Universidad de Alcalá

Departamento de Ciencias de la Computación

DOCTORAL THESIS

EMPIRICAL FOUNDATIONS FOR AUTOMATED QUALITY ASSESSMENT OF LEARNING OBJECTS INSIDE REPOSITORIES

Author:

Cristian Cechinel

Supervisors:

Dr. Salvador Sánchez-Alonso

Dr. Elena Garcia-Barriocanal

May 2012

Abstract

Learning objects can be defined as small units of knowledge that can be used and reused in the process of teaching and learning. They are considered by many as the cornerstone for the widespread development and adoption of e-learning initiatives over the globe. Most of the current learning objects repositories (systems where learning objects are published so users can easily search and retrieve them) have been adopting strategies for quality assessment of their resources which are normally based on the opinion of the community of experts and users around them. Although such strategies can be considered successful at some extent, they rely only on human-work and are not sufficient to handle the enormously amount of resources existing nowadays. Such situation has raised the concern for the development of methods for automated quality assessment inside repositories.

The present dissertation approaches this problem by proposing a methodology for the development of models able to automatically classify LOs stored on repositories according to groups of quality. The basic idea of our dissertation is to use the existing on-line evaluations (evaluative metadata) of the repositories in order to divide learning objects on groups of quality (e.g., *good* and *not-good*), thus allowing us to search for intrinsic features of the resources that present significant differences between these groups. These features (metrics) are called by us "highly-rated learning object profiles" and are considered potential indicators of quality that can be used by classification algorithms as input information to create models for automated quality assessment.

In order to test our proposal, we analyzed 35 metrics of a sample of learning objects refereed by the Multimedia Educational Resource for Learning and Online Teaching (MERLOT) repository, and elaborated profiles for these resources regarding the different categories of disciplines and material types available. We found that some of the intrinsic metrics present significant differences between highly rated and poorly-rated resources and that those differences are dependent on the category of discipline to which the resource belongs and on the type of the resource. Moreover, we found that different profiles should be identified according to the type of rating (peer-review or user) under evaluation.

Based on these findings, we decided to restrain the generation and evaluation of models to the three intersected subsets - considering the categories of discipline, material type and the peer-reviewers' perspective of quality) - with the higher number of occurrences in the repository. For those subsets we generated and evaluated models through the use of Linear Discriminant Analysis and Data Mining Classification Algorithms, and we found preliminary results that point out the feasibility of such approach for these specific subsets. The dissertation ends by presenting two possible usage scenarios for the developed models once they are implemented inside a repository.

The initial results of this work are promising and we expect that they will be used as the foundations for the further development of an automated tool for contextualized quality assessment of learning objects inside repositories.

Keywords: Learning Objects, Repositories, Automated Quality Assessment, MERLOT

Resumen

Los Objetos de Aprendizaje pueden ser definidos como pequeñas unidades de Conocimiento que son utilizadas y reutilizadas en el proceso de enseñanza y aprendizaje. Ellos son considerados por muchos cómo las piedras angulares para la amplia diseminación y adopción de las iniciativas de e-learning por todo el globo. Gran parte de los repositorios de objetos de aprendizaje (sistemas donde los objetos de aprendizaje son publicados para que los usuarios puedan buscar y recuperarlos fácilmente) han adoptado estrategias para la evaluación de la calidad de sus recursos que están normalmente basadas en la opinión de la comunidad de expertos y usuarios del repositorio. Aunque dichas estrategias pueden de algún modo ser consideradas exitosas, ellas dependen solamente del trabajo humano y no son suficientes para encargarse de la enorme cantidad de recursos existentes hoy en día. Esa situación ha hecho aumentar la preocupación sobre el desarrollo de métodos para la evaluación automática de la calidad dentro de los repositorios.

La presente disertación aborda ese problema proponiendo una metodología para el desarrollo de modelos capaces de clasificar automáticamente los objetos de aprendizaje disponibles en repositorios en distintos grupos de calidad. La idea básica es utilizar las evaluaciones on-line existentes (metadatos evaluativos) en los repositorios para dividir los objetos de aprendizaje en grupos de calidad (e.g., buenos y no-buenos), permitiéndonos así obtener medidas intrínsecas de los recursos que presenten diferencias significativas entre esos grupos. Denominaremos a esas características (medidas) "perfiles de objetos deaprendizaje altamente puntuados" que son consideradas como potenciales indicadores de calidad y por tanto pueden ser utilizados cómo variables de entrada por algoritmos de clasificación enfocados en la creación de modelos de evaluación automática de la calidad.

Con el fin de examinar nuestra propuesta, hemos analizado35 medidas de una muestra de objetos de aprendizaje referenciados por el repositorio Multimedia Educational Resource for Learning and Online Teaching (MERLOT) y elaboramos perfiles para esos recursos teniendo en cuenta las distintas categorías temáticas así como los tipos de materiales disponibles. Durante nuestra investigación,

descubrimos que algunas de las medidas intrínsecas presentan diferencias significativas entre los recursos con puntuaciones altas y el resto y que esas diferencias son dependientes de la categoría temática a la que el recurso pertenece, así como del tipo de recurso. Además, nosotros observamos que los diferentes perfiles deben ser identificados teniendo en cuenta el grupo que realizó la evaluación (expertos o usuarios).

Basándonos en esos hallados, decidimos restringir la generación y la evaluación de los modelos para los tres subconjuntos cruzados (considerando las categorías tema, tipo de material y perspectiva de calidad de los expertos) pues son éstas las que contienen el mayor número de ocurrencias en el repositorio. Para esos subconjuntos generamos y evaluamos modelos a través de la utilización de Análisis Discriminante Lineal y Algoritmos de Minería de Datos para Clasificación, y encontramos resultados preliminares que indican la viabilidad de tal enfoque para esos subconjuntos específicos. La disertación finaliza presentando dos posibles escenarios de utilización de los modelos desarrollados una vez que estén implementados en un repositorio.

Los resultados iniciales de ese trabajo son prometedores, por lo que esperamos que los mismos sean utilizados como fundamentos para el futuro desarrollo de una herramienta automática para la evaluación contextualizada de la calidad de los objetos de aprendizaje dentro de repositorios.

Palabras clave: Objetos de Aprendizaje, Repositorios, Evaluación Automática de la Calidad, MERLOT

Acknowledgements

Several are those who helped me to go through the development of the ideas for this dissertation and its final writing. First of all, I am deeply grateful to the Information Engineering Research Unit of the University of Alcalá for welcoming me as part of their team. My most profound thanks go to Salvador Sánchez-Alonso, for his wise guidance during the preparation of this thesis, and to Miguel-Ángel Sicilia for the many times he enlightened me with his fresh and innovative research ideas. I would also like to especially thank Elena García-Barriocanal, for co-supervising my work and for the innumerous occasions when she assisted me with my administrative requests. Crossing my path with them was an exciting and fruitful experience that I will always carry along with me. I want also to thank to the current members of the IERU that have made me part of the group: Daniel Rodríguez-García, Miguel Refusta, Leonardo Lezcano, Ramiro Cano Goméz, Alberto Albián, Maria-Cruz Valiente, Rosmary Calzadilla González, Raquel Rebollo and Elena Mena Garcés.

My many thanks also go to Cleo Sgouropoulou and Nikos Manouselis who received me in Athens and shared with me their inspiring considerations about my dissertation.

Moreover, I would like to acknowledge the financial aid granted to me by Fundación Carolina. Through their mobility program for Brazilian Professors of Public Universities, they supported my stay in Spain for two distinct periods, in 2009 and 2011.

Finally, I thank to my wife, Fran, for her patience and emotional support during the hard periods where we stayed far from each other.

Summary

List of Acronyms

DCM	Dublin Core Metadata
DMCA	Data Mining Classification Algorithms
eLERA	E-Learning Research and Assessment Network
IEEE	Institute of Electrical and Electronics Engineers
LDA	Linear Discriminant Analysis
LMS	Learning Management System
LO	Learning Object
LOM	Learning Object Metadata
LOR	Learning Object Repository
LORI	Learning Object Review Instrument
LTSC	Learning Technology Standards Committee
MAE	Mean Absolute Error
MERLOT	Multimedia Educational Resource for Learning and Online Teaching
SLOR	Semantic Learning Object Repository
OCW	OpenCourseWare
OECD	Organization for Economic Co-Operation and Development
OER	Open Educational Resources
PRR	Peer Reviewer Rating
UNESCO	United Nations Educational, Scientific and Cultural Organizations
UR	User Rating

List of Figures

List of Tables

Chapter 1 Introduction

Learning objects (LOs) are often defined as digital entities that can be used and reused in the process of learning and education, and are considered by many as the cornerstone for the widespread development and adoption of e-learning initiatives. Several initiatives and proposals for LO quality evaluation have been discussed in the last years (Díaz, Sicilia & Aedo, 2002; Kay & Knaack, 2009; Nesbit, Belfer & Leacock, 2003; Nesbit, Belfer & Vargo, 2002; Williams, 2000) nevertheless, there is still no consensus on what constitutes a good quality LO, neither which is the best way of conducting the process of evaluation. In part, this can be attributed to the heterogeneous and multi-faceted nature of these resources. As they can differ in several aspects (size, granularity, technology used, type, metadata standard, instructional design, duration, etc.) (Churchill, 2007), it is reasonable to assume that the quality criteria and the ways of measuring these criteria will also differ accordingly to many aspects. Moreover, the different evaluation approaches also reflect the many particular contexts of usage for the learning objects, as each one of them usually measures quality from the perspective of "*a given repository, a country or a community of users*" (Vuorikari, Manouselis & Duval, 2008). As a matter of fact, the continuous growth of educational resources on the internet has turned impractical to rely only on human effort to classify good quality learning materials, and has raised the concern about the development of new automated techniques and tools that could be used to complement the existing approaches in order to relieve manual work.

1.1 Problem statement

The actual abundance of resources inside repositories (Ochoa & Duval, 2008, 2009) and the availability of contextual evaluations in some of them have opened the possibility of seeking for intrinsic metrics of learning objects that could be used as indicators of quality. This means to say that learning objects could be "*mined*" and quantitative measures of good and not-good resources could be compared in order to discover intrinsic attributes associated with quality. These attributes could then allow the creation of statistical profiles of good and poor resources that serve as the basis for quality prediction. In fact, such approach was previously successfully applied to automatically analyze the usability of websites by Ivory & Hearst (2002b).

It is known that learning object quality can be considered a more complex construct than usability as the latter is included in existing instruments as Learning Object Review Instrument (LORI) (Nesbit et al., 2003) as one out of the several attributes considered. Therefore, we cannot take for granted that the same correlations found by them are still applicable to ratings of learning objects (even though it may be hypothesized that the former affects the latter to some extent). So a first step in finding statistical profiles for highly rated learning objects is exploring evidence on potential intrinsic measures which contribute to the classification of learning object quality, taking as a point of departure some of the ones that were identified for usability as well as others also found in related literature. This was initially done in the specific context of learning objects by García-Barriocanal & Sicilia (2009), where the authors preliminarily explored statistical profiles of highly-rated learning objects referenced on Multimedia Educational Resource for Learning and Online Teaching (MERLOT - www.merlot.org) repository. In that work, García-Barriocanal & Sicilia contrasted four basic metrics (*number of links*, *size in bytes*, *number of images* and *number of personal collections*; the last one as a factor of contrast) against the main categories of disciplines available in MERLOT (*Arts*, *Business*, *Education*, *Humanities*, *Mathematics and Statistics*, *Science and Technology*, and *Social Sciences*) and found initial evidence that the *number of images* is normally associated with the ratings of a learning object, and could consequently be considered as a possible intrinsic measure that could be used to assess quality.

Even though automated analysis cannot replace traditional inspection techniques, it carries the potential of offering an inexpensive and time saving mechanism to *a priori* explore the quality of materials, therefore complementing other existing approaches. This thesis aims to offer the very first foundations for the development of such tool. The deployment of such automated tool would certainly improve the general quality of the services provided by the repository regarding the processes of searching, selecting and recommending good quality materials. Contributors could, for instance, benefit of such new feature by evaluating beforehand the quality of their resources, which would allow their improvement through the use of the quality metrics referenced by the tool. We believe this would positively affect their intention to contribute to the repository with new resources. Moreover, it is known that many resources included by teachers inside their virtual courses are links to external websites (González-Videgaray, Hernández-Zamora & del-Río-Martínez, 2009), so such a tool would allow educators to have a complementary perspective of quality of the resources before adding them into their courses.

1.2 Thesis Objectives

The main goal of this thesis is to provide the very first foundations for the automated quality assessment of learning objects inside repositories based on the intrinsic features of the resources. By intrinsic features we mean "*characteristics that can be automatically extracted or derived from the resources themselves*". The main objective will be break down into the following sub-objectives:

1. Determine how the different materials inside a repository are associated to quality.
2. Determine if the different groups of evaluators inside a repository have the same impressions about the quality of learning objects.
3. Determine if it is possible to create statistical profiles of highly-rated learning objects based on their intrinsic features.
4. Determine if it is feasible to generate models for automated quality assessment of learning objects inside repositories based on the intrinsic features of the resources

The final result will be the profiles of highly-rated learning objects and the models for automatically assess quality of learning objects inside repositories. So far, there are still no studies reporting the existence of models for automated quality assessment of learning objects inside repositories that are strictly based on the intrinsic features of the resources. The profiles and the models that will be developed during this thesis will further allow the construction of tools that can be incorporated inside repositories so that preliminary automated quality assessment of learning objects can be performed before more time-consuming manual work is carried out.

1.3 Research Method

The methodology followed for this research is composed by the following steps:

1. Study of the state-of the art on learning objects and automated quality assessment.
2. Selection of one repository to be used as reference for developing automated models for quality assessment and collection of data from the selected repository.
3. In depth study of the characteristics of the repository and its strategies for assurance of quality.
4. Creation of statistical profiles of highly-rated learning objects based on the intrinsic features of the resources.
5. Development and evaluation of models for automated quality assessment of the resources.

1.4 Outline of the Thesis

The first step of this research will be to study the state-of-the art on learning objects which is presented in Chapters 2 and 3. In Chapter 2 we will explore the existing definitions of the term learning objects, as well as the main properties and features normally associated to these resources and that helped to scaffold the field. We will also review some of the main approaches to assess learning object quality. Moreover, we will present the different types of learning object repositories (LORs) and how the existing repositories assure the quality of their materials. In Chapter 3 we will extend the state-of-the art by tackling the existing approaches for automated quality assessment of learning objects, and by reviewing scattered work about quantitative and measurable aspects of learning objects that can be associated to the quality of the resources. Some of these aspects will further be used as metrics associated to quality during the process of creating the statistical profiles of highly-rated resources.

The second step of the research will be to study one repository to be used as the reference for the creation of learning object profiles. For that, we have selected MERLOT. The decision of choosing MERLOT lays mainly on the following two reasons:

1. MERLOT is one of the most successful and recognized learning object repositories currently in use, it has one of the largest amount of registered resources, and one of the biggest community of users so far;
2. The repository implements a system for quality assurance that works with evaluations given by experts and users of the repository and that serve as baseline for the creation of the learning object profiles. This is especially important in this thesis as a validation mechanism as we will detail later on.

Once data is collected from MERLOT, learning objects will be analyzed in a descriptive way in order to better understand their characteristics, as well as how they have grown over the years in the repository, and which are the relations between their features and the quality associated to them. Moreover, as MERLOT mainly uses two distinct approaches for quality assessment of the resources (users reviews and peer reviews), we will evaluate whether these approaches are complementary or not. This analysis will help us understand how we can establish the baseline for quality comparison inside the repository. All these studies are the presented in Chapter 4.

After understanding learning objects in MERLOT, in Chapter 5 we will define highly-rated profiles of learning objects stored in the repository. Here, a new step for collect data will be required and described. The statistical profiles will be created by contrasting intrinsic metrics of good and not-good learning objects and by identifying which are the metrics that presented significant differences between these groups. These metrics are considered to be associated to quality and potential indicators to be used in the development of models for automated classification.

The last step of the study is to evaluate the feasibility of creating models for automatically assessing quality of learning objects inside the repository. Chapter 6 will show the results of models generated through the use of statistical methods and data mining classification algorithms (DMCA). Moreover, the chapter will also points out the limitations and some possible applications for the use of the created models.

The work will conclude in Chapter 1 presenting the answers for the proposed research questions, as well as describing the main limitations of the results achieved here and the possible directions for future work.

Chapters 2, 3, 4, 5 and 6 are based in whole or in part on the following materials already published: Cechinel, Sánchez-Alonso, Sicilia & Mattos (2010a), Cechinel, Sánchez-Alonso, Sicilia & Mattos (2010b), Cechinel, Sánchez-Alonso & Sicilia (2010) Cechinel, Sánchez-Alonso & García-Barriocanal (2011), Cechinel & Sánchez-Alonso (2011), Cechinel, Sánchez-Alonso, Sicilia & Amador (2011), Cechinel, Sánchez-Alonso, Sicilia & Simões (2011) and Cechinel, Rebollo & Sánchez-Alonso (2012).

Chapter 2 Learning Objects

2.1 Historical Background

It is difficult to precise when and by whom the term learning object (LO) was firstly coined, but widespread credit is attributed to Wayne Hodgins, a recognized e-learning expert and strategic futurist at Autodesk, Inc. According to himself (Hodgins, 2000), he had an "*epiphany moment*" watching his children playing with **LEGO blocks** meanwhile he was pondering about some problems related to learning strategies. Hodgins observed that despite the fact that his kids had very different needs in terms of what they wanted to create using LEGO building blocks and how they wanted to do that, all those different needs were equally satisfied with the same very simple blocks of plastic they had at their disposal. He then realized that the same concept of blocks could be applied to the different needs people have regarding their learning processes. In other words, contents could be organized into small and independent pieces of instruction, called learning objects, which could be in turn assembled into larger instructional structures according to the different learning needs people have. In doing so, people and organizations would then be served by blocks of content that could be easily used and reused in the creation and adaptation of new forms of learning. It is been almost twenty years since the LEGO metaphor appeared and it is still presented to those who are new in the field as an easy way of understanding the main principles involved in this area of work.

However, as the field grew and evolved, this first analogy began to be no longer accepted as self sufficient to encompass all features and nuances behind the concept, and, in order to better redefine and characterize the field, new definitions of learning objects started to be proposed. As pointed by David Wiley, it seems that the number of definitions for learning objects is as huge as the number of people employing it (Wiley, 2000); in fact, even the terminologies used to describe the field vary according to the context in which they are inserted. Just to mention some of them, it is possible to find in the literature references to learning objects as: reusable learning objects (Polsani, 2003), knowledge objects (Merrill, 1999), educational objects (Friesen, 2001), reusable information objects (Barritt, Lewis & Wieseler, 1999), and learning resources, among others. According to (Saum, 2007), sometimes some of the different terms are used interchangeably and other times they are independent from each other, but the term learning objects remains as the most widely recognized one.

This chapter revises some of the different definitions of learning object as well as the main properties and features that, according to these definitions, a learning object must have to be considered one. The chapter will also briefly talk about other intrinsic and related issues that which normally are associated to the concept of learning objects.

2.2 Definition and Characterization

In 1999, David Wiley Wiley first questioned the LEGO metaphor claiming that some of the properties inherent to LEGO building blocks could not be applied to learning objects, and that this could reduce learning objects to mere information objects, or in other words, pieces of information instead of pieces of instruction (Wiley, 1999). Wiley pointed out that, regardless LEGO blocks:

1. A given learning object could not necessarily combine with any other learning object;
2. Learning objects could not be assembled disregarding their structure; and
3. The assembling process of learning objects could not necessarily be made by anyone without any previous training and understanding of the matter.

According to him, "*the combination of learning objects in the absence of any instructional theory will result in larger units that fail to be instructionally useful*". Then, in order to have a reference object which could better fit into a useful learning object system, Wiley proposed shifting from the LEGO metaphor to the **atom metaphor** stating that the atom is also a "*small thing*" that can be combined with other atoms to form large structures, although:

1. The atom does not necessarily combine with every other atom;

2. The atom can only be assembled according to certain ways described in their structure; and
3. It is necessary to have some training and understanding in order to assemble atoms.

Finally, one year later, David Wiley came up with the definition of learning object as "*any digital resource that can be reused to support learning*" (Wiley, 2000). Wiley's contribution is particularly important because it broke a vicious cycle of reducing learning objects to simple blocks of content and introduced the importance of allying learning objects with instructional theories. Moreover, it is necessary to highlight that his contribution excluded non-digital objects from the learning objects spectrum, differently from other definitions, such as, for instance, that one given by the Institute of Electrical and Electronics Engineers (IEEE) Learning Technology Standards Committee (LTSC), which considers a learning object as "*any entity, digital or non-digital, that may be used for learning education or training*" (IEEE-LOM, 2002).

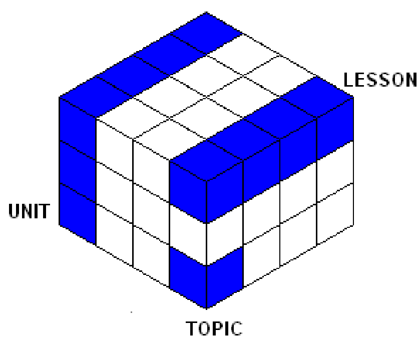


Figure 1. Course structure as a matrix , adapted from L'Allier (1997))

Another definition of learning objects was that given by L'Allier (1997), who stated that a learning object can be defined as "*the smallest independent instructional experience which contains an objective, a learning activity and an assessment*". In this sense, a learning object consists of a component which can serve as basis for a course, a unity or a lesson, and which can be reused to create other unique instructional structures (see Figure 1). The **learning objective** is characterized by a statement of the expected results that students must achieve after (while) using the learning object. The **learning activity** is the component which teaches the contents the students must learn so that they can achieve the proposed objective. To develop the learning activity, one must choose the most suitable technique for each learning content in a way that the learning object is able to catch the students' attention, demonstrate the content relevance, and allow them to apply the acquired knowledge in distinct situations. At last, **the assessment** measures if the students succeeded in achieving the learning objective in terms of synthesis, analysis, application and understanding of the learning object contents. Besides reinforcing the **reusability** property normally associated with the concept of learning objects, L'Allier's definition is important because it proposes an "*internal architecture*" of components which are, in his view, considered mandatory in order to compose a learning object (see Figure 2).

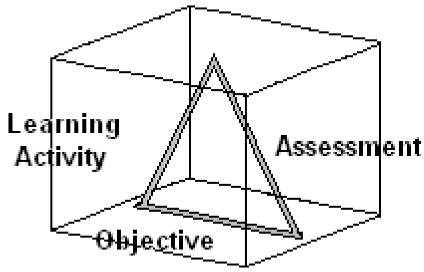


Figure 2. Learning Object elements, adapted from L'Allier (1997)

Still pursuing the concept of learning objects, Polsani (2003) recognizes L'Allier's definition as well articulated, nevertheless he criticizes that to stipulate beforehand the intention of usage, the method and the measuring mechanism of a LO restrain its reusability, since "*the methodology, the intention and the assessment are determined by the instructional situation and not by the LO itself*". In addition, the author defines learning object as "*an independent and self-standing unit of learning that is predisposed to reuse in multiple instructional contexts*". According to Polsani, there are two main principles that must be respected in order to consider any media as a learning object: learning intention and reusability. Regarding learning intention, the media must clearly present its instructional goals, which are represented by two aspects: form and relation. Form is the environment in which the learning object is embedded, its context and how it can be manipulated; and relation is the path that will guide the student during the learning process. Regarding reusability, the learning object must be developed independently from instructional methodologies in order to facilitate its exchange among developers, organizations and institutions. This last consideration is clearly a counterpoint of L'Allier's definition which proposes that the instructional methodologies must be integrated into the learning object.

The problem of defining learning objects was also approached by McGreal (2004), who has called attention to the fact that learning objects exist and interoperate at different levels of **granularity**, varying from simple modular components to lessons, modules, courses and entire programmes. Besides highlighting the importance of granularity for the concept of learning objects, McGreal also offers an interesting perspective examining the differences of opinion in the literature about learning objects definition. He identified four general types of meaning for learning objects that range from general to particular, and which are:

1. Objects that can be anything and everything,
2. Objects that can be anything digital,
3. Digital objects that have been designed for learning purposes, and
4. Objects specific to some approach or proprietary standard.

Figure 3 presents the diagram that resumes these different views.

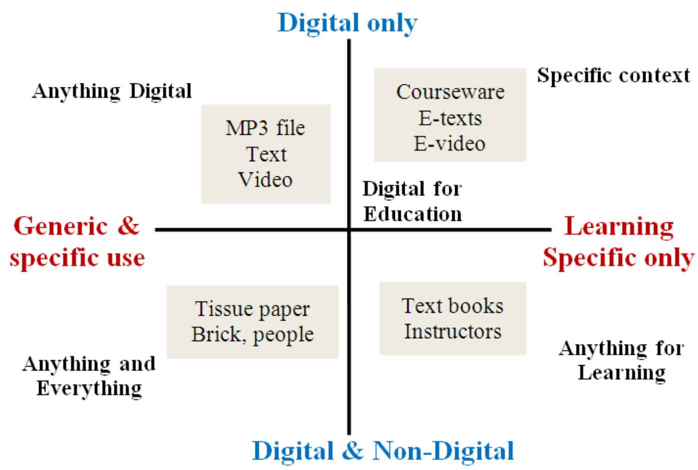


Figure 3. Learning object definitions diagram, adapted from McGreal (2004)

McGreal's diagram and considerations give a panoramic overview of the spectrum in which some of the most important definitions of learning objects are located as well as summarizes the main divergences among these definitions, which are normally related to:

1. The acceptance or not acceptance of non-digital objects as learning objects,
2. The existence or not existence of instructional and educational intention for the object,
3. The minimum and maximum sizes for some object to be considered a learning object.

The author claims that the object only becomes useful to learners when it has an expressed learning purpose, and he proposes to define learning objects as "*any reusable digital resource that is encapsulated in a lesson or assemblage of lessons grouped in units, modules, courses, and even programmes*" whereas the lesson is defined as "*a piece of instruction, normally including a learning purpose or purposes*". His definition encompasses the objects belonging to the top right quadrant of his diagram.

Trying to solve this myriad of definitions, but mostly, the permanent confusion about what a learning object can and cannot be considered, Churchill (2007) defines learning objects as "*a representation designed to afford uses in different educational contexts*" and proposes to classify learning objects into one of the following categories:

1. Presentation objects – resources focused on the achievement of a specific learning objective;
2. Practice objects - resources that allow the practice of certain procedures and incorporate some level of interactivity;
3. Simulation objects – resources that represent a real system and allow the learner to interact with it and to investigate operational and functional aspects of such system;
4. Conceptual models – resources that represent mind models that people are normally able to mentally manipulate, i.e., resources that represent conceptual knowledge and ideas rather than just information;
5. Information objects – resources that use visualization to give educational information; and
6. Contextual objects – resources that allow the learners to explore authentic problems though the use of their own collected data.

Figure 4 shows some examples for each one of these categories.

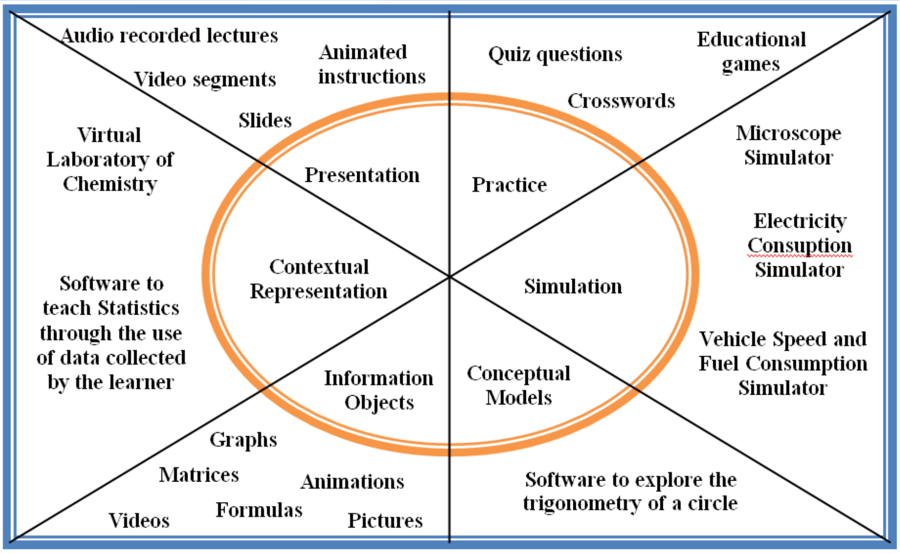


Figure 4. Classification of learning objects according to Churchill (2007) and some possible examples.

As it can be noticed, Churchill's typology excludes non-digital objects and is mostly focused on the structure used to represent knowledge. Similar approach for classifying learning objects has been successfully utilized by one of the most recognized on-line learning objects repositories, MERLOT. Finally, Churchill has also made comments about the difficulty of defining learning objects, which according to him, comes from the combination of the intrinsic ill-defined nature of that concept and a series of resulting issues that help to confuse the concept even more, such as, for instance:

1. The degree of reusability that a learning object must have;
2. The required metadata to describe the learning object;
3. The differences between the content and the structure of a learning object; and
4. How learning objects from different developers must be integrated.

Some of these issues will be briefly discussed in section 2.3.

As mentioned before, the studies carried out on this thesis will focus on learning objects of MERLOT repository. As it will be shown on Chapter 4, even though MERLOT is one of the most successful and recognizable initiatives on the field, resources available on MERLOT do not exclusively (or entirely) belong to any of the definitions presented on this Chapter, but to a combination of many of them. This highlights the difficulties faced by the community to establish a consensual definition that can be applied to real and existing scenarios.

This section has presented some existing definitions of learning objects (see Table 1). As it was discussed so far, and as it can be noticed by the given definitions, there is still no agreement among the community of work about what exactly can be considered a learning object. In any case, and despite all the disagreement around the theme, it seems clear that the concept of learning objects always brings together the implicit expectation of **reusability** (Duncan, 2009). In fact, this property, defined by Sicilia & Garcia (2003) as "*the possibility and adequacy for the object to be usable in prospective educational settings*", is the cornerstone of learning objects concept and one of the main responsible for the widespread success of such technologies. At last, most of the adjacent issues that arise when one deals with learning objects (metadata, granularity, interoperability, standards) are in one way or another related to this ideal level of reusability which community wishes to achieve when working with learning objects.

Table 1. Some learning object definitions

Author	**Definition**
Wiley (2000)	Any digital resource that can be reused to support learning
IEEE-LOM (2002)	Any entity, digital or non-digital, that may be used for learning education or training
L'Allier (1997)	The smallest independent instructional experience which contains an objective, a learning activity and an assessment
Polsani (2003)	An independent and self-standing unit of learning that is predisposed to reuse in multiple instructional contexts
McGreal (2004)	Any reusable digital resource that is encapsulated in a lesson or assemblage of lessons grouped in units, modules, courses, and even programmes
Churchill (2007)	A representation designed to afford uses in different educational contexts

2.3 Issues Related to Learning Objects

As mentioned before, several issues that naturally emerge when one is dealing with the concept of learning objects influence into the degree of reusability the resource is able to have, such as: **granularity**, **metadata** and more recently, **openness.** These issues in their turn have also helped to popularize the term learning objects and contributed to the consolidation of the field.

