OpenScout: Harvesting Business and Management Learning Objects from the Web of Data

Ricardo Kawase, Marco Fisichella
L3S Research Center
Leibniz Universität Hannover,
L3S, Appelstr. 9a, 30167
Hannover, Germany
{kawase, fisichella}@L3s.de

Katja Niemann
Fraunhofer FIT
SchloßBirlinghoven, 53754
Sankt Augustin, Germany
katja.niemann@fit.fraunhofer.de

Vassilis Pitsilis, Aristides Vidalis
NCSR DEMOKRITOS - Patriarchou Gregoriou and Neapoleos str, 153 10 Aghia Paraskevi, Greece
{ vpitsilis, avidal}@dat.demokritos.gr

Philipp Holtkamp
University of Jyväskyla, Mattilanniemi 2, Agora Building, Jyväskylä, Finland
philipp.holtkamp@jyu.fi

Bernardo Pereira Nunes
Department of Informatics - PUC-Rio
Rio de Janeiro, RJ - Brazil
nunes@inf.puc-rio.br

ABSTRACT

Already existing open educational resources in the field of *Business and Management* have a high potential for enterprises to address the increasing training needs of their employees. However, it is difficult to act on OERs as some data is hidden. In the meanwhile, numerous repositories provide Linked Open Data on this field. Though, users have to search a number of repositories with heterogeneous interfaces in order to retrieve the desired content. In this paper, we present the strategies to gather heterogeneous learning objects from the Web of Data, and we provide an overview of the benefits of the OpenScout platform. Despite the fact that not all data repositories strictly follow Linked Data principles, OpenScout addressed individual variations in order to harvest, align, and provide a single end-point. In the end, OpenScout provides a full-fledged environment that leverages on the Linked Open Data available on the Web and additionally exposes it in an homogeneous format.

Categories and Subject Descriptors

H.5.m [**Information Interfaces and Presentation**]: Miscellaneous—*Classification, Navigation*

General Terms

Performance, Experimentation, Standardization, Verification

Keywords

Linked Data, metadata, sharing, open content

WWW 2013 Companion, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2038-2/13/05.

1. INTRODUCTION

Over the past years we have witnessed the Web becoming an established channel for learning. Thus, educational institutions around the world have begun a major effort to adapt their offline learning materials into digital learning content appropriate for the Web. Due to the high development costs of learning materials, learning objects that can be reused have drawn attention in the e-learning community [12]. Consequently, new standards and specifications emerged in order to describe and handle learning objects, such as Dublin Core[1], IEEE Learning Object Model (LOM) and ADL SCORM[2], and interface mechanisms such as SQI[3] or OAI-PMH[4] to describe, store and retrieve LOs from repositories. Thus, reusability and interoperability became key concerns on most online repositories, especially those for learning purposes.

Nowadays, hundreds of repositories are freely available on the Web aiming at sharing and reusing learning objects, but lacking in interoperability. To alleviate the interoperability issue, the adoption of Linked Data principles[5] to expose data on the Web have been widely adopted. However, despite of its large adoption, in the educational field, publishing learning objects using linked data is still taking the first steps. As described by Dietze et al. [1], the main research problems to ensure Web-scale interoperability in educational respositories are to (a) integrate distributed data from heterogeneous repositories; (b) deal with continuous change; (c) metadata

[1] `http://dublincore.org/documents/dces/`

[2] Advanced Distributed Learning (ADL) SCORM: `http://www.adlnet.org`

[3] Simple Query Interface: `http://www.cen-ltso.net/main.aspx?put=859`

[4] Open Archives Protocol for Metadata Harvesting `http://www.openarchives.org/OAI/openarchivesprotocol.html`

[5] `http://www.w3.org/DesignIssues/LinkedData.html`

mediation and transformation; and (d) enrichment and interlinking of unstructured metadata.

Following this directions, we describe OpenScout[6] architecture, which is based on a query mediation approach that facilitates the retrieval of learning objects from autonomous heterogeneous learning repositories. Our approach aims at mapping different metadata schemas that enables a creation of single access point for accessing learning objects from multiple repositories and provide a means to publishing learning content harvested from different resources as linked data.

Briefly, the OpenScout learning platform is the outcome of the OpenScout project efforts, which stands for 'Skill based scouting of open user-generated and community-improved content for management education and training'.

The project had a dual aim at hand, first to examine existing standards and solutions and second to review how learning objects on the field of *Business and Management* can be easily embedded, shared and reused. The main achievements of OpenScout take leverages on the Web of Data, where different repositories with heterogeneous data descriptions (some of them following the Linked Data principles and some not) are gathered together in one aligned description for a single end-point query.

