Study on the use of metadata for digital learning objects in university institutional repositories (MODERI)[1]

Gema Bueno-de-la-Fuente**
Tony Hernández-Pérez
David Rodríguez-Mateos
Eva M. Méndez-Rodríguez
Bonifacio Martín-Galán*

ABSTRACT

Metadata is a core issue for the creation of repositories. Different institutional repositories have chosen and use different metadata models, elements and values for describing the range of digital objects they store. Thus, this paper analyzes the current use of metadata describing those Learning Objects that some open higher educational institutions' repositories include in their collections. The goal of this work is to identify and analyze the different metadata models being used to describe educational features of those specific digital educational objects (such as audience, type of educational material, learning objectives, etc.). Also discussed is the concept and typology of Learning Objects (LO) through their use in University Repositories. We will also examine the usefulness of specifically describing those learning objects, setting them apart from other kind of documents included in the repository, mainly scholarly publications and research results of the Higher Education institution.

KEYWORDS

Institutional Repositories, Learning Objects, Educational Metadata, Dublin Core, OAI-PMH.

1. INTRODUCTION

By definition, institutional repositories can house all kinds of material originating from the intellectual production of the members of the institution concerned. Thus, in the repositories of higher educational institutions, in addition to material typical of scientific production (articles, reports, conference papers, etc.) all kinds of resources can be stored, most importantly those related to the educational function of the institution concerned: digital learning objects, or simply learning objects, as they are widely known.

In this context, those in charge of the repositories confront the difficulty of describing, with the same metadata schema, different types of resources that require specific meta-information to

[1] MODERI is an acronym that stands for *Metadatos sobre Objetos Digitales Educativos en Repositorios Institucionales universitarios,* the Project name in Spanish.

** Corresponding author: gbueno@bib.uc3m.es, FPU Grant 2005-1425, Universities Office, Ministry of Science and Innovation, Spain.

* All authors are teaching staff from the *Departamento de Biblioteconomía y Documentación* ("Department of Librarianship and Information Science") of the *Universidad Carlos III de Madrid*, Spain.

identify all their particular characteristics. If such heterogeneous resources are grouped together in one repository and described with the same metadata schema, on the one hand there is the advantage of homogeneity and therefore interoperability, but on the other hand, there is the risk of losing a great deal of specific information which could be shared with other systems in the same particular domain.

Generally, open access digital repositories have implemented the Open Archives Initiative - Protocol for Metadata Harvesting (OAI-PMH) [1] as a mechanism to achieve interoperability in the exchange of meta-information with other systems, as the metadata harvesters. To do so, they have to use and display their records in the unqualified Dublin Core metadata schema or DC-Simple (Dublin Core Metadata Element Set, ISO 15836) [2]. However, if the software used permits it, each repository is free to use any additional metadata schema to describe its resources, as long as they also use DC-Simple or their metadata records are mapped to `oai_dc` format [3]. Furthermore, OAI-PMH protocol allows exposure of records in other formats based on XML Schema. In any case, general union catalogs like OAIster [4] usually only harvest DC-Simple records.

The capacity to use multiple metadata schemas, which both the OAI-PMH protocol and digital repositories have, would be the obvious answer to the difficulty in question: how to describe different types of material (scientific, educational, administrative) with different metadata application schemas or profiles which permit a more accurate description of each documentary area or typology of content. Thus, an institutional university repository could make a combined use of: the metadata application profile SWAP (Scholarly Works Application Profile) [5] for its collection of articles and preprints, together with ETD-MS [6] for its collection of theses and dissertations, and IEEE LOM [7] or the application profile for education DC-Ed [8] for the educational resources.

In this article, we will discuss the particular case of institutional repositories with learning objects in their collections, and attempt to answer the following questions: to what extent do institutional repositories use this option of mixing educational and scientific resources? How much interoperability does the OAI-PMH protocol provide? How has the description of such

resources using a basic, general schema like DC-S been resolved? Have any other educational metadata schema been adopted for this purpose? And, in general, which metadata models are institutional repositories adopting to resolve this situation?

