A Latent Semantic Analysis approach to improve semantic Learning Object annotation

M. Gentile, D. Taibi, G. Fulantelli, D. La Guardia and M. Allegra

Italian National Research Council - Institute for Education Technology

Key words: Semantic metadata, LSA, Learning Object, Semi-automatic Annotation

Abstract:

In this paper we describe an approach to improve the indexing and searching mechanisms in a learning object repository using semantic web technologies. The proposed solution is based on the Latent Semantic Analysis technique for a semi-automatic annotation of learning objects. The concepts presented in this paper are based on the experience of using a Learning Object Management System called FreeLOms developed within the framework of the EU-funded SLOOP Project, Sharing Learning Objects in an Open perspective.

1 Introduction

Over the last few years we have witnessed a rapid evolution of ICT-based solutions in education. In this scenario, specific issues concerning the production and sharing of learning resources and, more in general knowledge management, have been extensively analyzed. The proposal to structure learning contents according to the model of the Learning Object (LO) has evolved out of this context. The characteristics of durability, interoperability and reusability and the related standardization process needed to achieve these goals, have played an important role in the diffusion of the Learning Object model. In particular the standardization process has focused on two main aspects: the description of LOs to provide efficient search and retrieve mechanisms; the model of LO to guarantee the reusability and interoperability of educational resources in the hundreds of learning platforms and learning systems available worldwide.

Amongst others, the IEEE Learning Object Metadata (IEEE LOM) [1] and the Shareable Content Object Resource Model (SCORM) standard [2] are rapidly being adopted by Learning Management System (LMS) constructors as well as by digital content developers. The LOM standard facilitates description, search, evaluation, acquisition and use of LOs, while the SCORM provides the technical specifications in order to guarantee learning content interoperability and reusability.

At the same time it is necessary to consider the doubts, raised by different authors, about the lack of pedagogical aspects in the description of LOs [3], [4]; for example there are no references to the educational context or to the didactic process in which the learning objects can be used [5]. Moreover, the use of technical standards complicates the task of developing learning objects for the teacher. But it is particularly due to its shortcomings as far as the reusability issue is concerned that the LO model has failed to live up to expectations [6]

One reason for this is that Learning Objects are often compared to Lego building blocks [7] which can be combined to create new Lego objects; similarly learning objects could be
aggregated to create new lessons or new courses; this vision is one of the aspects of the learning object model which has come under most criticism, as mentioned in [8], since it has been proved that the combination only works in a few cases. Different studies, [9][10][11] amongst others, suggest that the introduction of a semantic layer can improve the management of learning resources and allow users to search, assemble and reuse learning resources in a semantically valid way. The idea of semantic information is strictly connected to the concept of ontologies. An ontology is a specification of a conceptualization[12]. It describes the concepts and relationships of some phenomenon in the world. By using ontologies it is possible for computer agents to process data semantically since there is a common knowledge base, made up of terms and the relationships between these terms. This paper describes how the metadata management in the FreeLOms platform can be enhanced by using semantic web technologies in order to improve the indexing and searching processes of learning objects. After describing the learning technology standards which are closely related to the semantic web concepts, the state of the art about tools and methodologies used to automatically generate LOM is presented. Then the metadata management in the FreeLOms platform, with special focus on the OpenLO model [13] and on the Learning Object Management System (LOMS) [14], is reported. Finally, the semantic annotation of learning objects through Latent Semantic Analysis (LSA) technique is described.

2 Learning technology standards and the semantic web

2.1 Learning Object Metadata

The efficient retrieval of learning resources is a well-known problem. For this reason it was immediately decided that an efficient system was needed for classifying LOs. For example, IEEE LOM and Dublin Core (DC) [15] are two initiatives specifying a standardized set of metadata that facilitates retrieval of learning resources. Unlike the DC model which describes the essential references of any Web resource, the IEEE LOM aims to communicate, in addition, a whole set of information which will identify the resource for educational purposes. The IEEE LOM uses a common pre-defined vocabulary to describe the content of LOs. There are nine categories in the model including General, Lifecycle, Meta-metadata, Technical, Educational, Rights, Relation, Annotation and Classification. The description of Learning Objects through metadata [9] facilitates the creation of cataloguing and classification mechanisms for learning resources. This is extremely important for improving search procedures, especially considering the large number of learning objects available on-line that makes the traditional search procedures difficult.