The remaining sections are described as follows. Section 2.1 introduces the architecture of OpenScout and the problem of mapping multiple metadata schemas. Section 3 present the strategies followed harvest educational repositories. Finally, we present our results in Section 4 and discuss our method and future directions in Section 5.

2. FROM THE WEB OF DATA TO LEARNING ENVIRONMENTS

Metadata play an important role in online repositories with learning resources. Metadata are in general used for describing the properties of information resources, in order to facilitate their categorization, storage, search and retrieval in digital collections. Storing the metadata in a structured and standardized manner supports the automation of search and retrieval mechanisms, the comparison between descriptions of different resources, the reusability of descriptions in different applications, as well as the interoperability between different storage systems [5].

Metadata are associated to resources and consist of various metadata elements. Metadata schemas (or metadata models) are sets of metadata elements designed for a specific purpose, such as describing a particular type of resource [8]. Metadata specifications are well-defined and widely agreed metadata schemas that are expected to be adopted by the majority of implementers in a particular domain or industry. When a specification is widely recognized and adopted by some standardization organization (such as ISO - the International Standardization Organization), it becomes a metadata standard. However, there is no single metadata standard that can be used in all application domains. Rather, there are various metadata standards or specifications that can be adapted or 'profiled' to meet application specific needs. This requirement for specific adaptations has brought up the concept of application profiles. An application profile is a collection of metadata elements selected from one or more metadata schemas, and its purpose is to adapt or combine existing schemas into a package that is tailored to the functional requirements of a particular application, while retaining interoperability with the original base schemas [2].

Metadata are in particular important for the description of learning objects stored in learning repositories. (Educational) Metadata associated with learning objects make search, retrieval and access faster, easier and more effective. For the description of the metadata related to Learning Objects, various standards exist. Using a recognized metadata standard is important for a variety of reasons: metadata descriptions (records) of learning resources may be exchanged among different Learning Object Repositories (LORs); search queries may be propagated among different (and interconnected) LORs; and generally the integration of data from different sources is facilitated (Web of Data).

2.1 Architecture

The different layers of the architecture in the OpenScout approach are exposed in Figure 1. Based on a shared technical infrastructure for federated access to the repositories, metadata harvesting, content enrichment, web services for metadata manipulation and retrieval and metadata-based content access are provided. The approach aims at making learning objects from different repositories jointly searchable and retrievable.

Services in OpenScout connect the presentation layer with data sources. They process user queries and return results, handle user management and provide means for gathering and manipulating metadata. Some services provide simple functions while others are more complex and can even aggregate functionality. Besides metadata and content retrieval, OpenScout services allow users to annotate contents with own metadata, track activities and generate metadata from user actions. Examples for basic services are:

'Subscribe' which allows users to become notified as soon as relevant content is added or changed; 'CompetencySearch' which makes competencies searchable by connecting competencies, contents and context; and so on. Based on these basic services, more complex services can be realized in order to aggregate and combine various functionalities, e.g. services to enable adaptation and localization of content to the culture and language of the European countries.

To ensure full interoperability, all services are based on open standards, such as the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) for metadata harvesting and SOAP for remote web service connectivity. The search service is enabled through the Simple Query Interface (SQI) [10] in order to be able for OpenScout to join Learning Object Repositories (LOR) federations like Globe[7] and Ariadne[8]. SQI can be combined with any query language [9].

The open content repositories federation is based on the exchange and combination of metadata. The real learning objects are not exchanged between the different components in the architecture, only the metadata description is processed during the progression of the federation. First, the content repositories provide accessible metadata describing the learning objects. The harvester component accesses this information and stores it in the centralized repository. Next, a SQI service grants a connector component (Enterprise Service Bus - ESB) access to the centralized metadata. Finally, the ESB component processes the metadata and provides