2. OBJECTIVES

The general objective of this study is to analyze how metadata for the description of digital learning objects is currently being used in academic institutional repositories worldwide. Specifically, this study examines a sample of selected repositories to determine which metadata schemas are being used, whether institutions have limited themselves to using DC-Simple, whether they are using other metadata formats (and, if so, which of them expose their records in conformity with the OAI-PMH protocol), and also if they have adopted their own schemas or application profiles, especially those designed to describe learning objects. With regard to the latter, how each repository has adapted the DC-Simple metadata schema has been analyzed: whether new elements have been added or whether the DC-Simple elements have been refined by means of element qualifiers. In any case, it is interesting to know whether these elements or qualifiers come from educational metadata application profiles or standards, other metadata schemas, or on the contrary, they are specific to a particular repository.

Another fundamental aspect of this study is to analyze the values of the metadata elements in the records, whether specific vocabularies for elements of educational interest are being applied and if these are controlled or not. Likewise, the values of the element `dc:type` have been analyzed, especially those related to learning objects. Where no educational metadata element or qualifier has been added, we have tried to find out if other DC-Simple elements are used to describe any kind of educational information.

3. LEARNING OBJECTS IN HIGHER EDUCATION: STUDY SCOPE

According to the widely accepted definition of the IEEE LTSC committee [8], a *learning*

object is “any entity, digital or non-digital, which can be used, re-used or referenced during technology supported learning.” To be more precise, a learning object is an educational information unit used as a constituent element of content in an e-learning system. These objects are the smallest instructional or educational units that can stand alone and still be significant for the student. For IEEE LTSC, “examples of Learning Objects include multimedia content, instructional content, learning objectives, instructional software and software tools, and persons, organizations, or events referenced during technology supported learning”.

However, the generally accepted concept of learning object only includes digital objects, and therefore excludes physical objects like people and organizations. Furthermore, the IEEE LTSC’s definition does not only cover those entities specifically created to be used in an educational process but also all those created for another purpose but which can be used to transmit any kind of knowledge and thus form part of a learning process. This occurs very often in Higher Education (HE), in which the typology of teaching support materials covers a very wide spectrum. As well as the material written by the teaching staff themselves, there are also the traditional library documents (monographs, manuals, journals, etc.) and press documents, audiovisual and multimedia material, software applications and even those produced by scientific activity (articles, research reports, conference papers, etc.) which allow students to access the original sources. Nevertheless, this study has considered only those objects that could be unequivocally identified as HE study material, and that, although they may have been utilized in different contexts for different purposes, have in fact been used for explicitly educational purposes and this is in some way captured in the corresponding metadata records.

Consequently, particular attention has been paid to differentiating between those kinds of documents which, although they may be considered as specific types of learning objects in some vocabularies (presentations, videos, images, software, etc), may or may not actually be so, depending on the usage context. Records of these document types, if not expressly accompanied by the term “learning objects” or its equivalents, or if not found in a collection of

an obviously educational nature have not been included in this study.

Finally, the concept of learning objects discussed in this work excludes any material, in whatever medium or format, produced by students as part of their learning process (individual or group works submitted for summative or formative assessment, including essays, and of course projects or doctoral theses).

4. CURRENT STATUS AND RELATED WORKS

Recently there has been a proliferation of studies on metadata evaluation, mainly in order to assess quality (e.g. *Cataloging & Classification Quarterly*, Vol. 46, no. 1, 2008). It is possible to cite various quantitative studies on the use of Dublin Core metadata in OAI repositories, for example, Ward, 2002 [9], Dushay and Hillmann, 2003 [10], Efron, 2007 [11], or Shreeves, Kaczmarek and Cole, 2003 [12]. Similarly, there are also several works focusing on IEEE LOM educational metadata in learning object repositories like ARIADNE in Najjar, 2003 [13]. However, none of these studies specifically refers to institutional repositories for educational content. Ward and Efron categorize different types of repository but do not differentiate between them on the basis of the kind of object they store; Dushay and Hillman focus their study on the repositories of the National Science Digital Library (NSDL) [14]; and Shreeves, Kaczmarek and Cole on the providers of cultural heritage data collected for a University of Illinois project [15]. Finally, the analysis of metadata in ARIADNE, while very interesting, is distanced from our area of study because it examines a different kind of digital repository.

Also of interest is the study by Barton, Currier and Hey, 2003 [16] which analyzes various problems associated with the creation of quality metadata, comparing the areas of e-prints and learning objects, and particularly elements like author, title, subject and date. Finally, the works of Tennant, 2004 [17] and Chumbe et al., 2006 [18] confirm some of the conclusions reached with regard to the problems of the data collection with OAI-PMH.