2.3.1 Granularity

The granularity (or the aggregation level) of a learning object refers to the size of the resource and its decomposability, i.e., the extent to which the resource is intended to be used in order to form larger resources. The concept of granularity is intrinsically attached to the idea of reusing learning objects, and as it is normally the case when we are dealing with learning objects, there is no universal agreement about the size a learning object must have. According to Thompson & Yonekura (2005), the decisions related to granularity are mainly ruled by the organizational context in which the materials are constructed. Sicilia & Garcia (2003) stated that the desirable granularity of a learning material is determined by the imposed reusability requirements, and that, to make possible for learning objects to be coupled and decoupled from each other, granularity must be limited to describe just a small number of related concepts or even a single educational objective. This last view is consistent with the one supported by Wiley, Gibbons & Recker (2000), to whom the "*ways in which learning objects can be combined with one another to facilitate learning are entirely dependent upon their structure*".

Balatsoukas, Morris & O'Brien (2008) have compared different content aggregation levels for learning objects and the constituent parts of their hierarchical structures and relations, and have identified two main points of view that summarize the ways granularity is approached: the **objectivist** and the **relativist** perspectives. According to the authors, the objectivist perspective defines explicit and measurable parameters to delimitate the spectrum of learning objects regarding their size, type and/or structure, whereas the relativist approach represents the "*open approaches that treat any level of granularity as a learning object, for example, from raw data to a whole course certification*".

An example of the **objectivist perspective** is the definition of learning object given by L'Allier (1997), which restrains learning objects to a certain pre-defined structure that must **be small** and contain an objective, a learning activity and an assessment. Other examples of this perspective can be found in the work of Downes (2003), which states that learning objects are "*typically small, consisting of no more than the equivalent of an hour or two of instructional time*", in the work of Mortimer (2002), who establishes a fifteen minute limit of time to complete a task proposed by a learning object. At last, Hodgins (2004) approached this issue showing a five-level content hierarchy, where the units of objects representing each level of the hierarchy are formed by assembling the units of the previous level. This hierarchy is composed by:

1. Data or raw media elements – the smallest level of the taxonomy consisting of pictures, texts, animation and illustrations;
2. Information Objects – a set of raw data elements put together in order to create a reusable piece of information which must be "*media independent*", such as a concept, an overview or a summary;
3. Application Specific Objects – a combination of information objects that form a structure focused on teaching a common topic or goal (the term Learning Object belongs to this level of hierarchy);
4. Aggregate Assemblies – a larger structure with a terminal objective, such as a lesson, a chapter, a brochure, or a unit; and
5. Collection – the biggest level of the taxonomy consisting of entire books, courses, or even a whole curricula.

The **relativist perspective** has a broader and more open approach regarding the structure, size and type of a learning object, where materials presenting all possible and existing levels of granularity are considered to be learning objects. For instance, according to this approach, the hierarchy of learning objects proposed by Hodgins could be seen as a continuum where all the elements belonging to all five levels can be regarded as learning objects. According to (Balatsoukas et al., 2008) this last approach has arisen within learning technology standardization communities such as the IEEE LTSC Learning Object Metadata (LOM), but also in corporate training departments and among individual researchers. In the case of IEEE LOM standard (IEEE-LOM, 2002), it is particularly easy to notice this trend in their generic and vague proposition for classifying learning objects according their granularity, which divides learning objects into the following four aggregation levels:

- *"Level 1: the smallest aggregation level, such as raw media data or fragments.*
- *Level 2: a collection of Level 1 learning objects, such as a lesson.*
- *Level 3: a collection of Level 2 learning objects, such as a course.*
- *Level 4: the largest level of granularity, such as a collection of courses."* (IEEE-LOM, 2002)

As it can be noticed, this definition does not explicitly describe or specify the various existing aggregation levels and it lets the spectrum of the classes in which learning objects can be put in wide open to the metadata annotator.

In any case, and despite these divergences, it seems consensus among the community that, since the reusability property of a learning object is context dependent, the grade of reusability of a learning object will be inversely proportional to the size and the level of aggregation a learning object has. This means to say that "*as the level of granularity-aggregation increases, the learning object content becomes more context dependent and its potential reusability decreases*" (Balatsoukas et al., 2008). Figure 5 graphically represents this idea.

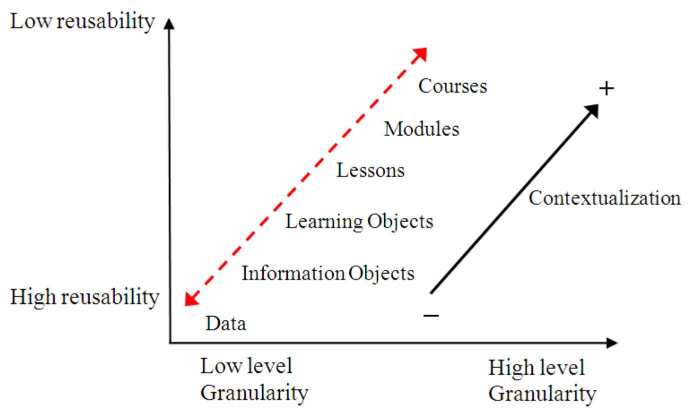


Figure 5. The learning object spectrum, adapted from Balatsoukas et al. (2008)

According to Hodgins (2004), the first two levels of the content hierarchy (raw data and information objects) are context-independent, of universal use and applicable to multiple kinds of applications, whereas the other levels are context dependent and specific to certain application profiles. In simple words, the larger and more complex the learning objects, the more difficult is to reuse them in different contexts.

Although the granularity of a learning object can be directly related to its level of reusability, there is no empirical evidence about how granularity affects the quality of a learning object, or even if it has any influence on their quality at all. This thesis does not directly address such issue, but provides the first foundations for further studies on this direction.

2.3.2 Metadata

It is possible to define metadata as "*structured data about other data, created and managed to describe it in some way that can be used for a function different from the functions of the original data object*" (Lytras & Sicilia, 2007). In the context of learning objects, metadata could be defined as records that describe the most important features of the resources. These descriptions may consider several aspects of the technology, such as authorship, technical and educational information (Sicilia & García-Barriocanal, 2003) with the purpose of supporting indexation and search of the materials inside retrieval systems. Providing **descriptive metadata** is key to succeed on discovering and selecting desirable and relevant material on any retrieval system, and as stated by Currier, Barton, O'Beirne & Ryan (2004) "*poor quality metadata can mean that a resource is essentially invisible within the repository and remains unused*". This statement makes clear the important role metadata plays on the reusability of learning objects. For the case of our study, for instance, metadata attached to learning objects inside LORs are essential for the classification of resources and the creation of classes of quality that will be used for the generation of models for automated quality assessment.

The known metadata **standard** for learning objects is the IEEE LOM standard (the IEEE LOM is based on the metadata work done by the Dublin Core Metadata Initiative – DCM - dublincore.org) (IEEE-LOM, 2002). The standard has a total of 45 data elements divided into the following nine categories:

1. General: it describes the resource as a whole, with information about its title and language, for instance.

2. Life Cycle: it contains the features about the history and current state of the resource, such as the name of the authors and the date of creation.
3. Meta-metadata: it describes the metadata record itself.
4. Technical: it groups the technical characteristics and requirements to properly run the learning resource.
5. Educational: it contains information about educational and pedagogic aspects of the material, such as the level of interactivity and the type of the resource.
6. Rights: it describes the conditions of use and the intellectual property rights of the material.
7. Relation: it refers to the relationships the learning object has with other related learning objects, such as, for instance, if the resource is part of another learning object, or if it requires any other resource in order to be used.
8. Annotation: it contains comments about the educational use of the learning object.
9. Classification: it describes the learning object according to a specific classification system.

Despite of some criticisms (Farance, 2003), the IEEE LOM standard provides a well-known and reliable structure for interoperable descriptions of learning resources and has significantly contributed to scaffold and popularize the field of learning objects, as well the learning objects terminology itself.

At a first glance, it may appear simple to supply retrieval systems with good metadata. However, creating metadata is generally costly (Lytras & Sicilia, 2007), and as long as the number of resources rapidly increases over the internet and inside LORs, it becomes impossible to rely only on **human work** to describe every existing resource. Such situation, has led researchers in the field to search for alternatives for providing **automatic extraction** and generation of good LO metadata, such as, for instance in Cardinaels, Meire & Duval (2005). Brasher & McAndrew (2004) address the problem of automatically generating metadata, and highlight that particular aspects of the resource being described are from two distinct categories of sources: ***intrinsic sources*** and ***extrinsic sources***.

According to the authors, the *intrinsic sources* refer to the information contained in the resource itself (such as the title, or the format of a resource) that can be automatically extracted to generate metadata (considering that the resource is textual). The *extrinsic sources* is information with a strong subjective component that is not explicitly available inside of the resource (such as the expected use of the material, or the amount of the time required for a certain audience to use it), and consequently cannot be obtained through the use of software tools, thus depending on direct human involvement.

2.3.3 Openness

According to Wiley (2009) the initial metaphors about learning objects have limited the ways we understand the concept of reuse, which is normally associated to the idea of assembling and decomposing existing resources in order to form new ones. This author points out that resources are just combined, but not changed and adapted. The author claims that such vision has prevented us from "*seeing reuse as the possibility of changing the learning object itself in substantive ways*" and calls the community to shift directions towards a new way of approaching and defining reuse, where developers and users of learning objects are allowed to **adapt existing resources according to their needs**. Such paradigm has emerged in the last few years named as open educational resources (OER) which are understood as resources that "*comprise content for teaching and learning, software-based tools and services, and licenses that allow for open development and re-use of content, tools and services*" (Geser, 2007).

The first step towards this new paradigm consists on overcoming current restrictions of copyright and intellectual property that are intrinsically attached to the developed learning objects. Learning object producers are invited to free their creations through the use of **copyright licenses that allow usage rights to others** without any cost and without the need to ask for permission when they want to change or to adapt these resources. The most known license covering these aspects is the Creative Commons (creativecommons.org) which allows authors to share their resources according to some predefined conditions such as authorizations for the commercial use of the work and for the modification of the work.

David Wiley (Wiley, 2007) describes four distinct ways open content can be used for free, called by him as "*the 4Rs*", which are:

1. Reuse - using and copying the same way the learning object is put available;
2. Revise - altering the resource according to our needs;
3. Remix - mixing the learning resource with other available resources according to our needs; and
4. Redistribute - sharing with others the work derived from the three previous ways of use.

This level of usage will depend on the **license attached to the resource** and also on the **type (or format) of the resource**. For instance, learning objects in flash format could be put available together with a license allowing revision and remixing, however, such tasks would be difficult to perform once the format is closed and do not allow modifications (Wiley, 2010).

The importance of open educational resources to disseminate knowledge over the globe has already been acknowledged by international organizations, such as the UNESCO (www.unesco.org/new/en/unesco) and the OECD (Geser, 2007).

2.4 Learning Object Life-Cycle

Several works approach the different stages of a learning object life-cycle. For instance, Dalziel (2002) highlights that between the steps of **creating** and **storing** a learning object in a given database (repository), issues such as licenses and right management must be handled during an **intermediate** stage. The author points out that this will further be used during the stage of **search and retrieval/delivery of resources**, where the described licenses and usage conditions will have to be accepted if one wishes to use the retrieved materials. He also describes five different actors involved in the learning object life-cycle:

1. Authority – responsible for prescribing learning objectives and outcomes;
2. Creator – the author of the learning object and responsible for submitting it for publication;

3. Arranger – responsible for designing learning activities and reviewing licenses and usage copyrights;
4. Infoseeker – which has the role of searching for the resources according to the provided metadata; and
5. Learner – the one who is going to use the resources and making the assessments.

Collis & Strijker (2004) and Strijker (2004) state the learning object life-cycle goes through six distinct stages that are shown in Figure 6. According to the authors the first three stages are primarily related to providers while the last three stages are focused on the users of the resources.

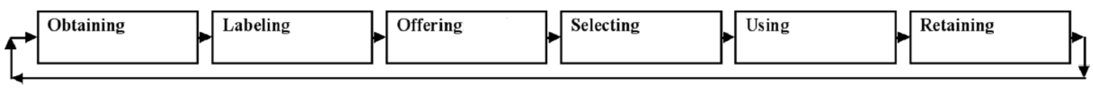


Figure 6. Stages within the learning-object lifecycle, taken from Collis & Strijker (2004)

According to Collis & Strijker (2004) the first stage of the cycle is **obtaining** or **creating** a learning object. Learning objects are developed in digital form from the scratch or through the use of templates that help to create structured and consistent materials. If a learning object already exists, it can be adapted to different contexts and scenarios (e.g. a different language, level of difficulty, or platform). The reasons behind the creation of the resource vary depending on the possible learning contexts. For instance, in companies, learning objects can be created for offering distance training courses with the aim of reducing travel expenses of the target audience; or a teacher can develop a learning object that simulates some equation behavior in order to better explain that subject for his students.

The second stage is **labeling** the learning object with some related information, a step which consists in providing metadata about the learning object. This can be done by different ways. Metadata can be provided by just given very basic information about the resource (such as the title or the subject) and without using any specific metadata standard, or by providing a complete description and using a standard form of metadata such as IEEE LOM. Moreover and as previously mentioned in Section 2.3.2, metadata can be provided manually or automatically extracted. Ochoa (2008) highlights that even though Collis & Strijker (2004) propose labeling as a "*finite and separate*" step, this stage could be considered as constant process, since that information about the object can be added every time the resource is accessed and used. In fact, Cardinaels (2007) proposes a repositioning of the traditional life-cycle stages where metadata is made dynamically and in parallel with all other phases. Labeling is essential for sharing the resource among the community of users given that all information provided will be further used by the mechanisms of search and retrieval.

The third step in Collis & Strijker's life-cycle is **offering**, and consists in storing and publishing the learning object so that the target audience is able to access it. In this step rests the decision of how and where this learning object should be put available. This normally depends on the original intentions and on the context of the organization (or person) which developed the resource. For instance, a learning object can be published for free and for all inside a social Learning Object Repository (LOR), or for a restricted group of people inside a Learning Management System (LMS).

Selecting is the fourth stage of the LO life-cycle. In this step LOs are searched inside repositories and selected according to the needs of the users. Several factors can influence the decision of which learning object is worth to be selected, for instance: recommendations of colleagues, advertisements, costs and ownership, granularity, among others. This stage is critical for the learning object life-cycle since if a given learning object is never retrieved this would mean the end of its life-cycle. Here is also where strategies such as the provision of information about the quality of the resources and the implementation of tools for personalized recommendations play an important role.

The fifth stage is **using** the LO. Learning objects can be retrieved and used as "they are" or be adapted to attend specific needs of the target audience. Adaptation can be done from different perspectives. For instance, LO can be adapted to a specific language and culture (Cechinel & Camargo, 2011; Velázquez et al., 2010), to work in a different platform, or even repurposed for a different discipline area (Gunn, Woodgate & O'Grady, 2005). Both forms of using the LO (*as "it is"* or *adapted*) will depend on the licensing policy set on the resource, and whether the user has full access to the source code of the material.

The last step of the life-cycle is **retaining**. After its use a learning object can be become outdated or even unnecessary, and therefore be discarded for future use. Resources can be revised and new versions and new learning objects based on it can be created. According to Collis & Strijker (2004) the quality control can be carried out through the use of tools that help to provide annotations about the usability and quality of the resources.

The models for automated quality assessment of learning objects evaluated in this thesis are intended to be used in different stages of the learning object life-cycle as we will present latter on Section 1.1.

2.5 Learning Object Repositories

After their production, LOs must be published in a place where users can easily search and retrieve them for future use, a phase defined in the LO life-cycle by Collis & Strijker (2004) as **offering**. LORs are the software systems that provide the functionalities for that. A repository could be simply defined as a digital collection where resources are stored for further retrieval. Heery & Anderson (2005) point out that in order for repositories to differ from other digital collections (such as catalogues, directories, or databases) they should present the following characteristics:

1. To allow the deposit of content by the creator, the owner or by a third party;
2. Their architecture should be able to manage content and metadata;
3. To offer services for putting, getting, searching and controlling access to the resources; and
4. To be trustable and well-supported and well-managed.

One of the most prominent attempts to classify repositories in the context of learning objects was done by McGreal (2008) who defines LORs as "*databases used for storing and or/enabling the interoperability of LOs*". According to this author LORs can be categorized based on:

1. The locality of the LOs - some repositories store the LOs contents locally, while others just store metadata with links to LOs stored elsewhere.

There are also hybrid LORs, i.e., those which store both the LOs and metadata to external resources. Repositories that store only metadata can be seen as portals to other sites and are normally defined as Learning Object Referatories (LORFs) (Hart & Albrecht, 2004).

2. The specificity of the subject area – there exist general repositories which cover a huge variety of disciplines, while others are focused only on specific subject areas.
3. Provision of full courses – some repositories provide materials covering entire courses and/or programmes. This is the case, for instance, of many OpenCourseWare (OCW) initiatives, where universities and organizations freely share the knowledge produced by them (Taylor, 2007).
4. Requirements for participation and usage – this aspect refers to the restrictions imposed by LORs in order to allow the access of the users to the materials. While most of the LORs are open to all users, some of them require subscription of the user to view or use their resources (sometimes the user has even to pay to use the LO). For instance, a lecture/instructor from the private sector who shares inside a LMS the materials produced for some discipline will normally require the subscription of his students to allow access to that course.

For each type of LOR included in his typology, McGreal (2008) has described the following characteristics:

1. The level of the target audience – whether the repository focuses on a specific educational level or not;
2. The granularity of the materials – components, courses, lessons, assets;
3. The size of the repository in terms of the number of resources stored;
4. The type of the materials – if there was a predominant type (applets, videos, e-books, etc) or if the materials were in a varied format; and
5. The type of metadata used – DCM, IEEE-LOM, CanCore, a taxonomy, or none.

Another important class of repositories are the semantic learning object repositories (SLORs)(Soto, Sánchez-Alonso & Sicilia, 2005). SLORs are a type of repository in which "*metadata is expressed in reference to or as part of formal* *ontologies, aimed at providing advanced search capabilities*” (Sánchez-Alonso, Rodriguez, Abián, Arroyo & Sicilia, 2008). This kind of repository provides metadata oriented to machine understandability thus allowing, for instance, software agents querying and searching information in an easy and automated way through the use of reasoning and inference tasks performed over the ontologies.

2.5.1 Examples of LORs

Below we briefly describe some of the most popular LORs available:

- **MERLOT** – One of the most known and recognized repository existing nowadays. It is developed by the California State University Center for Distributed Learning and stores metadata of over 30,000 materials (MERLOT, 2011) distributed in several areas (Arts, Business, Humanities, among others). Its community of users is formed by about 100,000 members. As MERLOT does not store LOs locally, it can be considered as a LORF.
- **Massachusetts Institute of Technology’ OCW** (ocw.mit.edu) – A huge collection of materials developed by MIT covering entire courses and programmes in several areas. Each available course is composed by: the description of the outline of the course and its contents; recommended readings, lecture notes of each content, the exams applied to the students, problem propositions to be solved through projects, and discussion lists. All materials in MIT OCW are shared under a Creative Commons license that allows using the resources for non-commercial purposes, and adapting and sharing the new adapted resources under the same license and conditions. Since its opening in 2001, the MIT OCW has published more than 2,000 courses that have been visited more than 100 million times (Miyagawa, 2010).
- **eLERA** (www.elera.net) – Stands for E-Learning Research and Assessment Network. It is a small LORF (with approximately three hundred resources), however, its importance rests on the fact that it was originally created for research purposes. The main focus of the repository is to provide mechanisms and tools for the collaborative and participative assessment of learning objects through the use of LORI.

- **Connexions** (cnx.org) - Probably the most recent successful initiative in terms of LORs. A repository that allows users to collaborative create and share learning materials, it has presented an exponential growth of contributors in the last few years. According to Ochoa (2010), such success can be attributed to the fact that, differently from the traditional learning objects repositories, Connexions functions through the "*social interaction for the creation of materials*", where all materials are created by its own community of members that can develop them in two formats: modules (small pieces of knowledge) and collections (groups of modules structured into course notes). In Connexions every material available is free for using, reusing and sharing with others under a Creative Commons license.
- **Organic.Edunet** (portal.organic-edunet.eu) – It is a federation of repositories funded by the European Union and focused on contents exclusively related to Organic Agriculture and Agroecology. Even though it is a very recent repository (launched in 2009) it has already approximately 2,500 users and 11,000 resources. The importance of Organic.Edunet also lays on the fact that this repository is a SLOR thus allowing users to perform a semantic search for the materials.

2.6 Quality and Evaluation of Learning Objects

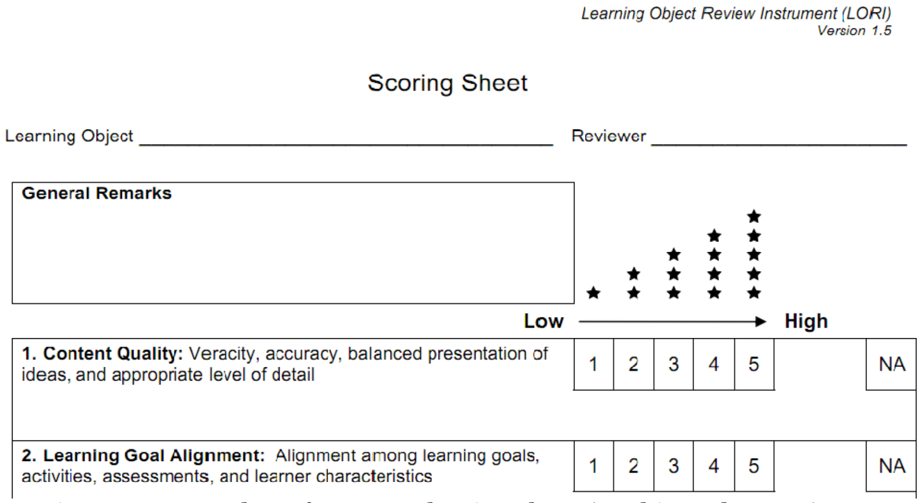
Assessing quality of learning resources is a difficult and complex task that often revolve around multiple and different aspects that must be observed. In fact, the very definition of quality is not straightforward. Vargo, Nesbit, Belfer & Archambault (2003) state that, even though LO evaluation can be considered a relatively new field, it has roots with an extensive body of prior work on the evaluation of instructional software. As stated by Bethard, Wetzer, Butcher, Martin & Sumner (2009) quality is contextual and it will depend on "*the alignment between the user constituency being served, the educational setting where deployed, and the intended purpose of the resource*". Vuorikari et al. (2008) highlights that existing evaluation approaches could be differentiated based on the process they focus. Among others, they mentioned two characteristic examples of approaches, those which focus on the process of creating resources, and those who focus on ready resources and their evaluation.

According to Williams (2000), what a LO ought to be is related to the perspectives of different opinions of those who are the actual users of the resource. So, in order to evaluate quality, it is necessary to consider the particular spectrum of users and the particular set of criteria used by these users to value the resource. Williams (2000) proposes a participant-oriented model (involving different users and stakeholders) composed by four types of LO evaluation that should be made simultaneously, repeatedly and sequentially during various stages of the LO development. This approach covers the whole process of creating resources, and the four types of LO evaluation proposed are:

1. Context Evaluation – It tries to establish if there is a need of some LO according to the needs and expectations of the possible users of this LO;
2. Input Evaluation – It compares alternative inputs focusing to meet the needs identified in the previous step. The main goal here is to evaluate the alternative learning objects that could attend the established needs.
3. Process Evaluation – It assesses the planning, the design and the development of the selected inputs, e.g., how well the instructional strategy and LO were implemented.
4. Product Evaluation – It assesses if the LO is attending the initial outcomes expected for its usage.

Each type of evaluation should consider who are the people which care about the LO (the audience of the LO), and what do they care or have interest about. The people who care about the LO could be, for instance, students, teachers, instructional designers, an organization, among others. These audiences can have different understandings and expectations about the LO, and thus can use distinct criteria and values to judge the quality of the LO (for instance, reusability, quality of the metadata, the instructional approach, among others). According to (Williams, 2000), the combination of these information would then define how should one conduct the process of evaluation of a LO.

Besides Williams (2000), other authors have also claimed that concerns about quality normally focus on different criteria. For instance, in the context of digital libraries, Custard & Sumner (2005) stated that the main issues related to quality are: Accuracy of content, Appropriateness to Intended Audience, Effective Design, and Completeness of Metadata Documentation. In the specific field of learning multimedia resources, the so far most recognized instrument for quantitatively measuring quality is the Learning Object Review Instrument (LORI) (Nesbit et al., 2003). This instrument is intended to evaluate the final and "*ready for use*" LO. In LORI quality is evaluated according to nine different criteria which are rated in a 1 to 5 Likert scale (see Figure 7).


Learning Object Review Instrument (LORI)
Version 1.5

Scoring Sheet

Learning Object ______________________ Reviewer ______________________

General Remarks

Low → **High**

1. Content Quality: Veracity, accuracy, balanced presentation of ideas, and appropriate level of detail	1	2	3	4	5	NA
2. Learning Goal Alignment: Alignment among learning goals, activities, assessments, and learner characteristics	1	2	3	4	5	NA

Figure 7. Screenshot of LORI evaluation sheet (Nesbit et al., 2003)

Leacock & Nesbit (2007) provide some explanations about how each one of the nine dimensions of LORI and how they should be interpreted to evaluate LOs:

1. Content quality – one of the most important aspects of LO quality. This dimension deals with the level of accuracy and reliability of the content, as well as the existence of biases, errors and omissions.
2. Learning goal alignment – it is focused for LOs with a moderate level of granularity, and containing a combination of content, learning activities, and assessments. It intends to evaluate whether the learning activities are aligned with the goals of the LO, and if these activities provide the required knowledge for the users successfully answer the assessments.
3. Feedback and adaptation – it measures the capability of the LO to provide feedback and adapt itself according to the user needs. Such adaptation can be related to the localization of the LO for a specific culture or language, or even to change the LO presentation and content according to a certain preferred user learning style, for instance.
4. Motivation – it evaluates the ability of the LO in retaining users attention, i.e. if the LO is relevant to the learners' goals and in accordance to their

level of knowledge. According to Leacock & Nesbit (2007) learner's expectations about their success or failure on performing a given task using the LO will also impact on motivation.

5. Presentation design – This refers to the quality of exposition (clearness and conciseness) of all items in a LO (text, video, animations, graphics). Aspects such as the font size, or the existence of distracting colors should also be taken into consideration.
6. Interaction usability – this criteria evaluates how easy is for a learner to navigate the LO. Good usability will present consistent layout and structure thus avoiding overloading the user with misleading responses and information. Problems with navigation could also be caused, for instance, by broken links or long delays during the usage.
7. Accessibility – it refers to accommodation of issues of accessibility of people with disabilities in the design of the LO. For instance, a LO with only textual information would exclude blind learners if no audio voice-over is included.
8. Reusability – This aspect was largely explored in the previous sections of this chapter, and it deals with the potential of the LO to be used in different courses and contexts. Issues as the granularity of the LO and openness will influence its portability to different scenarios.
9. Standards compliance - Whether the metadata fields associated to the LO follow the international standards and are complete in a way that allow others to effectively use that information to search and evaluate the LO relevance.

Even though Leacock & Nesbit (2007) provide structural and theoretical foundations for assessing and understanding these many aspects involving quality, they still are all broadly interpreted dimensions that can be subject of divergence from different evaluators. Moreover, different evaluators can also give more importance to one specific dimension than to the others. In order to soften this situation, Nesbit et al. (2002) propose applying LORI through the use of a convergent model, where several evaluators from distinct areas (instructors, instructional designers, and multimedia developers) collaborate to achieve a single and unique quality rating for a given resource. In fact, this concept is currently under application in eLera as it will be shown in next section.

2.6.1 Evaluation inside Repositories

LORs are potential aggregators of communities of practitioners (Brosnan, 2005; Han, Kortemeyer, Krämer & von Prümmer, 2008; Monge, Ovelar & Azpeitia, 2008), i.e. people who share interests and concerns about something they do and learn through their interactions. Due to that, they tend to harness the features of such social environments through the adoption of strategies for the establishment of quality that rely on the impressions of usage and evaluations given by regular users and experts that are members of the repository community. These strategies rely on the hypothesis of transactive memory systems (Wegner, 1986), i.e., systems that store individuals memories, impressions or/and information about a subject in order to form a universal and collective body of knowledge that can serve as an external memory aid for other individuals.

Vuorikari et al. (2008) address this kind of information as evaluative metadata. According to the authors, "*evaluative metadata has a cumulative nature, meaning that annotations from different users accumulate by the time, as opposed to having one single authoritative evaluation*". Inside repositories, evaluative information are normally used as the basis for quality assurance of the resources, but also for properly rank and recommend them for users. Examples of usage of evaluative metadata can be found in some of the most important LORs existing nowadays, such as: eLera, Connexions and MERLOT.

2.6.1.1 eLera

In eLera, members can create reviews of learning objects by using LORI, and experienced members can moderate teams of members in a collaborative online review process where reviewers discuss and compare their evaluations (Nesbit & Li, 2004) (see Figure 8). Besides, members can also add some resource to their personal bookmarks, allowing eLera to recommend materials not only by using their associated ratings, but also by using their popularity.

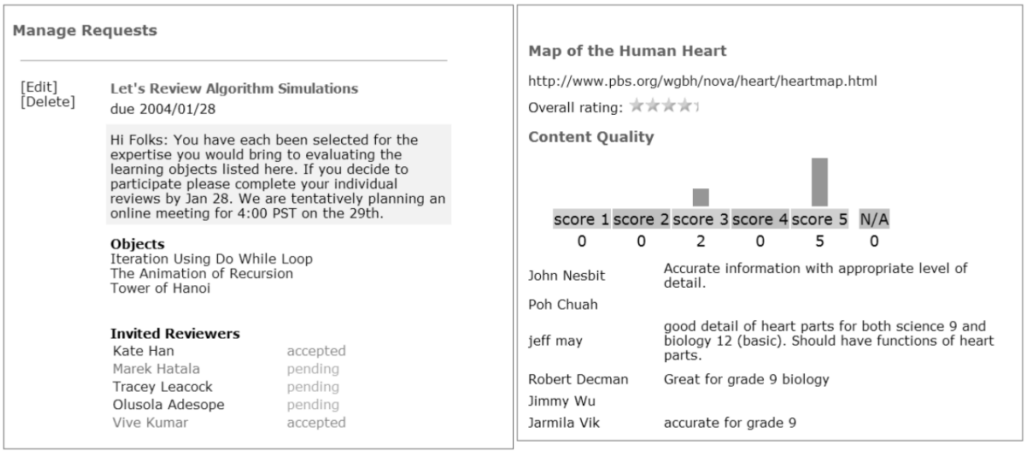


Figure 8. An eLera request for review (left) and distribution of ratings on a LORI item (right), taken from Nesbit & Li (2004)

2.6.1.2 Connexions

Quality in Connexions is approached by a system called *lenses* (see Figure 9) that arranges resources according to evaluations provided by individuals and organizations (Kelty, Burrus & Baraniuk, 2008). In this context, resources are explicitly endorsed by third parties, and gain higher quality assurance as they start to accumulate more endorsements (lenses) from others. Moreover, Connexions also provides mechanisms to sort materials considering their number of accesses over the time and considering the ratings given by users. Recently, Connexions has also integrated in the repository plugins of two popular and well succeeded tools for social interaction (Facebook and Twitter) thus allowing the community of users to recommend and disseminate materials across these social platforms.