[6] `http://learn.openscout.net`

[7] Globe - `http://globe.edna.edu.au/`

[8] Ariadne - `http://www.ariadne-eu.org`

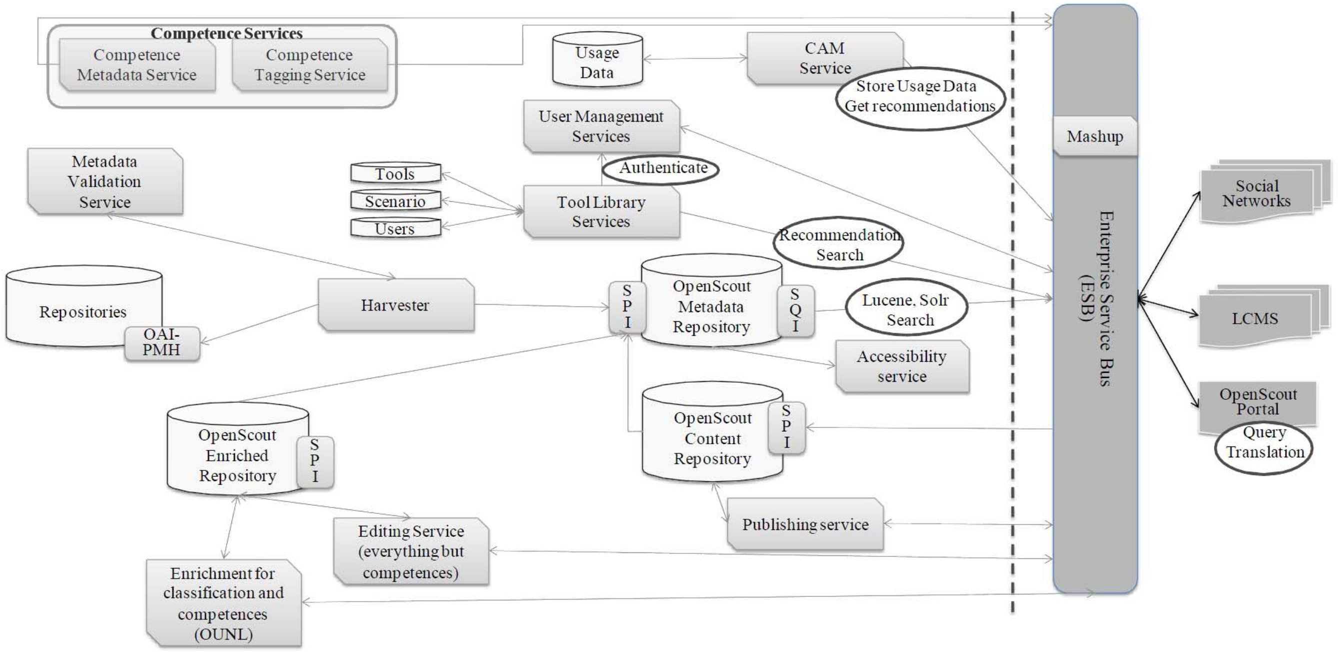

Figure 1: Openscout architecture overview.

high level services to the interfaces (Webportal, widgets, etc) accessed by the users.

To gather relevant learning objects metadata on the Web of data, we used three different approaches (as explained in Section 3), however our data exposure (OAI-PMH target) is available solely in LOM format.

2.2 Educational Metadata

Dublin Core (DC) is a widespread metadata description standard. Similarly to LOM, it consists in a defined set of terms to describe Web and physical resources. The main goal of the DC format is to support data discovery and provide interoperability for metadata vocabularies across linked data and semantic Web infrastructures. In terms of content, DC does not differ so much from LOM description; however, the differences in the format need to be addressed in order to enable OpenScout's metadata harvester to gather these resources.

In this subsection, we present the reasons that motivated us to chose LOM instead of Dublin Core. Actually, the simple version of the Dublin Core schema consists of a set of 15 independent elements, including for example: Title, Identifier, Language, Description, etc. Qualified Dublin Core employs additional qualifiers to further refine the description of a resource.

The conceptual schema for Dublin Core defines the semantics of the DC elements and their qualifiers, such as: 'An element is a property of the resource being described', 'An element refinement is a property of a resource that shares the meaning of a particular DCMI[9] element but with narrower semantics', 'An encoding scheme provides contextual information or parsing rules that aid in the interpretation of a value string'.

It should be noted that the Dublin Core schema is encoded in terms of RDF. LOM, by contrast, uses a completely different schema enconded in XML. LOM describes resources using a set of more than 70 attributes, divided into nine categories: General, Lifecycle, Meta-Metadata, Technical, Educational, Rights, Relation, Annotation and Classification.

The descriptors are organized in a tree-like structure under these categories. This tree makes it possible to organize the information in a consistent way, grouping information into related pieces. The LOM schema is thus based on a recursive container model. However, it can be seen that it is not compatible with the DC schema [6]. As a simple example, the 2.3.3 Date element (found in the LOM standard [3]) is not a property of the resource being described, but can be seen as a property of the 'Contribution' it belongs to. Similarly, the elements in the 'Metametadata' categories are not properties of the resource being described, but of the metadata document itself.

