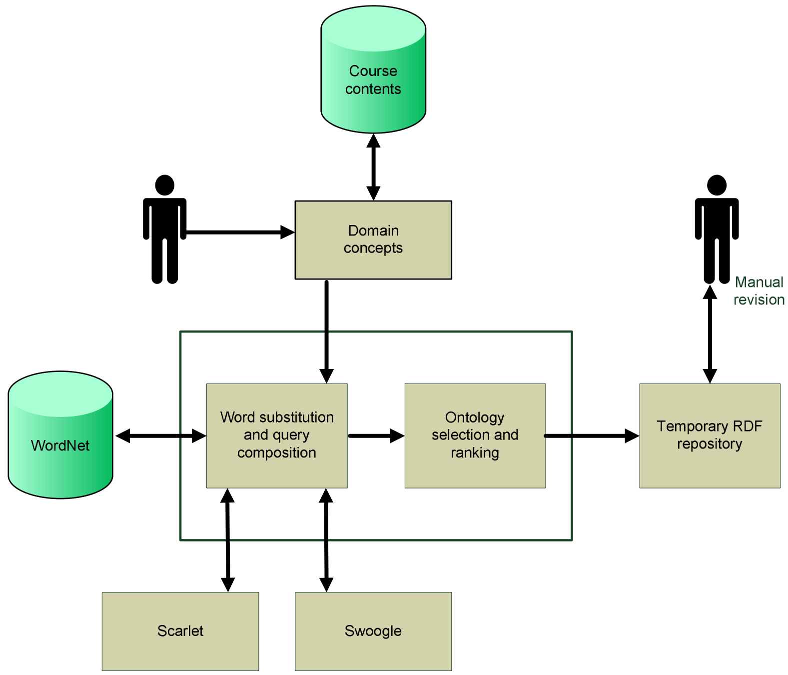

Fig. 1. Architecture of the semi-automated ontology reuse module.

For each desired concept, we apply a singularisation algorithm, and '`Distributed Databases`' is changed into '`Distributed Database`', for instance. That approach augments the number of perfect correspondences later. Then, using those concepts we compose a query that is submitted to Swoogle REST based service, which returns a RDF/XML document containing a short summary for each of them, which is used to predict relevancy.

Having C_1, C_2, … C_N as the singularized concepts for a course, we harmonize them, namely replacing white spaces with dashes. Then we first try queries with all the concepts, after disregarding those not covered by any online ontology. Thus, we might have a query with N terms, and then the system tries N-1 terms, and so on. Having all concepts covered (by at least a predefined number of ontologies) or having tried each concept alone, the process is finished. Often ontologies with only one single concept are not related to the domain of interest. Also, a query using only one search string might result in excessive results to be analysed. For instance, a query with 'table' as search string provides 935 results.

For the ontology retrieval we consider labels of instances, classes and properties in ontologies lexically matching the query terms. As some other relevant concepts could remain unnoticed, some related projects try a more conceptual matching [13] [14] and expand the initial concepts using WordNet [15], an electronic lexical database created by the Cognitive Science Laboratory of Princeton University. However, we found that approach more useful when applied in a later step, discussed in section 3.2.2.

There are three possibilities when trying to find adequate ontologies:

- One or more ontologies are found with all the desired concepts;
- The system returns different ontologies with most of the concepts, but the set of ontologies covers all of terms;
- It is impossible to find an ontology with one or more of the concepts. When this happens, we try to find relevant synonyms in the field under analysis and we repeat the process again.

Before carrying out the process of ontology selection, the ontologies whose URIs correspond to dead links are disregarded and the others are checked to verify if they are syntactically correct, and then they are ranked.

Once the found ontologies are fully analysed, the chosen ones are stored and manipulated using Jena (http://jena.sourceforge.net) and Sesame (http://sourceforge.net/projects/sesame/) frameworks.

3.2.1 Ontology ranking and selection

With the growing demand for reuse ontologies, there has been a need for criteria and standards to find out their quality. Once a system is reusing ontologies, their quality might affect the quality of the resulting ontologies, and the desirable functioning of the system. Deficiencies in modeling might remain unknown for a long time until unpredictable or poor query results catch the attention of someone.

Applying evaluation methods to estimate the quality of ontologies and rate them, the likelihood of that problem can be reduced, but not entirely eliminated.

However, evaluating a single ontology is different from evaluating a set of ontologies to decide on the one(s) to reuse, as an intensive evaluation of all candidates is not feasible.