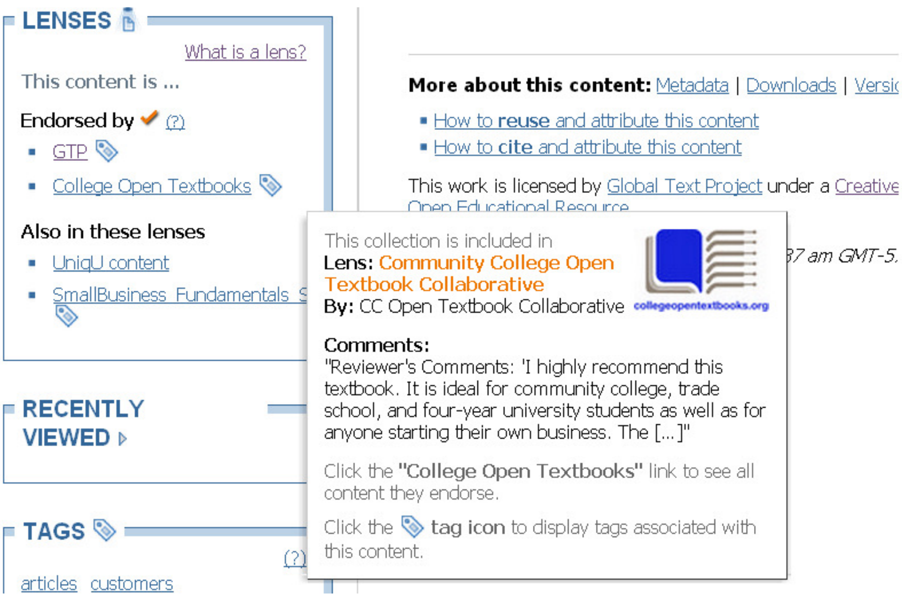


Figure 9. Connexions repository – lenses display

2.6.1.3 MERLOT

The MERLOT repository introduced a post-publication peer-review model in order to assure the quality of its catalogued resources (Cafolla, 2006). The materials catalogued in MERLOT are peer-reviewed by different experts in the discipline domain according to a formal and pre-defined evaluation criterion that addresses three different aspects:

1. Quality of Content;
2. Potential Effective as a Teaching Tool; and
3. Ease of use.

After peer-reviewers report their evaluations, the chief-editor composes one single report which is published in the repository with the authorization of the authors.

In addition to peer-reviewers evaluations, MERLOT also allows the community of users to provide comments and ratings for the materials, complementing its strategy of evaluation with an alternative and more informal mechanism. The ratings of both (users and peer-reviewers) range from 1 to 5, with 5 as the best rating.

Moreover, MERLOT also allows users to bookmark the resources in the so-called *Personal Collections*, providing them a way of organizing their favorite materials according to their individual interests (Sicilia, Sánchez-Alonso, García-Barriocanal & Rodriguez-Garcia, 2009). At last, MERLOT annually gives a special award (the MERLOT Classics Awards) for outstanding materials according to a program criterion of the disciplines (see Figure 10). All these evaluative metadata together are used to sort learning materials every time a user performs a search in the repository.

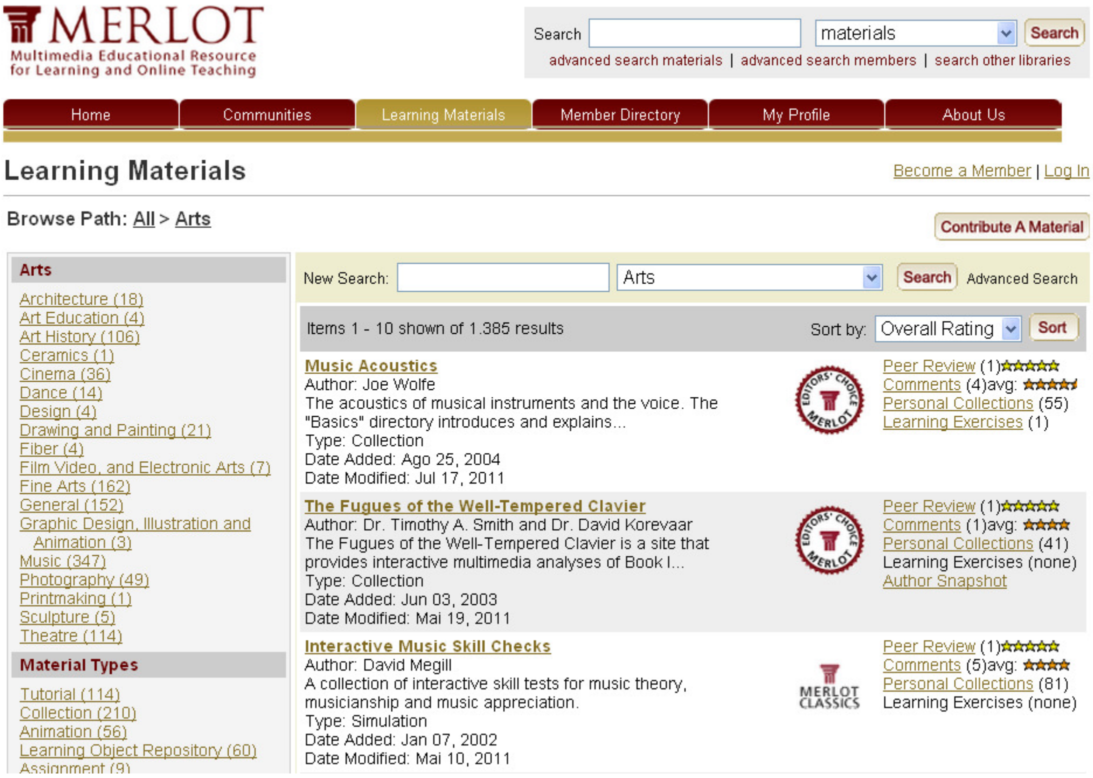


Figure 10. The MERLOT repository (Arts discipline learning materials)

MERLOT is particularly peculiar in the sense that ratings are gathered from two well defined and different groups[1] of people (general public and experts), which possibly come from distinct backgrounds and may have divergent opinions with respect to quality. In fact, these differences between reviewer's groups can be considered as a strong point of the adopted approach, which provides

[1] It becomes important to mention here that peer-reviewers in MERLOT are also members which may cause some overlap of individuals in these two groups.

complementary views about the same subject. Here we briefly describe the main characteristics and differences between these two approaches.

2.6.1.3.1 Peer-Review and Public-Review

Peer-review is conventionally known as the process of assessing a scientific paper or project idea by critical examination of third parties that are experts in the same work domain. This system is widespread in the process of publishing papers in journals and conferences, where the work under evaluation is submitted to a chief-editor which requests a group of fellow-experts to review it in order to obtain advices about whether or not the article must be accepted for publishing, and what further work is still required in the case of acceptance (Harnad, 2000). In the most widely adopted form of peer-review, the identity of the reviewers is hidden from the authors, as well as from the other reviewers. The defenders of peer-reviewing claim that this kind of professional approval serves as a way of assuring the quality of papers published. However, the system is not free from criticisms and issues such as: conflicts of interest, biases of the peers, unnecessary time delay, and the inability on detecting frauds, all mentioned as possible shortcomings of the peer-review process (Benos et al., 2007). In any case, and despite the controversies regarding its efficiency, the peer-review system remains as the cornerstone for quality assurance in the academic field, and has also entered in the scene of educational resources after its implementation in MERLOT.

On the other hand, public-review is widely diffused in areas such as online vendors (e.g. Amazon, eBay) and several communities of interest (e.g. IMDb, YouTube, RYM, etc). In these, users normally benefit themselves from comments and ratings given by the community through the use of recommender systems (such as collaborative filters) which, based on the comparison of user's profiles and the correlation of personal tastes, provide personalized recommendation of items and products that will probably be of their interest (Resnick & Varian, 1997). In this kind of social systems, the motivations and goals behind the users' participation vary significantly, from the desire and need of social interaction, to professional self expression and reputation benefits (Peddibhotla & Subramani, 2007). Table 2 explores some other aspects which normally differentiate standard peer-review and public-review systems.

Table 2. Different aspects involving peer-review and public-review

Aspects	**Peer-Review**	**Public-Review**
Evaluator background	Expert in the field domain	Non-expert
Existence of official criteria or metrics	Yes	No/Sometimes
Community of evaluators	Restricted	Wide opened
Common models	Pre-publication	Post-publication
Domain	Scientific field, journals and funding calls	Online vendors, communities of interest
Motivation	Prestige, fame, to determine the quality and direction of research in a particular domain, obligation	Desire and need of social interaction, professional self expression, reputation
Communication among evaluators	Not allowed	Encouraged
Selection of evaluators	Editor responsibility	None
Financial compensation	Normally none	None
Time taken for the evaluation	Typically slow	Typically fast
Level of formality	Formal process for editing and revision	Informal
Author's identity	Masked	Non-masked
Requirements to be a reviewer	To be an expert in the field and to be invited	Creation of a member's account

Being MERLOT the repository chosen for the studies carried out in this thesis, the associations between the ratings given by users and peer-reviewers are explored in-depth in Section 4.4.

Chapter 3 Problem Formulation

This chapter presents the state of the art on automated quality assessment of learning objects and identifies the most important measurable aspects of learning objects that were found as associated to quality in the literature. In the light of the concepts presented in Chapter 2 and in Sections 3.1 and 3.2, we summarize in Section 3.3 the most important detected shortfalls that we will address on the rest of this thesis and we describe in section 3.4 our proposed approach for the studies we will carry out.

3.1 Quality and Automated Assessment

Existing approaches for LO quality assessment are time consuming and rely solely on human effort, and therefore cannot fully attend the overwhelming increase of learning resources available nowadays in the existing repositories. An alternative solution for the problem of manually reviewing LOs is trying to identify and to establish lower level measures that are related to some of the LO quality dimensions and that are easily quantifiable and consequently free from ambiguities and misinterpretations, thus allowing a faster and automatic way of assessing quality in the spectrum of such dimensions. Examples of such strategy can be already observed to assess quality of learning object metadata (Ochoa, 2008) and more recently to estimate and measure learning objects reusability (Sanz-Rodriguez, Dodero & Sánchez-Alonso, 2010a).

Another tangible initiative towards automated evaluation of LOs is the work of Ochoa & Duval (2008) which proposes the development of a set of metrics for ranking the results of learning objects searches inside repositories. Here, the authors address the concept of quality as the relevance that some learning object present in some specific context of usage (Duval, 2006) using as a source of inspiration methods currently applied to rank other types of objects, such as books (collaborative filtering) and scientific papers (impact factor). They grouped the proposed metrics according to three dimensions of relevance: topical (what is the interest of the user), personal (why the user is interested in that topic) and situational (where the instructional resource is going to be used); and contrasted their performance against the text-based ranking traditional methods, finding significant improvements in the final ranking results. In order to calculate these metrics Ochoa & Duval (2008) use information obtained from the learning objects metadata, from the user queries, and from other external sources such as the records of historical usage of the resources.

All these approaches for automatically measuring quality according to specific dimensions depend either on the existence and availability of metadata attached to the resources (or inside the repositories), or on measures of popularity about the resources that are obtained only when the resource is publicly available after a certain period of time. Some of them also face real problems of scalability because, as the number of resources increases, it becomes impossible to provide evaluative metadata for every single resource in the repository. Such situation leaves many resources of current repositories without any measure of quality at all. For instance, a recent study Cechinel & Sánchez-Alonso (2011) has shown that in MERLOT, from the total amount of resources, approximately 12% were rated by users or peer-reviewers, and only 3% presented at least one peer-review and one user rating at the same time.

Considering that metadata may in some cases be incomplete (Sicilia, García-Barriocanal, Pages, Martinez & Gutierrez, 2005) or even contain inaccurate descriptions (Cechinel, Sánchez-Alonso & Sicilia, 2009), and that automated analysis will be typically used for objects that have not been shared to the public yet and are under preliminary assessment, we propose a complementary approach that relies only on the data that can be directly extracted from the learning resources themselves. The main advantage of such proposal is to allow the development of models to assess quality of new resources inserted in the repository without the need of annotations about them.

The very first step towards the development of the foundations for such approach is to identify intrinsic metrics of learning resources that could serve as potential indicators of quality. Such metrics, in turn, can be further used to build and test models and tools that can automatically perform quality assessment inside LORs.

3.2 Quantitative and Measurable Aspects of Learning Objects

To the best of our knowledge, there is no empirical evidence of intrinsic metrics that are indicators of LOs quality, however there are some works in adjacent fields which can serve as a source of inspiration. For instance, empirical evidence of quality indicators has been found by Custard & Sumner (2005) in the field of educational digital libraries. In that work, the authors identified and computed 16 metrics for quality and trained a support vector machine model to assess resources quality using these indicators. Their experiments demonstrated the models were "*sensitive to detect differences in the quality of resources catalogued*". These findings were further used by Bethard et al. (2009) who confirmed the feasibility of decomposing the concept of quality for educational resources into smaller pieces of measurable dimensions, opening the way for the automated characterization of quality of resources inside educational digital libraries. On that work, the authors were able to identify 5 quality indicators which could be automatically observed and measured inside learning resources through the use of natural language processing and machine learning techniques.

Meyer, Hannappel, Rensing & Steinmetz (2007) developed a model for classifying the didactic functions of a learning object based on measures about the presence of interactivity and information contained in the HTML code (lists, forms, input elements); while Mendes, Hall & Harrison (1998) identified evidence in some measures to evaluate sustainability and reusability of educational hypermedia applications, such as the type of link, and the structure and size of the application. In a wider context, Stefani, Vassiliadis & Xenos (2006) proposed a set of metrics for the quality of services provided by a Virtual Campus Infrastructure. These metrics were divided in four types (Functionality, Reliability, Usability and Efficiency) and contained quantitative aspects such as the number of video applications, the number of broken links, and the number of internal links, among others.

In the context of information quality inside communities-based encyclopedia (Wikipedia), Blumenstock (2008) has found the length of an article (measured in words) as a predictor of quality. Moreover, Stvilia, Twidale, Smith & Gasser (2005) have been able to automatically discriminate high quality articles voted by the community of users from the rest of the articles of the collection. In order to do that, the authors developed profiles by contrasting metrics of articles featured as best articles by Wikipedia editors against a random set. The metrics were based on measures of the article edit history (total number of edits, number of anonymous user edits, for instance) and on the article attributes and surface features (number of internal broken links, number of internal links, number of images, for instance). The study has shown that the median values of the measures varied widely between the two sets, and that these metrics could be used to construct a benchmark for article information quality assessment.

In the field of usability, Ivory & Hearst (2002b) found that "good" websites often contain more words and links than the "not so good" and "bad" ones. In this work the authors developed statistical profiles of highly-rated websites evaluated for the 2000 Webby Awards (International award recognition of excellence given by The International Academy of Arts and Sciences to websites in several categories - http://www.webbyawards.com/index.php). They classified the websites under study into three classes (good, average, poor) using the terciles of the ratings given by the judges of the Webby contest, and then contrasted 157 quantitative measures among these classes. The generated profiles served as the basis for the development of classification models able to discriminate pages among good, average and poor with high levels of accuracy. At last, the methodology used in the work of Ivory & Hearst (2002b) was further applied by García-Barriocanal & Sicilia (2009) in the context of learning objects inside repositories, where the authors have found correlations between the number of images and the ratings given by peer-reviewers and users of the MERLOT repository.

Table 3 highlights some of the quality indicators that we found in the literature as presented earlier, as well as the context in which these indicators were encountered (or tested) and the dimensions of quality that the authors were exploring while using such metrics.

Table 3. Overview of quality indicators in the reviewed literature

Measurable aspects used as quality indicators	***Prior Research***	***Context***	***Observed quality dimensions***
Multimedia, element count, link count, site domain, cognitive authority, metadata currency, alignment, word count, cost, functionality, number of annotations associated to the resource, resource currency, among others.	Custard & Sumner (2005)	Educational digital libraries	Provenance, Metadata Description, Content, Social Authority, Availability
Existence of interaction elements in the HTML code (lists, forms, input and choice elements). Number of JavaScript functions. Existence of flash animations embedded	Meyer et al.(2007)	E-learning resources	Not applicable. In here the metrics were used in the automatic classification of didactic functions
Link representation (whether the link are embedded or not), link type, highlighting of anchors, structure of the application (sequential, hierarchical, network), size of the application, compactness (the intrinsic connectedness of an application), and stratum (the degree of organization of a hypertext).	Mendes et al.(1998)	Educational hypermedia applications	Reusability of information, maintainability of applications
Has sponsor, has prestigious sponsor, indicates age range, identifies learning goals, organized for goal.	Bethard et al.(2009)	Educational digital libraries	Appropriate inclusion of graphics, readability of text, focuses on key content.
Length of an article (number of words)	Blumenstock (2008)	Community-based encyclopedia (Wikipedia)	Information quality (article quality)
Number of video applications, number of broken links, number of internal links, number of simulations.	Stefani et al.(2006)	Virtual Campus Infrastructure	Functionality, reliability, usability and efficiency
Article length (number of characters), number of internal links, number of external links, number of internal broken links, number of images, readability scores (Flesch-Kincaid), Information-to-noise (ratio of the total length of index items	Stvilia et al.(2005)	Community-based encyclopedia (Wikipedia)	Information quality (article quality)

over the page size)			
Number of words, number of images, number of links, minimum font size, number of times a color is used, number of graphical ads, number of interactive objects.	Ivory & Hearst (2002b)	Websites	Website usability, website information, navigation, and graphic design.
Number of images, Number of Personal Collections (bookmarks)	García-Barriocanal & Sicilia (2009)	LORs (MERLOT)	No explicit dimension

3.3 Detected Needs and Goals

As presented earlier in Chapter 2, LOs are multifaceted artifacts that can encompass many varied definitions according to the different contexts in which they are being used. Therefore, to assess LOs quality, it is mandatory to take into account these contexts and their specificities in order to be successful in the process of evaluation. Nowadays, LORs are the common place where authors put available their LOs; hence, such environments constitute on reliable contexts to carry out empirical experiments about quality assessment of the resources. As the MERLOT repository is one of the most successful initiatives in this field and it implements a robust system for quality assurance, we selected it as the reference repository for the studies carried out on this thesis. In here, MERLOT is the context where resources are inserted; therefore we are dealing with the definition of learning objects as any resource or website refereed by MERLOT as a "*Learning Material*". As it will be presented on Section 4.1, MERLOT has its own taxonomy for the classification of LOs, using different types of materials, categories of disciplines, and audiences. In addition, some of the metadata for the classification of the materials on MERLOT are mandatory whilst others are optional and not always available. As it can be perceived, all these features leave wide open the spectrum of LOs that we are dealing with and demand initial studies to characterize and better understand such resources inside the repository. Such studies will be presented on section 4.3 and are intended to attend the following goal:

Goal 1 - Determine how the different materials inside a repository are associated to quality.

Moreover, as presented on subsection 2.6.1.3, the quality assurance system implemented on MERLOT is mainly composed by ratings given by experts in the subject areas, and complemented by ratings and comments given by the community of users of the repository. Considering that these ratings belonging to the two groups are intended to serve in our study as the basis for categorizing LOs into classes of quality, it becomes important to understand the existing relations between them. This will allow us to evaluate how the classes of quality could/should be created and whether it is possible to suppress one group of ratings in favor of the other inside the methodology we are proposing. This study will be presented on Section 4.4 and addresses the following specific goal:

Goal 2 - Determine if the different groups of evaluators inside repositories have the same impressions about the quality of learning objects.

Furthermore, the core problem we are focusing on this thesis is the automated evaluation of LOs inside repositories. Although the existing approaches for quality assessment inside repositories (ratings, awards, etc) can be considered efficient (at some level) to ensure the quality of the evaluated resources, they are time-consuming and rely entirely on human-work. Because of that, a huge amount of resources is left without any quality evaluation for long periods and thus are not showed to the users (as they could or should) during the process of search and retrieval. As mentioned earlier in Section 3.1, the most important current strategies for automated quality assessment rely on metadata that are not always available in the repository, or that are only available for those LOs that are already stored in the repository for some period. Considering that we are proposing that automated quality assessment should also be used for those LOs which were just catalogued in the repository, it is necessary to adopt an alternative approach that relies only on information that can be directly extracted from the LOs (intrinsic features). Chapter 5 will then present the studies carried out to evaluate whether this is feasible, thus addressing the following goal:

Goal 3 - Determine if it is possible to create statistical profiles of highly-rated learning objects based on their intrinsic features.

The automated quality assessment of LOs inside repositories must be developed through the use of models that are able to represent the knowledge behind the decision that is made during the regular process of evaluation. Once we are able to determine if it is possible to identify intrinsic features of LOs that are associated to quality, it is still necessary to evaluate whether these features can be used to generate reliable models that can be implemented inside repositories in order to perform the task of quality evaluation. This need raises the following goal of our thesis which is addressed on Chapter 6:

Goal 4 - Determine if it is feasible to generate models for automated quality assessment of learning objects inside repositories based on the intrinsic features of the resources.

3.4 The Proposed Approach

Our approach is exclusively related to those aspects of learning objects that are displayed to the users and that are normally associated to the dimensions of presentation design and interaction usability (included in LORI), and the dimension of information quality (normally mentioned in the context of educational digital libraries). However, even though the intrinsic metrics we are evaluating are mentioned by other works as being related to these quality dimensions, we are still in the process of identifying the existence of association between these metrics and the quality of the learning resources, and we cannot state that they are representative of such dimensions at this point. Such statement would require further controlled studies focused on understanding why the encountered associations are occurring, as well as if they are representative of cause and consequence relations between quality and the metrics.

In our thesis, the starting point for the establishment of learning objects quality are the ratings given by the community of users and peer-reviewers of the MERLOT repository. Such evaluative information constitute good social knowledge that can serve us as a reliable baseline for comparison of features between good and not-good resources, thus helping in the process of identifying intrinsic metrics that are possibly associated to quality. For the context of this study, resources are considered to be of good quality if they are highly-rated by users and peer-reviewers of MERLOT.

Chapter 4 will describe two studies conducted to better understand the MERLOT repository and that will guide us during the development of the models for automated quality assessment inside the repository. Chapter 5 will present the methodology for the creation of statistical profiles of LOs based on their intrinsic features, and will describe how the quality of the resources can differ depending on these metrics. At last, Chapter 6 will use Linear Discriminant Analysis and Data

Mining Classification Algorithms to generate models for automated quality assessment.

Chapter 4 Studies for a Solution

This chapter is intended to provide a better understanding of the MERLOT repository. Here, we present the studies we carried out before starting to develop the solution for the problem of automatically assessing quality inside the repository. These studies are basically organized into two parts:

1. A descriptive study of MERLOT; and
2. The analysis of associations between the ratings given by users and peer-reviewers.

The first part (presented in sections 4.1, 4.2 and 4.3) consists on briefly contrast some of the main issues related to LOs and LORs presented in the previous chapters against the features encountered in MERLOT. Moreover, we also applied descriptive statistics on a sample collected from MERLOT. Here, besides performing an analysis about the growth of the different types of material in the repository over the years, we have also described them according to: 1) the MERLOT Classics Awards; 2) the ratings given by peer-reviewers and users, 3) the different categories of discipline in MERLOT and 4) the number of personal collections in which they are bookmarked (included).

The second part (section 4.4) analyzes the existence of associations between the two dimensions of ratings in the repository: users and peer-reviewers ratings. For that we have used the same sample collected for the descriptive analysis of the repository.

4.1 Characteristics of MERLOT

As mentioned on subsection 2.5.1, MERLOT is a referatory. This means that it stores only metadata about resources stored elsewhere, but not the resources themselves. Only registered users can contribute materials in MERLOT. The process of inserting a new material in the repository consists on informing essential (and optional) metadata about the resource in the following 5 steps:

1. Title and URL – This is basic information necessary to locate and identify the resource.
2. Description – it consists on a textual description about the resource, as well as information on the material type, its primary audience and some keywords. In MERLOT, materials can be classified into just one of the following types: *Animation*, *Assessment Tool*, *Assignment*, *Case Study*, *Collection*, *Development Tool*, *Drill and Practice*, *Learning Object Repository*, *Online Course*, *Open Journal-Article*, *Open Textbook*, *Lecture/Presentation*, *Reference Material*, *Simulation*, *Social Networking Tool*, *Tutorial*, *Quiz/Test*, *Workshop and Training Material*. Moreover, resources can be focused in the following audiences: *Grade School*, *Middle School*, *High School*, *College General*, *College Lower Division, College Upper Division*, *Graduate School* and *Professional*.
3. Category – Here the user selects the general and specific subject areas that are related to the material. Resources in MERLOT are organized in seven different disciplines: *Arts*, *Business*, *Education*, *Humanities*, *Mathematics and Statistics*, *Science and Technology*, and *Social Sciences*, which in turn are also subdivided in several subcategories. It is important to mention that a learning resource can be related to more different areas and the repository allows the insertion of all of them. In this step the user also informs the languages in which the resource is available.
4. Author – Information about the author of the resource (name, last name, email and organization).
5. Optional Information – Here the user can inform the technical format of the resource (Applet, HTML/Text, Image, pdf file, video, and flash, among others) and the technical requirements for its usage. Moreover, there are fields to provide the version of the resource, the costs involved and copyright information.

Besides the evaluative information, the present study has considered just the classification information which is mandatory during the process of contributing a material in the repository, i.e., material type and category of discipline. Table 4 summarizes the characteristics of MERLOT according to some of the main aspects of LOs and LORs presented in Chapter 2.

Table 4. Features of MERLOT according to some aspects of LOs and LORs

	Aspect	*Comments*
Learning Objects Features	Granularity	It varies significantly. MERLOT adopts a relativist perspective, where materials presenting all possible and existing levels of granularity are considered to be learning objects.
	The specificity of the subject area	It covers a variety of disciplines, as follows: Arts, Business, Education, Humanities, Mathematics and Statistics, Sciences and Technologies, and Social Sciences.
	Material Types	MERLOT classifies the materials according to different types (Animation, Simulation, Case Study, etc). However, all these materials are normally webpages.
	Openness and licenses	As materials in MERLOT are stored elsewhere, this is independent from the repository and will vary according to each material. MERLOT adopts a Creative Commons license for all the information stored by them (reviews, material metadata, personal collections, etc).
Learning Object Repository Features	Type	Referatory.
	Size of the repository	Over 30,000 materials
	Metadata	It is not possible to retrieve metadata. It uses a simplified metadata standard. According to Q4R (2007), MERLOT also uses a modified version of IEEE LOM to export xml files.
	The level of the target audience	All levels. It does not focus on a specific educational level. It allows to classify resources according to the following target audiences mainly based on the American Educational System: Grade School, Middle School, High School, College General, College Lower Division, College Upper Division, Graduate School and Professional
	Requirements for participation and usage	The repository is open for search. In order to contribute a material, users must register in the repository.
Quality Assurance on the Repository	Who evaluates	Users can rate resources, provide comments about them, and add favorite resources to their personal collections. Experts in the subject area rate the resources and provide explanation about the given rates. The repository also conceives awards to outstanding materials.
	When the evaluation takes place	After the publication of the materials. Post-publication review model.

	What is evaluated	The experts review resources according to three well defined criteria: Quality of Content, Potential Effective as a Teaching Tool, and Ease of use. Users do not address any specific criteria.
	Type of Evaluation	Evaluation of the product, it does not consider the design process.

4.2 Data Collection

Data from a total of 20,267 learning objects was gathered on September 2009 through a web crawler developed by Fernández (2011) for that purpose. However, even though the Material Type metadata is a mandatory field during the MERLOT Learning Object registration process, 33 resources from the original set did not have this information and were disregarded. Most of the resources did not have any peer-review information (PRR) or user rating (UR) and, from the total amount of collected data, only 3.42% presented at least one peer-reviewer and one user rating at the same time (Table 5).

Table 5. Sample sizes for peer-reviewed and user-reviewed resources

Sample Size Containing Material Type Information	**PRR > 0**		**UR > 0**		**PRR > 0 ∩ UR > 0**	
	Size	**%**	**Size**	**%**	**Size**	**%**
20,234	2,586	12.78	2499	12.35	694	3.42

4.3 Descriptive Analysis of the Materials in MERLOT

Table 6 shows the amount and percentage of each type of material[2] in the data base for: overall sample (All); cases where learning resources have at least one peer

[2] The classes of material types in MERLOT have been altered since the beginning of the repository in 1997. Because of that, the sample collected for the present study contains some types of materials that are not mentioned in this section.

reviewer rating (PRR > 0); cases where learning resources have at least one user rating (UR > 0); and cases where learning resources have at least one peer reviewer rating and user rating at the same time (PRR > 0 ∩ UR > 0).

Table 6. Amount and percentage for the types of material

Material Type	**Size (Freq %)**			
	All	**PRR > 0**	**UR > 0**	**PRR > 0 ∩ UR > 0**
3D Learning Object	1 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)
Animation	875 (4.32)	126 (4.87)	163 (6.52)	45 (6.48)
Case Study	491 (2.43)	74 (2.86)	35 (1.40)	11 (1.59)
Collection	2,953 (14.59)	367 (14.19)	288 (11.52)	86 (12.39)
Drill and Practice	953 (4.71)	156 (6.03)	110 (4.40)	38 (5.48)
Learning	17 (0.08)	0 (0.00)	1 (0.04)	0 (0.00)
Learning Material	173 (0.85)	22 (0.85)	50 (2.00)	6 (0.86)
LO Repository	424 (2.10)	14 (0.54)	24 (0.96)	2 (0.29)
Lecture/Presentation	2,200 (10.87)	265 (10.25)	190 (7.60)	42 (6.05)
News or Event	6 (0.03)	0 (0.00)	0 (0.00)	0 (0.00)
Online Course	153 (0.76)	3 (0.12)	5 (0.20)	2 (0.29)
Open Textbook	81 (0.40)	0 (0.00)	0 (0.00)	0 (0.00)
Other Resource	23 (0.11)	4 (0.15)	4 (0.16)	3 (0.43)
Professional Paper	1 (0.00)	1 (0.04)	1 (0.04)	1 (0.14)
Quiz/Test	878 (4.34)	84 (3.25)	60 (2.40)	29 (4.18)
Reference Material	5,018 (24.80)	443 (17.13)	546 (21.85)	131 (18.88)
Simulation	2,963 (14.64)	528(20.42)	607 (24.29)	164 (23.63)
Technical Tool	16 (0.08)	0 (0.00)	1 (0.04)	0 (0.00)
Tutorial	2,907 (14.37)	496 (19.18)	413 (16.53)	134 (19.31)
Workshop and Training	101 (0.50)	3 (0.12)	1 (0.04)	0 (0.00)

Total	**20,234 (100)**	**2,586 (100)**	**2,499 (100)**	**694 (100)**

As it can be noticed in Table 6, the five types of material which have the highest number of occurrences in descending order are:

1. For the overall sample: *Reference Material*, *Collection*, *Simulation*, *Tutorial* and *Lecture/Presentation*;
2. For the PRR > 0 sample: *Simulation*, *Tutorial*, *Reference Material*, *Collection*, and *Lecture/Presentation*;
3. For the UR > 0 sample: *Simulation*, *Reference Material*, *Tutorial*, *Collection*, and *Lecture/Presentation*;
4. For the PRR > 0 ∩ UR > 0 sample: *Simulation*, *Tutorial*, *Reference Material*, *Collection*, and *Animation*.

All five types of material that most occur in the overall sample remain the ones which presented the highest number of occurrences in all other data sets of Table 6 (exception is made for the case of *Lecture/Presentation* versus *Animation* for the PRR > 0 ∩ UR > 0 sample). However, as it can be noticed above, the order of occurrences among the material types varies over the samples. For instance, the *Simulation* type, which is the third in number of occurrences in the overall sample, is the first one in the last three data sets. This kind of information can be useful if we consider the act of picking and evaluating a learning resource as some sort of evaluator's preference for that kind of material. In other words, and for our current data samples, this would mean that peer-reviewers and users tend to prefer *Simulation* learning resources over the others. Figure 11 helps to better illustrate this idea.