For both IEEE LOM and DC an XML binding is defined; the XML structure of LO metadata facilitates the indexing process; in fact it is possible to use the typical XML search mechanisms, such as XPath and XQuery, to retrieve annotated learning resources. However, the challenge is to overcome the limitations of text based research mechanisms. From this point of view the XML metadata are not sufficient; XML defines the structure of objects without explain the semantic of these structures. To represent the semantic relations between LOs and the related concepts we need an effective language such as the Resource Description Framework (RDF) [16]; RDF can formalize knowledge through subject-predicate-object expressions; terms and constraints to
compose these expression can be defined in RDF-Schema documents. RDF is the basis for the so-called semantic web.

The IEEE Learning Object Metadata RDF binding[17] provides an RDF representation of the IEEE LOM standard. The idea is to provide a mechanism which allows the RDF language to be used to describe the LOMs so they can then be integrated into a more complete ontological system. This means that a didactic resource can be linked to a domain ontology which represents the semantic relations of the concepts connected to the field of study.

Although RDF is a powerful language for representing knowledge, on its own it has no way of making inferences or deductions. It is necessary to go to a higher level where concepts can be associated with logical rules by means of specific ontologies.

As in the case of XML and RDF, the W3C has defined a standard for ontologies: ontology web language (OWL)[18]. The use of ontologies makes it possible to establish correspondences and relations between different dominions of information.

2.2 Shareable Content Object Resource Model

As far as interoperability is concerned, the adoption of standards like SCORM allows teachers to create learning materials that can be used in different Learning Management Systems. In particular, SCORM is a set of standards that defines a model to create, share, and re-use learning resources.

First of all the SCORM defines the content model, that is the set of components that are used in a learning experience, like:

- Asset, the most basic form of learning resource; it can be a single media, or a text, an assessment object.
- Sharable Content Object (SCO), the smallest unit that can be used as a self-contained LO, in fact it represents the lowest level of granularity of a learning resource that can communicate with an LMS using the SCORM Run-Time Environment; moreover, it can be a collection of one or more Assets.
- Content Organization; it provides a complete learning experience, introducing Activities as structured units of instruction; the Activities represented in a Content Organization may consist of other activities (there is no set limit to the number of levels of nesting for Activities).

SCORM defines also how such elements must be organized to obtain the intended behaviors of the learning experience. Moreover, SCORM uses the IMS Content Packaging Specification to provide a standardized way to exchange learning content between different systems or tools. The Content Package also provides a place for describing the organization and the intended behaviour of a collection of learning contents.

A Content Package contains two major components: the manifest file and the physical resource files.

In particular the manifest file is an XML document that describes the content structure and the associated resources of the package. The standard requires that a manifest must be present at the root of the content package. The manifest is composed of four major sections:

- Meta-data: data describing the content package as a whole.
- Organizations: used to describe how the content is organized in the content package. The content organization should not be confused with the physical structure of the content package, or with the structure of the manifest itself. For example, the files in a content package are often organized in a hierarchy of folders, but that structure in itself cannot tell the user of a content package how to use the content of the package.
- Resources: defines the learning resources bundled in the content package, which are assumed to be physically located in the package.
- (sub)Manifest(s): describes any logically nested units of instruction (which can be treated as stand-alone units).
The organization of resources specifies in which order the resources must be shown to students accessing the LO through an LMS. Moreover, the organization specifies the relationships between activities and the resources. In particular, the organization represents a tree structure where each node is called an item; each item can be related to a resource or it can contain other items in a nested structure.

