# Semantic Annotation of Multilingual Learning Objects Based on a Domain Ontology

Petr Knoth

*Knowledge Media Institute, The Open University*
p.knoth@open.ac.uk

**Abstract.** One of the important tasks in the use of learning resources in e-learning is the necessity to annotate learning objects with appropriate metadata. However, annotating resources by hand is time consuming and difficult. Here we explore the problem of automatic extraction of metadata for description of learning resources. First, theoretical constraints for gathering certain types of metadata important for e-learning systems are discussed. Our approach to annotation is then outlined. This is based on a domain ontology, which allows us to annotate learning resources in a language independent way. We are motivated by the fact that the leading providers of learning content in various domains are often spread across countries speaking different languages. As a result, cross-language annotation can facilitate accessibility, sharing and reuse of learning resources.

## 1 Introduction

This work is being undertaken within the context of the Eurogene project, which is supported by the Commission of the European Communities (CEC) and its objective is to enhance reuse of multilingual learning resources in the field of human genetics. The consortium includes 16 academic providers of learning materials based in 11 European countries and 2 partners specialized in machine translation. The role of the Knowledge Media Institute in the project is to support the annotation and retrieval of learning resources.

In practise, we usually refer to a piece of educational content as a learning object (LO). Learning Technology Standards Committee of the IEEE Computer Society defines a learning object as "any entity, digital or non-digital, that may be used for learning, education, or training" [1]. Our approach is focused on processing learning objects containing text in different languages, such as slide presentations or textbooks.

In order to allow retrieval and reuse of LOs from online Learning Object Repositories (LORs), it is necessary to annotate them with appropriate metadata. An essential requirement for the metadata is that they are described in a machine readable way, thereby allowing interoperability on the Semantic Web. Currently, the most widely used metadata scheme for the description of LOs is the IEEE Learning Object Metadata (LOM) standard [1]. Broadly speaking, LOM describes the metadata fields that can be used to describe a LO. We have classified the most important metadata fields into three types:

1) Metadata fields describing the content of a learning object. This type of metadata specifies concepts, such as the name of an author, title of a LO or a set of keywords used.
2) Metadata fields classifying a LO using a taxonomy. This type is used to associate a LO with a a coarse grained structure of the subject domain.
3) Metadata fields connecting two LOs usually by a semantic relation.

Manual provision of type 1 metadata requires an annotator to have only a knowledge about a given LO. Metadata of type 2 requires to understand the domain (i.e. to know where the LO fits), and providing type 3 metadata requires understanding of LOs available and checking whether a semantic relation holds.

If metadata about LOs are stored in a language independent way, type 1 metadata should allow for example, to search for a LO using a set of keywords in one language, while retrieving LOs in a specified set of languages. Type 2 should allow to browse learning objects according to topics they discuss regardless of their language, and type 3 allows to relate two LOs using a relation, such as that one LO is summarizing another or that one object is a prerequisite of another.

The rest of the paper is organized as follows. In section 2, we first explore the theoretical constraints in the annotation of LOs. In section 3, we discuss the core component of our approach - the multilingual domain ontology, which is applied to keyword extraction (type 1 metadata). Section 4 explaines how keywords in a language independent representation can be used for classification, thus provision of type 2 metadata, and for generation of semantic links referring to type 3 metadata.

## 2 Theoretical constraints

We will now explore the time constraints of the above metadata provision tasks by investigating their complexity. Let $t_1$ denote the maximal time needed to access, view and broadly understand a LO. Let $h$ denote the number of nodes/topics in a classification taxonomy and $t_2$ denotes the maximal time needed to check whether a given LO should be associated with the node in a taxonomy. Finally, let $n$ be the number of LOs available. Then, the maximal time $t_{max}$ needed to provide type 1 metadata is:

$$t_{max} = t_1.n \Rightarrow t(n) = O(n) \tag{1}$$

thus the time complexity is linear with respect to the number of LOs available. The maximal time needed to generate type 2 metadata is:

$$t_{max} = (t_1 + t_2.h).n \ \Rightarrow t(n) = O(n) \tag{2}$$

The maximal time is given by the time of understanding a LO plus the time of associating the LO to a taxonomy times the number of LOs available. The complexity is still linear with respect to the number of LOs available, but the actual time required for annotation rises quickly with the size of the taxonomy.

Finally, maximal time spent in deriving type 3 metadata is given by the following expression:

$$t_{max} = (t_1.n).[t_1.(n-1)] \Rightarrow t(n) = O(n^2) \qquad (3)$$

This equation states that for the creation of links specifying binary semantic relations it is necessary to access all LOs and to take into account all remaining LOs. Thus, the time complexity is quadratic with respect to the number of LOs stored in the repository.

