Frameworks for the Automatic Indexation of Learning Management Systems Content into Learning Object Repositories

Xavier Ochoa, Kris Cardinaels, Michael Meire, Erik Duval
Escuela Superior Politécnica del Litoral, Ecuador
Katholieke Universiteit Leuven, Belgium
xavier@cti.espol.edu.ec
{krisc, michael, erikd}@cs.kuleuven.ac.be

Abstract
Learning Management Systems (LMS) generally offer big repositories of learning material, but the lack of publicly available metadata to describe the learning objects makes it very difficult to share and reuse these objects. If metadata could be generated for these objects in order to index them in Learning Object Repositories (LOR), the gained amount of resources could solve the sub-critical mass problem of most existing Learning Object Repositories. This work proposes two orthogonal frameworks that could facilitate the analysis, design and implementation of Automatic Indexers for LMS content. One of them focuses on the methodology needed to pass from a LMS to a LOR, while the other focuses on the technological aspect of the Automatic Metadata Generation. Prototype implementations of Automatic Indexing Systems for real LMSs (SIDWeb and Toledo-Blackboard) show that the amount of effort needed to construct them was small compared with the benefits that this kind of systems could have in educational technologies. While the metadata seems to be good enough at first sight, further research of the quality of automatic generated metadata is needed.

1. Introduction

LMSs have been a relative successful technology used in education. According to Green, 83% of post-secondary institutions in the US use some kind of LMS (2003). Similar studies in Europe (Keegan 2002, and Paulsen 2003) show the same trend. Although Learning Management System (LMS) is a broad term that is used for a wide range of systems used in e-learning, they usually include some type of web-publishing functionality (Paulsen 2002) that enables teachers to upload the content of their courses online. Thanks to this functionality, LMSs end up being big repositories of learning material. From thousands to ten thousands learning objects are typically contained in a mid-sized LMS.

Unfortunately, LMSs use a rather simplistic approach to manage learning objects: URLs to documents stored in a web-server. While this works well for the objectives of the LMS, being to give access to the learning object to the students of the course, the lack of more descriptive information about the object (e.g.: author, summary, duration, etc.) makes it difficult to implement more advanced functionality, notably the sharing and reuse of learning objects in a context different to the one of the course where they were published.

Learning Objects Repositories (LORs) evolved from this need to share and reuse learning material across a community of users. They treat the learning object as an entity which is described by several fields of information (metadata). This metadata could contain information about different aspects of the object: description, recommended usage, technical specifications, relation with other objects, etc. Each repository establishes which fields should be stored and which values are possible for each field. International standards exist to allow the interoperability between several repositories, e.g.: Learning Object Metadata (LOM) (IEEE 2001). MERLOT (www.merlot.org), SMETE (www.smete.org) and ARIADNE (www.ariadne-eu.org) (Duval et al 2001) are real examples of LORs.

As favorable as it may seem for education, Learning Objects reuse through LORs is not currently in mainstream use among teachers (Duval & Hodgins 2003). While most LORs have been in operation for several years (MERLOT: 7 years, SMETE: 5 years, ARIADNE: 7 years), the amount of learning objects indexed in any one of them is small and it is comparable in number with the amount of learning objects that a single medium-sized LMS contents (See Table 1). One of the main barriers for the growing of LORs is the lack of critical mass of teachers and content. Duval et al pointed that this is a "chicken or egg" problem where there are not enough teachers that use LORs because there are not enough content and there is not enough content because there are not enough teachers that use LORs (Duval & Hodgins 2003).

Table 1. Comparison between number of Learning Objects in LORs and a mid-size LMS

	MERLOT	SMETE	ARIADNE	SIDWEB
Number of LOs	≈11000	≈11000	≈ 5000	≈ 6000

This cycle can be broken if the LOR contents enough LOs to render the repository appealing to teachers. This work tries to find a solution to this problem through the Automatic Indexation of the content of existing LMSs into a LOR. Section 2 explains why this indexation has to be done automatically. Sections 3 and 4 propose a methodological and a technical framework, respectively, to create an indexation procedure and system. Section 5 presents study cases where the frameworks have been applied. Section 6 discusses the results obtained with the study cases and Section 7 recommends further work in the area.

2. Manual or Automated Indexing

Two main alternatives exist to index the learning objects present in an LMS into a LOR. It could be done either by manual or by automatic indexation. In the first alternative, an expert, after reviewing the learning object, generates the metadata values. In the second alternative, some kind of information extraction system tries to deduce the value of the metadata fields based on the information available about the object. We will compare the advantages and disadvantages of these two alternatives to the problem at hand.