Sequencing information is external to the learning resources associated with those learning activities. In fact, the LMS is responsible for launching learning resources associated with these activities. This is important both at a conceptual and a practical level, because the reuse of learning resources is limited if a learning resource has embedded sequencing information that is context-specific to a specific learning experience.

According to different authors [19][39] SCORM compliant resources contain relevant information, beyond metadata, that allow a semantic description of LO.

3 OpenLO and semantic metadata: the reusability issue

In [9] authors analyze the relations of LOs and the Semantic Web, identifying several kinds of ontologies related to LOs: domain ontologies, e-learning ontologies, teaching and learning strategy ontologies, and learning object structure ontologies.

Some authors[19] suggest adding a context ontology in order to describe the educational context where a learning object has been used; others propose the addition of an ontology related to the student, thus guaranteeing a personal content customization depending on a student’s previous knowledge[11].

In our opinion, just using a semantic level to describe the resources does not allow the evolution of the existing contents and does not guarantee a real reusability of learning objects. Actually, the application of semiautomatic procedures to reorganize closed contents (contents that cannot be modified) generates hybrid objects that may exhibit coherence at a semantic level, but in practice are difficult for teachers to reuse easily.

For this reason we sustain [13][14] that, in order to guarantee effective pedagogical reusability of the educational materials in a constructivist approach [20][21], it is necessary to provide content authors with mechanisms to enter the learning object at an appropriate level; in such a way, the author can modify and evolve the content according to his real pedagogical needs.

To overcome this problem, we need to rethink the current model of LOs, moving to a new model that we call Open Learning Object (OpenLO)[13]. Only through this new model can users edit LOs created by different authors, and customize the LOs according to their own pedagogical needs.

The OpenLO model reconsiders the role and opportunities offered by the use of LO metadata. In the OpenLO model, metadata assume a key role as essential tools in supporting the evolution of learning resources. Metadata could act as descriptors of the evolving process of the resources. These considerations reveal the need for new tools which can treat metadata not only as static data but as information in constant evolution, thus supporting effectively the development of educational resources.

3.1 FreeLOms: a Learning Object Management System

Traditional Learning Object Repositories are not sufficient to implement the OpenLO model, since they do not manage the evolution of open contents; so we need new environments that we have called Learning Object Management Systems (LOMS). LOMS must have innovative features to manage learning objects: they have to support versioning mechanisms for LOs, both at metadata and content levels; they must provide an environment for sharing resources...
and tools to support collaborative work; they must support e-learning standards to guarantee interoperability between learning environments; and, finally, they should permit the development of learning resources following an open licence (e.g. creative commons) in order to guarantee their editing and their effective reuse.

Within the framework of the EU-funded SLOOP Project[38], Sharing Learning Objects in an Open perspective, we have developed FreeLOms, a Learning Object Management System aimed at managing learning objects according to the OpenLO model[14].

![Fig. 1: A folder from the FreeLOms repository](image)

The main objective of FreeLOms is to provide a community of teachers with an on-line platform to share and produce learning resources collaboratively: new learning objects can be developed as the evolution of LOs already stored on the platform. FreeLOms is inspired by the philosophy behind the Open and Free Source Movement, and is based on the idea that not only can the software be freely developed by a community of practice, but also the educational digital contents can be developed by a community of teachers and educational experts. This vision is reinforced by the adoption of the new OpenLO model.

FreeLOms includes functionalities for:

- uploading digital educational resources into a repository (LOs in SCORM terminology: Assets, SCOs or Content Aggregations);
- editing LO Metadata (IEEE Standard for Learning Object Metadata); editing of metadata can occur at any stage of the LO lifecycle, not only when it is uploaded onto the platform;
- searching for LOs shared by the users; specialized and personalized searches can also be defined in order to meet the needs of authors who usually apply the same search criteria (e.g. to search for some specific topics in their field);
− managing existing LOs in a SCORM vision, by allowing users to edit Assets, SCOs and Content Aggregations (CAs);
− creating Content Aggregations by using the resources available in the repository;
− managing the changes made to the didactic contents through versioning and differencing, both at metadata and content levels; more precisely, these features will make it possible to handle the contributions supplied by each user on the same LO, also through RSS feed, thus guaranteeing the “collaborative evolution” of LOs;
− allowing end users to access the repository as though they were accessing a shared drive through CIFS;
− transforming digital contents developed in technical formats unsuitable for learning platforms, into contents compliant with the SCORM standards;
− communicating asynchronously and/or synchronously with other users in order to support group processes; this reflects the typical functionalities available in a Computer Supported Collaborative Work system, providing an efficient environment for the collaborative management of didactic re-sources.