The container-based model used by LOM is thus not compatible with the model used by Dublin Core. Binding LOM to RDF is the obvious example in this context, as the schema of RDF is based on a property-value model and not containment. In general, it leads to difficulties when trying to combine terms from two metadata standards into the same system. When the schemas are compatible, such a combination or mapping can be realized by a simple translation. If the schemas are incompatible, the translation must be done on an idiosyncratic, element-by-element basis.

This schema incompatibility is the main source of the challenges in binding LOM to RDF [7]. Furthermore, LOM is gradually becoming the reference standard for educational systems managing learning objects of many kinds, besides

[9] DCMI Abstract Model - `http://www.ukoln.ac.uk/metadata/dcmi/abstract-model/`

Table 1: Mapping of required LOM fields.

DC-field	LOM-field	LOM-hierarchy
identifier	identifier	header
datestamp	datestamp	header
setSpec	setSpec	header
title	title\string	metadata\lom\general
creator	entity	metaMetadata\contribute
subject	keyword\string	metadata\lom\general
description	description\string	metadata\lom\general
date	datetime	metadata\lom\lifeCycle\contribute
type	learningResourceType\value	metadata\lom\educational
format	format	metadata\lom\technical
identifier	location	metadata\lom\technical
rights	description	metadata\lom\rights

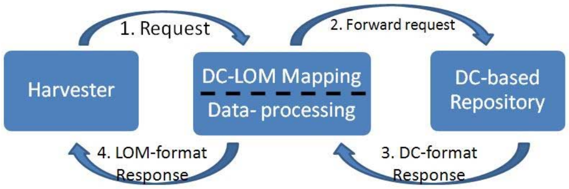

Figure 2: LOM DC mapping.

that it is one of most important standard for interoperability. Also, LOM is part of SCORM (Sharable Content Object Reference Model) which is the standard to package learning resources; it is used by most LMS and consequently it is a de facto standard. We therefore support LOM. In addition, due to its full coverage of learning objects metadata description, the IEEE 1484.12.1 - 2002 Standard for Learning Object Metadata has been chosen as the schema model for the OpenScout centralized repositories.

2.3 Connecting Learning Objects

The federated OpenScout repository does not only allow users to search through a lot of learning repositories at the same time, it also offers the possibility to add metadata to the LOMs describing the LOs, e.g. social metadata as tags and comments as well as classifications and competences from pre-defined taxonomies .

On the one hand, this metadata is used to offer the users more possibilities to find suitable learning objects, on the other hand it also establishes connections between learning objects, e.g. when they share the same tags or classifications that was not possible when the LOMs were stored in separate repositories. Additionally, the users' activities are tracked in the portal and the learning objects can be associated with each other when they share similar users or usage contexts.

2.4 Metadata Harvester

The harvesting component is the foundation of the OpenScout effort. In our case, *harvesting* goes beyond a simple data gathering. The whole harvesting process in OpenScout encompasses crawling, content metadata analysis of Learning Objects (LO) from different Learning Object Repositories (LOR) and storage in a centralized repository. This is not a one-time import action, it is an event repeated in a regular basis or triggered by updates. Once harvested, the LO is described through an application profile described by LOM standard [3]. The result of the harvesting processes provides a centralized repository where metadata of learning objects of all repositories are federated, thus providing means to uniformly query and retrieve the learning objects. It is important to remark that the learning objects remain on the content provider's repositories; only the metadata is transferred and indexed.

As we have seen, the infrastructure provides means to enrich the LO metadata so users are able to acquire knowledge and contribute sharing additional inputs. The central repository offers an OAI-PMH interface so that enriched metadata can be retrieved by the content providers, thus augmenting their content. Supporting this integration facilitates the goal of finding LOs and enables a full extension of operational possibilities over the LOs, albeit each LO belongs to a different repository that possess different metadata schema [9]. The OpenScout's harvesting extends the reach of knowledge gathering by providing flexible means to collect valuable LOs. The harvesting model gathers content metadata by collecting information from repositories that offer an OAI-PMH. After the data is harvested, it is validated using the OpenScout Application Profile and then stored in the centralized repository.

3. HARVESTING

In total, we apply three different methods for harvesting LOs, namely, plain OAI-PMH harvesting [4], DC to LOM mapping and Web crawling.