If the indexation process is to be done manually there are also two options. The teacher that uses the LMS generates the metadata for their own LOs or alternatively, a group of indexers generate the metadata values for all the objects present in the LMS. The first approach scales because typically the average number of learning objects that an individual teacher indexes keeps relatively small no matter the size of the LMS. Unfortunately, this approach doesn't work in "real life" (Duval & Hodgins 2004). It requires more work from teachers that do not perceive any immediate benefit in filling 10 to 20 fields of metadata. The second approach, using a group of dedicated indexers, makes the indexing process transparent to the teachers, but the solution is not scalable. Bigger LMSs need a bigger group of indexers, and educational institutions are not willing to contract more personal whose only work is to index teachers' LOs. Voluntary review and indexing works but does not scale either. Example of this is MERLOT's reviewing system. Human review accounts for less than 15% of the published objects.

Automatic indexing seems to be the only feasible way for the task of indexing existing LMSs. For any automatic indexing scheme to work we need some kind of already available information about the LOs to index. Fortunately, LMSs implicitly store a great amount of information in the context where the LO is published: usually we have information about the course and lesson where the LO is, the description of the task to be performed with the LO, the navigation structure of previous and following material, creator and users information, etc. All this information could help in the design and implementation of information extraction systems.

There exists the perception in the LO research community that the metadata information automatically generated could not be as good as the one manually generated by the teacher. However, the generated metadata does not need to be perfect, just good enough to enable sharing (Duval & Hodgins 2004). The main goal of adding metadata to a LO is to render the object easier to search and retrieve in a repository, being this retrieval done by human or automatic searchers. If the metadata generated by the automatic indexer makes the objects "findable" it will fulfill its purpose. There is a broad spectrum of "non-perfect" but "extremely useful" applications that rely on automatic generated metadata: Google (www.google.com), Citeseer (www.citeseer.com), Ask Jeeves (www.ask.com), spam filters, etc.

The question must not be whether we should create automatic indexers to generate metadata from existing LMSs, but how to do it. The next two sections try to describe this "How" through the creation of two interrelated frameworks. One presents the methodological steps that should be followed to index an LMS, while the other focus on the technological structure to implement an automatic metadata generator.

3. Methodological Framework

We outline a methodological framework to automatically generate metadata information for Learning Objects present in an existing LMS. This is a generic framework that could be applied to different situations and technologies. The framework is divided in seven consecutive steps; each one details a process and an output. The output of the framework will be the metadata instances for all the learning objects published in the LMS. These instances could be stored in a local database (creating a local LOR) or could be added to an existing LOR (decision to be made in the first step). These instances could be manually checked later by willing teachers or could be automatically reviewed when more usage information is available.

3.1. Definition of objectives and policies

The first step in the road to create an automatic indexer for an LMS is to clearly define which will be the objectives and policies that will guide the following steps. During this step the following questions should be answered:

- *Which are the benefits that we look for with the indexing of the LMS content?*
  Whether it is to enable the sharing of resources between teachers to promote reuse, catalog learning content to use it in an intelligent tutoring system, let students to browse free through the institution content or any combination of the above, a clear answer to this question will help to guide the whole designing process.
- *A new repository is to be created or an existing one will be used?*
  Because this framework controls the whole process of creation of metadata, and the amount of learning objects in an LMS would create a decent-sized repository, the creation of a new repository is an alternative. While this framework is repository-neutral, it is recommended to aggregate the produced learning objects metadata to an existing LOR in order to obtain a critical mass of material that could fuel reuse.
- *What will be the sharing scope/copyright policy of the resulting metadata?*
  Who will have access to what metadata and what copyright rules should he/she follow to use the object is the answer to this question. It is not easy to answer and it will deeply depend on the actual policies of each institution. The rules of sharing have to be defined. It could be as simplistic as "free" or "paid" or could be regulated by copyright licenses like GPL, Creative Commons, etc. A guide to answer this question is beyond the scope of this framework, but a right answer is needed for the resulting system to be viable and accepted by the teachers and the institution.