5. METHODOLOGY

This study included various phases of data collection and analysis, from the selection of the repositories to be analyzed, to metadata collection and the quantitative/qualitative analysis of

the results.

5.1. Selection of the sample repositories

For the selection of institutional repositories with digital learning objects in their collections, we used the repository directory OpenDOAR [19]. It allows selecting by repository type (institutional) and by type of content (learning objects). Of the 1128 repositories registered by OpenDOAR as of 22 April 2008, 124 fulfilled both conditions.

From this group, the repositories in Asian languages were excluded, as well as those that were really aggregators rather than source repositories. After, we filtered by the software used to create the repository, choosing the most widely used on a global level (DSpace, GNU EPrints, Fedora and Opus).

The technical problems encountered during the retrieval of records served as an additional filter. Several repositories could not be entirely harvested for various reasons (lack of a valid OAI URL, incomplete XML responses or, in the case of repositories with a large volume of records, incomplete retrieval when obtaining records). These were discarded for the final sample, which only included the 47 repositories fully harvested (Table 1).

Table 1: List of the 47 repositories analyzed

Repository name	Repository URL	Country	Software	No. of records
Infoscience - École polytechnique fédérale de Lausanne	http://infoscience.epfl.ch/	Switzerland	CDSWare	588
Athabasca University Library Institutional Repository	http://auspace.athabascau.ca/	Canada	DSpace	819
DSpace at the University of Guelph	http://dspace.lib.uoguelph.ca/	Canada	DSpace	1510
Dépôt Institutionnel Numérique	https://papyrus.bib.umontreal.ca/dspace/	Canada	DSpace	2174
QSpace at Queen's University	http://qspace.library.queensu.ca/	Canada	DSpace	779
Biblioteca Digital - Autonoma Virtua	http://bohr.cuao.edu.co/dspace/	Colombia	DSpace	19
Repositorio Académico de la Universidad de Chile	http://captura.uchile.cl/dspace/	Chile	DSpace	4565
DHanken	http://openax.shh.fi:8180/dspace	Finland	DSpace	181
DSpace an der Universität Kassel	https://kobra.bibliothek.uni-kassel.de/	Germany	DSpace	954
SSPAL.doc	http://doc.sspal.it/	Italy	DSpace	459
DSpace a Parma	http://dspace-unipr.cilea.it/	Italy	DSpace	234
OpenstarTs	http://www.openstarts.units.it/	Italy	DSpace	2375
ARMIDA@UniMi	http://armida.unimi.it/	Italy	DSpace	309
DSpace at University Leiden	http://openaccess.leidenuniv.nl/	Netherlands	DSpace	11753
e-Learning Repository	http://e-repository.tecminho.uminho.pt/	Portugal	DSpace	408
Universidade do Minho: RepositoriUM	https://repositorium.sdum.uminho.pt/	Portugal	DSpace	6888
DADUN	http://dspace.unav.es/	Spain	DSpace	1163
Diposit Digital de la Universitat de Barcelona	http://diposit.ub.edu/	Spain	DSpace	256
Repositorio Institucional de la	http://rua.ua.es/	Spain	DSpace	4937