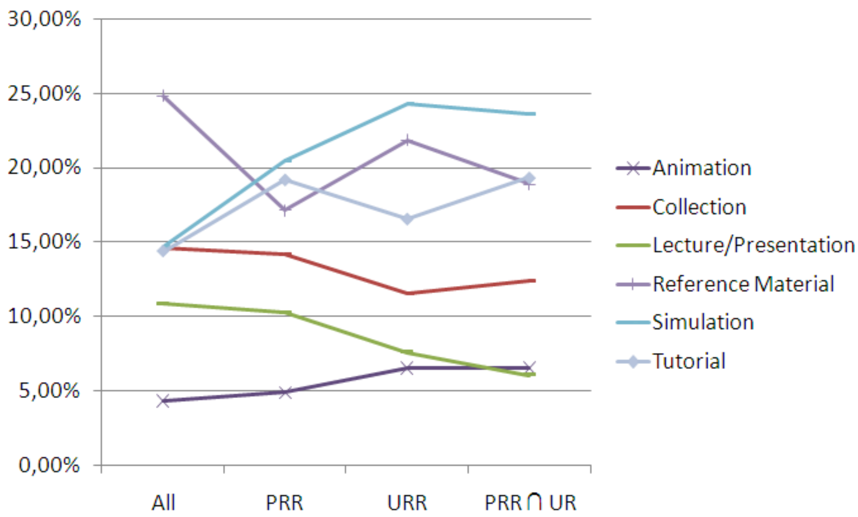


Figure 11. Peer-reviewers and users' preferences when choosing resources to evaluate

The same reasoning can be used to compare the preferences between peer-reviewers and users. For example, it is possible to see in Figure 11 that peer-reviewers tend to select more resources of *Lecture/Presentation* and *Collection* than users, and users tend to select more resources of *Animation* and *Reference Material* than peer-reviewers.

4.3.1 How material types grow over time

The types of materials in MERLOT have grown differently over the years. Table 7 shows the number of occurrences of each type of material at three distinct periods in time.

Table 7. Amount of each type of material over the years

Material Type	**Year Count (Data Base Percentage %)**		
	2000	**2004**	**2009**
3D Learning Object	0 (0.00)	0 (0.00)	1 (0.00)
Animation	119 (4.21)	364 (3.39)	875 (4.32)
Case Study	5 (0.18)	194 (1.81)	491 (2.43)
Collection	276 (9.76)	1549 (14.42)	2953 (14.59)
Drill and Practice	67 (2.37)	488 (4.54)	953 (4.71)
Learning	0 (0.00)	0 (0.00)	17 (0.08)
Learning Material	173 (6.12)	173 (1.61)	173 (0.85)
LO Repository	0 (0.00)	3 (0.03)	424 (2.10)

Lecture/Presentation	187 (6.61)	1011 (9.41)	2200 (10.87)
News or Event	0 (0.00)	6 (0.06)	6 (0.03)
Online Course	1 (0.04)	2 (0.02)	153 (0.76)
Open Textbook	0 (0.0.00)	3 (0.03)	81 (0.40)
Other Resource	23 (0.81)	23 (0.21)	23 (0.11)
Professional Paper	1 (0.04)	1 (0.01)	1 (0.00)
Quiz/Test	82 (2.9)	316 (2.94)	878 (4.34)
Reference Material	384 (13.58)	2609 (24.29)	5018 (24.80)
Simulation	1234 (43.65)	2340 (21.79)	2963 (14.64)
Technical Tool	16 (0.57)	16 (0.15)	16 (0.08)
Tutorial	259 (9.16)	1642 (15.29)	2907 (14.37)
Workshop and Training	0 (0.00)	0 (0.00)	101 (0.50)
Total	**2827**	**10740**	**20234**

As it can be seen in Table 7, *Reference Material* had the most significative growth in the repository, from 13.58% of the resources in 2000, to 24.80% in the last period. The second most significant growth is that of *Tutorial* (from 9.16% to 14.37%), followed by *Collection* (from 9.76% to 14.59%) and *Lecture/Presentation* (from 6.61% to 10.87%). It is also important to highlight that the growth of *Learning Object Repository* is concentrated in last period; this could be considered an indicator of significant increase in the number of repositories available on the internet in the last few years. At last, Table 7 also shows a significative decline of the *Simulation* resources (from 43.65% to 14.64%). Figure 12 helps to visualize the growth and decline of the mentioned types of material.

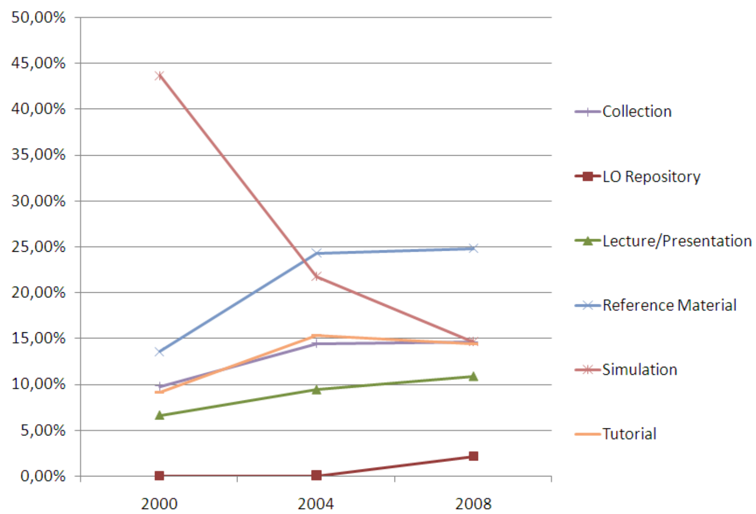


Figure 12. Evolution of resources in MERLOT over the years

4.3.2 MERLOT Classics Awards versus Material Types

As mentioned before, MERLOT gives an award (called MERLOT Classic Award) for the best materials stored in the repository. From the 20,234 learning resources of our overall sample, only 85 (0.42%) received the recognition of excellence from MERLOT. Figure 13 shows the percentage of each type of material that received this recognition. As it can be seen in Figure 13, the type of material which most received the award was *Simulation* (with 25.9%), followed by *Tutorial* (23.5%), *Collection* (16.5%) and *Lecture/Presentation* (9.4%).

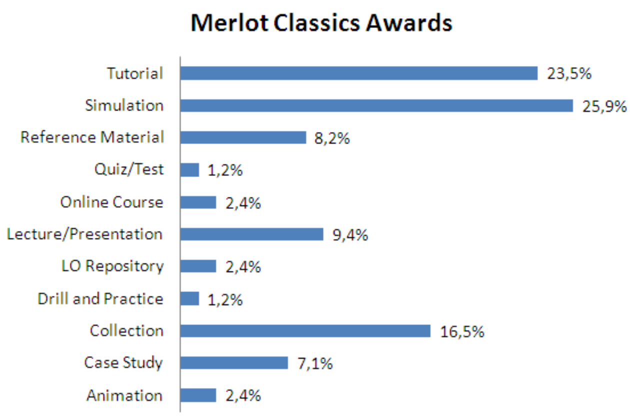


Figure 13. Percentage of types of materials that received the MERLOT Classics Awards

4.3.3 Material types and the ratings given by peer-reviewers and users

In order to allow a cross-tabulation between the different types of material and their associated quality, ratings from peer-reviewers and users were categorized into the following classes: *very poor* (ratings from 0 to 2); *poor* (ratings from 2.1 to 3); *average* (ratings from 3 to 3.9); *good* (ratings from 4 to 4.9) and *excellent* (ratings equal to 5). Table 8 and Table 9 provide the percentages of each type of material in each one of the classes of quality for the ratings given by peer-reviewers and by users respectively. This analysis was performed for the sample where learning resources were rated at least one time by peer-reviewers and users at the same time ($PRR > 0 \cap UR > 0$).

As it can be seen in Table 8, most of the resources are considered by peer-reviewers as *good* (47.55%) or *excellent* (35.45%) and only a small part of them is considered *poor* (2.59%) or *average* (14.41%). Regarding the different types of material, it is possible to observe that *Animation* is the one with the worst evaluations, with 33.33% of its materials classified as *average*, followed by *Lecture/Presentation* with 19.05% classified as *average* and 4.76% classified as *poor*. The types of materials that received the best evaluations are: *Case Study* with 54.55% of excellent materials and 36.36% of *good* materials; *Collection* with 44.19% of *excellent* materials and 45.35% of *good* materials; and *Drill and Practice* with 36.84% of *excellent* materials and 57.89% of *good* materials (in fact, the types of materials with the best evaluations are *LO Repository*, *Online Course*, *Other Resource* and *Professional Paper*, however, as they have very few elements in the data set, this information was disregarded in our analysis)

As it can be seen in Table 9, 54.32% of the resources rated by users are considered *good* and 32.28% *excellent*. Regarding the types of materials, the one which presented the highest evaluations is *Case Study* with 72.73% of its materials rated as *excellent* and 18.18% rated as *good*, followed by *Drill and Practice* with 47.37% of *excellent* materials and 47.37% of *good* materials, and *Collection* with 33.72% of *excellent* materials and 59.30% of *good* materials (in fact, the types of materials with the best evaluations are *LO Repository*, *Online Course* and *Other Resource*, however as they have very few elements in the data set, this information was disregarded in our analysis).

The materials which presented the worst evaluations are *Quiz/Test* with 13.79% of *average* and 10.34% of *poor*, *Animation* with 12.32% of *average* and 2.22% of *poor* and *Simulation* with 12.20% of *average*, 1.49% of *poor* and 0.75% of *very poor* (in fact, the type of material with the worst evaluations is *Professional Paper*, however, as they have very few elements in the data set, this information was disregarded in our analysis).

4.3.4 Material types and categories of discipline

Table 10 presents the cross-tabulation that was conducted between material types and the categories of discipline existing in MERLOT. As it can be seen in the table, the types of materials present different percentages of occurrence in the different disciplines, as follows: in *Arts*, we observe more resources of the *Collection* type; in *Business* and *Social Sciences*, resources are most concentrated on *Tutorial*, *Reference Material* and *Simulation*; in *Humanities*, we find more resources of *Collection*, *Reference Material* and *Drill and Practice*; in *Mathematics and Statistics*, resources of the *Simulation* type predominate over the others, followed by *Reference Material*; and in *Science and Technology*, those with the highest number of occurrences are *Simulation* and *Tutorial*.

Table 8. Percentage of Type of Material per classes of quality according to peer reviewers ratings

Material Type	**Total Percentage (Row Percentage)**				
	poor	**average**	**good**	**excellent**	**Total**
Animation	0.00 (0.00)	2.16 (33.33)	2.16 (33.33)	2.16 (33.33)	6.48
Case Study	0.00 (0.00)	0.14 (9.09)	0.58 (36.36)	0.86 (54.55)	1.59
Collection	0.43 (3.49)	0.86 (6.98)	5.62 (45.35)	5.48 (44.19)	12.39
Drill and Practice	0.00 (0.00)	0.29 (5.26)	3.17 (57.89)	2.02 (36.84)	5.48
Learning Material	0.00 (0.00)	0.00 (0.00)	0.29 (33.33)	0.58 (66.67)	0.86
LO Repository	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.29 (100.0)	0.29
Lecture/Presentation	0.29 (4.76)	1.15 (19.05)	3.17 (52.38)	1.44 (23.81)	6.05
Online Course	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.29 (100.0)	0.29
Other Resource	0.00 (0.00)	0.00 (0.00)	0.43 (100.0)	0.00 (0.00)	0.43
Professional Paper	0.00 (0.00)	0.00 (0.00)	0.14 (100.0)	0.00 (0.00)	0.14
Quiz/Test	0.00 (0.00)	0.58 (13.79)	2.31 (55.17)	1.30 (31.03)	4.18
Reference Material	0.72 (3.82)	3.17 (16.79)	8.36 (44.27)	6.63 (35.11)	18.88
Simulation	0.43 (1.83)	3.31 (14.02)	12.54 (53.05)	7.35 (31.10)	23.63
Tutorial	0.72 (3.73)	2.74 (14.18)	8.79 (45.52)	7.06 (36.57)	19.31
Total	**2.59**	**14.41**	**47.55**	**35.45**	**100.00**

Table 9. Percentage of material type per classes of quality according to users ratings

Material Type	**Total Percentage (Row Percentage)**					
	very poor	**poor**	**average**	**good**	**excellent**	**Total**
Animation	0.14(2.22)	0.14(2.22)	0.86(13.32)	3.75(57.78)	1.59(24.44)	6.48
Case Study	0.14(9.09)	0.00(0.00)	0.00(0.00)	0.29(18.18)	1.15(72.73)	1.59
Collection	0.00(0.00)	0.00(0.00)	0.86(6.98)	7.35(59.30)	4.18(33.72)	12.39
Drill and Practice	0.00(0.00)	0.00(0.00)	0.29(5.26)	2.59(47.37)	2.59(47.37)	5.48
Learning Material	0.00(0.00)	0.00(0.00)	0.29(33.33)	0.43(50.00)	0.14(16.67)	0.86
LO Repository	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.14(50.00)	0.14(50.00)	0.29
Lecture/Presentat	0.00(0.00)	0.00(0.00)	0.72(11.90)	3.75(61.90)	1.59(26.19)	6.05
Online Course	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.29(100.0)	0.29
Other Resource	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.14(33.33)	0.29(66.67)	0.43
Professional Paper	0.00(0.00)	0.00(0.00)	0.14(100.0)	0.00(0.00)	0.00(0.00)	0.14
Quiz/Test	0.00(0.00)	0.43(10.34)	0.58(13.79)	2.31(55.17)	0.86(20.69)	4.18
Ref. Material	0.00(0.00)	0.14(0.76)	2.16(11.45)	9.80(51.91)	6.77(35.88)	18.88
Simulation	0.29(1.22)	1.01(4.27)	2.88(12.20)	12.82(54.27)	6.63(28.05)	23.63
Tutorial	0.14(0.75)	0.29(1.49)	1.87(9.70)	10.95(56.72)	6.05(31.34)	19.31
Total	**0.72**	**2.02**	**10.66**	**54.32**	**32.28**	**100.0**

Table 10. Percentage of material type per category of discipline

Material Type	Total Percentage (Row Percentage)							
	Arts	**Business**	**Education**	**Humanities**	**Math & Statistics**	**Science& Tech**	**Social Sciences**	**Total**
Animation	0.43(12.00)	0.43 (5.08)	0.72 (3.23)	0.29(1.77)	0.86 (10.71)	3.75(10.74)	0.00 (0.00)	6.48
Case Study	0.00 (0.00)	0.43 (5.08)	0.72 (3.23)	0.29 (1.77)	0.00 (0.00)	0.00 (0.00)	0.14 (2.27)	1.59
Collection	1.30 (36.00)	0.58 (6.78)	3.46 (15.48)	4.03 (24.78)	0.58 (7.14)	1.87(5.37)	0.58 (9.09)	12.39
Drill and Practice	0.00 (0.00)	0.58 (6.78)	0.29 (1.29)	3.03 (18.58)	0.29(3.57)	1.30 (3.72)	0.00 (0.00)	5.48
Learning Material	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.86 (2.48)	0.00 (0.00)	0.86
LO Repository	0.00 (0.00)	0.00 (0.00)	0.29 (1.29)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.29
Lecture/Presentation	0.29 (8.00)	0.72 (8.47)	1.30 (5.81)	1.59 (9.73)	0.00 (0.00)	1.73 (4.96)	0.43 (6.82)	6.05
Online Course	0.00 (0.00)	0.14 (1.69)	0.14 (0.65)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.29
Other Resource	0.00 (0.00)	0.00 (0.00)	0.43 (1.94)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.43
Professional Paper	0.00 (0.00)	0.00 (0.00)	0.14 (0.65)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.14
Quiz/Test	0.00 (0.00)	0.43 (5.08)	1.87 (8.39)	0.29 (1.77)	0.14 (1.79)	0.86 (2.48)	0.58 (9.09)	4.18
Ref. Material	0.43(12.00)	1.73 (20.34)	5.91 (26.45)	3.46 (21.24)	1.87 (23.21)	3.75 (10.74)	1.73 (27.27)	18.88
Simulation	0.72 (20.00)	1.30 (15.25)	2.88 (12.90)	1.44 (8.85)	3.89 (48.21)	11.53 (33.06)	1.87 (29.55)	23.63
Tutorial	0.43 (12.00)	2.16 (25.42)	4.18 (18.71)	1.87 (11.50)	0.43(5.36)	9.22 (26.45)	1.01 (15.91)	19.31
Total	**3.60**	**8.50**	**22.33**	**16.28**	**8.07**	**34.87**	**6.34**	**100.00**

These concentrations of certain material types in specific disciplines can indicate some sort of suitability of these materials to the subject area of these disciplines.

4.3.5 Material types and personal collections

The number of personal collections assigned to a learning resource is a significative indicator of users' preferences and was also previously identified as a good predictor of quality in MERLOT (García-Barriocanal & Sicilia, 2009). Table 11 shows the average of personal collections assigned to each one of the types of material in the analyzed data base.

Table 11. Average and standard deviations of personal collections (PC) by material type

Material Type	*Average (Std Dev)*	*Material Type*	*Average (Std Dev)*
Animation	14.77 (27.90)	Online Course	46.00 (14.14)
Case Study	17.54 (17.55)	Other Resource	13.33 (9.01)
Collection	13.65 (22.00)	Professional Paper	-
Drill and Practice	11.57 (11.23)	Quiz/Test	13.44 (20.07)
Learning Material	7.66 (4.22)	Reference Material	12.53 (28.95)
LO Repository	4.50 (3.53)	Simulation	12.02 (24.85)
Lecture/Presentation	15.71 (34.89)	Tutorial	12.17 (16.46)

As can be seen in Table 11, *Case Study* is the type of material which presented the highest average of personal collections per resource (17.54) (in fact, the type of material with highest average in personal collections is *Online Course*, however, as this material has only 2 resources, we disregarded it in our analysis), followed by *Lecture/Presentation* (15.71), *Animation* (14.77) and *Collection* (13.65). It is important to highlight here that *Case Study* and *Collection* were also highly rated by peer-reviewers and users. The high number of personal collections associated to resources of the *Animation* type is another important information observed here, since that corroborates the tendency previously observed that users have in selecting this kind of material. However, the same association does not apply to the case of *Lecture/Presentation*, where the tendency observed earlier (i.e. that users do not tend to select this kind of material) is not confirmed since this material has a high average number of personal collections associated to it. An in-depth look inside personal collections regarding material types and the different categories of disciplines could help to better understand why this contradictory situation is happening.

4.4 Associations between ratings from the two groups of evaluators in MERLOT

This section analyzes the existence of associations between the ratings given by these two groups in MERLOT. This serves us to discover whether or not they diverge about the quality assessment of the same materials, as well as to explore the usefulness of these two complementary evaluations towards the assurance of quality inside the repository.

4.4.1 The Analysis

A non-parametric analysis was performed using the Spearman's rank correlation (r_S) to evaluate whether or not there was an association between the ratings of the two groups, i.e., whether the raters agree or not about the quality of the resources. Considering that MERLOT divides resources into different categories of disciplines, we split the collected sample according to these categories. This allowed us to observe potential differences according to the background of the evaluators. As MERLOT allows users to catalogue the materials in more than one category of discipline, some part of the split sample was overlapped. However, we decided to maintain the learning objects classified in more than one discipline due to the fact we considered this overlap relatively small (16%). Table 12 presents the results of this analysis.

Table 12. Correlation between users ratings and peer-reviewers ratings considering the categories of disciplines

Discipline	**Sample Size**	**PRR Average (std)**	**UR Average (std)**	r_S	**P**	**S**
All	694	4.34(0.70)	4.29(0.70)	0.19	0.00	Y
Arts	25	4.14(0.74)	4.43(0.58)	0.20	0.33	N
Business	59	4.22(0.79)	4.15(0.94)	0.06	0.66	N
Education	167	4.41(0.68)	4.36(0.72)	0.16	0.04	Y*
Humanities	133	4.60(0.51)	4.40(0.67)	0.19	0.03	Y

Mathematics & Statistics	66	4.67(0.52)	4.25(0.69)	0.17	0.31	N
Science & Technology	284	4.21(0.71)	4.25(0.72)	0.26	0.00	Y
Social Sciences	73	4.20(0.75)	4.38(0.60)	0.2	0.09	Y^+

In Table 12, the column P stands for the p-value obtained in the analysis. In the column S, N stands for no significant association between the ratings given by the two groups of evaluators, Y represents significant association at 99% level, Y* means significant association at 95% level, and Y^+ at 90% level. The correlation coefficient (r_s) indicates the strength of the association between the two groups of ratings varying from -1 to 1, where 0 means no association (no agreement). The closer the correlation coefficient is to 1, the better is the association. As can be seen in Table 12, the disciplines of *Arts*, *Business*, and *Mathematics and Statistics* did not present any association between the ratings given by users and peer-reviewers. However, the ratings for the disciplines of *Education*, *Humanities*, *Science and Technology* and *Social Sciences*, and for the overall sample did presented an association. But, even though these associations exist, they are not too strong, as their coefficients of correlation are relatively small. Figure 14 better illustrates the weakness of the association for the discipline of *Science and Technology* (we selected this discipline due to the fact that this was the one which has presented the highest coefficient of correlation of all).

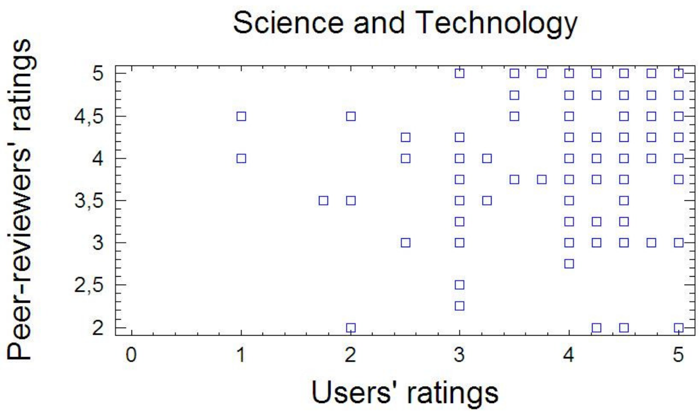


Figure 14. Scatter plot matrix of ratings for the discipline of Science and Technology

As can be seen in Figure 14, it is not possible to observe patterns indicating concordance between the ratings of the two groups. (A strong correlation between the ratings could be suggested by a formation of a diagonal line, or the agglomeration of dots in some region of the matrix, for instance). In fact, we can observe several cases where users and peer-reviewers strongly disagree about the ratings. The weakness of the associations is also confirmed when we perform a linear regression analysis in order to explore more deeply the relationships between users and peer-reviewers ratings in the discipline of *Science and Technology*. Despite the fact that it is possible to generate a linear prediction model at a 99% level of significance, the coefficient of correlation remains small (0.28) and the model is able to represent only 7.94% of the entire population.

At first glance, this exploratory analysis indicates that both groups of reviewers have different impressions about the quality of the learning objects catalogued in MERLOT, thus serving as complementary views of assessment inside the repository.

4.4.2 Users ratings against the three criteria of peer-reviewers ratings

As mentioned before, the peer-reviewer rating is composed of three main criteria: 1) *Quality of Content*, 2) *Potential Effectiveness as a Teaching Tool*, and 3) *Ease of use*. As the existence (or the absence) of associations could be specifically related to one of these dimensions, we ran the same analysis to evaluate the associations between the users' ratings and each one of the criteria of peer-reviewers' ratings. Table 13 presents the results of this analysis.

Table 13 . Correlation between users ratings and the three criteria of peer-reviewers ratings

Discipline	**Quality of content**			**Potential effective as a teaching tool**			**Ease of Use**			**Total Rating**
	r_S	**P**	**S**	**r_S**	**P**	**S**	**r_S**	**P**	**S**	**S**
All	0.21	0.00	Y	0.25	0.00	Y	0.19	0.00	Y	Y
Arts	0.21	0.30	N	0.14	0.50	N	0.41	0.04	Y*	N
Business	0.07	0.61	N	0.04	0.73	N	0.13	0.32	N	N
Education	0.11	0.15	N	0.30	0.00	Y	0.15	0.05	Y*	Y*
Humanities	0.27	0.01	Y	0.26	0.00	Y	0.15	0.09	Y+	Y
Mathematics & Statistics	0.23	0.06	Y*	0.02	0.87	N	0.02	0.85	N	N
Science & Technology	0.25	0.00	Y	0.32	0.00	Y	0.25	0.00	Y	Y
Social Sciences	0.25	0.03	Y*	0.27	0.05	Y*	0.11	0.33	N	Y^+

Table 13 shows that in some disciplines the associations encountered before do not persist for all three evaluation criteria. For instance, in the *Education* discipline, users' ratings are associated with only two of the three evaluation criteria, more precisely *Potential effectiveness as a teaching tool* and *Ease of use,* and in the discipline of *Social Sciences* users' ratings are associated only to the *Quality of content* and *Potential effectiveness as a teaching tool* criteria. Moreover, other disciplines that did not present any association between the two groups of ratings now present association between users' ratings and some specific evaluation criteria. This is the case of the discipline of *Mathematics and Statistics*, which presented association between the users' ratings and the *Quality of content* criteria, and of the discipline of *Arts*, which presented association between users' ratings and the *Ease of use* criteria.

It is interesting to highlight that the associations encountered in this analysis are slightly stronger than the associations encountered before (with just two exceptions). However, the coefficients of correlations found are still weak, which reinforces the initial conclusion that users and peer-reviewers have different impressions about quality. Such weakness is again confirmed when we perform a linear regression analysis between users ratings and the *Ease of use* criteria in the discipline of *Arts* (these are the data sets which presented the highest coefficient of correlation, with $r_s = 0.45$). Here, the prediction model is generated at a 95% level of significance, however, the coefficient of correlation is small (0.46) and the model represents only 21% of the entire population.

As the evaluations given by users do not follow any pre-defined criteria, it is difficult to precisely understand what they are referring to without an in-depth look at the textual information (comments) attached to them. Bearing this in mind, one of the two following situations may be occurring:

1) The impressions of quality that users have are not related to any one of the criteria evaluated by peer-reviewers; or
2) If users evaluate the same aspects as peer-reviewers, they do not agree with the ratings given by the experts about these aspects.

From our point of view, these two situations alternate depending on the category of the discipline and on the peer-review criteria under evaluation. Moreover, the fact that some associations observed in the first analysis now present a stronger coefficient of correlation in this second analysis may be an indication that, for these cases, the situation number 2 is occurring. However, these are all assumptions that still require more investigation to be confirmed.

In the present Chapter we have described the characteristics of MERLOT according to the main aspects of LOs earlier presented on Chapter 2. These descriptions served to the purpose of contextualizing the scope of the studies carried out on this thesis. Moreover, in this Chapter we have also observed that the types of materials inside MERLOT have grown differently over the years, which might indicate preferences of the community of the repository for certain types of materials. At last, it was possible to conclude that both communities of evaluators inside the repository are communicating different impressions about the quality of the LOs. In the light of what was observed on this Chapter, it was possible to deduce the following conclusions that will help/guide us through the process of creation of the learning objects profiles:

1. The development of learning objects profiles must consider the different types of materials and the different categories of disciplines attended by them; and
2. As both communities of evaluators in MERLOT are communicating different impressions about quality, learning object profiles will most likely differ depending on the perspective of quality we are using as reference (peer-reviewers or users). Therefore, it is necessary to create profiles considering both perspectives.

Chapter 5 Statistical Profiles of Highly-Rated Learning Objects

The methodology used for the present study was the development of highly-rated learning object profiles of the MERLOT repository. The study described in this chapter is based on the methodology applied by Ivory & Hearst (2002b) , as well as on the methodology described on García-Barriocanal & Sicilia (2009) and Cechinel, Sánchez-Alonso & García-Barriocanal (2011). The created profiles were then further used in Chapter 6 to generate models for automated quality assessment of learning objects. Figure 15 gives a general idea of the methodology applied here.

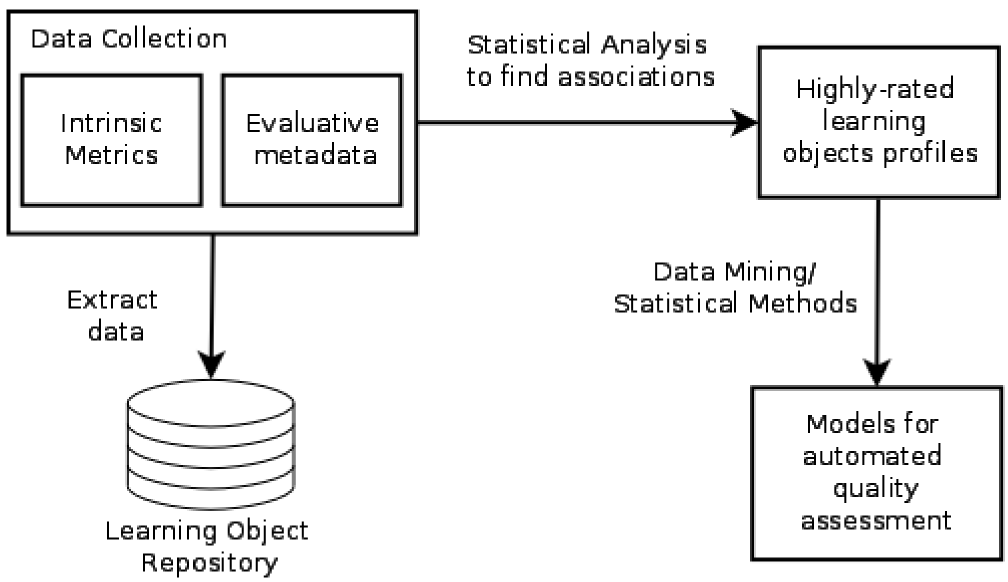


Figure 15. Methodology for generating models for automated quality assessment

The process for creating models for automated quality assessment starts with the collection of data about learning objects stored on a given repository. These data are constituted of two parts: 1) the LOs intrinsic metrics, and 2) their evaluative metadata. Once we have such information, it is possible to split the LOs into groups of quality according to the quality measures contained in the evaluative metadata, and contrast the intrinsic features of these groups of quality in order to encounter associations. The identified associations form the profiles of the learning objects and will serve as input for data mining algorithms that will generate the models for automated quality assessment. Sections 5.1 and 5.2 will present the data collected for the present study and describe the methodology used for the establishment of the classes of quality, respectively. On Section 5.3 the created learning objects profiles are presented and discussed more in-depth, and on Section 5.4 we will present the most important conclusions of these studies.

5.1 Data Collection

A second database from the MERLOT repository was gathered on January 2010 by using a web crawler that systematically traversed the web pages of the repository[3]. Information of a total of 35 metrics was extracted from 6,470 learning objects that were incrementally traversed according to the numerical identifiers of learning objects used by MERLOT. It is important to mention that the crawler has actually accessed information of 8,250 learning objects addressed in MERLOT; however a large amount of the pages refereed by the links (1,780) were not available anymore and consequently their metrics were not computed.

As the catalogued learning objects varied enormously in their number of pages (from those with just one single page to those with thousands of pages), a limit of 2 levels of depth from the root node was established for the crawler (where the root has depth 0). The number of pages refers to the number of websites that form the learning object as a whole, whereas the depth level of a certain web page refers to the number of links the user must click to reach that web page starting from the main home page (or the root node).