Although more specific for evaluation of a single ontology at a time, some of the more automatic approaches for ontology evaluation are:

- Use a certain ontology and realize its performance in a real environment with interest, allowing a functional evaluation. When evaluating a number of ontologies, the best is the one that provides the optimal fulfillment for an application. However, this technique is not often a feasible way to test and evaluate ontologies.
- Use a Gold Standard ontology. The most similar to this one is considered the most suitable one and should be adopted. Some drawbacks of this approach are the lack of availability of the optimum ontology to compare with others, and the comparison itself, namely how to perform it.
- Realize the coverage of the domain concepts. The ontology that includes most of the concepts is the one to be used, but that analysis is usually based on matching at lexical level, as stated before.

We evaluate and rank the discovered ontologies considering three different aspects, each one leading to a different number. These considered characteristics are: their concepts coverage (considering the distance between the provided terms and those in the ontology), popularity (ontoRank value provided by Swoogle) and knowledge richness (the number of classes and properties, considered all triples).

During the concepts coverage analysis, some extraneous matching can occur between the terms provided and those in the ontology when carried out at lexical level, but applying a similarity metric that possibility can be avoided to certain extend, but not completely.

After computing the Levenshtein metric to estimate the distance between strings, only the terms that match some class, property or instance label greater than a defined threshold, are regarded.

For any given ontology, we calculate the ratio between the sum of similarities of occurrences and the number of occurrences for each domain concept.

Once a list of possible appropriate ontologies is ranked, an ontology engineer analyses it and manually selects the most suitable ones, occasionally with the assistance of an expert from that area, namely when two or more ontologies have very similar ranking values. In that case, competency questions might be considered to decide on the most appropriate ontology. They have also proved valuable to detect missing parts of knowledge. For example, considering the database domain, an ontology extract might have to address the following competency questions:

- Which are the Normal Forms?
- Which are the subsets of SQL statements?

A final ontology is obtained from the merging from one or more ontologies, and then assigned to a collection in the TREE repository.

3.2.2 Additional strategies adopted to find online ontologies

When no ontology is found covering a certain concept provided, and yet relying on the information provided by the search engine Swoogle, we verify if the ontologies already found for others have some variations of them. For instance, finding ontologies with the term '`relational_model`' for the Database field is unsuccessful. But analyzing the ontologies already found, we can discover a similar one: '`relational_data_model`', as explained before.

It is worth to note that many semantic search engines do not satisfactorily consider wildcards, although accepting them. Thus, our first attempt to submit queries using wildcards like 'relational*model', for example, was unsuccessful.

In addition, we try word substitution through WordNet, after some ontologies have possibly been found. WordNet is structured semantically, containing nouns, verbs, adjectives and adverbs. Words that are synonymous are grouped together in synsets - synonym sets. Polysemous words in WordNet are included in more than one synset and each synset leads a different sense. For instance, the concept '`data modelling`', which appears in a database course description, is not contained in any online ontology. Considering sister terms, hypernyms (words from ancestor synsets) and hyponyms (words from descendant synsets), there many possible substitutions for '`modelling`' from different synsets, such as '`modeling`', '`molding`', '`moulding`', '`model`', '`pattern`', and so on. When we have already found some ontologies for the other concepts supplied, we check if one of them applies the terms provided by WordNet, as a class, a property or instance, and in this case we consider that new concept as a substitute for the one initially specified.

When no ontology was already found using the terms provided by WordNet, we exploit Scarlet java API (http://scarlet.open.ac.uk/) to verify if there are some relations between the probable substitutes supplied by WordNet and the other terms provided, choosing then the one with most relations found or found in the same ontologies. Scarlet (SemantiC relAtion discoveRy by harvesting onLinE onTologies) is “a technique for discovering relations between two concepts by harvesting the Semantic Web” [14], whose usage was first envisaged for ontology matching. We regard all type of relations, more than one ontology and inheritance depth equals to 5.

When all other attempts have proved unsuccessful to find ontologies related to a concept, we ask for a set of documents containing that concept, which is used for ontology extraction, with some parts subsequently incorporated into the ontology already obtained earlier, after some revision. However, this last approach is significantly more time-spending than the others.

4 Conclusions

In this paper we presented a case study about ontology reuse. Based on the literature review and this particular case study, we have identified some issues that should be attended in order for the reuse of ontologies to increasingly become a reality.

Even though the TREE repository was designed for engineering resources, the adopted approaches can be used for other areas as well.