As a result of this, it can be seen that when $t_1$ is small, providing type 1 metadata may be feasible for human annotators. Generating type 2 metadata may be still possible when $t_2$ and especially $h$ are small. However, specifying type 3 metadata can be performed by humans only for a very limited number of LOs. For example, if we assume that accessing and understanding a LO takes one minute, interlinking of a repository of 100 LOs can take up to 165 hours. Furthermore, binary linking of LOs requires constant maintenance of the metadata fields as the amount of LOs changes. Multilingual environment makes it even more difficult for humans to perform such a task. On the other hand, computer systems are capable of generating links in repositories containing up to one million of LOs [4]. [1]

## 3 Multilingual ontology

In the Eurogene project, we have developed an English monolingual domain ontology of genetics by merging 6 genetic glossaries[2] that contained a descriptive, but not too extensive, terminology for our domain. The terminology currently contains about 1,700 concepts. These concepts were translated by providers of educational content with the help of machine translation into 6 languages (English, French, Spanish, German, Italian and Lithuanian). The providers were instructed to provide all possible versions (terms) of a concept being used in their target language. The ontology currently contains more than 12,000 terms.

The terminology is represented in a Simple Knowledge Organization System (SKOS) like structure. Using SKOS, concepts can be easily labeled with lexical strings in one or more natural languages. In particular, SKOS defines for a resource property `skos:prefLabel` and `skos:altLabel`. The former can be used to specify a preferred string label for a concept in a particular language while the later is used to specify an alternative string label for a concept. In this way, SKOS helps us to connect different representations of the same concept in multiple languages. SKOS also allows to specify relations between concepts, such

[1] This applies in the case when all possible pairs of LOs are checked, thus algorithms with $O(n^2)$ complexity are used. For even larger repositories it would be necessary to compute approximations by algorithms with lower complexity (see [4]).

[2] Published by the University of Washington in Seattle, National Institute of General Medical Sciences in Bethesda, Elsevier, Oracle ThinkQuest, University of Michigan and Centre for Genetics Education in Sydney

as `skos:broader`, `skos:narrower` and `skos:related`, that are used to create *isa* hierarchies and to refer to related concepts in a vocabulary.

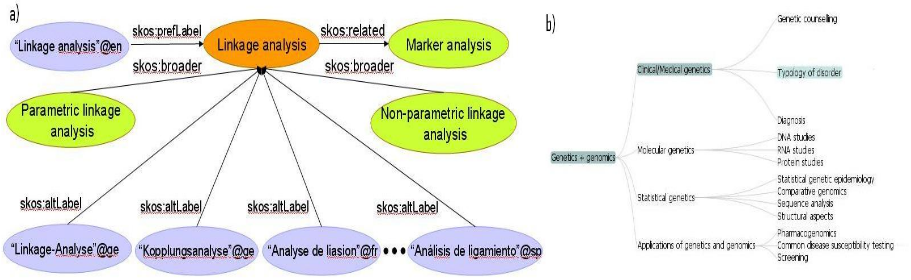

**Fig. 1.** a) Representation of a concept *linkage analysis* in the multilingual ontology b) top part of a topic hierarchy developed in the Eurogene project

Figure 1 a) shows how a genetic concept *linkage analysis* is represented in our ontology. The preferred label of this concept is the English version *Linkage analysis*. The concept has a two alternative representations in German (*Linkage-Analyse* and *Kopplungsanalyse*). The representation in French is *Analyse de liasion* and in Spanish *Análisis de ligamiento*. The concept *Linkage analysis* is a broader concept for *Parametric linkage analysis* and *Non-parametric linkage analysis*, and it is related to a concept *Marker analysis*.

## 4 The annotation process

**Extraction of type 1 metadata** - This type of metadata is provided in a semi-automatic way. Authors describe metadata which can be provided quickly and easily, such as name of the author or a LO's title, and the system automatically extracts a set of keywords.

The multilingual ontology is used to annotate textual content of LOs using its concepts. The source language of a LO can be detected automatically and a language specific stemmer [5] can then be applied on a LO's text. When the stemming is finished, the terminology of the detected language is loaded and applied to find all of the occurrences of the terms present in the source text.

As the ontology connects different syntactic representations of a concept, it allows us to abstract to a language independent representation, i.e. from terms to concepts. We assume that the main carrier of information in a domain specific LO is the terminology the LO contains. Based on this assumption, a learning object may be represented using a Vector Space Model (VSM) where dimensions of the vector correspond to concepts. This means that each LO is represented by a vector of length $n$, where $n$ is the number of concepts in the multilingual domain ontology. Non-zero values of the vector correspond to concepts acquired by abstracting from terms found in the LO's text.

**Extraction of type 2 metadata** - In order to logically organise the content within the LOR, LOs are usually associated to a certain node in a hierarchy of topics. This is particularly helpful for students who are not experts in the field and are currently unable to form a good query (because they do not know what they should search for). The LOM standard suggests to use the `classification` element for this purpose. Based on this type of metadata, an e-learning system may allow to browse a hierarchy to retrieve LOs relevant to a given topic.