3.1 LOM Harvesting Strategy

In this work, we use the Ariadne Harvester for harvesting the integrated repositories [11]. The Ariadne Harvester has been developed to manage the process of gathering all contents from existing metadata repositories by making use of a unified interface called OAI-PMH target which is based on the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH).

For each OAI-PMH target, the harvester maintains basic parameters, such as the base URL, an enumeration of the harvested OAI-PMH sets, the metadata prefix that identifies the standard or application profile, or the metadata provider. Every time new data are retrieved, the harvester publishes the metadata to the centralized OpenScout Metadata Repository using its publishing interface.

3.2 Mapping DC to LOM

The mapping of DC to LOM consists in a PHP implementation that provides an OAI-PMH end-point to the harvester, and forwards each call to the original content and metadata provider. The 'request' steps do not require any processing since the OAI-PMH calls are standardized (Figure 2, steps 1 and 2). The answer from the repository is then according to the DC standards. At this point the Mapping service parses the DC response and translates it to LOM format. Since there is no guarantee of the consistence of data in the external repositories and its availability we prioritized the mapping of the LOM required fields of the OpenScout application profile. The LOM required fields that needed a direct mapping from LOM are listed in Table 1. Currently, eight repositories have been integrated in the OpenScout portal through the DC-LOM mapping strategy (see repositories marked with a * in Table 2).

3.3 Crawling Data from Website

A lot of websites offer open and relevant learning content in the business domain but no metadata instances that can easily be harvested. However, as websites often come with a fixed structure, they can be parsed and analyzed to create those metadata instances automatically.

For instance, Moneyterms[10] website as example. Moneyterms' website offers explanations for business terms, e.g. 'Absolut Return' or 'Income Effect'. The site provides five index pages containing all terms with a short explanation and a link to the full explanation. The index pages can easily be retrieved with an http request and parsed using an html parser. A LOM instance is created automatically for each term with the term as title, the short explanation as description and the link as URL. The creator of the page is mentioned as author of the learning object, whereas the OpenScout project is named as creator of the LOM instance. Additionally, the Terms of Use of the learning object are noted in the LOM's right section.

After the LOM instances are created, they are stored in a separate repository that is harvested by OpenScout. The crawling and LOM creation process is repeated regularly to cover changes in the web pages.

4. RESULTS

In this section, we preset the outcomes of our infrastructure. Our success indicator is given by the number of repositories integrated and the hours of learning material available.

4.1 Number of integrated Repositories

As a total 24 repositories have been integrated to the federated infrastructure (see Table 2). From these, six repositories are provided by the OpenScout Partners, one repository includes the published material of OpenScout users and 18 repositories are integrated from external sources. It is worth noting that all repositories are focused on the field of *Business and Management* but not all of them are compliant with linked data principles.

4.2 Hours of Content

The hours of content for the learning objects in the OpenScout federation were calculated using simple mechanisms depending on the content type. Courses and lectures usually have clear learning hours connected which made the calculation for this type easy and allowed an exact estimation of the learning hours. For other types, experts from the management domain with strong teaching experiences estimated the learning hours. For this purposed a set of materials of different types (presentation, pdf file, picture, video etc.) was analyzed and an average learning time was established. Table 2 gives an overview of the repositories, their content and the estimated learning hours.

5. CONCLUSION

In this paper, we presented the main approaches developed in OpenScout to provide a unified and aligned access to learning material in the field of *Business and Management*. The work described here centers on the problem of collecting relevant records from the Web of Data and converting it into the *Standard for Learning Object Metadata* (LOM).

As a result of such efforts, OpenScout portal is the top repository of open educational data in the field of *Business and Management* with over 38,000 learning objects in English language and thousands more in other languages. Thanks to the Web of Data, we were able to gather together a great focused repository. Additionally, all LOM data in OpenScout is exposed to the Web through an OAI-PMH target.

Aligned with the discussion of the main research problems to ensure Web-scale interoperability in educational repositories by Dietze et al. [1], Openscout is a step towards to publish learning objects repositories as Linked Open Data: (a) Our methods retrieves learning objects metadata and match the information using learning metadata standards; (b) The process implemented by OpenScout continuously check for new information available in the repositories and the new content is incorporated for further query; (c) As a means to search for learning objects, we also provide a mediation that makes the metadata transformation transparent. Further, OpenScout also addresses some of the principles of linked data by providing all the data collected from the repositories on the Web in a machine-readable structure in a non-proprietary format. In this work, our contribution lies on the creation of the necessary infra-structured to, our future goals: enrich the metadata in the learning context, transform the data into the same standards from a common vocabulary, publish the data on the Web and link to other resources.