In other words, the crawler counts the information available at the main home page (root node or depth 0), at the web pages linked by the main home page (depth 1), and at the web pages linked by those web pages linked by the main home page (depth 2). The computation of all the metrics was made in a two-step process: 1) The intrinsic metrics were computed observing the overall counting of some attribute for

[3] The first database described on Chapter 4 was only related to metadata; here we also collected the intrinsic metrics of the resources.

all the pages traversed, and 2) average metrics were derived from the previous group by dividing each counting by the total number of pages. We also computed the number of bookmarks in *Personal Collections* for each LO as referential data for contrasting with the metrics extracted from their contents. Table 14 shows the metrics computed for this study.

Table 14. Description of the metrics collected for the study

Class of Measure	**Metric**
Link Measures	Number of Links, Number of Unique[a] Links, Number of Internal Links[b], Number of Unique Internal Links, Number of External Links, Number of Unique External Links
Text Measures	Number of Words, Number of words that are links[c]
Graphic, Interactive and Multimedia Measures	Number of Images, Total Size of the Images (in bytes), Number of Scripts, Number of Applets, Number of Audio Files, Number of Video Files, Number of Multimedia Files
Site Architecture Measures	Size of the Page (in bytes), Number of Files for downloading, Total Number of Pages[+]
Evaluation Metadata (contrast metric)	Number of Personal Collections[+]

[a] The term Unique stands for "non-repeated"
[b] The term internal refers to those links which are located at some directory below the root site
[c] For these metrics the average was not computed or does not exist

5.2 Definition of the Learning Objects Classes

The method followed for this study was the development of statistical profiles of learning objects aiming to classify them into three classes (i.e. *good*, *average*, and *poor*) considering their quantitative measures. Learning objects with at least one user or one peer-review rating were selected from the gathered database to perform the analysis, and the rest was discarded. The classes were established by the ratings thresholds (considering peer-reviews and users ratings) for categories of disciplines and also for categories of material types. The thresholds are the terciles of the ratings, so that learning objects with ratings below the first threshold belong to the *poor* category (it is important to highlight here that the label "*poor*" does not necessarily means that some resource is of bad quality), objects with ratings in between the two thresholds belong to the *average* category, and objects with ratings equal or higher to the second threshold belong to the *good* category.

On a previous study carried out with data collected from MERLOT, García-Barriocanal & Sicilia (2009) observed that the distribution of ratings "*tend to have increasing relative frequency histograms*", i.e., most ratings tend to be positive (above the intermediate rating three). In consequence, they stated that the focus of the creation of LOs profiles should be on distinguishing the attributes of highly-rated LOs, and not "*establishing categories of good and bad resources*". Following the same steps of García-Barriocanal & Sicilia (2009), we also focused our study on highly-rated learning objects and conducted our analysis contrasting *good* resources with *not-good* resources, where the *not-good* category is formed by the joint of the *average* and *poor* categories.

From the 6,470 learning objects gathered from MERLOT, 1,765 (27.27%) had at least one peer review or one user rating and were used in the analysis, and the rest were discarded. As MERLOT allows users to classify resources in more than one discipline, some of the data regarding the breakdown in categories of disciplines overlapped. This means that some learning objects appear more than once in the sample, for instance, a learning object which was tagged to the Arts discipline and to the Business discipline at the same time will appear duplicated in the sample. We decided to retain the objects classified into more than one discipline due to the fact that the overlap is small (8.3% of the rated resources). Table 15 presents the amount of learning objects, peer-reviews and users ratings, and the average thresholds that divide each category of discipline in three subsets.

Table 15. Sample used and breakdown in categories per discipline

Discipline	**Size**	**Peer Reviewer Rated**	**User Rated**	**Thresholds Peer reviews**	**Thresholds Users**
Arts	69	57	24	4 | 4.5	4 | 5
Business	245	201	60	3.75 | 4.5	4 | 5
Education	227	174	100	4.25 | 5	4 | 5
Humanities	286	225	102	4.5 | 5	4 | 5
Mathematics and Statistics	213	158	83	4 | 5	4 | 4.25

Science and Technology	758	354	524	4 | 4.75	4 | 4.25
Social Sciences	114	60	73	4 | 4.5	4 | 4.5
Total	1,912	1,229	966	-	-

Table 16 provides the breakdown of reviewed and user commented/rated resources regarding their material type and also presents the thresholds that divide material types in *poor*, *average* and *good*. From the 1,765 learning objects associated to ratings, 4 were not classified in any category of material type and were disregarded. Moreover, some categories of material types (*Learning Material*, *Learning Object Repository*, *Online Course*, and *Technical Tool*) presented too few learning objects and were also disregarded. Other categories that do not appear in Table 16 did not have any rated material in our sample.

Table 16. Sample used and breakdown in categories per material type

Material Type	**Size**	**Peer Reviewed Rated**	**User Rated**	**Thresholds Peer reviews**	**Thresholds Users**
Animation	99	64	58	4 | 4.75	3 | 4
Case Study	51	42	14	4 | 4.5	4 | 5
Collection	206	156	73	4.25 | 4.75	4 | 4.5
Drill	80	58	33	4 | 4.5	4 | 5
Lecture/Presentation	183	127	71	4 | 4.5	4 | 4.75
Quiz	34	28	14	4 | 4.25	4 | 4.5
Reference Material	344	178	202	4 | 4.75	4 | 4.5
Simulation	400	232	238	4 | 4.75	3.75 | 4
Tutorial	342	231	167	4 | 4.75	4 | 5
All	1,739	1,116	870	-	-

As it can be noticed in the thresholds presented in Table 15 and Table 16, the ratings given by users and peer-reviews tend to be positive (above the intermediate rating 3) endorsing the conclusions of García-Barriocanal & Sicilia (2009) about this issue. Here, almost all thresholds for the *average* materials are higher or equal to 4 (there are only 3 situations where they are between 3 and 4).

5.3 Analysis and Discussion

The analysis was conducted to contrast the metrics against the group of *good* and *not-good* resources considering the peer-reviews and the users ratings thresholds for the categories of disciplines and for the categories of material types with the aim of deriving statistical profiles of highly-rated learning resources that could be further used in the process of automated quality assessment. As the samples did not follow a normal distribution, a Mann-Whitney (Wilcoxon) test was performed to evaluate whether or not the selected metrics presented significant difference in their medians between the groups, and a Kolmogorov-Smirnov test was performed to evaluate differences regarding the distributions. The good and not-good groups were considered to have different profiles regarding a specific metric when both the distributions and medians presented significant difference at 90% confidence level for the thresholds evaluated. When such a situation is observed it means that the evaluated metric is associated (or correlated) to learning object quality.

5.3.1 Analysis within Categories of Discipline

From the 35 metrics evaluated for the 7 categories of discipline, just the *Average Number of Files for Downloading* metric did not present any significant difference between *good* and *not-good* resources across all the disciplines. The other 34 metrics have presented significant differences in both distributions and medians for at least 1 category either in the users or in the peer-reviews ratings thresholds. Table 17 and Table 18 introduce the frequency and the percentage in which *good* and *not-good* materials presented distinct profiles for each one of the metrics across the 7 categories of discipline for peer-reviews and users ratings thresholds respectively

Table 17. Frequency and percentage in which good and not-good resources presented different profiles regarding the metrics in the context of categories of discipline for peer-reviews ratings thresholds

Amount	Metrics	Frequency	Percentage
1	Number of Personal Collections	6	85.7%
1	Number of Images	4	57.1%
4	Number of External Links, Average Number of Unique Internal Links, Average Number of External Links, Average Number of Images	3	42.9%

11	Number of Links, Number of Unique Links, Number of Unique Internal Links, Number of Unique External Links, Size of the Page (in bytes), Total Size of the Images (in bytes), Average Number of Unique External Links, Average Number of Unique Links, Average Number of Links, Average Size of the Pages, Average Size of the Images	2	28.6%
16	Number of Internal Links, Number of Scripts, Number of Applets, Number of Words, Total Number of Pages, Number of Files for downloading, Number of audio files, Number of video files, Number of multimedia files, Average Number of Internal Links, Average Number of Words, Average Number of Audio Files, Average Number of Video Files, Average Number of Multimedia Files, Average Number of Applets, Average Number of Scripts	1	14.3%
2	Number of Words that are links, Average Number of Files for downloading	0	0.0%

Table 18. Frequency and percentage in which good and not-good resources presented different profiles regarding the metrics in the context of categories of discipline for users ratings thresholds

Amount	Metrics	Frequency	Percentage
1	Total Size of the Images (in bytes)	4	57.1%
5	Number of Internal Links, Number of Applets, Number of Words that are links, Average Number of Applets, Number of Personal Collections	3	42.9%
15	Number of Links, Number of Unique Links, Number of Unique Internal Links, Number of External Links, Number of Unique External Links, Number of Images, Size of the Page (in bytes), Number of Scripts, Total Number of Pages, Number of video files, Average Number of Unique Internal Links, Average Number of Internal Links, Average Number of Links, Average Number of Words, Average Size of the Images	2	28.6%
7	Number of Words, Average Number of Unique External Links, Average Number of External Links, Average Number of Video Files, Average Size of the Pages, Average Number of Scripts	1	14.3%
8	Number of Files for downloading, Number of audio files, Number of multimedia files, Average Number of Unique Links, Average Number of Files for	0	0.0%

	downloading, Average Number of Audio Files, Average Number of Multimedia Files, Average Number of Images

According to our analysis, the metrics behave differently across the categories of disciplines depending on the thresholds evaluated (peer-reviews or users ratings). For instance, the *Number of Personal Collections* presents different profiles in 6 of the 7 categories of disciplines for the peer-reviews ratings thresholds (see Table 17) and only in 3 categories of disciplines for the users ratings thresholds (see Table 18). This also happens for many other metrics (see Table A. 1 of Appendix A) and indicates strong differences for the highly-rated learning object profiles between the two groups of evaluators (peer-reviewers and users). This general absence of correlation between peer-reviews and users ratings is consistent with the previous study presented on Section 4.4 which stated that these two groups of evaluators possess distinct impressions about the quality of learning objects catalogued in MERLOT, and thus tend to rate the resources differently.

In the case of the peer-reviews ratings thresholds, after *Personal Collections*, the *Number of Images* has appeared to be the strongest candidate to be used in the model for quality assessment (with different profiles in 4 of the 7 categories of disciplines), followed by *Number of External Links*, *Unique Internal Links*, *Average Number of External Links*, and *Average Number of Images* (with different profiles in 3 of the 7 categories of disciplines). For the users ratings thresholds, the *Total Size of the Images (in bytes)* was the metric most associated to the ratings, followed by the 5 metrics of the second line of Table 18.

The differences among the metrics and the groups of evaluators across the categories of disciplines are presented in Table 19. As it can be noticed, certain categories of disciplines are more correlated to metrics depending on the different groups of ratings thresholds. For instance, *Education* presents different profiles for 19 metrics when we analyze them against the peer-reviews thresholds, but it presents different profiles for only 4 metrics when we consider the users thresholds. On the contrary, the *Science and Technology* and *Mathematics and Statistics* disciplines present more different profiles for the metrics in the users ratings thresholds (23 and 21 respectively) than for metrics in peer reviews ratings thresholds (9 and 13 respectively). The discipline of *Arts* has shown the same amount of metrics in both ratings thresholds, and the discipline of *Humanities* did not have any distinct profile for metrics regarding the users ratings thresholds.

Table 19. Frequency and Percentage in which good and not-good materials presented different profiles regarding the disciplines

Discipline	Frequency in Peer-Reviews Threshold	Percentage in Peer-Reviews Threshold	Frequency in Users Threshold	Percentage in Users Threshold
Arts	3	8.6%	3	8.6%
Business	4	11.4%	2	5.7%
Education	19	54.3%	4	11.4%
Humanities	2	5.7%	0	0.0%
Mathematics and Statistics	13	37.1%	21	60.0%
Science and Technology	9	25.7%	23	65.7%
Social Sciences	9	25.7%	2	5.7%

The tendencies of the metrics, i.e. whether they present a positive or a negative correlation with *good* learning objects, also vary among disciplines (see Table A. 2 of Appendix A). A good example of this is the *Number of Images* metric which is positively associated to *good* materials in the disciplines of *Education*, *Mathematics and Statistics,* and *Science and Technology*, and negatively associated in the discipline of *Social Sciences* (for peer-reviews ratings threshold). If a certain metric presents different profiles in some discipline for both thresholds, it is most likely the tendency of this association (whether it is a positive or a negative association) will be the same in these two groups (we have found just one exception in the *Average Number of External Links* metric for the discipline of *Science and Technology*).

In the context of peer-reviews ratings thresholds, *good* learning objects from the discipline of *Arts* have more *Average Number of Unique Internal Links* and *Average Number of Images* than *not-good* resources. For the discipline of *Business* the *Number of Words* in *good* materials tend to be smaller than in *not-good* materials. For the discipline of *Education*, all the metrics with significative difference present positive association with *good* learning objects which normally present a higher *Number of Links*, *Pages*, *Audio, Video* and *Multimedia files* than the *not-good* ones. In the discipline of *Social Sciences*, *good* learning objects are negatively associated with most of the metrics; they tend to be smaller in their *Average size* and to have less *Number of External Links* and *Images* than *not-good* materials.

Considering the users ratings thresholds, the *Number of Applets* (and its average) associates positively with *good* resources in the discipline of *Business* and negatively in the disciplines of *Mathematics and Statistics* and *Science and Technology*. For these two last disciplines, all 6 *text measure metrics* (not considering the averages) are positively associated with *good* materials. The same is also valid for: *Number of Images, Size of the Page, Number of Scripts* and *Number of Words that are Links*. At last, the *Number of Personal Collections* has again demonstrated to be highly correlated with quality, since it is positively associated with *good* materials for every category of discipline and rating threshold in which it has presented significant difference.

5.3.2 Analysis within Material Type Categories

The number of metrics that presented different profiles for *good* and *not-good* resources is smaller for the material types than for the categories of discipline. Here, from the 35 evaluated metrics, six of them (*Number of Audio Files*, *Number of Multimedia Files*, *Average Number of Unique Internal Links*, *Average Number of Internal Links*, *Average Number of Unique Links*, *Average Number of Audio Files*) did not present significative difference between *good* and *not-good* learning objects in none of the categories of material type for both ratings thresholds (see Table A. 3 of Appendix A). Even though, it was possible to observe that the metrics do also behave differently across the types of materials and according to the considered threshold. Table 20 and Table 21 show the amount of metrics and the frequency in which they presented different profiles in the context of peer-reviews and users ratings thresholds respectively. As it can be noticed, the *Number of Personal Collection* is the metric most associated with quality in both thresholds. In the case of the peer-reviews ratings thresholds, the *Number of Scripts* (and *Average*) appears as the second metric that has presented more different profiles across the types of material, followed by *Number of Images*, *Number of Words that are links*, *Total Number of Pages*, and *Number of Files for downloading*. Regarding the users ratings thresholds, the *Number of Applets* (and *Averages*) presents different profiles for 3 of the 9 categories.

Table 20. Frequency and percentage in which good and not-good resources presented different profiles regarding the metrics in the context of material types for peer-reviews ratings thresholds

Amount	Metrics	Frequency	Percentage
1	Number of Personal Collections	7	77.8%
2	Number of Scripts, Average Number of Scripts	3	33.3%
4	Number of Images, Number of Words that are links, Total Number of Pages, Number of Files for downloading	2	22.2%
17	Number of Links, Number of Unique Links, Number of Internal Links, Number of Unique Internal Links, Number of External Links, Number of Unique External Links, Size of the Page (in bytes), Total Size of the Images (in bytes), Number of Applets, Number of video files, Average Number of Unique External Links, Average Number of External Links, Average Number of Files for downloading, Average Number of Video Files, Average Number of Applets, Average Number of Images, Average Size of the Images	1	11.1%
11	Number of Words, Number of audio files, Number of multimedia files, Average Number of Unique Internal Links, Average Number of Internal Links, Average Number of Unique Links, Average Number of Links, Average Number of Words, Average Number of Audio Files, Average Number of Multimedia Files, Average Size of the Pages	0	0.0%

Table 21. Frequency and Percentage in which good and not-good resources presented different profiles regarding the metrics in the context of material types for users ratings thresholds

Amount	Metrics	Frequency	Percentage
1	Number of Personal Collections	4	44.4%
2	Number of Applets, Average Number of Applets	3	33.33%
7	Number of External Links, Number of Unique External Links, Size of the Page (in bytes), Total Size of the Images (in bytes), Number of Scripts, Average Number of Images, Average Number of Scripts	2	22.22%
11	Number of Links, Number of Words , Number of Files for downloading, Number of video files, Average Number of Unique External Links, Average Number of External Links, Average Number of Links, Average Number of Words, Average Number	1	11.1%

	of Multimedia Files, Average Size of the Pages, Average Size of the Images		
14	Number of Unique Links, Number of Internal Links, Number of Unique Internal Links, Number of Images, Number of Words that are links, Total Number of Pages, Number of audio files, Number of multimedia files, Average Number of Unique Internal Links, Average Number of Internal Links, Average Number of Unique Links, Average Number of Files for downloading, Average Number of Audio Files, Average Number of Video Files	0	0.0%

Table 22 presents the frequency and percentage in which metrics have shown different profiles for the types of materials. Considering the peer-reviews ratings thresholds, the two types of material which have more metrics with different profiles are *Lecture/Presentation* and *Simulation* (with 11 of 35 metrics). Considering the users' ratings thresholds, *Simulation* followed by *Animation* are the types of material that presented a higher number of metrics with different profiles (11 and 10 metrics respectively).

Table 22. Frequency and percentage in which good and not-good materials presented different profiles regarding the material types

Discipline	**Frequency in Peer-Reviews Ratings Threshold**	**Percentage in Peer-Reviews Ratings Threshold**	**Frequency in Users Ratings Threshold**	**Percentage in Users Ratings Threshold**
Animation	1	2.9%	10	28.6%
Case Study	1	2.9%	1	2.9%
Collection	3	8.6%	0	0.0%
Drill	3	8.6%	2	5.7%
Lecture/Presentation	11	31.4%	4	11.4%
Quiz	3	8.6%	0	0.0%
Reference Material	2	5.7%	6	17.1%
Simulation	11	31.4%	11	31.4%
Tutorial	3	8.6%	1	2.9%

Similarly to the analysis of the categories of disciplines, here the metrics also present different tendencies depending on the type of material and the ratings thresholds (see Table A. 4 of Appendix A). For instance, in the context of the users ratings thresholds, the *Number of Applets* (and its average) is positively associated with *Lecture* materials and negatively associated with *Animation* and *Simulation* materials; and in the context of peer-reviews ratings threshold the *Number of Words* is positively associated with *Lecture/Presentation* material type, and negatively associated with *Tutorial* material type.

Regarding the peer-reviews thresholds, *good Lecture* materials tend to have a higher *Number of Links, Internal Links, Unique Internal Links, Images, Scripts, Pages* and *Files for downloading* than *not-good* materials. *Good Collection* materials have more *Video Files* than *not-good* materials, and *good Quiz* materials present a higher *Number of Pages, External and Unique External Links* than the *not-good* ones. At last, *Simulation* materials are positively correlated with the *Number of Images*, *Total Size of the Images* and *the Number of Scripts* (and their respective averages) and negatively correlated with the *Averages Numbers of Unique External Links, External Links, Unique Links* and *Applets*.

Considering the users ratings thresholds, *Animation good* materials have more *Links, External Links, Unique External Links, Scripts,* and *Pages* than *not-good* materials; they also tend to be larger in *Size*. *Reference Materials* belonging to the *good* category tend to be smaller in *Size*, and to have less *Words* (and average) and more *Average Images* than *not-good* materials.

5.3.3 Intersecting Categories of Disciplines and Material Types

Interesting insights can be gained when the previous analyses is again performed for a specific type of material belonging to a particular category of discipline. We have analyzed whether or not the metrics which present different profiles in the discipline of *Science and Technology* and in the *Simulation* material type maintain their correlations with ratings when these two groups are intersected in the context of peer-reviews ratings threshold. Table 23 provides the results of this comparison.

Table 23. Significant discriminators and tendencies of the metrics for the good category of Science and Technology, Simulation and the intersection of both

Metric	**Science and Technology**	**Simulation**	**Simulation ∩ Science and Technology**
Number of External Links	N*	N*	Y↓
Number of Unique External Links	N*	N*	Y↓
Number of Images	(Y)↑	Y↑	(Y)↑
Total Size of the Images (in bytes)	N*	Y↑	Y↑
Number of Scripts	Y↑	Y↑	Y↑
Number of Applets	(Y)↓	Y↓	Y↓
Average Number of Unique External Links	(Y)↓	Y↓	Y↓
Average Number of External Links	Y↓	Y↓	Y↓
Average Number of Applets	(Y)↓	Y↓	Y↓
Average Number of Images	Y↑	Y↑	Y↑
Average Size of the Pages	N	N	Y↑
Average Size of the Images	N*	Y↑	Y↑
Average Number of Scripts	Y↑	Y↑	Y↑
Number of Personal Collections	Y↑	Y↑	Y↑
Total	**9**	**11**	**14**

In Table 23, N represents no significant difference for the median of the two samples, N* represents no significant difference for the medians but significant difference in distribution, N+ stands for significant difference between the distributions (Kolmogorov-Smirnov) but significant difference between medians (Mann-Whitney), and Y stands for both differences at the same time. The overall analysis was conducted for a 95% confidence level; information in parenthesis means the results are significant at the 90% level. A hyphen means that the analysis was impossible to because all fields have the same information. Moreover, ↑ stands for a positive contribution, ↓ stands for negative contribution.

As it can be seen in Table 23, all metrics which showed significant difference in both *Science and Technology* discipline and *Simulation* material type are also associated to quality in the intersection group, as well as they have preserved the same tendencies. This also happens for the 2 metrics (Total Size of the Images and average) which have shown significant differences only for the *Simulation* group. In this last case, it is possible to conclude that significant differences for these metrics do not appear in the *Science and Technology* discipline due to the influence of all other types of materials, once that none of them have presented significative difference in these metrics for predicting good resources (see Table 24).

Table 24. Significant discriminators of the good category considering peer-reviews ratings in material types (extracted from Table A. 3 Appendix A)

Metric	**Total Size of the Images (in bytes)**	**Average Size of the Images (in bytes)**
Animation	N*	N*
Case Study	N	N
Collection	N*	N*
Drill	N	N
Lecture/ Presentation	N	N
Quiz	N	N
Reference Material	N*	N
Simulation	**Y**	**Y**
Tutorial	N*	N*

A similar kind of reasoning could be applied for the other 3 metrics (*Number of External Links*, *Number of Unique External Links*, and *Average Size of the Pages*) which have not presented significant differences in neither of the two initial groups but now present significant differences in their intersection. This may be happening because other disciplines can be pushing down the association of these metrics inside *Simulation* material type, or/and other material types can be pushing down the association of these metrics inside *Science and Technology*.

In other words, this means that not only the different categories of disciplines and material types present distinct highly-rated learning objects profiles depending on the evaluated thresholds, but that the many intersections between these groups can also present different profiles.

5.4 Conclusions

During the analysis carried out on Section 5.2 of this Chapter, it was possible to confirm a previous observation of García-Barriocanal & Sicilia (2009) about the fact that ratings on MERLOT tend to be positive (above the intermediate rating three), leading us to follow the steps of the authors and conducting the creation of profiles focused on distinguishing characteristics of highly-rated LOs.

On subsections 5.3.1 and 5.3.2, the process of creation highly-rated learning objects profiles revealed that the evaluated metrics present different associations with the quality of the resources depending on: 1) the categories of disciplines, 2) the perspective of quality used (peer-reviewers or users ratings), and 3) the different types of the materials. These differences are also true for the tendencies of such metrics, i.e., whether they present a positive or a negative correlation with the quality of the LOs. As the experiments of the subsections 5.3.1 and 5.3.2 were carried out in separate for categories of discipline and for the types of the materials, we continued exploring the creation of profiles on subsection 5.3.3, but now testing the profiles for a single subset formed by the intersection of a specific category of discipline (*Science and Technology*) and a specific type of material (*Simulation*). On that experiment it was observed that the association of the metrics can also depend on the intersection among the different categories of disciplines, material types.

All these findings suggest that the development of models for automated quality assessment must be carried out through the development of highly-rated learning objects profiles for each one of the possible intersections among categories of discipline, material types and perspectives of quality of MERLOT. However, as the number of possible intersections is considerably high (7 categories of discipline times 9 types of materials times 2 perspectives of quality = 126 subsets), we decided to restrain the generation and evaluation of models to the three intersected subsets with the higher number of occurrences and from the quality perspective of the peer-reviewers. Chapter 6 will present the details of such study.

Chapter 6 Predicting Learning Object Quality Classification

In Chapter 5, we developed statistical profiles of highly-rated learning objects of MERLOT based on their intrinsic metrics, and we observed that the creation of such profiles must be done for all possible subsets formed by the intersections among categories of discipline, material types and the perspectives of quality (peer-reviewers or users) existing in the repository. Moreover, we also concluded that the process of creating models for predicting LOs quality must be carried out for each one of these possible subsets

The present Chapter presents the creation and evaluation of models for automated quality assessment of LOs inside MERLOT through the use of Linear Discriminant Analysis (LDA) and Data Mining Classification Algorithms (DMCA). The models were generated using as parameters the intrinsic metrics identified as potential indicators of quality for the target subsets (according to the methodology described on Sections 5.1, 5.2 and 5.3). Due to the fact that the number of possible subsets is considerably high, we carried out the experiments of the present Chapter for the three subsets that presented the highest amount of resources inside the repository (considering the perspective of peer-reviewers quality). The Chapter is structured as follows: Section 6.1 we describe the data used on the study and the potential indicators of quality for each one of the selected subsets. Section 6.2 presents and discusses the results found, and Section 1.1 proposes two possible usage scenarios for the models.

6.1 Data Description

The collected sample contained LOs classified into 7 different disciplines and 9 distinct types of material, thus totalizing 63 different classes of possible learning object profiles. As mentioned in Chapter 4, learning objects on MERLOT present very different percentages of occurrences among these classes, with some subsets containing much more resources than others. Table 25 shows the occurrences of peer-reviewed learning objects for each intersected group[4].

Table 25. Number of occurrences of learning objects in each subset considering only peer-reviewed resources

Material Type/ Discipline	Arts	Business	Education	Humanities
Animation	6	4	7	9
Case Study	1	17	11	5
Collection	17	24	36	68
Drill	2	21	6	24
Lecture/ Presentation	8	28	16	25
Quiz	3	7	8	7
Reference Material	18	40	43	31
Simulation	12	16	13	16
Tutorial	12	47	37	39
Total	79	204	177	224
Material Type/ Discipline	**Mathematics and Statistics**	**Science and Technology**	**Social Sciences**	**Total**
Animation	18	32	2	78
Case Study	3	9	2	48
Collection	12	25	9	191
Drill	5	6	1	65
Lecture/ Presentation	7	45	11	140
Quiz	0	4	3	32
Reference Material	12	48	8	200
Simulation	83	97	12	249
Tutorial	24	83	12	254

[4] As some learning objects are classified in more than one discipline, there is some overlap in the groups. This is why the total sum of Table 25 is not equal to the sums presented in Table 15 and Table 16.

Total	164	349	60	1,257

We have selected the three subsets with the highest number of occurrences to generate and evaluate models for automated quality assessment in the context of peer-reviews thresholds. The selected subsets are: *Simulation ∩ Science and Technology*, *Simulation ∩ Mathematics and Statistics*, and *Tutorial ∩ Science and Technology*. Table 26 shows the size and thresholds for the three subsets and

Table 27 presents the metrics associated to quality and their tendencies.

Table 26. Thresholds of the subsets

Subset	Size	Thresholds Peer reviews
Simulation ∩ Science and Technology	97	3|4
Simulation ∩ Mathematics and Statistics	83	4|4,75
Tutorial ∩ Science and Technology	83	4|4,75

Table 27. Significant discriminators and tendencies of the metrics for the good category of the selected subsets

Metric	Simulation ∩ Science and Technology	Simulation ∩ Mathematics and Statistics	Tutorial ∩ Science and Technology
Number of Links	-	Y↑	Y↓
Number of Unique Links	-	Y↑	(Y)↓
Number of Internal Links	-	(Y)↑	Y↓
Number of Unique Internal Links	-	(Y)↑	(Y)↓
Number of External Links	Y↓	-	(Y)↓
Number of Unique External Links	Y↓	-	-
Size of the Page (in bytes)	-	Y↑	(Y)↓
Number of Images	(Y)↑	Y↑	-
Total Size of the Images (in bytes)	Y↑	Y↑	-
Number of Scripts	Y↑	Y↑	-
Number of Words	-	-	(Y)↓
Number of Words that are Links	-	-	Y↓
Number of Applets	Y↓	-	-

Average Number of Unique Internal Links	-	-	(Y)↓
Average Number of Internal Links	-	-	Y↓
Average Number of Unique External Links	Y↓	-	-
Average Number of External Links	Y↓	-	(Y)↓
Average Number of Unique Links	-	(Y)↑	Y↓
Average Number of Links	-	-	Y↓
Average Number of Applets	Y↓	-	-
Average Number of Images	Y↑	-	-
Average Size of the Pages	Y↑	-	-
Average Size of the Images	Y↑	Y↑	-
Average Number of Scripts	Y↑	(Y)↑	-
Total	**13**	**11**	**13**

On Table 27, Y stands for significant difference between the distributions (Kolmogorov-Smirnov) and medians (Mann-Whitney) of the two samples (*good* and *not-good*), thus meaning association between that metric and the LO quality for the given subset. The overall analysis was conducted for a 95% confidence level; information in parenthesis means that the results are significant at the 90% level. Besides, ↑ stands for a positive contribution of that metric to the quality of the LO and ↓ stands for negative contribution.

As it can be seen from Table 27 , the quality of the resources from *Simulation* ∩ *Science and Technology* subset tends to be positively influenced by the *Number of Images* (and average), *Total Size of the Images* (and average), *Number of Scripts* (and average) and *Average Size of the Pages*. The quality of these same resources tends to be negatively influenced by the *Number of External Links*, the *Number of Unique External Links*, the *Number of Applets* and their respective averages. Moreover, all metrics associated to good quality resources on the subset *Simulation* ∩ *Mathematics and Statistics* present positive tendencies. These metrics are: *Number of Links*, *Number of Unique Links* (and average), *Number of Internal Links*, *Number of Unique Internal Links*, *Size of the Page* (and average), *Number of Images*, *Total Size of the Images (in bytes)*, *Number of Scripts*. At last, all metrics associated to quality on the *Tutorial* ∩ *Science and Technology* subset exert a negative influence on the quality of the resources. These metrics are: *Number of*

Links (and average), *Number of Unique Links* (and average), *Number of Internal Links* (and average*), Number of Unique Internal Links* (and average), *Number of External Links* (and average), *Size of the Page* (in bytes), *Number of Words*, and *Number of Words that are Links*.

Although there are not many coincidences on the metrics among the three subsets, it is interesting to highlight that when one metric is associated to quality on two datasets that have one of the sets in common (category of discipline or material type), this metric maintain the same tendency for both subsets. For instance, the *Number of Images*, *Total Size of the Images* (and average), and *Number of Scripts* present positive tendencies in both *Simulation* ∩ *Science and Technology* and *Simulation* ∩ *Mathematics and Statistics*, and the *Number of External Links* (and average) maintain the same negative tendency for both *Simulation* ∩ *Science and Technology* and *Tutorial* ∩ *Science and Technology*.