The basic version of the repository was evaluated using resources from four different courses, whose responsible people complemented the concepts gathered from the course description.

The initial results were very encouraging. The empirical evaluation revealed the practical usefulness of the discussed approaches and the users were generally satisfied with the efficiency of the information retrieval. However those responsible for courses were not so pleased with the time spent in all the activities, but a full automation of the whole process of discovering, selecting and revising ontologies is impossible. Also, some performance issues were highlighted but, as most of the process is carried offline, that point is not a main concern for the upcoming months.

Some detected errors have been corrected and we plan to further evaluate the whole system via quantitative measurements in the next months.

Finally, our approach to reuse ontology has some similarities with the one described in [13], but we have achieved more automation in some steps, namely we do not ask for user feedback before using WordNet for finding suitable substitutions for domain concepts, relying on information available on the Semantic Web.

Acknowledgments. The work described in this paper has being developed under the project CASPOE - Characterization Semantics and Pragmatics of Educational Objects (PTDC/EIA/65387/2006), with funding from the Portuguese Foundation for Science and Technology.

References

1. Cardoso, J.: The Semantic Web Vision: Where Are We? IEEE Intelligent Systems (2007) 84-88

2. Davenport, T.H., Jarvenpaa, S.L., Beers, M.C.: Improving knowledge work processes. Sloan Management Review 37 (1996) 53-65
3. Alavi, M., Leidner, D.E.: Review: Knowledge Management and Knowledge Management Systems: Conceptual Foundations and Research Issues. MIS Quarterly 25 (2001) 107-136
4. Nechs, R., Fikes, R., Finin, T., Gruber, T., Patil, R., Senator, T., Swartout, W.: Enabling Technology for Knowledge Sharing. AI Magazine 12 (1991) 36-56
5. Markus, M.L.: Toward a theory of knowledge reuse: Types of knowledge reuse situations and factors in reuse success. Journal of Management Information Systems 18 (2001) 57--93
6. Pinto, H.S., Gomez-Perez, A., P.Martins, J.: Some Issues on Ontology Integration. Proc. of IJCAI99's Workshop on Ontologies and Problem Solving Methods: Lessons Learned and Future Trends, Stockholm, Sweden (1999)
7. Ding, L., Finin, T., Joshi, A., Pan, R., Cost, R.S., Peng, Y., Reddivari, P., Doshi, V., Sachs, J.: Swoogle: A Search and Metadata Engine for the Semantic Web. Proceedings of the Thirteenth ACM Conference on Information and Knowledge Management (2004)
8. d'Aquin, M., Sabou, M., Dzbor, M., Baldassarre, C., Gridinoc, L., Angeletou, S., Motta, E.: Watson: A Gateway for the Semantic Web. European Semantic Web Conference (ESWC 2007), Innsbruck, Austria (2007)
9. Oren, E., R.Delbru, Catasta, M., Cyganiak, R., H. Stenzhorn, Tummarello, G.: Sindice.com: A document-oriented lookup index for open linked data. International Journal of Metadata, Semantics and Ontologies 3 (2008)
10. Buitelaar, P., Eigner, T., Declerck, T.: OntoSelect: A Dynamic Ontology Library with Support for Ontology Selection. The International Semantic Web Conference (Demo Session), Hiroshima, Japan (2004)
11. Alani, H., Brewster, C.: Metrics for Ranking Ontologies. Fourth International Evaluation of Ontologies for the Web Workshop held as part of the 15th International World Wide Web Conference, Edinburgh, United Kingdom (2006)
12. IEEE: IEEE Standard 1484.12.1-2002, Learning Object Metadata (LOM). (2002)
13. Cantador, I., Fernández, M., Castells, P.: Improving Ontology Recommendation and Reuse in WebCORE by Collaborative Assessments. Proceedings of the 1st International Workshop on Social and Collaborative Construction of Structured Knowledge (CKC 2007), at the 16th International World Wide Web Conference (WWW 2007), Vol. 273, Banff, Canada (2007)
14. Brewster, C., Alani, H., Dasmahapatra, S., Wilks, Y.: Data-driven Ontology Evaluation. Proceedings of the 4th International Conference on Language Resources and Evaluation, Lisbon, Portugal (2004)
15. Miller, G.A.: WordNet: A Lexical Database for English. New Horizons in Commercial and Industrial Artificial Intelligence. Communications of the ACM 38 (1995) 39-41