Universidad de Alicante				
Göteborgs universitets publikationer - e-publicering och e-arkiv	http://gupea.ub.gu.se/dspace/index.jsp	Sweden	DSpace	6316
DSpace at Bromley College	http://vle.bromley.ac.uk/dspace/	UK	DSpace	38
Minds @ University of Wisconsin	http://minds.wisconsin.edu/	U.S.A.	DSpace	6764
ScholarsArchive@OSU	http://ir.library.oregonstate.edu/dspace/	U.S.A.	DSpace	8223
Scholarly Materials And Research @ Georgia Tech	http://smartech.gatech.edu/dspace/	U.S.A.	DSpace	18637
Humboldt Digital Scholar	http://dScholar.humboldt.edu:8080/dspace	U.S.A.	DSpace	274
DSpace at Drexel University Library	http://dspace.library.drexel.edu/	U.S.A.	DSpace	2340
DSpace at Rice University	http://dspace.rice.edu/	U.S.A.	DSpace	13397
INtellectual property in DIGital form available online in an Open environment	http://indigo.lib.uic.edu/	U.S.A.	DSpace	426
DSpace University of New Mexico	https://repository.unm.edu/	U.S.A.	DSpace	3766
IUScholarWorks	https://scholarworks.iu.edu/dspace/	U.S.A.	DSpace	2814
Digital Repository at Texas A&M University	https://txspace.tamu.edu/	U.S.A.	DSpace	8080
Illinois Digital Environment for Access to Learning and Scholarship Repository	http://www.ideals.uiuc.edu/	U.S.A.	DSpace	7325
eArchives	http://archives.iupui.edu/	U.S.A.	DSpace	563
University of Zimbabwe Institutional Repository	http://ir.uz.ac.zw/	Zimbabwe	DSpace	219
Almae Matris Studiorum Campus	http://amscampus.cib.unibo.it/	Italy	EPrints	809
ISS Library	http://eprints.isofts.kiev.ua/	Ukraine	EPrints	646
New Bulgarian University Scholar Electronic Repository	http://eprints.nbu.bg/	Bulgaria	EPrints	60
St Andrews Eprints	http://eprints.st-andrews.ac.uk/	U.K.	EPrints	305
STOÀ e-PRINTS	http://eprints.stoa.it/	Italy	EPrints	148
Universität München: Elektronische Publikationen	http://epub.ub.uni-muenchen.de/	Germany	EPrints	3337
Minority Health Archive	http://minority-health.pitt.edu/	U.S.A.	EPrints	773
National Aerospace Laboratories Institutional Repository	http://nal-ir.nal.res.in/	India	EPrints	2724
Swinburne Research Bank	http://researchbank.swinburne.edu.au/	Australia	Fedora	7365
Bielefelder Server für Online-Publikationen - Universität Bielefeld	http://bieson.ub.uni-bielefeld.de/	Germany	OPUS	964
Hochschulschriftenserver der Universität Stuttgart	http://elib.uni-stuttgart.de/opus/	Germany	OPUS	3293
OPUS Digitale Hochschulschriften an der FH Düsseldorf	http://fhdd.opus.hbz-nrw.de/	Germany	OPUS	268
Kaiserslauterer uniweiter elektronischer Dokumentenserver	http://kluedo.ub.uni-kl.de/	Germany	OPUS	1964

5.2. Retrieval of metadata records via OAI-PMH

The collection of content was performed using the OAIHarvester2 java tool, developed by OCLC [20]. The tool was configured to use only “ListRecords” with the metadata prefix `oai_dc`, automatically taking the successive values of the “ResumptionToken” attribute for every repository, in order to retrieve all the metadata of each source in a single XML file. In four cases the OAIHarvester2 failed to retrieve the metadata of some repositories, thus other harvesting methods were used (the wget Linux command, or the script available at [21]). Data collection was carried out on 28 May 2008.

5.3. Transformation of XML documents

The XML files obtained from each repository were transformed into tabular HTML documents, applying two consecutive XSLT stylesheets. The output tables were later transferred onto MS Excel spreadsheets, in order to do quantitative studies and review the content of the records. During this process, the empty records (displaying the header element but no metadata element) were excluded.

5.4. Results tables; obtaining indicators and graphics

Various quantitative and qualitative analyses were carried out using the data collected, which will be presented in the next section. We did consider the total sum of occurrences of elements per repository, where an element appeared in each record, but we did not enter the number of occurrences of that element. For some specific metadata elements, the values of all the occurrences in the records were recorded and analyzed.

Some generic data for each repository was also considered, such as geographical location, software used, total number of records, the number of records containing a determined DC-Simple element, etc., which turned out to be extremely useful for contextualizing the results.

5.5. Direct observation of the repositories

As well as analysing the results of the metadata record retrieval, direct observation of the repositories being studied was effected, analyzing their organization system, the search and browsing options and other questions that, in short, would help to locate the educational objects and check the metadata model used to describe them. In particular, direct observation was performed on repositories with a larger quantity of educational material.

6. ANALYSIS AND EXAMINATION OF THE RESULTS

6.1. Data on the selected repositories

Of the 47 repositories that comprised the final sample, 18 countries are represented. The

United States is the most common location, where one fourth of the cases were found (28%), followed by Germany and Italy with six repositories each.