Next Section will present the results of the use of these metrics in the creation of models for automated quality assessment of LOs belonging to the three mentioned subsets.

6.2 The Models

6.2.1 Linear Discriminant Analysis

We used Linear Discriminant Analysis (LDA) to build models in order to distinguish *good* from *not-good* resources, *good* from *average* resources, *good* from *poor* resources, and *good* from *average* and *poor* resources for the three mentioned subsets. This method is suitable to classify objects into one or more groups based on features that describe the objects. In order to build these models we used the intrinsic metrics identified as presenting correlations with the quality of resources

Table 28, Table 29 and Table 30 show the accuracy of the predictions of the four developed models for the 3 subsets. In the tables, the *squared canonical correlation* represents the percentage of variance in the metrics explained by the discriminant function, and the *classification accuracy* represents the percentage of successful categorization achieved by the model for each group.

As it can be seen in Table 28, all four models developed for the subset *Simulation ∩ Science and Technology* were able to classify resources at a significant level. On the other hand, just one model (model 2) developed for the subset *Simulation ∩ Mathematics and Statistics* (Table 29) was able to classify resources at a 90% significant level. The worst case was for the subset *Tutorial ∩ Science and Technology* (Table 30), were none of the models was able to classify resources at a significant level.

For the *Simulation ∩ Science and Technology* subset (Table 28), the model number 1, which discriminates *good* and *average* materials, has the smallest percentage of variance explained, the highest p-value (0.0983) and the worst overall accuracy (71.08%). The model number 2 presents the second best results, with an overall accuracy of 72.16% and statistical significance at the 95% level. The third model presents the best results among the 4 models when predicting between *good* and *poor* resources, achieving 91.49% of overall accuracy. This model has also the best squared canonical correlation and it is statistically significant at the 99% level. It is interesting to highlight here that as the classes included in the models get far from each other in qualitative terms (e.g. *good* and *poor* are the most distant groups, and *good* and *not-good* are the second most distant), the generated discriminant models represent higher percentage of variance in the metrics, are more accurate and more statistical significant. For instance, the model to predict between *good* and *poor* materials has an overall accuracy of 91.49%, a squared canonical correlation of 0.81, and it is statistical significant at a 99% level, whereas the model to predict between *good* and *average* has an overall accuracy of 71.08% (approximately 20% less than the first one), a squared canonical correlation of only 0.46 (almost half of the first one), and it is statistical significant only at a 90% level. This illustrates how the metrics present different strengths depending on the different groups that are contrasted and the qualitative distance among them.

For the *Simulation ∩ Mathematics and Statistics* subset (Table 29), the model number 2 was able to classify resources between good and not-good with an overall accuracy of 73.49% at a 90% significant level. Here, it is interesting to highlight the similarity of this model with the model number 2 for the *Science and Technology ∩ Simulation* subset. Both models are intended to classify resources between good and not-good, and both presented an overall accuracy between 72 and 74%, and a squared canonical correlation around 0.47. Moreover, in both cases the accuracy of the models for classifying *not-good* resources is pretty much higher than for classifying *good* resources.

For the *Tutorial* ∩ *Science and Technology* (Table 30), none of the models achieved significant levels.

Table 28. Results of linear discriminant analyses for *Simulation* ∩ *Science and Technology* in the context of peer-reviews ratings thresholds

N	Classes in the model	Num of Metrics	Squared Canonical Correlation	P-Value	Classification Accuracy				
					Good	Average	Poor	Not-Good	Overall
1	Good and Average	12[a]	0.4688	0.0983	72.00%	69.70%	-	-	71.08%
2	Good and Not-Good	13	0.47700	0.0435	63.64%	-	-	76.56%	72.16%
3	Good and Poor	13	0.81130	0.0001	96.97%	-	78.57%	-	91.49%
4	Good, Average and Poor	13	0.54510	0.0016	57.58%	62.00%	64.29%	-	60.82%

[a] The Average Size of the Pages was not included in this model

Table 29. Results of linear discriminant analyses for *Simulation* ∩ *Mathematics and Statistics* in the context of peer-reviews ratings thresholds

N	Classes in the model	Num of Metrics	Squared Canonical Correlation	P-Value	Classification Accuracy				
					Good	Average	Poor	Not-Good	Overall
1	Good and Average	11	0.44661	0.1060	58.06%	82.61%	-	-	72.73%
2	Good and Not-Good	11	0.46900	0.0656	58.06%	-	-	82.69%	73.49%
3	Good and Poor	11	0.34624	0.9762	67.74%	-	66.67%	-	67.57%
4	Good, Average and Poor	11	0.46971	0.5911	54.84%	28.26%	83.33%	-	42.17%

Table 30. Results of linear discriminant analyses for *Tutorial* ∩ *Science and Technology* in the context of peer-reviews ratings thresholds

N	Classes in the model	Num of Metrics	Squared Canonical Correlation	P-Value	Classification Accuracy				
					Good	Average	Poor	Not-Good	Overall
1	Good and Average	13	0.43866	0.5391	89.29%	41.67%	-	-	62.50%
2	Good and Not-Good	13	0.40412	0.4261	92.86%	-	-	43.64%	60.24%
3	Good and Poor	13	0.1795	0.60369	92.86%	-	52.63%	-	76.60%
4	Good, Average and Poor	13	0.50222	0.1915	57.14%	47.22%	47.37%	-	59.60%

6.2.2 Data Mining Classification Algorithms

Due to the fact that we were not able to generate LDA models for all three selected subsets, we have also attempted to create models for automated quality assessment of the resources through DMCA. According to Goldschmidt & Passos (2005), classification algorithms aim to construct models capable of associating each record of a given dataset to a labeled category. Precisely, we have used WEKA software (Hall et al., 2009) to generate and test models for the classification of resources between *good* and *not-good*, and among *good*, *average* and *poor* resources through the following classification algorithms: J48, SimpleCart, PART, Multilayer Perceptron Neural Network and Bayesian Network. Table 31, Table 32 and Table 33 present the results of these tests. For all tests we have used the same metrics previously identified as potential indicators of quality (Table 27).

There are several possible criteria for evaluation the good prediction of classification models (Cichosz, 2011). Here we selected a few of them to present the results of our analysis. In the tables, the column “metrics used by the model” presents the number of metrics that were included in the model generated by the given algorithm. The mean absolute error (MAE) measures the average deviation between the predicted classes and the true classes of the resources. The closer to 0 is the MAE, the lower is the error of prediction and the better is the model. The K stands for “Kappa statistic’ which is a coefficient that measures the overall agreement between the observed and the expected data. The coefficient varies from -1 to 1, where 1 means total agreement, 0 means no agreement, and -1 means total disagreement. At last, the tables also present the overall accuracy of the model and the specific accuracies for each one of the classes in the dataset. We adopted the MAE measure as the main reference of quality for the models, i.e., when we mention that a given model is the best for a given subset, we mean that this model has presented the minimum MAE among all.

As it can be seen in the tables, apparently there is no best classification algorithm that fits for all subsets for the generation of good models. The results vary significantly depending on the algorithm used, the subset from which the models were generated and the classes of quality included in the datasets.

Table 31. Results of DMCA for *Simulation* ∩ *Science and Technology* in the context of peer-reviews ratings thresholds

Classification Algorithm	**N**	**Classes in the model**	**Metrics used by the model***	**Number of Leaves Number of rules**	**Size of the tree**	**MAE**	**K**	**Classification Accuracy**				
								Good	**Average**	**Poor**	**Not-Good**	**Overall**
J48	1	Good and Not-Good	2	3	5	0.31	0.38	33.33%	-	-	98.43%	76.29%
	2	Good, Average and Poor	11	19	37	0.1	0.83	96.96%	84.00%	92.85%	-	89.69%
Simple Cart	3	Good and Not-Good	2	3	5	0.30	0.53	57.57%	-	-	92.18%	80.41%
	4	Good, Average and Poor	8	14	27	0.15	0.76	90.90%	86.00%	71.40%	-	85.57%
PART	5	Good and Not-Good	5	4	-	0.28	0.38	33.33%	-	-	98.43%	76.29%
	6	Good, Average and Poor	8	11	-	0.16	0.74	97.00%	72.00%	92.9%	-	83.51%
Multilayer Perceptron	7	Good and Not-Good	13	-	-	0.29	0.58	60.60%	-	-	93.75%	82.47%
	8	Good, Average and Poor	13	-	-	0.26	0.53	60.60%	92.00%	42.90%	-	74.23%
Bayesian Network	9	Good and Not-Good	3	-	-	0.30	0.37	84.84%	-	-	57.81%	67.01%
	10	Good, Average and Poor	5	-	-	0.30	0.41	60.60%	48.00%	100%	-	59.79%

*All models were tested with 13 metrics.

Table 32. Results of DMCA for *Simulation* ∩ *Mathematics and Statistics* in the context of peer-reviews ratings thresholds

Classification Algorithm	N	Classes in the model	Metrics used by the model*	Number of Leaves Number of rules	Size of the tree	MAE	K	Classification Accuracy				
								Good	Average	Poor	Not-Good	Overall
J48	1	Good and Not-Good	2	4	7	0.36	0.44	58.1	-	-	84.6%	74.70%
	2	Good, Average and Poor	4	8	15	0.26	0.47	64.5	89.1%	0	-	73.49%
Simple Cart	3	Good and Not-Good	1	2	3	0.40	0.37	48.4	-	-	86.5%	72.29%
	4	Good, Average and Poor	1	2	3	0.32	0.32	48.4	87.0%	0	-	66.26%
PART	5	Good and Not-Good	5	5	-	0.30	0.55	54.8	-	-	96.2%	80.72%
	6	Good, Average and Poor	5	-	-	0.23	0.55	77.4	87.0%	0	-	77.11%
Multilayer Perceptron	7	Good and Not-Good	11	-	-	0.42	0.17	16.1	-	-	98.1%	67.47%
	8	Good, Average and Poor	11	-	-	0.34	0.13	16.1	97.8%	0	-	60.24%
Bayesian Network	9	Good and Not-Good	0	-	-	0.47	0	0	-	-	100%	62.65%
	10	Good, Average and Poor	0	-	-	0.37	0	0	100%	0	-	55.42%

*All models were tested with 11 metrics.

Table 33. Results of DMCA for *Tutorial* ∩ *Science and Technology* in the context of peer-reviews ratings thresholds

Classification Algorithm	**N**	**Classes in the model**	**Metrics used by the model***	**Number of Leaves Number of rules**	**Size of the tree**	**MAE**	**K**	**Classification Accuracy**				
								Good	**Average**	**Poor**	**Not-Good**	**Overall**
J48	1	Good and Not-Good	3	6	11	0.25	0.62	60.7%	-	-	96.4%	84.34%
	2	Good, Average and Poor	2	4	7	0.37	0.21	0	97.2%	47.4%	-	53.01%
Simple Cart	3	Good and Not-Good	0	1	1	0.45	0	0	-	-	100%	66.26%
	4	Good, Average and Poor	5	10	19	0.24	0.64	82.1%	83.3%	57.9%	-	77.11%
PART	5	Good and Not-Good	4	6	-	0.24	0.66	67.9%	-	-	94.5%	85.54%
	6	Good, Average and Poor	5	3	-	0.35	0.25	0	100%	52.6%	-	55.42%
Multilayer Perceptron	7	Good and Not-Good	13	-	-	0.40	0	0	-	-	100%	66.26%
	8	Good, Average and Poor	13	-	-	0.38	0.20	10.7%	86.1%	47.4%	-	51.81%
Bayesian Network	9	Good and Not-Good	0	-	-	0.45	0	0	-	-	100%	66.26%
	10	Good, Average and Poor	0	-	-	0.43	0	0	100%	0	-	43.37%

*All models were tested with 13 metrics.

6.2.2.1 Simulation ∩ Science and Technology

As occurred with LDA classifiers, here the models also presented (in general) the best results for the *Simulation ∩ Science and Technology* subset. For this subset, the best model was a decision tree generated by a J48 algorithm (model number 2 of Table 31) which was able to correctly classify resources among *good*, *average* and *poor* with an overall accuracy of 89.69%, and presented a Kappa coefficient of 0.83, and a MAE of just 0.1. The percentages of accuracy of this model for classifying resources in the specific categories of quality are considerably similar. *Good* resources are classified with 96.96% of accuracy, while *average* and *poor* resources are classified with accuracy of 84% and 92.85% respectively. The second and third best models for this subset were also focused on classify resources among *good*, *average* and *poor*. The second best model was a decision tree generated by a Simple Cart algorithm with an overall accuracy of 85.57% (model number 4 of Table 31) and the third best model was a set of if-then-rules generated by the PART algorithm with an overall accuracy of 83.51% (model number 6 of Table 31). The main difference between these two models (in terms of accuracy) is that the former presented the worst accuracy percentages for classifying *poor* resources (71.40%), where the latter presented the worst accuracy percentages for classifying *average* resources (72%). At last, the best results for classifying resources between *good* and *not-good* were achieved by the PART algorithm and by a Multilayer Perceptron Neural Network. The PART model achieved an overall accuracy of 76.29 a MAE of 0.28 and Kappa Statistic of 0.38. Moreover, it classified *not-good* resources with an accuracy of 98.43%, and *good* resources with accuracy of only *33.33%*. The Multilayer Perceptron presented an overall accuracy of 82.47%, a MAE of 0.29 and a Kappa coefficient of 0.58. The drawback of these two models is the very low accuracy for classifying *good* resources.

6.2.2.2 Simulation ∩ Mathematics and Statistics

For the *Simulation ∩ Mathematics and Statistics* subset the best model was generated by the PART algorithm (model 5 of Table 32) for classifying resources between *good* and *not-good*. This model contains a set of 5 if-then-rules that uses 5 from the 11 metrics identified as possible indicators of quality. It achieved an overall accuracy of 80.72%, a MAE of 0.30 and a Kappa coefficient equals to 0.55. Even though the overall results can be considered good, the model presents a serious limitation for the classification of *good* resources, with only 54.8% of accuracy. Coincidently, the earlier attempt to generate models for this subset through the use of LDA only achieved significant results for the classification of resources between *good* and *not-good* (see model 2 of Table 29). In fact, both models presented very similar accuracy percentages. The second best model for this subset is a decision tree generated by the J48 algorithm to classify resources between *good* and *not-good* (model 1 of the Table 32). Here the model achieved an overall accuracy of 74.70, a MAE of 0.36, and a Kappa coefficient of 0.44. The main problem with this model is the fact that it uses just 2 of the 11 possible indicators of quality. For this subset, all models for classifying resources among *good*, *average* and *poor* have completely failed on the classification of the *poor* category (presenting 0% of accuracy). It is also possible to see that the accuracies for classifying *good* and *average* resources in these models are very similar to the accuracies for classifying *good* and *not-good* resources on the other models.

6.2.2.3 Tutorial ∩ Science and Technology

The best model for the subset *Tutorial ∩ Science and Technology* was generated by the PART algorithm to classify resources between *good* and *not-good* (model 5 of Table 33). The model presents an overall accuracy of 85.54%, a MAE of 0.24 and a Kappa coefficient of 0.66. From the 13 metrics identified as quality indicators, the model has included only 4 in the 6 if-then-rules generated. Moreover, the model has a high accuracy for classifying *not-good* resources (94.5%), but a low accuracy for classifying *good* resources (67.9%). The second best model for this subset is a decision tree generated by a Simple Cart algorithm that classifies resources among *good*, *average* and *poor* (model 4 of Table 33). Here the model uses 5 from the 13 metrics identified as quality indicators; it has an overall accuracy of 77.11%, a MAE of 0.24, and a Kappa coefficient of 0.64. The model is able to classify *good* resources with 82.1% of accuracy, *average* resources with 83.3% of accuracy, and *poor* resources with 57.9% of accuracy. The third best model is a decision tree generated by a J48 algorithm (model 1 of Table 33). This model classifies resources between *good* and *not-good* with an overall accuracy of 84.34%, a MAE of 0.25, and a Kappa coefficient of 0.62. The model uses only 3 from the 13 metrics identified as quality indicators. Moreover, similarly to the best model for this subset, this model also has a high accuracy for classifying *not-good* resources (96.4%) and a low accuracy for classifying *good* resources (60.7%).

6.2.3 General considerations about the results

Table 34 shows the models for the classification between *good* and *not-good* resources ordered by MAE. As it can be from the table the models normally exclude several of the metrics previously identified as indicators of quality. For instance, from the top 10 best models of Table 34, only one has used all metrics included in the dataset (a Multilayer Perceptron for the *Simulation* ∩ *Science and Technology* subset). The rest of the models have used from just one to five metrics. It is also interesting to highlight that it was possible to generate models for all three subsets. Moreover, practically all models presented a higher accuracy for the classification of *not-good* resources than for *good* resources. Figure 16 presents this last observation more clearly. As it can be seen on the figure, from the 10 best models, 9 of them presented better accuracies for classifying *not-good* resources and just one of them - a Bayesian Network for the *Simulation* ∩ *Science and Technology* subset – presented a higher accuracy for classifying *good* resources than *not-good* ones.

The best models generated for classifying resources among *good*, *average* and *poor* achieved lower MAEs and higher Kappa coefficients than the models for classifying resources between *good* and *not-good* (see Table 35). Moreover, as it can be seen in Figure 17, the models here also tend to use more indicators of quality. The main problem found for this set of models is the fact that it was not possible to create good models for the subset of *Simulation* ∩ *Mathematics and Statistics* (all models presented 0.0% of accuracy for classifying *poor* resources). Another important thing to highlight is that the best 3 models presented more balanced accuracies for the classification among the different classes. However, it is still possible to observe all kind of models, i.e., those which classify more accurate *good* resources, those which classify more accurate *average* resources, and those which classify more accurate *poor* resources (see Figure 17).

The results found here point out the possibility of generating models for automated quality assessment of learning resources inside repositories based on their intrinsic metrics. However, as the models are very heterogeneous (different MAEs, Kappa coefficients, number of metrics used, classification accuracies), the decision of which one is the best will depend on the combination of some facts such as: the specific scenario to which the model is going to be applied, the specific subset (category of discipline versus material type) to which they are being generated for, and the classes of quality included in the dataset. Next section will briefly discuss two possible scenarios of usage for the automated quality assessment models.

Table 34. Models for classifying resources between *good* and *not-good* ordered by MAE

N	**Classification Algorithm**	***Subset***	**Metrics used by the model***	**MAE**	**K**	**Classification Accuracy**		
						Good	**Not-Good**	**Overall**
1	PART	Tutorial ∩ Science and Technology	4	0.24	0.66	67.9%	94.5%	85.54%
2	J48	Tutorial ∩ Science and Technology	3	0.25	0.62	60.7%	96.4%	84.34%
3	PART	Simulation ∩ Science and Technology	5	0.28	0.38	33.33%	98.43%	76.29%
4	Multilayer Percept.	Simulation ∩ Science and Technology	13	0.29	0.58	60.60%	93.75%	82.47%
5	PART	Simulation ∩ Mathematics and Statistics	5	0.30	0.55	54.8%	96.2%	80.72%
6	Simple Cart	Simulation ∩ Science and Technology	2	0.30	0.53	57.57%	92.18%	80.41%
7	Bayesian Network	Simulation ∩ Science and Technology	3	0.30	0.37	84.84%	57.81%	67.01%
8	J48	Simulation ∩ Science and Technology	2	0.31	0.38	33.33%	98.43%	76.29%
9	J48	Simulation ∩ Mathematics and Statistics	2	0.36	0.44	58.1%	84.6%	74.70%
10	Simple Cart	Simulation ∩ Mathematics and Statistics	1	0.40	0.37	48.4%	86.5%	72.29%
11	Multilayer Percept.	Tutorial ∩ Science and Technology	13	0.40	0	0%	100%	66.26%
12	Multilayer Percept.	Simulation ∩ Mathematics and Statistics	11	0.42	0.17	16.1%	98.1%	67.47%
13	Simple Cart	Tutorial ∩ Science and Technology	0	0.45	0	0%	100%	66.26%
14	Bayesian Network	Tutorial ∩ Science and Technology	0	0.45	0	0%	100%	66.26%
15	Bayesian Network	Simulation ∩ Mathematics and Statistics	0	0.47	0	0%	100%	62.65%

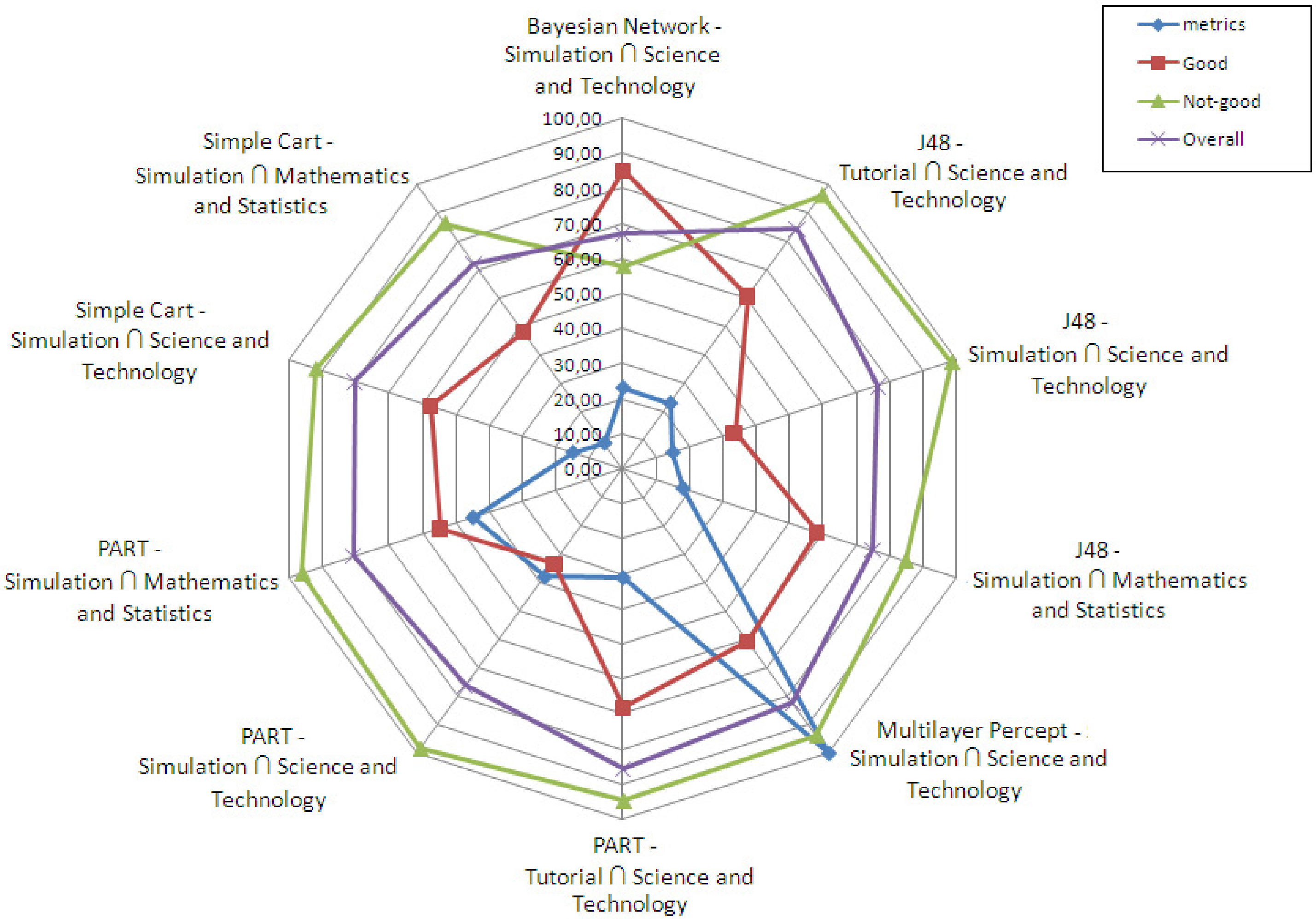


Figure 16. Radar graph for the 10 best models for classifying resources between *good* and *not-good*

Table 35. Models for classifying resources among *good*, *average* and *poor* ordered by MAE

N	Classification Algorithm	*Subset*	Metrics used by the model*	MAE	K	Classification Accuracy			
						Good	Average	Poor	Overall
1	J48	Simulation ∩ Science and Technology	11	0.1	0.83	96.96%	84.00%	92.85%	89.69%
2	Simple Cart	Simulation ∩ Science and Technology	8	0.15	0.76	90.90%	86.00%	71.40%	85.57%
3	PART	Simulation ∩ Science and Technology	8	0.16	0.74	97.00%	72.00%	92.9%	83.51%
4	PART	Simulation ∩ Mathematics and Statistics	5	0.23	0.55	77.4%	87.0%	0%	77.11%
5	Simple Cart	Tutorial ∩ Science and Technology	5	0.24	0.64	82.1%	83.3%	57.9%	77.11%
6	Multilayer Percept.	Simulation ∩ Science and Technology	13	0.26	0.53	60.60%	92.00%	42.90%	74.23%
7	J48	Simulation ∩ Mathematics and Statistics	4	0.26	0.47	64.5%	89.1%	0%	73.49%
8	Bayesian Network	Simulation ∩ Science and Technology	5	0.30	0.41	60.60%	48.00%	100%	59.79%
9	Simple Cart	Simulation ∩ Mathematics and Statistics	1	0.32	0.32	48.4%	87.0%	0%	66.26%
10	Multilayer Percept.	Simulation ∩ Mathematics and Statistics	11	0.34	0.13	16.1%	97.8%	0%	60.24%
11	PART	Tutorial ∩ Science and Technology	5	0.35	0.25	0%	100%	52.6%	55.42%
12	Bayesian Network	Simulation ∩ Mathematics and Statistics	0	0.37	0	0%	100%	0%	55.42%
13	J48	Tutorial ∩ Science and Technology	2	0.37	0.21	0%	97.2%	47.4%	53.01%
14	Multilayer Percept.	Tutorial ∩ Science and Technology	13	0.38	0.20	10.7%	86.1%	47.4%	51.81%
15	Bayesian Network	Tutorial ∩ Science and Technology	0	0.43	0	0%	100%	0%	43.37%

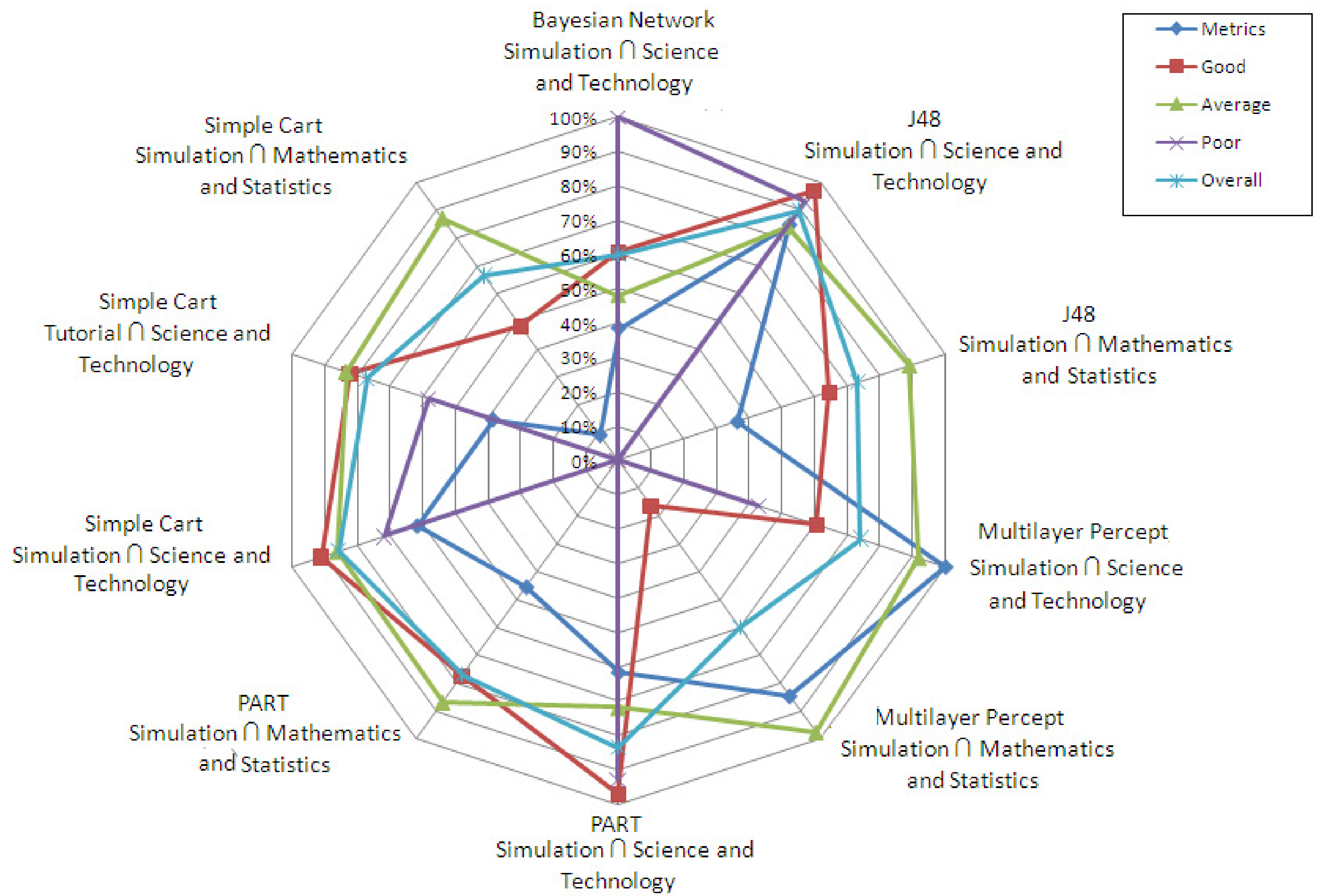


Figure 17. Radar graph for the 10 best models for classifying resources among *good*, *average* and *poor*

6.3 Examples of the Model Application

As mentioned on Chapter 1, the models for automated quality assessment are intended to be used for a preliminary analysis of the resources in order to relieve manual work. In this section we briefly describe two possible scenarios for the utilization of the models (once they are properly implemented inside the repository). Each one of the scenarios involves a different actor and takes place in a particular stage of the LO life-cycle. Table 36 summarizes the main aspects of each scenario that are presented below.

6.3.1 Scenario One – Improving Learning Objects Design

Models and statistical profiles of the learning resources may contain useful information for LO creators who wish to have a preliminary assessment of their materials in order to improve them. During the stage of **obtaining** (or creating) the LO, the author could provide to the tool, information related to the target discipline and the type of his learning resource, and have access to a preliminary evaluation of the resource. For such scenario it would be essential to use models that are able to explain the reasoning behind the resultant classification, such as the ones represented by decision trees and if-then-rules. Once preliminary assessment takes place, the creator could consult the model (and the statistical profiles of the resources) to understand which intrinsic metrics of his resource are influencing the classification. For this scenario, models involving a great number of quality indicators would certainly enrich the results from the perspective of the creator.

As most of the models present very different accuracy percentages depending on the classes used in the dataset, the different categories of discipline and material types, it would also be necessary to select the most suitable model taking into account such aspect. A model which is good for the classification of *good* resources and bad for the classification of *not-good* resources has great chances of classifying as *good* a resource that is *not-good*. Such situation would mislead the creator to believe that his resource is already of a high quality level, thus missing the opportunity to evolve it and improve it before publication. On the other hand, the classification of a *good* resource as *not-good* would probably make the creator waste time on searching for problems that do not necessarily exist.