3.2. Inventory of already available Information

Any attempt to construct an automatic indexer needs to identify available information about published learning objects. During this step, a list of all electronic sources information is made. These sources could be found inside and outside the LMS and define the context of the LO. The structure of the courses, the fields filled by teachers during the publication process, usage logs could be found inside the LMS. Curriculum, courses syllabus, teacher/students information (LDAP directories), university information and statistics could be found outside the LMS. All this information creates the context of the LO.

3.3. Selection of metadata fields and values

In this step, the information that would be stored about the learning object is selected. This selection is extremely related to the answers to the different questions of the first step (Section 3.1). There are three issues to deal with: Which metadata fields will be added, with which vocabulary these fields will be filled and which of the fields will be mandatory. For example, if the goal of the indexing process is to be able to share content with a world-wide community of teachers, a language specification field is highly important. On the other hand if the objects will only be shared across teachers in an institution the field language will not have special utility. The vocabulary used to fill these fields should also be selected based on the objective and scope of the indexation. If the repository will be used only in a Belgium institution, the language vocabulary will include French, Dutch, German and English, but if it will be use only in Ecuador, Spanish and English will suffice. Depending also on the importance of the field, it could be declared as mandatory, that is that it has to be filled for each learning object, or optional, that could be filled only if the indexer thinks that it is appropriate.

There is no restriction to which information could be stored, but if sharing with an enlarged group of institution is in mind, it is wise to choose a widely accepted metadata standard like LOM. Even if it is not possible (or desirable) to generate all the fields of the standard, a profile of it can be created. A profile is a subset of the standard fields that could allow extra fields not present in the standard. One requisite of this profile is that it can be easily converted to and from the standard representation in order to allow interoperability with other repositories (Najjaf et al 2003).

The output of this stage should be a specification of the metadata fields. Examples of this specification could be the Ariadne Metadata profile (ARIADNE Profile) or the LOM standard specification (IEEE 2001).

3.4. Classification of Learning Objects in different Levels

To generate metadata at the level of a course is different than to generate metadata at the level of a document. These different levels are made explicit inside the context of an LMS. As example, an LMS can be considered as an aggregation of learning objects at different levels. Usually the largest unit of knowledge in an LMS is the course. The course is normally divided into smaller parts like chapters. Each chapter is formed by one or more lessons. Lessons could have one or more documents associated.

Although easy, this step will generate a classification schema that could be used inside the automatic indexers to facilitate the metadata extraction (as described in Sections 3.5 and 3.7).

3.5. Mapping available information to different elements in the standard

Based on the inventory made in the Section 3.2, each metadata field is related to the information available to fill it. This mapping should be done for each metadata field and for each one of the elements of the classification made in the previous section. For example, Table 2 shows the mapping for the metadata field Duration and the Classification Course-Activity-Document. The table also shows that the information to calculate the value for an object in one level (Activity) could be the extracted form the information already calculated for objects in other level (Document).

Table 2. Mapping of available information to Pedagogical Duration Metadata Field

Pedagogical Duration	
Level	**Source**
Course	Syllabus (Hours per Week * Number of Weeks)
Activity	Addition of duration of its documents
Document	Size of the document, log registries of LMS

During the mapping there are three possible scenarios for the mapping of fields:

- *There is only one source of information:* The information only needs to be converted from the source format to the metadata value. The confidence of the extracted value will be tied to the confidence of the source information and the extraction method.
- *There are several sources of information:* Having several sources to fill a field always results in more robust results. Even if the sources are contradictory, conflict resolution methods could be used to obtain the value for the field. For example to determine the science type of a course we could have 3 sources of information: The faculty of the course, the faculty of the teacher and the output of an automatic classifier based on keywords. We could use a Bayesian Network (Heckerman 1996) to weight all the outputs and obtain a more robust value that if only one of the sources is used. The conversion step is analyzed in more detail in the next section.
- *There are no sources of information:* There is also a possibility that none of the available information could easily determine the value of a field. Two options arise: If there are no sources for a field we could leave that field unfilled (if it is optional); or we could try to find an average or best-guess value. As an example, there is no easy way to transform context information into the Difficulty value of a document. Based on statistics or empirical analysis (Najjaf 2003) we can assign the value Medium as the default value for all objects.

The output of this step is a mapping like the one described in Table 2 for each one of the metadata fields and each one of the classification levels.