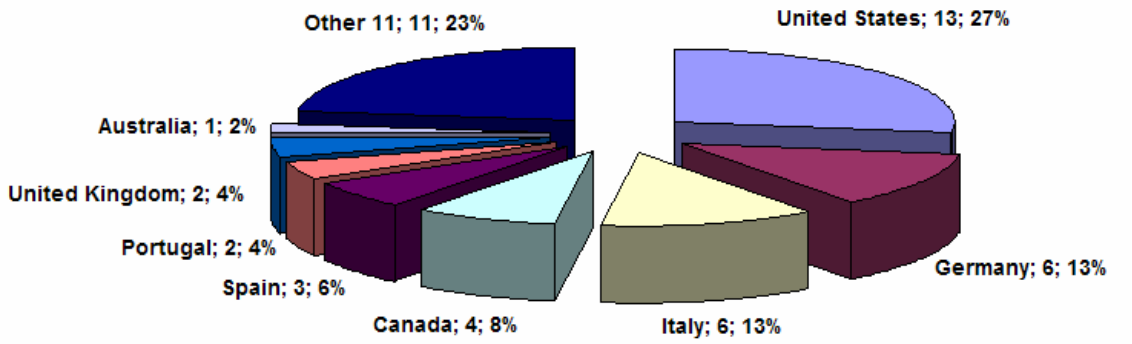

Fig. 1. Repositories per country.

With regard to linguistic aspects, although the number of repositories from Anglo-Saxon countries barely amounts to 40% of the cases, English is the main language in 75% of the archives. Furthermore, a third of the repositories have content in more than one language.

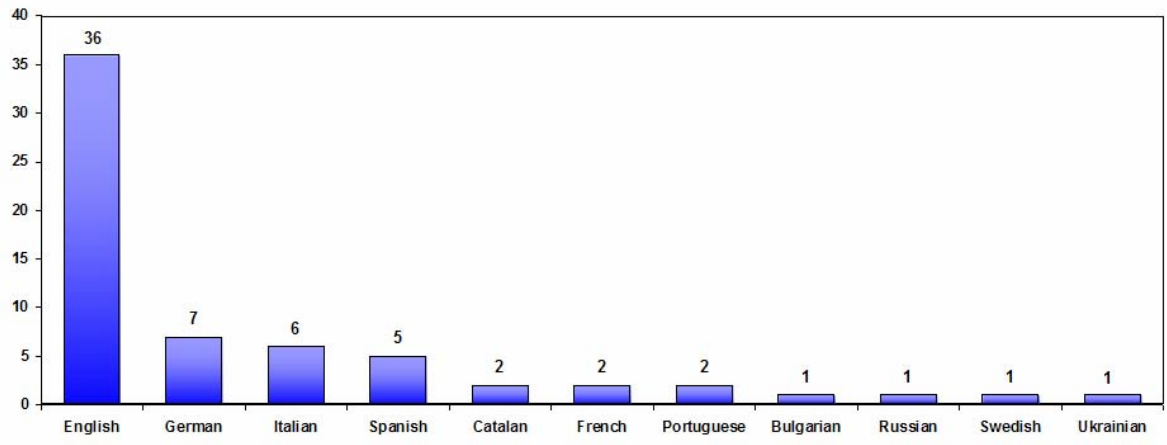

Fig. 2. Languages of the repositories analyzed (>10% content).

In terms of software, the great majority of the repositories use DSpace (three out of every four cases), while the GNU EPrints, OPUS or Fedora repositories are greatly inferior (Fig. 3). In practice, this aspect has a great influence on the metadata model adopted for the repositories studied. For example, DSpace offers options enabling the modification and addition of namespaces and the personalization of *inputForms*. Similarly, the other repository systems feature different particularities regarding the metadata schemas used, and even utilize their own version of DC-Simple for internal repository use, regardless of the

compliance with the OAI-PMH.