Considering that several good models can exist (some most suitable for classifying *good* resources and some for classifying *not-good* resources), it is also possible to offer to the creator the possibility of choosing the ones he/she wishes to use based on what he/she considers most suitable for his/her needs.

6.3.2 Scenario Two – Providing Hidden Internal Ratings

It is known that LORs normally use evaluative information to rank resources during the process of search and retrieval. However, the amount of resources inside LORs increases more rapidly than the number of contributions given by the community of users and experts. Because of that, many LOs that do not have any quality evaluation receive bad rank positions even if they are of high-quality, thus remaining unused (or unseen) inside the repository until someone decides to evaluate it.

The models developed here could be used to provide internal ratings for those LOs still not evaluated, thus helping the repository in the stage of **offering** resources. Resources recently added to the repository would be highly benefited by such model since that they hardly receive any assessment just after their inclusion. Once the resource finally receives a formal evaluation from the community of the repository, the initial implicit rating given by the model could be disregarded. Moreover, both evaluations (implicit and real ratings) could be contrasted in order to observe possible correlations between them, allowing the evaluation of the model usefulness.

As the models will be used inside repository and the classifications will serve just as input information for searching mechanisms, it is not necessarily required that the models provide explanations about their reasoning. Models constituted of neural networks, mathematical functions, or Bayesian networks could perfectly be used in this scenario. Moreover, different models or ratings given by different models could be combined in order to form a single and unified rating.

Finally, the accuracies of the models should also be taken into consideration when selecting the most suitable one. The less damaging situation seems to occur when the model classify as *not-good* a *good* material. In this case, *good* materials would just remain hidden in the repository, i.e., in bad ranked positions (a similar situation to the one of not using the model). On the other hand, if the model classifies as *good* a resource that is *not-good*, it is most likely that this resource will be put at a higher rank position, thus increasing its chances of being accessed by the users. This would mislead the user towards the selection of a “not-so-good” quality resource, and it would probably put in discredit the ranking mechanism.

Table 36. Summary of aspects for each scenario of models application

Aspect	**Scenario 1** **Improving the Learning Object Design**	**Scenario 2** **Providing Hidden Internal Ratings**
Main actor	LO Creator (Author)	Repository
Life-Cycle stage	Obtaining (or Creating)	Offering
Suitable models	Those which are able to explain the reasoning behind the results of the classification, e.g. decision trees and if-then-rules	There is no need to explain the classification made. It is possible to use models represented by neural networks, Bayesian networks and mathematical functions.
Purpose and Advantages	To offer a preliminary assessment of quality so that the author can improve the resource	To provide internal ratings for non-rated learning objects already stored in the repository in order to have more evaluative information during the search and retrieval (ranking) process.
Problems due to misclassification of *good* resources	The creator would probably waste some time trying to improve a resource that is already good.	The resource would remain “invisible” inside the repository, i.e., a bad ranking position after searching and retrieval processes.
Problems due to misclassification of *not-good* resources	The creator would believe that his resource is already good and would miss the opportunity to improve it.	The repository would increase the chances of a not-good resource be selected by the user. The ranking mechanism could fall into discredit.
Alternative usage	It is possible to offer several different models so that the creator could compare the indicators of quality used, and the results presented by each model.	It is also possible to combine different models in order to form a single and unified rating.

Chapter 7 Conclusions

On this dissertation we approached the problem of testing a methodology for the development of models able to automatically classify learning objects according to their quality and based on their intrinsic features. For that, we selected MERLOT as the target repository, collected a data sample from it, and described the resources of the repository according to their categories of disciplines, types of materials and the mechanisms of quality assurance used by the repository. Moreover, we developed statistical profiles of highly-rated learning objects in order to discover which were the intrinsic features of the resources associated to quality and that could be used to generate models for automated quality assessment. Finally, we used Linear Discrimant Analysis and Data Mining Classification Algorithms to generate and test models automatically classify the quality of the resources available on MERLOT. During these experiments we were able to accomplish the goals we defined on Section 3.3 as presented next.

Goal 1 - Determine how the different materials inside a repository are associated to quality.

Chapter 4 described the types of material existing in MERLOT from different perspectives. It was possible to see that material types in the repository have grown differently over the years, and that the most significative growth came from the *Reference Material* type, followed by *Tutorial*, *Collection* and *Lecture/Presentation*. Chapter 4 also shown that materials are rated differently depending on their type and that the three best rated types of materials in both groups of raters (peer-reviewers and users) are the same: *Case Study*, *Drill and Practice* and *Collection* (with small differences in the order among them). Another important observation is that different types of materials tend to concentrate in certain categories of discipline, indicating some sort of suitability of those materials types for some specific subject areas. This is the case, for instance, of *Simulation* material type for the disciplines of *Science and Technology* and *Mathematics and Statistics*, and *Collection* and *Reference* material types for the discipline of *Humanities*. At last, the study has shown the average number of personal collections associated to the different material types. Here, it was possible to observe that *Case Study*, the type of resource with the best ratings, is also the one with the highest number of personal collections, followed by *Lecture/Presentation*, *Animation* and *Collection*. These findings helped us to better understand users' preferences for different types of material in the repository, and can serve as initial material for those developers of learning resources who are looking for the best format and pedagogical strategy for their materials.

Another important discovery of the dissertation is the confirmation of preliminary findings of previous work by García-Barriocanal & Sicilia (2009), which revealed the inclusion of learning objects in the bookmarks collections of the users as an important metric associated to quality. Such finding can be explained or justified because, since MERLOT ranks their materials (mainly) according to the ratings given by users and peer-reviewers, it is reasonable to assert that users which enter in MERLOT have a higher probability to select those resources with higher ratings. Therefore, these selected materials will also have a high probability of being added to personal collections.

Goal 2 - Determine if the different groups of evaluators inside repositories have the same impressions about the quality of learning objects.

Both communities of evaluators in MERLOT (users and peer-reviewers) are communicating different views regarding the quality of the learning objects refereed in the repository. Even though we have found associations between the users' ratings and the peer-reviewers' ratings in some disciplines, such associations are relatively weak and cannot confirm that users and experts agree about the quality of the evaluated learning resources. This reinforces the idea that peer-review and public-review approaches can be adopted in learning objects repositories as complementary strategies of evaluation that can both serve for the assurance and the establishment of quality parameters for further recommendation of materials. Most important, the pursuit of models for automated quality assessment must take into consideration these two perspectives. This finding has leaded us to conduct our experiments for both groups in separate. Future work can try to combine (integrate) into one single formula these different quality indicators, in a similar way of what has been proposed by Sanz-Rodríguez, Dodero & Sánchez-Alonso (2010b).

Goal 3 - Determine if it is possible to create statistical profiles of highly-rated learning objects based on their intrinsic features.

According to what we discovered, yes. In Chapter 5 we found that the tested metrics present different profiles and tendencies between *good* and *not-good* materials depending on the category of discipline and the type of material to which a resource belongs. In addition, the experiment where we intersected the discipline of *Science and Technology* and the *Simulation* material type (Section 5.3.3) has demonstrated that most of the metrics which presented different profiles in the original sets preserved their tendencies and remained associated to quality in the intersected set, but other metrics that were not correlated in the original sets, presented association to the ratings in this second moment. All these findings indicate that the pursuit for automated models for the quality evaluation of learning objects must consider the development of rated learning object profiles taking into account the intersection of the categories of disciplines and material types, as well as the distinct groups of raters (see Figure 18).

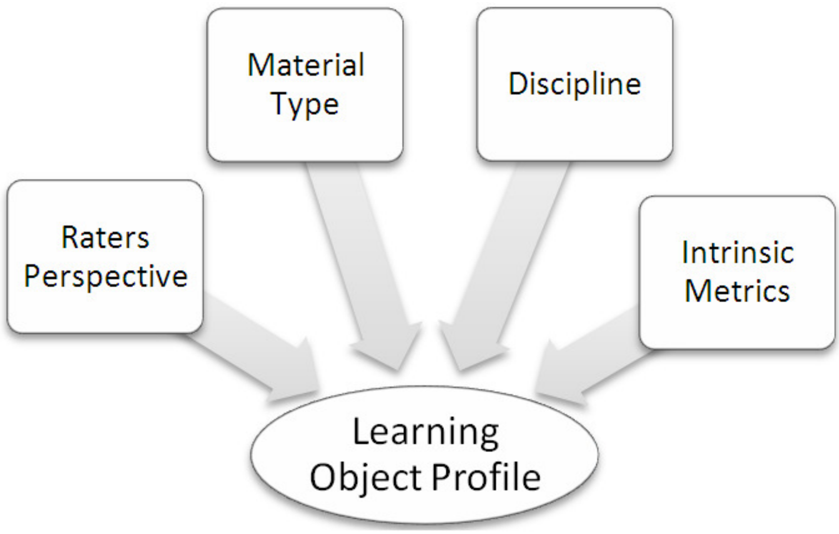


Figure 18. Aspects involved in the search for rated learning objects profiles

Goal 4 - Determine if it is feasible to generate models for automated quality assessment of learning objects inside repositories based on the intrinsic features of the resources.

According to our findings, it is. The results found during the dissertation initially point out for the possibility of creating models for automated quality assessment of learning objects inside repositories. However, the accuracy of the models will depend on the specific method used to generate them, the specific subsets to which they are being generated for, and the classes of quality included in the dataset. Moreover, it is important to highlight that most of the models tend to exclude some of the metrics identified as possible indicators of quality in the learning object profiles. Finally, the feasibility of the models will also depend on the specific scenarios of usage where automated quality assessment will take place.

7.1 Limitations and Future Work

Despite of all the achievements, our present work has several limitations. First, we are assuming that “poor” resources are those that received low ratings, but perhaps, rating a resource is an endorsement of quality by itself if we consider that only good resources are worth the effort of reviewing. In such case, the real “poor” category may be lying in non-rated objects (which were discarded in our study). However, the absence of ratings can also be attributed to other causes (e.g. objects that have been contributed recently are still not rated), which makes that analysis difficult. Another limitation is the uneven distribution of ratings among objects. While this is not an issue for peer reviewers (as it is uncommon that an object has more than a couple of them), the distribution shows a long tail of objects with few reviews and some of them having much more. Figure 19 shows the distribution of number of comments related to the rank (being the lower rank the object with more comments) for one of the categories. Another limitation is the heterogeneity of the format of learning objects, as some are simpler, low granularity elements, while others are actually link collections. This limitation might be overcome dividing the collections of objects using the granularity metadata element, if this metadata element would be available and accurate for a significant number of objects. But again, this would imply the availability of such metadata information.

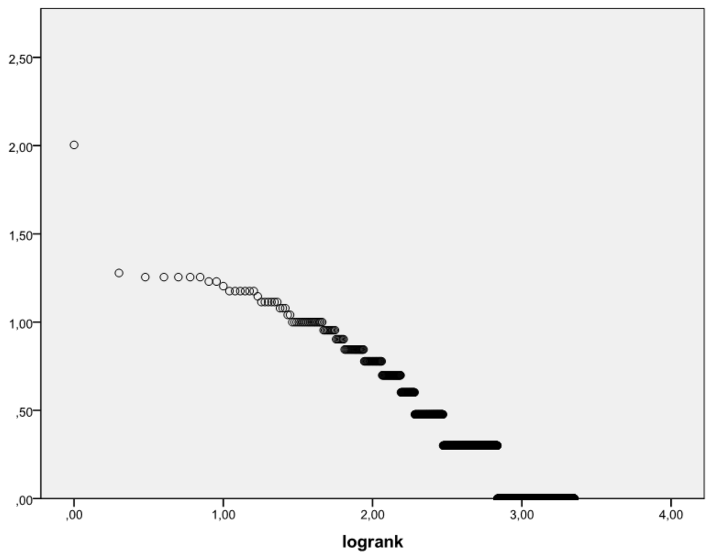


Figure 19. Scatter plot of rank of learning objects and number of comments in log scale for the Science and Technology category, taken from García-Barriocanal & Sicilia (2009)

Moreover, the crawler does not make any distinction among the resources that are web pages from those that are not, computing the metrics equally for all different possible learning resources scenarios. This is an important limitation, because even though learning resources referenced in MERLOT are accessed by a web page, some of them are just scripts, applets, or flash animations, for instance. In such cases, the crawler will disregard the content available inside the "real" resource, and it will consider the web page where the learning resource is embedded as the learning resource itself, computing the resource as one of the measures that are being collected, such as, for instance, the number of scripts or the number of applets. This is a difficult situation to solve since the information contained inside such resources formats is often closed and inaccessible.

The references for the establishment of learning objects quality in this study are the ratings given by the community of users and peer-reviewers of the MERLOT repository. This means that, for the context of this study, resources are considered to be of good quality if they are highly-rated by users and peer-reviewers of MERLOT.

Having that in mind, it is important to consider our quality indicators as indicators of the preferences that users and peer-reviewers show when selecting learning resources of a certain type and category of discipline. In other words, when we observe a correlation between some metric and the ratings of a given set of learning objects, we assume that this association can be an indicator of preference of that feature by the group who rated the resources, and therefore can be further tested in order to predict the preferences of this group in the future. Regarding this, it is important to highlight that the associations between the metrics and the ratings of learning resources we have encountered do not guarantee that these metrics are predictors of quality of learning resources, since correlation does not necessarily means causation (Holland, 1986). However, even though discovering correlations between the metrics and good resources does not imply we are encountering cause and consequence relationships, correlation is a necessary condition for causality, and once it is observed it can be considered as "*a good indication that some underlying causal relation exists*" (Lagnado, Waldmann, Hagmayer & Sloman, 2006). Edward R. Tufte (Tufte, 2003) has resumed this idea using the following expression: "*correlation is not causation but it sure is a hint*". In resume, finding the correlations is the very first step needed to identify variables that can be used in the development of such automated analysis tools, but controlled studies for these independent variables are still needed to answer some open issues, such as:

- Why some metrics are associated to quality, or which are the causes of such correlations?
- Are the associations representatives of cause and consequence relations regarding quality?
- Are the associations being generated by hidden causal events?
- What are the strengths of these relations?

At this point of our study we do not have the answers for such questions, even though we can try to venture providing some thoughts about the first one. For instance, it is claimed that students learn differently according the ways they receive information – visual (sights, pictures, diagrams), auditory (sounds, words) or kinesthetic (taste, smell), and that they have distinct mental process (and preferences) to convert the perceived information into knowledge (Tang & Austin, 2009) – some feel more comfortable with an active experimentation while other with reflective observations (Felder & Silverman, 1988; Kinshuk, Liu & Graf, 2009; Klasnja-Milicevic, Vesin, Ivanovic & Budimac, 2011). Considering that, positive correlations found between the *Number of Images* and highly rated learning resources in the disciplines of *Education*, *Mathematics and Statistics*, and *Science and Technology* could be an indicative that users coming from these disciplines tend to prefer receiving information by visual means rather than others. Moreover, positive correlations found between the *Number of Applets* in the *Business* discipline could indicate that users from this field work better with active experimentation rather than reflective observations.

Besides trying to answer these questions, future work will expand the present study to cover more metrics that are still under implementation in our crawler, such as, for instance: the number of colors and different font styles, the existence of adds, the number of redundant and broken links, and some readability measures (e.g. Gunning Fog index and Flesch-Kincaid grade level). In the case of multimedia content, it is still possible to acquire more advanced features, such as color histograms and thickness of lines.

We also intend to work with a full sample of MERLOT repository; however, in order to do so, we will need to restrain even more the number of pages that the crawler computes for each material[5]. Moreover, it is still possible to test the development of profiles considering other information about the resources, such as, for instance, the target audience, and the technical format of the material.

Once significant profiles and models are ultimately found, the next step would be that of constructing an analysis tool that uses the models to get a priori assessment of the quality of learning objects. Such tool could be integrated with LORs and could

[5] In order to acquire the sample used in this study, the crawler kept running uninterruptedly for 2 full months.

be developed with an architecture similar to the one described by Ivory & Hearst (2002a) for the WebTango system (a system that automatically assess the quality of websites).

As the present work is context focused, whether or not these findings can be extrapolated for other repositories is still a subject for further investigation and research. In here, we rely on information (categories of discipline, types of materials, peer reviewers and users ratings) that are not (necessarily) available in other learning resources repositories. In the cases where some of these information is not available, alternative ways of searching for LOs quality must be found in order to contrast with the metrics for the establishment of these profiles, such as, for instance, the use of ranking metrics (Ochoa & Duval, 2008) or other kinds of evaluative metadata (Vuorikari et al., 2008) available in such repositories. In fact, we have recently tested the methodology proposed on this thesis to develop models for automated quality assessment of learning resources available in Connexions (Cechinel, Sánchez-Alonso, Sicilia & Simões, 2011). On this study we used the number the endorsements (lenses) given for a learning resource as the quality parameter to create the classes of quality and the profiles of the resources. For that, we divided the resources into the two following groups of quality: 1) resources with just one endorsement and 2) resources with two or more endorsements. Although the developed models achieved very high overall accuracies, these good performances were limited to classify only one group of resources (those resources with just one lens). From our point of view, the poor performance of the models for classifying resources with two or more endorsements was caused by the small amount of these resources in the sample (just 3.33% of the sample). We concluded that was still needed to wait for the growth of endorsements in the repository in order to better evaluate the feasibility of our methodology for this specific context.

As we previously mentioned, we believe the deployment of such tool inside repositories would significantly improve the services provided by them in terms of searching, selecting and recommending good quality materials, and it would positively affect the intention of contributors to deliver new resources. However, it is difficult to precisely measure such impact without running specific controlled studies that would also have to take into account other factors, such as: 1) the type of repository (referatory, institutional repository, learning management system, etc) and whether it is driven by a community of volunteers or not – it is known, for instance, that the size of a repository and the number of objects published by their contributors are associated to the repository type (Ochoa & Duval, 2009); 2) the motivations and benefits that drive users to contribute to repositories – possible intrinsic motivations could be: academic recognition, bonus for staff/funding, create add value for sharing (incentives) (Margaryan, Currier, Littlejohn & Nicol, 2006), help others and influence others (Herlocker, Konstan, Terveen & Riedl, 2004); and 3) the current state of success of the repository and the attributes that contributed to its success.

7.2 Final comments

Learning Objects are rapidly growing over the globe. As the amount of resources increases, it becomes impossible to rely only on human work to assure the quality of this plethora of materials available inside the existing repositories. Even though several approaches tackle the problem of automated quality assessment of learning objects they normally depend on the metadata that describe the resources, or on measures of access and popularity of the materials. Such information is not always available or it can be sometimes inaccurate. This work is an attempt to complement the existing approaches by providing a methodology for automated quality assessment of resources that rely only on the intrinsic features of the resources, i.e., those measures that can be automatically extracted from the materials themselves. In this dissertation we analyzed intrinsic measures of learning objects refereed by the MERLOT repository, and developed statistical profiles for these materials taking their associated ratings as a baseline for quality comparison. These profiles were then used to generate models for automated quality assessment of resources.

Even though the study presents several limitations, it has achieved the initials goals established and has presented significant contributions that can further lead to the development of contextualized models for automated quality assessment of learning object inside repositories. The approach presented here does not mean to replace the existing traditional evaluation methods, but to complement them by providing a useful and inexpensive quality analysis before more time and effort-consuming evaluation is carried out. We hope the findings of this dissertation will serve as basis for future research of those who wish to improve the processes of quality assurance inside repositories through the offering of tools for automated quality assessment.

References

Balatsoukas, Panos, Morris, Anne, & O'Brien, Ann. (2008). Learning Objects Update: Review and Critical Approach to Content Aggregation. *Educational Technology & Society, 11*(2), 119-130.

Barritt, C., Lewis, D., & Wieseler, W. (1999). Cisco systems reusable information object strategy: Definition, creation overview, and guidelines. A CISCO Whitepaper.

Benos, Dale, Bashari, Edlira, Chaves, Jose, Gaggar, Amit, Kapoor, Niren, LaFrance, Martin, . . . Zotov, Alexander. (2007). The ups and downs of peer review. *Advances in Physiology Education, 31*(2), 145-152. doi: http://dx.doi.org/10.1152/advan.00104.2006

Bethard, Steven, Wetzer, Philipp, Butcher, Kirsten, Martin, James H., & Sumner, Tamara. (2009). *Automatically characterizing resource quality for educational digital libraries*. Paper presented at the Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries, Austin, TX, USA.

Blumenstock, Joshua E. (2008). *Size matters: word count as a measure of quality on wikipedia*. Paper presented at the Proceedings of the 17th international conference on World Wide Web, Beijing, China.

Brasher, A., & McAndrew, P. (2004). Human-Generated Learning Object Metadata. In R. Meersman, Tari Zahir & A. Corsaro (Eds.), *On the Move to Meaningful Internet Systems 2004: OTM 2004 Workshops* (Vol. 3292, pp. 723-730). Berlin: Springer.

Brosnan, Kevin. (2005). *Developing and sustaining a national learning-object sharing network: A social capital theory perspective.* Paper presented at the Proceedings of the ASCILITE 2005 Conference, Brisbane, Australia.

Cafolla, Ralph. (2006). Project MERLOT: Bringing Peer Review to Web-Based Educational Resources. *Journal of Technology and Teacher Education, 14*(2), 313-323.

Cardinaels, Kris. (2007). *A Dynamic Learning Object Life Cycle and its Implications for Automatic Metadata Generation.* Ph.D., Katholieke Universiteit Leuven, Leuven. Retrieved from https://lirias.kuleuven.be/bitstream/1979/882/2/

Cardinaels, Kris, Meire, M., & Duval, E. (2005). *Automating metadata generation: the simple indexing interface*. Paper presented at the 14th international Conference on World Wide Web, Chiba, Japan.

Cechinel, Cristian, & Camargo, Sandro da Silva. (2011). Localization of Learning Objects. In Esa Kujansuu, Luis A. Álvarez-González & Ion Mierlus-Mazilu (Eds.), *Analysis of State of the Art Solutions for Personalised Learning Support Workshop.* (pp. 12-21): IGUAL Project - Innovation for Equality in Latin-America University.

Cechinel, Cristian, Rebollo, Raquel, & Sánchez-Alonso, Salvador. (2012). Objetos de aprendizaje: Definición y caracterización In Francisco Alvarez Rodriguez & Jaime Muñoz Arteaga (Eds.), *Avances en objetos de aprendizaje: experiencias de redes colaboracion en México*. Aguascalientes: Editorial Universidad Autónoma de Aguascalientes e Universidad Veracruzana.

Cechinel, Cristian, & Sánchez-Alonso, Salvador. (2011). Analyzing Associations between the Different Ratings Dimensions of the MERLOT Repository. *Interdisciplinary Journal of E-Learning and Learning Objects 7*, 1-9.

Cechinel, Cristian, Sánchez-Alonso, Salvador, & García-Barriocanal, Elena. (2011). Statistical profiles of highly-rated learning objects. *Computers & Education, 57*(1), 1255-1269. doi: 10.1016/j.compedu.2011.01.012

Cechinel, Cristian, Sánchez-Alonso, Salvador, & Sicilia, Miguel-Ángel. (2009). Empirical Analysis of Errors on Human-Generated Learning Objects Metadata. In Fabio Sartori, Miguel Ángel Sicilia & Nikos Manouselis (Eds.), *Metadata and Semantic Research* (Vol. 46, pp. 60-70): Springer Berlin Heidelberg.

Cechinel, Cristian, Sánchez-Alonso, Salvador, & Sicilia, Miguel-Ángel. (2010). *Exploratory analysis of the correlations between peer-reviewers and users ratings on MERLOT repository*. Paper presented at the II Conferencia conjunta Iberoamericano sobre Tecnologías del Aprendizaje (CCITA 2010), Cadiz, Spain.

Cechinel, Cristian, Sánchez-Alonso, Salvador, Sicilia, Miguel-Ángel, & Amador, César Velazquez. (2011). Evaluating Models for Automated Quality Assessment of Learning Objects inside Repositories *6th Latin American Conference on Learning Objects and Technology Enhanced Learning – LACLO 2011* Montevideo - Uruguay.

Cechinel, Cristian, Sánchez-Alonso, Salvador, Sicilia, Miguel-Ángel, & Mattos, Merisandra Côrtes. (2010a). Descriptive Analysis of Learning Object Material Types in MERLOT. In Salvador Sánchez-Alonso & Ioannis N. Athanasiadis (Eds.), *Metadata and Semantic Research* (Vol. 108, pp. 331-341): Springer Berlin Heidelberg.

Cechinel, Cristian, Sánchez-Alonso, Salvador, Sicilia, Miguel-Ángel, & Mattos, Merisandra Côrtes. (2010b). *Estudo Exploratório sobre a Utilização de Algoritmos de Mineração de Dados na Predição e Caracterização da Qualidade de Objetos de Aprendizagem no Repositório MERLOT.* Paper presented at the XXI Simpósio Brasileiro de Informática na Educação (SBIE 2010) João Pessoa, Paraíba.

Cechinel, Cristian, Sánchez-Alonso, Salvador, Sicilia, Miguel-Ángel, & Simões, Priscyla Waleska Targino de Azevedo. (2011). Exploring the Development of Endorsed Learning Resources Profiles in the Connexions Repository. In Elena García-Barriocanal, Zeynel Cebeci, Mehmet C. Okur & Aydın Öztürk (Eds.), *Metadata and Semantic Research* (Vol. 240, pp. 12-21): Springer Berlin Heidelberg.

Churchill, Daniel. (2007). Towards a useful classification of learning objects. *Educational Technology Research and Development, 55*(5), 479-497. doi: 10.1007/s11423-006-9000-y

Cichosz, Paweł. (2011). Assessing the quality of classification models: Performance measures and evaluation procedures. *Central European Journal of Engineering, 1*(2), 132-158. doi: 10.2478/s13531-011-0022-9

Collis, Betty, & Strijker, Allard. (2004). Technology and Human Issues in Reusing Learning Objects. *Journal of Interactive Media in Education; May 2004: JIME Special Issue on the Educational Semantic Web, 2004*(1).

Currier, Sarah, Barton, Jane, O'Beirne, Rónán, & Ryan, Ben. (2004). Quality assurance for digital learning object repositories: issues for the metadata creation process. *ALTJ: Research in Learning Technology, 12*(1), 5-20 doi: citeulike-article-id:539808

Custard, Myra, & Sumner, Tamara. (2005). Using Machine Learning to Support Quality Judgments. *D-Lib Magazine, 11*(10).

Dalziel, J. (2002). *Reflections on the COLIS (Collaborative Online Learning and Information Systems) Demonstrator Project and the" Learning Object Lifecycle"*. Paper presented at the Winds of change in the sea of learning. Proceedings of the 19th Annual Conference of ASCILITE, Auckland, New Zealand. http://www.ascilite.org.au/conferences/auckland02/proceedings/papers/207.pdf.

Díaz, Paloma, Sicilia, Miguel-Ángel, & Aedo, Ignacio. (2002). *Evaluation of Hypermedia Educational Systems: Criteria and Imperfect Measures.* Paper presented at the International Conference on Computers in Education (ICCE'02), Auckland, New Zealand.

Downes, S. (2003). Design and reusability of learning objects in an academic context: A new economy of education. *USDLA Journal, 17*(1), 3-22.

Duncan, S. M. (2009). *Patterns of Learning Object Reuse in the Connexions Repository.* Retrieved from http://www.archive.org/details/PatternsOfLearningObjectReuseInTheConnexionsRepository

Duval, Erik. (2006). LearnRank: Towards a real quality measure for learning. In Ulf-Daniel Ehlers & Jan Martin Pawlowski (Eds.), *Handbook on Quality and Standardisation in E-Learning* (pp. 457-463): Springer Berlin Heidelberg.

Farance, F. (2003). IEEE LOM standard not yet ready for 'prime time'. *Learning Technology, 5*(1).

Felder, Richard, & Silverman, Linda. (1988). Learning and Teaching Styles in Engineering Education. *Engineering Education, 78*(7), 674-681. doi: citeulike-article-id:6101172

Fernández, Miguel Refusta. (2011). *Métricas en repositorios de objetos de aprendizaje.* Bachelor Final Project, University of Alcalá, Alcalá de Henares, Spain.

Friesen, N. (2001). What are educational objects? *Interactive Learning Environments, 9*(3), 219–230.

García-Barriocanal, Elena, & Sicilia, Miguel-Ángel. (2009). Preliminary Explorations on the Statistical Profiles of Highly-Rated Learning Objects. In Fabio Sartori, Miguel Ángel Sicilia & Nikos Manouselis (Eds.), *Metadata and Semantic Research* (Vol. 46, pp. 108-117): Springer Berlin Heidelberg.

Geser, Guntram. (2007). Open educational practices and resources: OLCOS Roadmap 2012. *Revista de Universidad y Sociedad del Conocimiento, 4*(1), 1-9.

Goldschmidt, R., & Passos, E. (2005). *Data Mining: um guia prático*. Rio de Janeiro: Elsevier.

González-Videgaray, MariCarmen, Hernández-Zamora, Gregorio, & del-Río-Martínez, Jesús H. (2009). Learning objects in theory and practice: A vision from Mexican University teachers. *Computers & Education, 53*(4), 1330-1338. doi: 10.1016/j.compedu.2009.06.015

Gunn, Cathy, Woodgate, Sheila, & O'Grady, Winnie. (2005). Repurposing learning objects: A sustainable alternative. *Association for Learning Technology Journal. Research in learning technology, 13*(3), 189-200.

Hall, Mark, Frank, Eibe, Holmes, Geoffrey, Pfahringer, Bernhard, Reutemann, Peter, & Witten, Ian H. (2009). The WEKA data mining software: an update. *SIGKDD Explor. Newsl., 11*(1), 10-18. doi: 10.1145/1656274.1656278

Han, Peng, Kortemeyer, Gerd, Krämer, Bernd, & von Prümmer, Christine. (2008). Exposure and Support of Latent Social Networks among Learning Object Repository Users. *Journal of the Universal Computer Science, 14*(10), 1717-1738. doi: citeulike-article-id:3558788

Harnad, S. (2000). The invisible hand of peer review. *Exploit Interactive, April*(5). Retrieved from http://www.exploit-lib.org/issue5/peer-review

Hart, J., & Albrecht, B. (2004). Instructional repositories and referatories. *Research Bulletin. EDUCAUSE Center for Applied Research, 2004.*(5), 1-12.

Heery, R., & Anderson, S. (2005). Digital repositories review. Project Report. Joint Information Systems Committee: UKOLN.

Herlocker, Jonathan L., Konstan, Joseph A., Terveen, Loren G., & Riedl, John T. (2004). Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst., 22*(1), 5-53. doi: 10.1145/963770.963772

Hodgins, H. W. (2000). The future of learning objects. In D. A. Wiley (Ed.), The Instructional Use of Learning Objects: Online Version. Retrieved from http://reusability.org/read/chapters/hodgins.doc.

Hodgins, H. W. (2004). The future of learning objects. *e-Technologies in Engineering Education: Learning Outcomes Providing Future Possibilities*, 11.

Holland, Paul. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association, 81*(396), 945-960.

IEEE-LOM. (2002). Learning Object Metadata Standard. *IEEE 1484.12.1-2002*. Available at: http://ltsc.ieee.org/wg12/files/LOM_1484_12_1_v1_Final_Draft.pdf. [Last Access 18th July 2011].

Ivory, M. Y., & Hearst, M. A. (2002a). Improving Web site design. *IEEE Internet Computing, 6*(2), 56-63.

Ivory, M. Y., & Hearst, M. A. (2002b). *Statistical profiles of highly-rated web sites*. Paper presented at the Proceedings of the SIGCHI conference on Human factors in computing systems: Changing our world, changing ourselves, Minneapolis, Minnesota, USA.