3.6. Extraction and Conversion of information to metadata field values

The information needs now to be extracted from its source and transformed in the values to fill metadata fields. The extraction methods could be as simple as reading a field in from a database or as complex as text mining documents. Given that LMSs follow a structure and are based on databases, the extraction mechanisms are generally simple. For example the author of a document (the publisher to be exact) could be easily extracted from the LMS database, on the other hand, the extraction of the keywords of a document require the calculation of word frequencies and word relevance inside the text of the document. The conversion methods also could range from using the extracted data to fill directly the metadata field to using Computational Intelligence techniques to deduce the value of that field based on the available information (specially useful when there are several sources of information as explained in previous section). The most important rule to remember in this step is that we do not need perfect metadata, just useful metadata. "Do it as simple as possible, but not simpler" seems to be a good guideline for this step.

To implement the software needed to do the extraction and conversion, the authors develop a technological framework for Automatic Indexation that is explained in detail in Section 4. The output of this step is metadata instances generated for each one of the LOs present in the LMS.

3.7. Sharing and Cross-validation of metadata of related LOs

During this step three main actions could be taken to improve the quality of the generated metadata: sharing of metadata values between LOs at different levels, creation of links between related LOs and cross-validation and conflict solution of common metadata generated for LOs at different levels. As stated in section 3.5 certain metadata fields could be easily extracted in a high level of the classification (e.g.: discipline, author) while others could be easier to extract at a lower level (e.g.: size, technical requirements). The metadata generated for objects of high level could be used to deduce the value of empty fields of lower level objects. For example the science type of a course is inherited by all the lessons and documents. Also metadata generated at low level could be used to infer a value of metadata fields in higher levels. For example the interactivity level of a course could be inferred based on the interactivity level of its documents.

To index a whole course together also enables the creation of parent-son and previous-next relationships in the metadata. All the lesson objects could be linked to the course where they belong. Documents could also be linked with the lessons where they are used. This kind of relationship is usually not present in manual generated metadata instances because of the amount of repetitive work that it represents. These links created between objects could add navigational capabilities to LOR searches. For example, if the teacher finds a document and wants to know how it has been used, he/she only needs to access the lesson where that document is used.

Finally, if in two or more related objects a common field has been calculated, it is possible to use computational techniques for conflict management and cross validation (as mentioned in section 3.5) to "perfect" the value. For example the keywords extracted for a lesson could be compared with the ones extracted from its documents. Repeating keywords will have a better probability of being right keywords that the ones that are only found in just one object.

4. Technological Framework

The technological framework is described in detail in (Cardinaels et al 2005). It has been developed in Java, and is made available as a Web Service. A demo can be reviewed in (http://piepau.cs.kuleuven.ac.be:8989/is/IndexingForm2.html). It uses three types of classes to generate metadata, as depicted in Figure 1.

a. Extractors

These classes basically enable the other classes to do their job by getting the contents and other information of a learning resource out of the document. For now, 2 main groups of extractors exist: TextExtractors and PropertyExtractors. The first one contains classes that get the text from a document in plain format, such as the plain text in a word/pdf document. The second one for example retrieves the Microsoft Office-specific document properties.

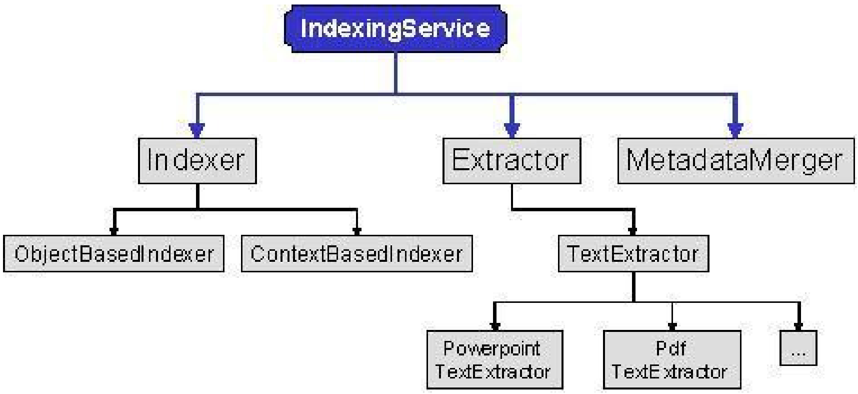

Figure 1: Overall Structure of the Automatic Indexing Framework

b. Indexers

The main metadata generation is done by the indexers. Indexers are classes that generate some specific part of metadata for a document. This type is split in two subtypes: ObjectBasedIndexers and ContextBasedIndexers. This distinction is based on the fact that metadata can be retrieved from 2 different categories of sources: the object itself and the contexts in which the document is used. One important category of ContextBasedIndexers is the one of LMSContextBasedIndexers. It contains classes that can extract information that the LMS offers. Possible sources of this kind of information are file system or database used by the LMS, or an API, offered to the end-user. For example the Blackboard system offers a Java API that allows you to retrieve information about authors, courses.