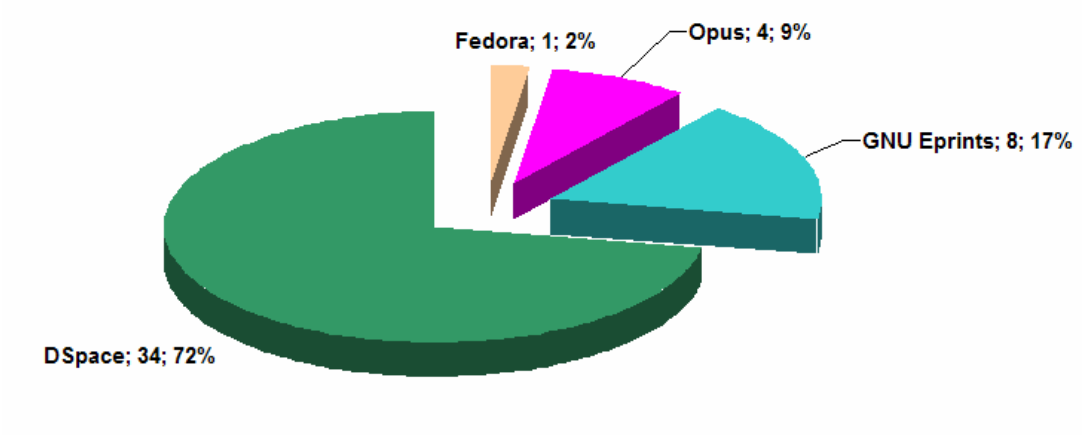

Fig. 3. Repositories analyzed by the software used to create them (DSpace, EPrints, OPUS and Fedora).

6.2. Data on the records collected

The data obtained in the collection phase from the 47 repositories is composed of 141,883 non-empty records, so the average is slightly over three thousand records (3019). However, the variability of record quantity per repository was very high, with a typical variance of 3900; we found repositories with less than 20 records, like the *Biblioteca Digital - Automa Virtual,* at the *Universidad Autónoma de Occidente* (Colombia), but also three cases with over 10,000 records, reaching 18,000 in the SMARTech repository at the Georgia Institute of Technology.

6.3. Use of metadata elements in the records retrieved

The harvesting of metadata elements was only carried out in the basic metadata format established by the protocol OAI-PMH: i.e. `oai_dc`, DC-Simple or ISO 15836, composed of the 15 unqualified elements (Dublin Core Metadata Element Set, DCMES). However, the harvesting process retrieved almost 10,000 records displaying extra metadata elements, some of them not recognized in the DC terms namespace [22].

The usage levels of all these elements have been quantified by record and by repository (Table 2, Fig. 4) with very similar relative results to those obtained by other quantitative studies on metadata in OAI repositories [9] [11].

Table 2. Usage of DC-S metadata elements in the repositories and records analyzed.

Element	No. records containing element	% Records containing element	No. repositories using element	% repositories using element
DC:CONTRIBUTOR	54460	38,38%	38	80,85%
DC:COVERAGE	7909	5,57%	10	21,28%
DC:CREATOR	126175	88,93%	46	97,87%
DC:DATE	141658	99,84%	47	100,00%
DC:DESCRIPTION	113278	79,84%	47	100,00%
DC:FORMAT	100809	71,05%	44	93,62%
DC:IDENTIFIER	139566	98,37%	47	100,00%
DC:LANGUAGE	119139	83,97%	40	85,11%
DC:PUBLISHER	95111	67,03%	42	89,36%
DC:RELATION	47569	33,53%	35	74,47%
DC:RIGHTS	39320	27,71%	26	55,32%
DC:SOURCE	22839	16,10%	12	25,53%
DC:SUBJECT	112225	79,10%	45	95,74%
DC:TITLE	141054	99,42%	47	100,00%
DC:TYPE	126944	89,47%	46	97,87%
Not **OAI_DC** Element				
DC:AUDIENCE	213	0,15%	2	4,25%
DC:MEDIASOURCE	160	0,11%	1	2,13%
DC:GUP	4311	3,04%	1	2,13%
DC:SETSPEC	2896	2,04%	1	2,13%
DC:SUBJECT-BROAD	22	0,02%	1	2,13%
DC:IDENTIFIER-STATIONID	1959	1,38%	1	2,13%

Based on these data it is possible to define three levels of usage for the metadata elements:

- Generalized usage: elements used in 98-100% cases. The elements `dc:date`, `dc:title` and `dc:identifier` fall into this category.
- Frequent usage: those used in 65-90% of the records. These are frequent elements in the repositories analyzed: `dc:type`, `dc:creator`, `dc:subject`, `dc:language`, `dc:description`, `dc:format` and `dc:publisher`.
- Minor or occasional usage: DC elements used in 5-40% of the records studied. The metadata elements `dc:contributor`, `dc:relation`, `dc:rights`, `dc:source` and `dc:coverage` fall into this category.