Kay, Robin, & Knaack, Liesel. (2009). Assessing learning, quality and engagement in learning objects: the Learning Object Evaluation Scale for Students (LOES-S). *Educational Technology Research and Development, 57*(2), 147-168. doi: 10.1007/s11423-008-9094-5

Kelty, C. M., Burrus, C. S., & Baraniuk, R. G. (2008). Peer Review Anew: Three Principles and a Case Study in Postpublication Quality Assurance. *Proceedings of the IEEE, 96*(6), 1000-1011.

Kinshuk, Liu, Tzu-Chien, & Graf, Sabine. (2009). Coping with mismatched courses: students' behaviour and performance in courses mismatched to their learning styles. *Educational Technology Research and Development, 57*(6), 739-752. doi: 10.1007/s11423-009-9116-y

Klasnja-Milicevic, Aleksandra, Vesin, Boban, Ivanovic, Mirjana, & Budimac, Zoran. (2011). E-Learning personalization based on hybrid recommendation strategy and learning style identification. *Computers & Education, 56*(3), 885-899. doi: DOI: 10.1016/j.compedu.2010.11.001

L'Allier, J.J. (1997). Frame of reference: NETg's map to the products, their structure and core beliefs. *NetG. Retrieved February, 21*, 2007.

Lagnado, David A., Waldmann, Michael R., Hagmayer, York, & Sloman, Steven A. (2006). Beyond covariation: Cues to causal structure. In Alison Gopnik & Laura Schulz (Eds.), *Causal Learning: Psychology, Philosophy, and Computation*: Oxfor University Press.

Leacock, Tracey L., & Nesbit, John C. (2007). A Framework for Evaluating the Quality of Multimedia Learning Resources. *Educational Technology & Society, 10*(2), 44-59.

Lytras, Miltiadis D. , & Sicilia, Miguel-Ángel (2007). Where is the value in metadata. *International Journal of Metadata, Semantics and Ontologies, 2*(4), 235-241. doi: 10.1504/ijmso.2007.019442

Margaryan, Anoush, Currier, Sarah, Littlejohn, Allison, & Nicol, David. (2006). CD-LOR Deliverable 1: Report on Learning Communities and Repositories *Community Dimensions of Learning Object Repositories*: University of Dundee.

McGreal, R. (2004). Learning objects: A practical definition. *International Journal of Instructional Technology and Distance Learning, 1*(9), 21-32.

McGreal, R. (2008). A Typology of Learning Object Repositories. In Heimo H. Adelsberger, Kinshuk, Jan M. Pawlowski & Demetrios G. Sampson (Eds.), *Handbook on Information Technologies for Education and Training* (pp. 5-28): Springer Berlin Heidelberg.

Mendes, Emilia, Hall, Wendy, & Harrison, Rachel. (1998). Applying Metrics to the Evaluation of Educational Hypermedia Applications. *Journal of Universal Computer Science, 4*(4), 382-403. doi: 10.3217/jucs-004-04-0382

MERLOT. (2011). MERLOT home page. Retrieved 25th July, 2011, from http://www.merlot.org/

Merrill, M. D. (1999). Instructional Transaction Theory (ITT): Instructional Design Based on Knowledge Objects. In C. M. Reigeluth (Ed.), *Instructional Design Theories and Models: A New Paradigm of Instructional Technology* (pp. 397-424): Lawrence Erlbaum Associates.

Meyer, Marek, Hannappel, Alexander, Rensing, Christoph, & Steinmetz, Ralf. (2007). *Automatic classification of didactic functions of e-learning resources*. Paper presented at the Proceedings of the 15th international conference on Multimedia, Augsburg, Germany.

Miyagawa, Shigeru. (2010). MIT OpenCourseWare: A Decade of Global Benefit. *MIT Faculty Newsletter, XXIII*.

Monge, Sergio, Ovelar, Ramón, & Azpeitia, Iker. (2008). Repository 2.0: Social Dynamics to Support Community Building in Learning Object Repositories. *Interdisciplinary Journal of E-Learning and Learning Objects, 4*, 191-204.

Mortimer, L. (2002). (Learning) objects of desire: Promise and practicality. *Learning Circuits. Available at http://www.astd.org/LC/2002/0402_mortimer.htm [Last Access 18th August 2010]*.

Nesbit, John C., Belfer, Karen, & Leacock, Tracey. (2003). Learning object review instrument (LORI). E-learning research and assessment network. Retrieved from http://www.elera.net/eLera/Home/Articles/LORI%20manual.

Nesbit, John C., Belfer, Karen, & Vargo, John. (2002). A Convergent Participation Model for Evaluation of Learning Objects. *Canadian Journal of Learning and Technology, 28*(3).

Nesbit, John C., & Li, Jerry Z. (2004). *Web-based tools for learning object evaluation*. Paper presented at the International Conference on Education and Information Systems: Technologies and Applications, Orlando, Florida.

Ochoa, Xavier. (2008). *Learnometrics: Metrics for learning objects*. Ph.D. Doctoral Thesis, KatholiekeUniversity Leuven, Leuven. Retrieved from https://lirias.kuleuven.be/bitstream/1979/1891/2/ThesisFinal.pdf (978-90-5682-982-7)

Ochoa, Xavier. (2010). Connexions: a Social and Successful Anomaly among Learning Object Repositories. *Journal of Emerging Technologies in Web Intelligence, 2*(1).

Ochoa, Xavier, & Duval, Erik. (2008). Relevance Ranking Metrics for Learning Objects. *Learning Technologies, IEEE Transactions on, 1*(1), 34-48. doi: http://dx.doi.org/10.1109/TLT.2008.1

Ochoa, Xavier, & Duval, Erik. (2009). Quantitative Analysis of Learning Object Repositories. *Learning Technologies, IEEE Transactions on, 2*(3), 226-238.

Peddibhotla, Naren, & Subramani, Mani. (2007). Contributing to Public Document Repositories: A Critical Mass Theory Perspective. *Organization Studies, 28*(3), 327-346.

Polsani, Pithamber R. (2003). Use and Abuse of Reusable Learning Objects. *Journal of Digital Information, 3*(4).

Q4R. (2007). Quality for Reuse Best Practices - Wiki. Merlot best practices. Retrieved 9th August 2011, from http://q4r.licef.ca/q4r/index.php?title=MERLOT_best_practices&printable=yes

Resnick, Paul, & Varian, Hal R. (1997). Recommender systems. *Commun. ACM, 40*(3), 56-58. doi: 10.1145/245108.245121

Sánchez-Alonso, Salvador, Rodriguez, Daniel, Abián, Alberto, Arroyo, Sinuhé, & Sicilia, Miguel-Ángel. (2008). *On Integrating Semantic Learning Object Repositories in the Open Knowledge Initiative (OKI) Architecture*. Paper presented at the Workshop Next Generation e-Learning: from Web 2.0 to Web 3.0, part of International Conference on Interactive Collaborative Learning (ICL2008). Villach, Austria.

Sanz-Rodriguez, Javier, Dodero, Juan, & Sánchez-Alonso, Salvador. (2010a). Metrics-based evaluation of learning object reusability. *Software Quality Journal*, 1-20. doi: 10.1007/s11219-010-9108-5

Sanz-Rodriguez, Javier, Dodero, Juan, & Sánchez-Alonso, Salvador. (2010b). Ranking Learning Objects through Integration of Different Quality Indicators. *IEEE Transactions on Learning Technologies, 3*(4), 358 - 363. doi: 10.1109/TLT.2010.23

Saum, R. R. . (2007). An Abridged History of Learning Objects. In Pamela Taylor Northrup (Ed.), *Learning Objects for Instruction: Design and Evaluation* (pp. 1-16). New York: Idea Group Publishing

Sicilia, Miguel-Ángel, & García-Barriocanal, Elena. (2003). On the Concepts of Usability and Reusability of Learning Objects. [Learning objects; reusability; usability evaluation; learning technology standards]. *The International Review of Research in Open and Distance Learning, 4*(2).

Sicilia, Miguel-Ángel, García-Barriocanal, Elena, Pages, Carmen, Martinez, Jose J., & Gutierrez, Jose M. (2005). Complete metadata records in learning object repositories: some evidence and requirements. *International Journal of Learning Technology (IJLT), 1*(4), 411-424. doi: http://dx.doi.org/10.1504/IJLT.2005.007152

Sicilia, Miguel-Ángel, Sánchez-Alonso, S., García-Barriocanal, Elena, & Rodriguez-Garcia, D. (2009, 4-6 Nov. 2009). *Exploring Structural Prestige in Learning Object Repositories: Some Insights from Examining References in MERLOT.* Paper presented at the International Conference on Intelligent Networking and Collaborative Systems, 2009. INCOS '09. .

Soto, J., Sánchez-Alonso, Salvador, & Sicilia, Miguel-Ángel. (2005). *Flexibility in semantic learning object repositories.* Paper presented at the Third International Conference on Multimedia and Information and Communication Technologies in Education (m-ICTE 2005), Cáceres, Spain.

Stefani, Antonia, Vassiliadis, Bill, & Xenos, Michalis. (2006). On the quality assessment of advanced e-learning services. *Interactive Technology and Smart Education, 3*(3), 237-250.

Strijker, Allard. (2004). *Reuse of Learning Objects in Context: Human and Technical Aspects.* Ph.D. Doctoral Thesis, University of Twente, Leuven. Retrieved from http://doc.utwente.nl/41728/

Stvilia, B., Twidale, M. B., Smith, L. C., & Gasser, L. (2005). *Assessing information quality of a community-based encyclopedia.* Paper presented at the Proceedings of the International Conference on Information Quality - ICIQ 2005.

Taylor, James C (2007). Open Courseware Futures: Creating a Parallel Universe. *e-Journal of Instructional Science and Technology (e-JIST), 10* (1), 1-9.

Thompson, K., & Yonekura, F. (2005). Practical guidelines for learning object granularity from one higher education setting. *Interdisciplinary Journal of Knowledge and Learning Objects, 1*(2005), 163-179.

Tufte, Edward R. (2003). *The Cognitive Style of Power Point*: Graphics Press.

Vargo, John, Nesbit, John C., Belfer, Karen, & Archambault, Anne. (2003). Learning Object Evaluation: Computer-Mediated Collaboration and Inter-Rater Reliability. *International Journal of Computers and Applications, 25*(3), 1-8.

Velázquez, César, Muñoz, Jaime, Álvarez, Francisco, Sicilia, Miguel-Ángel, Cechinel, Cristian, & Hernández, Yosly. (2010). *Estudio de los factores para el desarrollo y evaluación de objetos de aprendizaje multiculturales bajo la modalidad LOMA*. Paper presented at the Fifth Latin American Conference on Learning Objects and Technology Enhanced Learning – LACLO 2010, São Paulo.

Vuorikari, Riina, Manouselis, Nikos , & Duval, Erik. (2008). Using Metadata for Storing, Sharing and Reusing Evaluations for Social Recommendations: the Case of Learning Resources. *Social Information Retrieval Systems: Emerging Technologies and Applications for Searching the Web Effectively. Hershey, PA: Idea Group Publishing*, 87–107.

Wegner, D. M. . (1986). Transactive memory: A contemporary analysis of the group mind. In M. B. & G. G. R (Eds.), *Theories of Group Behavior* (pp. 185-205). New York: Springer-Verlag.

Wiley, David A. (1999). The post-LEGO learning object. Available at http://opencontent.org//docs/post-lego.pdf [Last Access 7th June 2010]

Wiley, David A. (2000). Connecting learning objects to instructional design theory: A definition, a metaphor, and a taxonomy. In D. A. Wiley (Ed.), The Instructional Use of Learning Objects: Online Version. Retrieved from http://reusability.org/read/chapters/wiley.doc.

Wiley, David A. (2007). Open education license draft. Retrieved from http://opencontent.org/blog/archives/355.

Wiley, David A. (2009). Impediments to Learning Object Reuse and Openness as a Potential Solution. *Revista Brasileira de Informática na Educação, 17,* 3-10.

Wiley, David A. (2010). Openness and Analytics: The Future of Learning Objects. Presentation at the *Fifth Latin American Conference on Learning Objects and Technology Enhanced Learning – LACLO 2010*. São Paulo.

Wiley, David A., Gibbons, A., & Recker, M. (2000). A reformulation of the issue of learning object granularity. Available at http://reusability.org/granularity.pdf [Last Access 1st September 2010].

Williams, David D. (2000). Evaluation of learning objects and instruction using learning objects. In David A. Wiley (Ed.), The Instructional Use of Learning Objects: Online Version. Retrieved from http://reusability.org/read/chapters/williams.doc.

Appendix A Statistical Profiles

Table A. 1 Significant Discriminators of the good category considering peer-reviews and users ratings (P/U) in Disciplines

Metric	*Arts*	*Business*	*Education*	*Humanities*	*Mathematics and Statistics*	*Science and Technology*	*Social Sciences*
Number of Links	N/N+	(N*)/N	(Y)/N	N/N	(Y)/Y	N*/Y	N/N
Number of Unique Links	N/N	(N*)/N	(Y)/N	N/N	Y/Y	N*/Y	N*/N
Number of Internal Links	N/Y	N*/N	Y/N*	N*/(N*)	N*/Y	N*/Y	N/N
Number of Unique Internal Links	N/(N+)	N*/N	Y/N*	N*/(N*)	(Y)/Y	N*/Y	N/N
Number of External Links	N/N	N*/N	N*N*	N*/(N*)	Y/Y	N*/Y	(Y)/N
Number of Unique External Links	N/N	N*/N	N*N*	N*/(N*)	(Y)/Y	N*/Y	(Y)/N
Number of Images	N/(N+)	N*/N	Y/N	N/N	(Y)/Y	(Y)/Y	(Y)/N
Size of the Page (in bytes)	N/(N+)	N/N	(Y)/N	N/N	Y/Y	N/(Y)	(N+)/N
Total Size of the Images (in bytes)	N/Y	N*/N	Y/(Y)	N*/N*	(Y)/(Y)	N*/Y	N*/N
Number of Scripts	N*/N	N*/N	N*/(N*)	N*/N*	N*/(Y)	Y/Y	N/N
Number of Applets	-/N*	N*/Y	-/-	N*/N*	N*/(Y)	(Y)/Y	N*/-
Number of Words	N/ (N+)	(Y)/N	N/N	N/N	N/Y	N*/N	N/N
Number of Words that are links	N/(Y)	(N*)/N	(N+)/N	N/N	(N*)/Y	N*/Y	N/N
Total Number of Pages	(N*)/N+	N*/N	(Y)/N*	N*/N*	N*/(Y)	N*/Y	N*/N*
Number of Files for downloading	N*/N*	N*/N*	N*/N*	(Y)/N*	N*/N*	N*/N*	N*/N*
Number of audio files	N*/N*	N*/-	(Y)/N*	N*/N*	-	-/-	-/N*
Number of video files	N*/-	N*/-	(Y)/(Y)	N*/N*	N*/N*	N*/Y	N*/ N*
Number of multimedia files	N*/N*	N*/-	(Y)/N*	N*/N*	N*/N*	N*/N*	N*/N*
Average Number of Unique Internal Links	(Y)/N	(Y)/N	Y/N*	N*/(N*)	N*/Y	N*/Y	N*/N
Average Number of Internal Links	N+/(N+)	(N*)/N	Y/N*	N*/(N*)	N*/Y	N*/Y	N/N
Average Number of Unique External Links	N/N	N*/N	N* /(Y)	N*/(N*)	N*/N*	(Y)/N*	Y/N
Average Number of External Links	N/N	N*/N	N*/N*	N*/(N*)	(Y)/N*	Y/(Y)	Y/N
Average Number of Unique Links	N/N	(N*)/N	(N*)/N	N/N	(Y)/N	N*/N*	Y/N
Average Number of Links	(N+)/N	(N*)/N	(Y)/N	N/N	(Y)/(Y)	N*/(Y)	N/N

Average Number of Words	N/N	(Y)/N	N/N	N/N	N/Y	N*/N	(N+)/Y
Average Number of Files for downloading	N*/N*	N*/N*	N*/N*	N*/N*	N*/N*	N*/N*	N*/N*
Average Number of Audio Files	N*/N*	-/-	(Y)/N*	N*/N*	-/-	-/-	-/-
Average Number of Video Files	N*/-	N*/-	(Y)/N*	N*/N*	N*/N*	N*/(Y)	N*/-
Average Number of Multimedia Files	N*/N*	N*/-	(Y)/N*	N*/N*	N*/N*	N*/N*	N*/-
Average Number of Applets	-/N*	N*/Y	-/-	N*/N*	N*/(Y)	(Y)/Y	N*/-
Average Number of Images	Y/N	N*/N	N/N	N/N	N*/N	Y/N*	Y/N
Average Size of the Pages	N/N	N/N	N/N	N/N	(Y)/Y	N/N	Y/N
Average Size of the Images	N/N	N*/N	(Y)/Y	N*/N*	N*/N	N*/Y	(Y)/N
Average Number of Scripts	N*/N	N*/N	N* /(N*)	N*/N*	N*/N*	Y/Y	N/(N+)
Number of Personal Collections	Y/N	Y/N	Y/N	Y/N	Y/Y	Y/Y	(N*)/Y
Frequency of significant (Y) metrics (P/U)	**3/3**	**4/2**	**19/4**	**2/0**	**13/21**	**9/23**	**9/2**

Note: N represents no significant difference for the median of the two samples, N* represents no significant difference for the medians but significant difference in distribution, N+ stands for significant difference between the distributions (Kolmogorov-Smirnov) but significant difference between medians (Mann-Whitney), and Y stands for both differences at the same time. The overall analysis was conducted for a 95% confidence level; information in parenthesis means the results are significant at the 90% level. A hyphen means that the analysis was impossible to perform because all fields have the same information.

Table A. 2 Tendency of the metrics for predicting good in the disciplines with significant differences for peer-reviews and users ratings (P/U)

Metric	*Arts*	*Business*	*Education*	*Humanities*	*Mathematic and Statistics*	*Science and Technology*	*Social Sciences*
Number of Links			↑/-		↑/↑	-/↑	
Number of Unique Links			↑/-		↑/↑	-/↑	
Number of Internal Links	-/↓		↑/-		-/↑	-/↑	
Number of Unique Internal Links			↑/-		↑/↑	-/↑	
Number of External Links					↑/↑	-/↑	↓/-
Number of Unique External Links					↑/↑	-/↑	↓/-
Number of Images			↑/-		↑/↑	↑/↑	↓/-
Size of the Page (in bytes)			↑/-		↑/↑	-/↑	
Total Size of the Images (in bytes)	-/↓		↑/↑		↑/↑	-/↑	
Number of Scripts					-/↑	↑/↑	
Number of Applets		-/↑			-/↓	↓/↓	
Number of Words		↓/-			-/↑		
Number of Words that are links	-/↓				-/↑	-/↑	
Total Number of Pages			↑/-		-/↑	-/↑	
Number of Files for downloading				↓/-			
Number of audio files			↑/-				
Number of video files			↑/↑			-/↑	
Number of multimedia files			↑/-				
Average Number of Unique Internal Links	↑/-	↑/-	↑/-		-/↑	-/↑	
Average Number of Internal Links			↑/-		-/↑	-/↑	
Average Number of Unique External Links			-/↑			↓/-	↓/-
Average Number of External Links					↑/-	↓/↑	↓/-
Average Number of Unique Links					↑/-		↓/-
Average Number of Links			↑/-		↑ /↑	-/↑	
Average Number of Words		↓/-			-/↑		-/↓
Average Number of Files for downloading							
Average Number of Audio Files			↑/-				
Average Number of Video Files			↑/-			-/↑	
Average Number of Multimedia Files			↑/-				
Average Number of Applets		-/↑			-/↓	↓/↓	
Average Number of Images	↑/-					↑/-	↓/-
Average Size of the Pages					↑/↑		↓/-
Average Size of the Images			↑/↑			-/↑	↓/-
Average Number of Scripts						↑/↑	
Number of Personal Collections	↑/-	↑/-	↑/-	↑/-	↑/↑	↑/↑	-/↑

Note: Here (↑) stands for a positive contribution, (↓) stands for negative contribution, and (-) means the metric did not presented significant difference for that threshold and its tendency was not evaluated.

Table A. 3. Significant Discriminators of the good category considering peer-reviews and users ratings (P/U) in material types

Metric	*Animation*	*Case Study*	*Collection*	*Drill*	*Lecture/ Presentation*	*Quiz*	*Reference Material*	*Simulation*	*Tutorial*
Number of Links	N*/(Y)	N/N	N/N	N/N	(Y)/N	N+/N	N/N	N*/N*	(N+)/N*
Number of Unique Links	N*/(N*)	N/N	N/N	N/N	Y/N	N/N	N/N	N*/N*	N/N*
Number of Internal Links	N*/N*	N/N	N*/N	N/N	(Y)/N	N+/N	N/N*	N*/N*	N*/N*
Number of Unique Internal Links	N*/N	N/N	N*/N	N/N	Y/N	(N+)/N	N/N*	N*/N*	N*/N*
Number of External Links	N*/Y	N/N	N*/N	N*/N	N*/N*	(Y) /N	N*/N*	N*/Y	N*/N*
Number of Unique External Links	N*/Y	N/N	N*/N	N*/N	N*/N*	(Y) /N	N*/N*	N*/(Y)	N*/N*
Number of Images Size	N*/(N*)	N/N	N/N	N/N	Y/N	N/N	N/N*	Y/N*	N*/N*
Size of the Page (in bytes)	N/Y	N/N	N/N	N/N	(N+)/N	N/N	N/(Y)	N/N	(Y)/N
Total Size of the Images (in bytes)	N*/(N*)	N/N	N*/(N*)	N/(Y)	N/(N*)	N/N	N*/N*	Y/(Y)	N*/N*
Number of Scripts	N*/Y	N*/N	N*/N	N*/(Y)	Y/N*	N/N	(Y)/N*	Y/N*	N*/N*
Number of Applets	N*/Y	-/-	N*/N*	-/-	N*/Y↑	-/-	-/N*	Y/Y	N*/N*
Number of Words	N/N	N/N	N*/N	N/N	N/N	N/N	N/Y	N/(N*)	(N+)/N
Number of Words that are links	N*/N	N/N	N/N	N/N	Y/N	N/N	N/(N*)	N*/N*	Y/N*
Total Number of Pages	N*/Y	N/N	N*/N*	N*/N*	Y/N*	(Y)/N	N*/N*	N*/N*	N*/N*
Number of Files for downloading	N*/N*	N*/(N*)	N*/N*	(Y)/N*	(Y)/N*	N*/-	N*/N*	N*/N*	N*/N*
Number of audio files	-/-	-/-	N*/-	-/-	N*/-	-/-	N*/N*	N*/-	N*/N*
Number of video files	N*/N*	-/-	(Y)/-	N*/-	N*/(Y)	-/-	-/N*	N*/N*	N*/N*
Number of multimedia files	N*/N*	-/-	N*/-	N*/-	N*/N*	-/-	N*/N*	N*/N*	N*/N*
Average Number of Unique Internal Links	N*/N	N/N	N*/N	N/N	N/N	N/N	N/N*	N*/N*	N*/N*
Average Number of Internal Links	N*/N	N/N	N*/N	N/N	N/N	N/N	N/N*	N*/N*	N*/N*
Average Number of Unique External Links	N*/N*	N/N	N*/N	N*/N	N*/N*	N*/N	N*/N*	Y/Y	N*/N*
Average Number of External Links	N*/N*	(N*)/N	N*/N	N*/N	N*/N*	N*/N	N*/N*	Y/Y	N*/N*
Average Number of Unique Links	N*/N	N/N	N/(N*)	N/N	N/N	N/N	N/N	N*/N*	N/N*
Average Number of Links	N*/N	N/N	N/N	N/N	N/N	N/N	N/N	N*/(Y)	(N*)/N*
Average Number of Words	N/N	N/N	N/N	N/N	N*/N	N/N	N/Y	N/N	(N+)/N
Average Number of Files for downloading	N*/N*	N/ (N*)	N*/N*	(Y)/N*	N*/N*	N*/-	N*/N*	N*/N*	N*/N*
Average Number of Audio Files	-/-	-/-	N*/-	-/-	-/-	-/-	N*/N*	N*/-	-/-
Average Number of Video Files	N/N*	-/-	(Y)/-	N*/-	N*/-	-/-	-/N*	N*/N*	N*/N*
Average Number of Multimedia Files	N/N*	-/-	N*/-	N*/-	N*/(Y)	-/-	N*/N*	N*/N*	N*/N*
Average Number of Applets	N/Y	-/-	N*/N*	-/-	N*/Y	-/-	-/N*	Y/Y	N*/N*

Average Number of Images	N*/(N*)	N/Y↓	N/N	N/N	N/N	N/N	N/(Y)	Y/N*	N*/N*
Average Size of the Pages	N/N	N/N	N/N	N/N	N/N*	N/N	N/Y	N/N	(N*)/N
Average Size of the Images	N*/(N*)	N/N	N*/(N*)	N/(N*)	N/(N*)	N/N	N/N*	Y/(Y)	N*/N*
Average Number of Scripts	(Y)/Y	N*/N	N*/(N*)	N*/N*	Y/N*	N/N	N*/N*	Y/(Y)	N*/N*
Number of Personal Collections	N/Y	Y/N	Y/N	(Y)/N	Y/N*	N/N	Y/Y	Y/Y	Y/Y
Frequency of significant (Y) metrics (P/U)	**1/10**	**1/1**	**3/0**	**3/2**	**11/4**	**3/0**	**2/6**	**11/11**	**3/1**

Table A. 4. Tendency of the metrics for predicting good in the material types with significant differences for peer-reviews and users ratings (P/U)

Metric	*Animation*	*Case Study*	*Collection*	*Drill*	*Lecture/ Presentation*	*Quiz*	*Reference Material*	*Simulation*	*Tutorial*
Number of Links	-/↑				↑/-				
Number of Unique Links					↑/-				
Number of Internal Links					↑/-				
Number of Unique Internal Links					↑/-				
Number of External Links	-/↑					↑/-		-/↑	
Number of Unique External Links	-/↑					↑/-		-/↑	
Number of Images					↑/-			↑/-	
Size of the Page (in bytes)	-/↑						-/↓		↓/-
Total Size of the Images (in bytes)				-/↑				↑/↑	
Number of Scripts	-/↑			-/↑	↑/-		↑/-	↑/-	
Number of Applets	-/↓			-	-/↑			↓/↓	
Number of Words							-/↓		
Number of Words that are links					↑/-				↓/-
Total Number of Pages	-/↑				↑/-	↑/-			
Number of Files for downloading				↑/-	↑/-				
Number of audio files									
Number of video files			↑/-		-/↑				
Number of multimedia files									
Average Number of Unique Internal Links									
Average Number of Internal Links									
Average Number of Unique External Links								↓/↑	
Average Number of External Links								↓/↑	
Average Number of Unique Links									
Average Number of Links								-/↑	
Average Number of Words							-/↓		
Average Number of Files for downloading				↑/-					
Average Number of Audio Files									
Average Number of Video Files			↑/-						
Average Number of Multimedia Files					-/↑				
Average Number of Applets	-/↓				-/↑			↓/↓	

Average Number of Images		-/↓				-/↑	↑/-	
Average Size of the Pages						-/↓		
Average Size of the Images							↑/↑	
Average Number of Scripts	↑/↑				↑/-		↑/↑	
Number of Personal Collections	-/↑	↑/-	↑/-	↑/-	↑/-	↑/↑	↑/↑	↑/↑

Note: Here (↑) stands for a positive contribution, (↓) stands for negative contribution, and (-) means the metric did not presented significant difference for that threshold and its tendency was not evaluated.

Publications

1. Cechinel, Cristian, Sánchez-Alonso, Salvador, & García-Barriocanal, Elena. (2011). Statistical profiles of highly-rated learning objects. Computers & Education, 57(1), 1255-1269. doi: 10.1016/j.compedu.2011.01.012
2. Cechinel, Cristian, & Sánchez-Alonso, Salvador. (2011). Analyzing Associations between the Different Ratings Dimensions of the MERLOT Repository. Interdisciplinary Journal of E-Learning and Learning Objects 7, 1-9.
3. Cechinel, Cristian, Sánchez-Alonso, Salvador, Sicilia, Miguel-Ángel, & Mattos, Merisandra Côrtes. (2010). Estudo Exploratório sobre a Utilização de Algoritmos de Mineração de Dados na Predição e Caracterização da Qualidade de Objetos de Aprendizagem no Repositório MERLOT. Paper presented at the XXI Simpósio Brasileiro de Informática na Educação (SBIE 2010) João Pessoa, Paraíba.
4. Cechinel, Cristian, Sánchez-Alonso, Salvador, Sicilia, Miguel-Ángel, & Mattos, Merisandra Côrtes. (2010). Descriptive Analysis of Learning Object Material Types in MERLOT. In Salvador Sánchez-Alonso & Ioannis N. Athanasiadis (Eds.), Metadata and Semantic Research (Vol. 108, pp. 331-341): Springer Berlin Heidelberg.
5. Cechinel, Cristian, Sánchez-Alonso, Salvador, & Sicilia, Miguel-Ángel. (2010). Exploratory analysis of the correlations between peer-reviewers and users ratings on MERLOT repository. Paper presented at the II Conferencia conjunta Iberoamericano sobre Tecnologías del Aprendizaje (CCITA 2010), Cadiz, Spain.
6. Cechinel, Cristian, Rebollo, Raquel, & Sánchez-Alonso, Salvador. (2012). Objetos de aprendizaje: Definición y caracterización In Francisco Alvarez Rodriguez & Jaime Muñoz Arteaga (Eds.), Avances en objetos de aprendizaje: experiencias de redes colaboracion en México. Aguascalientes: Editorial Universidad Autónoma de Aguascalientes e Universidad Veracruzana.
7. Cechinel, Cristian, Sánchez-Alonso, Salvador, Sicilia, Miguel-Ángel, & Azevedo Simões, Priscyla Waleska Targino. (2011). Exploring the Development of Endorsed Learning Resources Profiles in the Connexions Repository. In Elena García-Barriocanal, Zeynel Cebeci, Mehmet C. Okur & Aydın Öztürk (Eds.), Metadata and Semantic Research (Vol. 240, pp. 12-21): Springer Berlin Heidelberg.
8. Cechinel, Cristian, Sánchez-Alonso, Salvador, Sicilia, Miguel-Ángel, & Amador, César Velazquez. (2011). Evaluating Models for Automated Quality Assessment of Learning Objects inside Repositories 6th Latin American Conference on Learning Objects and Technology Enhanced Learning – LACLO 2011 Montevideo - Uruguay.

9. Lytras, M. D, Sicilia, M-Á, & Cechinel, C. (Forthcoming). The value and cost of metadata. In Miguel-Ángel Sicilia (Ed.), Handbook of Metadata, Semantics and Ontologies: World Scientific Publishing Company.

10. Sicilia, Miguel-Ángel, García-Barriocanal, Elena, Sánchez-Alonso, Salvador, & Cechinel, Cristian. (2010). Exploring user-based recommender results in large learning object repositories: the case of MERLOT. Procedia Computer Science, 1(2), 2859-2864. doi: DOI: 10.1016/j.procs.2010.08.011

11. Cechinel, Cristian, Sánchez-Alonso, Salvador, & Sicilia, Miguel-Ángel. (2009). Empirical Analysis of Errors on Human-Generated Learning Objects Metadata. In Fabio Sartori, Miguel Ángel Sicilia & Nikos Manouselis (Eds.), Metadata and Semantic Research (Vol. 46, pp. 60-70): Springer Berlin Heidelberg.