c. Metadata Mergers

When several indexers generate sets of metadata for one learning object, these sets must be combined into one resulting set that acts as the metadata record for the resource, which can be stored in the repository. As different indexers can generate different values for the same metadata element, it is important to have strategies to combine the records into one. Those strategies are implemented by the MetadataMerger classes. In some cases different values might be conflicting with each other. In other cases the different values all apply. The merger classes decide on how the correct value will be obtained.

5. Case Studies

5.1. SIDWeb Automatic Indexation

We implement the methodological and technical frameworks steps to automatically index Learning Objects published in SIDWeb, an LMS used in ESPOL, Ecuador. SIDWeb could be considered as a fair example of current LMSs: easy web publishing, user tracking, course structure, etc. It was taken as a test case also because one of the authors has a high degree of access to the implementation of the LMS. Table 3, Column 1 shows the decisions made during all the steps of the framework. The creation of an experimental demo of the Automatic Indexer for SIDWeb took 3 weeks of work of one programmer who was not involved in the creation of the technological framework. This demo generates metadata instances for courses and activities. The fields extracted are: Title, Authors, Language, Description, Keywords, Publication Date, Aggregation Level, Pedagogical Duration, Pedagogical Context, Source Documents and Relationships. All Computational Intelligence methods used were extracted from free (GPLed) available java libraries. The metadata generated with this experimental demo seems to be useful and gives ground to create a bigger implementation.

5.2. Toledo-Blackboard Indexation

In (Cardinaels et al 2005), we reported on a case study of the technical framework for the Toledo system, which is the name for the Blackboard Configuration at the K.U.Leuven. It was the first major test of the idea that a context in which a learning object is used, can provide us with useful metadata. For this particular system, we used the available Blackboard Java API, which allows you to retrieve information about authors, courses, etc. When investigating that API, we especially looked at what information could be used to fill the elements of the Ariadne LOM application profile, with most
attention going to the mandatory elements of that profile. The use of the Blackboard context indexer, together with the other parts of the framework, allowed us to generate a value for 17 out of 18 mandatory fields. Only for

pedagogical duration we couldn't generate a reasonable value so far. Table 3, Column 2 shows the decisions made during all the steps of the framework.

Table 3. Summary of Frameworks steps made by the two Study Cases

Framework Step		SIDWeb	Toledo BB
1. Definition of Objectives and Policies			
	Which are the benefits that we look for with the indexing of the LMS content?	To enable the sharing and reuse of learning objects	To enable the sharing and reuse of learning objects
	A new repository is to be created or an existing one will be used?	All the instances of LO metadata will be aggregated to ARIADNE repository	All the instances of LO metadata will be aggregated to ARIADNE repository if they are not already present there.
	What will be the sharing scope/copyright policy of the resulting metadata?	ESPOL is the owner of the material published on the LMS The metadata fields will be shared with the rest of the members of the ARIADNE Foundation under one of the Creative Commons licenses.	Not decided yet
2. Inventory of already available information			
	Inside LMS	- SIDWeb database (information published in the LMS) - Document content	- The Blackboard database and file system, made available by a Java API - Document content
	Outside LMS	-Word documents with courses syllabus - LDAP directory of teachers and students of ESPOL -Web sites of different faculties	- LDAP directory of teacher information
3. Selection of metadata fields and values		LOM standard, all fields	Ariadne Profile, Mandatory fields
4. Classification of Learning Objects in different Levels		Course has Chapters. Chapters have Lessons. Lessons have Documents.	Until now, we only consider course documents as such, so we only have 1 level.
5. Mapping available information to different elements in the standard		**Example:** *Metadata Field:* Pedagogical Context *Level:* Course *Available Information:* Type of institution, department, area, level in the curriculum, prerequisites and following courses, level of actual registered students.	**Example:** *Metadata Field:* author *Available Information:* Instructors of a course, LDAP directory of the university.
6. Extraction and Conversion of information to metadata field values		A class inherited from LMSContextIndexer is implemented. **Example:** Metadata Field: Pedagogical Context Level: Course *How to extract and convert:* Fixed values: the institution is a University. Easy text parsing from web pages or documents. Looking up information of students in the directory	A class inherited from LMSContextIndexer is implemented. **Example:** *Metadata Field:* author *Available Information:* First determine the author by using the Blackboard Java API, then determine the author details by using the LDAP directory of the university.
7. Sharing and Cross-validation of metadata of related LOs		Sharing among different levels. Link related LO Bayesian Network cross-validation	Sharing from higher to lower levels No structure constructed Fixed confidence value validation