In order to create a complete environment for managing, sharing and also using learning objects, we have developed a FreeLOms module that combines the learning objects repository features of FreeLOms with the Learning Management System features of Moodle [22]. This module has been developed by modifying the SCORM module used by Moodle in order to provide direct access to the learning objects stored within FreeLOms.

4 Automatic Learning Object Metadata Generation: an overview

In the work of [23] a framework for the automatic generation of metadata is proposed, that is based on a formal analysis of the nature of metadata defined in the IEEE LOM standard [1]. Starting from the analysis of the learning object life cycle, it is possible to classify metadata in order to identify the set of metadata that can be obtained from the information generated by each stage of the life cycle. In particular, these information can be collected from various sources of data, e.g. the Learning Management System where these resources are used. While some works in the definition of tools for the automatic generation of metadata are concentrated only on single issues, such as, for example, the context of use of resources, in [23][24][25] several possible sources of data from which it is possible to extract information useful for the description of the resource are analyzed.

Many works consider the authoring phase as a key element for the automatic generation of metadata. In this phase for example, the metadata can be analyzed in the light of the relationship that exists between the learning resource as a whole, and the parties that compose it.

In the works of [26][23] the metadata IEEE LOM are classified to highlight when these may be inherited by the parties or vice versa when the metadata of the parties may contribute to the definition of metadata of learning resource (accumulate metadata). Many researches insist on the relationship that exists between the learning resource and the users and on the information that may be obtained to describe the learning resource from the analysis of context of use or user profile. Some authors insist on subjective nature of metadata [27] and propose indeed an active role of users in the implementation of metadata. Brooks and McCalla [28] propose a so-called “ecological approach” in which it is by analysis of the user profile that is possible to derive information useful for the categorization of the resource. The central role of users is emphasized also in those works that analyze the mechanisms of collaborative filtering and recommendation for the retrieval and the evaluation of educational resources. These works exploit social relationships between users trying to first locate people who might “…share a great deal of interests with the searching person”[29].
From a technical point of view, different works highlight the limits of the XML LOM binding, indicating as essential points of criticism, on the one hand, the lack of a shared vocabulary and by other the impossibility to bind to the descriptions the different contexts in which these descriptions are created [28]. Forte proposes [30] a shared thesaurus of possible values in order to promote interoperability of the generated metadata. Other studies [31][28] suggest an ontological approach and the use of the RDF binding [32] allowing the insertion of these descriptions into a graph of concepts. Accordingly, different authors highlight that the use of semantic web technologies is useful to face different problems connected to the learning resources.

The process of annotating the learning resources semantically consists in associating a semantic description to the resources in order to link them to the corresponding concepts and properties as defined in the specific learning ontologies. Metadata become machine understandable information that can be used to improve the management of learning objects, such as the retrieval mechanisms in a repository. Moreover, some works look in detail how some ontological relationships between educational resources may be translated into relations between the values of the related metadata. In this sense, the ontological relationships between the learning resources may generate simple rules that may facilitate the automatic generation or the validation of metadata [31]. The ontological relationships between the learning resources may take an important role in other phases of the cycle of life as the design phase or stage of retrieval. Starting from a learning unit syllabus and analyzing the relationships between the resources that make it up, the LessonMapper Toolkit [33] allows to obtain a description of the resource that it intends to find or create ex-novo.