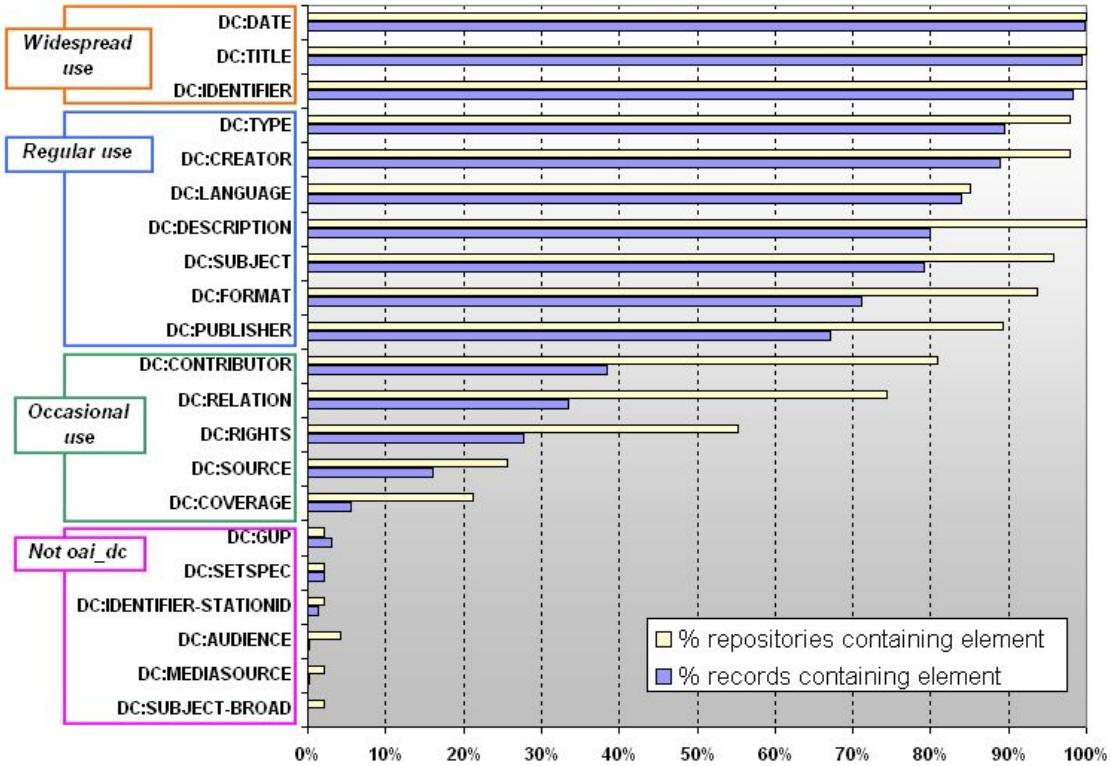

Fig. 4. Usage of DC-S metadata elements in the repositories and records analysed.

In addition, six elements not included in `oai_dc` were found, which one used in just one single repository but `dc:audience`, present in two (Fig. 4). The usage of each element is quite representative within the correspondent repositories; but it is residual with respect to the total records harvested (141,883) (Table 3).

Table 3. Usage of added metadata elements with relation to the repositories and records analyzed.

OAI_DC Element	No. of records containing element	Repository name	No. Records in repository	% of records in containing element
DC:AUDIENCE	153	ARMIDA@UniMi (University of Milan)	309	49.51
DC:AUDIENCE	60	SSPAL.doc (Scuola Superiore della Pubblica Amministrazione Locale)	459	13,07
DC:MEDIASOURCE	160	DSpace at the University of Guelph	1510	10.56
DC:GUP	4311	GUPEA (Göteborg's University)	6316	68,26
DC:SETSPEC	2896	GUPEA (Göteborg's University)	6316	45,85
DC:SUBJECT-BROAD	22	ScholarsArchive@OSU (Oregon State University)	8223	0,27
DC:IDENTIFIER-STATIONID	1959	ScholarsArchive@OSU (Oregon State University)	8223	23,82

The addition of these elements achieves specific goals for each repository:

- The element `dc:mediasource` is used by the University of Guelph in 10% of their

records to indicate the source of digital images, with the following values: *Scanned image, Scanned kodachrome,* and *Digital camera*.