6. Conclusions

Based on the experience we had in the study cases, it could be concluded that the frameworks proposed facilitate the creation of Automatic Indexing Systems. The methodological framework guides the steps needed to analyze and design these systems while the technological framework speed up the implementation. The amount of effort involved in the creation the test systems following the frameworks was insignificant compared with the amount of benefits that a mass migration of Learning Objects to existing repositories could give. Even if the automatic generated metadata is not as good as the one created by humans, the values obtained in the case studies seems to be "good-enough" and even content some features (like the link between related LOs) that could spark the creation of new search functionalities.

7. Further Work

This work could lead to further research and development work in the area of educational technologies. Among the most important future steps to be taken: Generate a complete implementation on several LMSs, evaluate how well the automatic generated metadata behaves in real retrieval tools and real searchers and compare automatically generated metadata against manual generated metadata. After this the first mass generation of metadata from an existing LMS (SIDWeb) into an LOR (ARIADNE) is planned. Also at this moment discussions are being initiated on this aspect of quality of metadata, questioning how the quality should be proven if we want to move toward recommender systems (Paul Libbrecht, ProLearn mailing list).

8. References

Green, K. (2003). **Campus Computing 2003. The 14th National Survey of Computing and Information Technology in American Higher Education.** The Campus Computing Project.

Keegan, D. (2002). **The use of Learning Management Systems in North Western Europe**. Web-Education Systems in Europe, Hagen: Zentrales Institut für Fernstudienforschung, FernUniversität, 58-81.

Paulsen, M. F. (2003). **Experiences with Learning Management Systems in 113 European Institutions**. Educational Technology & Society, 6 (4), 134-148.

Paulsen, M. F. (2002). **Online Education: Discussion and Definition of Terms**. Web-Education Systems in Europe, Hagen: Zentrales Institut für Fernstudienforschung, FernUniversität, 23-28.

Cardinaels, K., Meire, M., Duval, E. (2005). **Automating Metadata Generation: the Simple Indexing Interface**. preprint of an Article accepted for publication in ACM 1-59593-046-9/05/0005. International World Wide Web Conference Committee, WWW 2005, May 10-14, 2005, Chiba, Japan.

Najjar, J., Duval, E., Ternier, S., Neven, F. (2003).**Towards Interoperable Learning Object Repositories: The Ariadne Experience.** Proceedings of IADIS International Conference on WWW/Internet 2003, 219-226

Najjar, J., Ternier, S., Duval, E. (2003). **The actual use of Metadata in Ariadne: An Empirical Analysis**. Proceedings of Ariadne 3rd International Conference

Heckerman, D. (1996). **Bayesian Networks for Knowledge Discovery**. Advances in Knowledge Discovery and Data Mining, 1, 273-305

Duval, E., Hodgins, W., (2003). **A LOM Research Agenda**. WWW2003. Budapest, Hungary

Duval, E., Hodgins, W., (2004). **Making Metadata go away. Hiding everything but the benefits**. Dublin Core Conference 2004, Shangai, China.

Duval, E., Forte, E., Cardinaels, K., Verhoeven, B., Durm, R. V., Hendrikx, K., Forte, M. W., Ebel, N., Macowicz, M., Warkentyne, K., and Haenni, F. (2001). **The ariadne knowledge pool system**. Communications of the ACM, 44(5), 72–78.

MERLOT. Multimedia Educational Resource for Learning and Online Teaching. Available at: http://www.merlot.org/.

SMETE. The SMETE Digital Library. Available at: http://www.smete.org/.

IEEE, (2002). IEEE Standard for Learning Object Metadata. Available at: http://ltsc.ieee.org/doc/wg12/

ARIANDNE Profile, Available at: http://www.ariadne-eu.org/en/publications/metadata/ams_v32.html

SIDWeb, LMS System of Escuela Superior Politécnica del Litoral (ESPOL). http://www.cti.espol.edu.ec/sidweb