6. ACKNOWLEDGEMENT

This research has been co-funded by the European Commission within the eContentplus targeted project OpenScout, grant ECP 2008 EDU 428016 (cf. http://www. openscout.net).

7. REFERENCES

[1] S. Dietze, S. Sanchez-Alonso, H. Ebner, H. Yu, D. Giordano, I. Marenzi, and B. P. Nunes. Interlinking educational resources and the web of data - a survey of challenges and approaches. *Emerald Program: electronic Library and Information Systems*, 47(1), 2013.

[10] `http://moneyterms.co.uk`

Table 2: List of main repositories integrated and details on their contents. Repositories marked with a * are integrated through the DC-LOM mapping.

Repository	Objects and content type	Learning hours
OpenER	23 complete courses	575
OpenLearn Learning Space	47 full courses	458
12manage	2897 objects	1448
Consortia Academia*	15 research journal articles	30
EconStor	11089 scientific publications	22178
EMAJ Journal*	18 scientific journal articles	36
epub WU Vienna	1255 scientific publications	2510
FGV Fundação Getulio Vargas*	62 dissertations	620
digitalarchives@gsu*	88 scientific journal articles	176
IBSU*	93 articles of a scientific journal	186
INSEAD Knowledge	581 articles	290
INSEAD Library	2624 articles	5248
Jorum*	290 scientific publications and course	435
LACLO	9618 learning objects	11541
Leeds Metropolitan University Repository*	528 scientific publications an books	1056
ePrints LSE Research Online*	13498 research articles and books	26996
Money Terms	985 definitions	492
OILI	12 complete courses	440
OpenArchive@CBS	2052 scientific publications	4104
Reference for Business	4179 definitions and models	2089
The Open University iTunes U	1096 learning objects	22
The Times 100	105 full case studies	1575
UNED repository	32 multimedia learning objects	9
YouTube EDU Channels	426,74 video hours in 1096 learning objects	640
OpenScout Repository	13 learning objects	128

[2] E. Duval, W. Hodgins, S. A. Sutton, and S. Weibel. Metadata principles and practicalities. *D-Lib Magazine*, 8(4), 2002.

[3] W. Hodgins et al. Draft Standard for Learning Object Metadata. *W3C Candidate Recommendation IEEE P*, 1484:08855–1331, 2002.

[4] C. Lagoze and H. Van de Sompel. The open archives initiative: building a low-barrier interoperability framework. In *Proceedings of the 1st ACM/IEEE-CS joint conference on Digital libraries*, JCDL '01, pages 54–62, New York, NY, USA, 2001. ACM.

[5] N. Manouselis, K. Kastrantas, and A. Tzikopoulos. An ieee lom application profile to describe training resources for agricultural & rural smes. In *Proc. MTSR07-2nd International Conference on Metadata and Semantics Research*, 2007.

[6] M. N. Matthias, M. Palmér, and J. Brase. The lom rdf binding - principles and implementation.

[7] M. Nilsson, P. Johnston, A. Naeve, and A. Powell. The future of learning object metadata interoperability. In *Learning Objects: Standards, Metadata, Repositories, and LCMS.*, 2006.

[8] NISO. Understanding metadata. National Information Standards Organisation, NISO Press, 2004.

[9] C. R. Prause, S. Ternier, T. de Jong, S. Apelt, M. Scholten, M. Wolpers, M. Eisenhauer, B. Vandeputte, M. Specht, E. Duval, and E. Duval. Unifying learning object repositories in mace. In *Proceedings of the First International Workshop on Learning Object Discovery & Exchange (LODE'07)*, 2007.

[10] B. Simon, D. Massart, F. van Assche, S. Ternier, E. Duval, S. Brantner, D. Olmedilla, Z. Miklos, and Z. Miklos. A simple query interface for interoperable learning repositories. In *Proceedings of the 1st Workshop on Interoperability of Web-based Educational Systems*, pages 11–18, 2005.

[11] S. Ternier, K. Verbert, G. Parra, B. Vandeputte, J. Klerkx, E. Duval, V. Ordoez, and X. Ochoa. The ariadne infrastructure for managing and storing metadata. *Internet Computing, IEEE*, 13(4):18 –25, july-aug. 2009.

[12] D. A. Wiley. Connecting learning objects to instructional design theory: A definition, a metaphor, and a taxonomy. *Learning Technology*, 2830(435):1–35, 2000.