5 Latent Semantic Analysis technique for a semi-automatic annotation of learning objects

Semantic web technologies can be used in the FreeLOms platform to store both learning resources and metadata. In order to facilitate the semantic annotation of learning resources, the LSA technique can be adopted for the automatic classification of the resources. LSA technique analyses the text of the learning resources through statistic algorithms in order to produce a semantic representation of the resources. The resources which have been already annotated with the concepts of the UK Archival Thesaurus (UKAT) domain ontology[36] can be used as training data for the classification algorithm. The UKAT ontology is the OWL version of the Simple Knowledge Organization System (SKOS) model [37].

The efficacy of this iterative process increases with the number of the training data resources available in the repository. Starting from the learning resources stored into FreeLOms a word-documents co-occurrences matrix can be built. The obtained matrix is then processed with a truncated singular value decomposition (SVD), in order to get a proper approximation of the original matrix. This implies the emerging of latent information which leads to a sub-symbolical coding of the semantic of words[34][35].

From the SVD application three matrices are obtained:

\[ A = U \Sigma V^T \]

The obtained matrices provide compressed information about the learning resources. In particular the i-th row of the U matrix, multiplied by the square root of the \( \sigma_i \) element in the \( \Sigma \) matrix represents the i-th document, while the i-th row of the V matrix, multiplied by the square root of the \( \sigma_i \) element in the \( \Sigma \) matrix represents the i-th word.
At the end of the process a representative vector is associated to each learning resource. These vectors are used to train a single layer neural network. This network produces the connections between the resources and the concepts of the UKAT ontology. When a new resource is uploaded into the platform, the corresponding representative vector is calculated using the folding-in technique. This vector is used by the neural network to identify the concept associated with the uploaded resource. The user can accept the concept suggested by the system or he can select different concepts from the UKAT ontology. Finally, the annotated resource is added to the training data thus improving the classification procedure.

6 Conclusion and further work

Our experience in using a Learning Object Management System called FreeLOms has proved that:
- the classification of learning resources is still an unresolved problem;
- most of the e-learning systems offer to the user a single way to associate metadata to their resources: filling metadata fields by hand.

The possibility to semantically annotate the resources stored in a repository could provide users with effective knowledge organization procedures and search mechanisms. The application of a semantic level to educational resources developed according to the OpenLO model provides further advantages, since it represents an extremely important opportunity to activate semantic-based LO aggregation procedures and, at the same time, to modify the newly created resources. In order to exploit the potential of semantic annotated learning objects it is necessary to make simpler the description of learning resources by means of automatic generation of semantic metadata. In such a way, it is possible to guarantee effective reusability of the educational resources.

In this paper, we have proposed an approach based on Latent Semantic Analysis (LSA) techniques for the semantic annotation of learning objects. The proposed approach is strongly dependent on the textual content of the learning resources and does not consider other types of media. For these reasons we are evaluating different approaches to automatically annotate other types of media such as image, video and audio.

References:


Author(s):

Manuel, Gentile
Italian National Research Council
Institute for Educational Technologies
Via Ugo La Malfa 153 - 90146 Palermo - ITALY
e-mail: manuel.gentile@itd.cnr.it

Davide, Taibi
Italian National Research Council
Institute for Educational Technologies
Via Ugo La Malfa 153 - 90146 Palermo - ITALY
e-mail: davide.taibi@itd.cnr.it

Giovanni, Fulantelli
Italian National Research Council
Institute for Educational Technologies
Via Ugo La Malfa 153 - 90146 Palermo - ITALY
e-mail: giovanni.fulantelli@itd.cnr.it

Dario, La Guardia
Italian National Research Council
Institute for Educational Technologies
Via Ugo La Malfa 153 - 90146 Palermo - ITALY
e-mail: dario.laguardia@itd.cnr.it

Mario, Allegra
Italian National Research Council
Institute for Educational Technologies
Via Ugo La Malfa 153 - 90146 Palermo - ITALY
e-mail: mario.allegra@itd.cnr.it