In the Eurogene project a topic hierarchy of genetics has been developed by providers of the learning content. A small fraction of the hierarchy, consisting of about 200 topics, can be seen in figure 1 b). The association of LOs to the hierarchy is currently done manually with the goal to get a critical amount of training data. The next step is to use the concept vectors to classify incoming content automatically or to suggest at least a class in the topic hierarchy. A solution we are exploring is application of statistical machine learning classifiers, such as Support Vector Machines (SVMs). As the dimensions of concept vectors correspond to concepts rather than terms this approach allows to automatically categorize content regardless of its source language.

**Extraction of type 3 metadata** - Finally, a task in which providers of learning content urgently need to be supported is the generation of semantic relations. It is clear that there are more semantic relations that, if discovered, may help users of educational content to navigate over an e-learning system. We believe that links to similar and complementary content and links to content that discusses a topic in a more or less detail are particularly important.

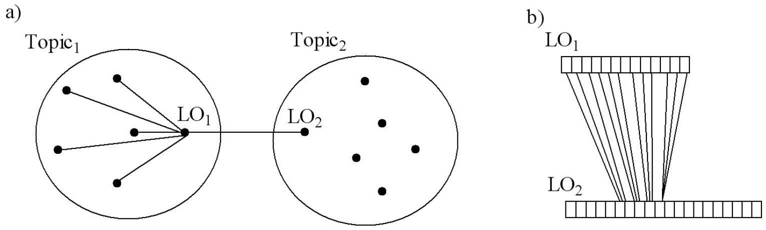

**Fig. 2.** a) $LO_1$ is semanticaly similar to other LOs, but the relation between $LO_1$ and $LO_2$ may express a complementary view on the same issue. b) $LO_2$ summarizes $LO_1$

There is a variety of criteria to measure the semantic similarity of two vectors including Pearson's product-moment correlation coefficient, cosine similarity and the Minkowski-measure. These techniques can be straightforwardly applied to our language independent concept vectors. However, for a learner it may also be useful to detect cases, where LOs share terminology, but address a topic from different perspectives. We believe that this type of relation can be detected by checking the association to the topic hierarchy. The most common case is probably that a given LO is similar to LOs associated with the same topic in the topic hierarchy, as we expect that association to the same educational topic implies similar use of the terminology. However, as shown in Figure 2 a), we hypothesize that if $LO_1$ is associated to a different topic than $LO_2$ and both are similar according to a selected similarity measure and a chosen threshold, then there may exist a complementary relation between $LO_1$ and $LO_2$, which expresses that they may discuss the same problem from different perspectives.

Another type of semantic relation can be discovered in a similar way as in [2]. Textual parts of multilingual LOs can be automatically divided into parts, such as slides or pages, or using topic segmentation methods [3]. Each of these parts can be annotated separately using a multilingual ontology. Similarities can then be measured between parts of two LOs. Then, according to figure 2 b), if a fraction of parts of $LO_2$ is similar to most parts of $LO_1$, it is probable that $LO_1$ discusses the same topic in more depth than $LO_2$ and that $LO_2$ summarizes $LO_1$.

The Eurogene system currently supports the automatic extraction of keywords, association of a LO with a taxonomy, and computation of semantic similarity between LOs across languages. This has been tested over a corpora of over $2,000$ LOs. Our plan is now to integrate and evaluate the extraction of more sophisticated semantic relations among LOs.

## 5 Conclusion

We have analyzed and discussed that there is a need to automate the extraction of different types of metadata to support current standards for describing LOs on the Semantic Web. If e-learning systems should be truly interoperable, it is also necessary to face the problem of annotation in multilingual settings. This is particularly important in Europe to enable learners and educators to share and reuse learning resources.

## References

1. *IEEE P1484.12.3/D8 - Draft Standard for Learning Technology - Extensible Markup Language Schema Definition Language Binding for Learning Object Metadata*, 2005.
2. Allan, J. Automatic hypertext link typing. In *HYPERTEXT '96: Proceedings of the the seventh ACM conference on Hypertext* (New York, NY, USA, 1996), ACM, pp. 42–52.
3. Dias, G., Alves, E., and Lopes, G. P. Topic segmentation algorithms for text summarization and passage retrieval: An exhaustive evaluation. In *Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence, July 22-26, 2007, Vancouver, British Columbia, Canada, AAAI 2007.* (07 2007), AAAI Conference on Artificial Intelligence, AAAI Conference on Artificial Intelligence, pp. 1334–1339.
4. Manning, C. D., Raghavan, P., and Schütze, H. *Introduction to Information Retrieval.* Cambridge, July 2008.
5. Porter, M. F. Java implementation of porter's algorithm, 2000.