- The Oregon State University adds two elements: `dc:subject-broad` for the main subject, and `dc:identifier-stationid`, for the unique identifier of the digitization workstation for the document. In the same way, the qualifier `digitization` is added to the element `dc:description`, describing the process carried out for each item.
- The repository of the University of Göteborg displays two elements: `dc:gup`, with multiple qualifiers for various purposes (thesis, reports, articles and presentations, or videos); and `dc:setspec` (`dc:setspec:uppsok` in the repository), that groups the undergraduate theses by areas of knowledge and then integrates them in the Uppsök portal [23]. This is a portal of Swedish universities undergraduate theses, sharing a common metadata model and a sets structure with semantic agreements on OAI-PMH protocol, harvested in a central service provider at the Swedish Royal Library [24]. Furthermore, this metadata model refines the `dc:type` element with its own qualifiers, `uppsok` (subject) and `degree` (level of students' works, e.g. *essays)*.

6.4. Content/values of metadata elements

One of the fundamental elements for assessing the quality of metadata or, in our case, their designatory coherence when reflecting the educational value of a digital object, is to analyze the content of the elements. Thus, we have analyzed the values assigned to some metadata elements of special interest for this study (`dc:type`, `dc:format` and `dc:audience`), and whether they use a specific vocabulary encoding scheme.

The `dc:type` element has been vital to this study in order to detect the metadata records for educational material. This was not an easy task given that the 47 repositories harvested together supply nearly 273 different values for this element alone (of which 49 are presumably educational). Although there are some content schemes for this element, such as DCMI Type [25] or EPRINTS Type Vocabulary [26], these are not used consistently;

rather the different repositories adapt them to their needs, using their own values to designate the various typologies of objects stored in their collections. As a result, the quantity and variety of values in `dc:type` is excessive and far from user-friendly. This set includes a significant number of equivalent terms, even in multiple languages, which demonstrates the inadequacy of the existing vocabularies, as well as the lack of consensus in resource description for institutional repositories.

However, the `dc:format` values are far more standardized than those of `dc:type`. The great majority of records use the terms established under the MIME type standard [27] (partly provided by repository software systems like DSpace, which generate the value automatically when uploading files). The most common format type in collected repositories is pdf (almost 80,000 objects) followed by text formats such as HTML, text_plain and msword; and also digital image formats like image_jpeg, image_tiff or image_x-djvu (Fig. 5).

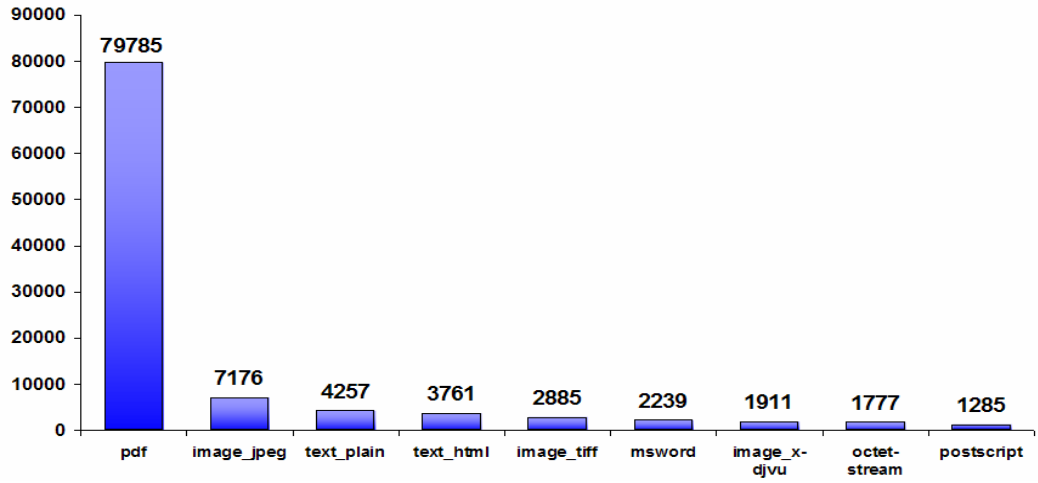

Fig. 5. Formats used in over 1000 records.

Finally, `dc:audience`, the only element collected that could have a built-in educational purpose, is used by two Italian repositories: *Armida* of the University of Milan, and that of the *Scuola Superiore Pubblica Amministrazione Locale* (SSPAL). The former, which can be considered as a learning object repository (LOR), uses `dc:audience` to encode subject codes, subject names, academic years or groups (e.g.: A04-041::2006-2007), generating a considerable amount of combinations. The SSPAL institutional repository, which has a large collection of educational materials, uses `dc:audience` quite differently, applying a more

restricted set of values referring to four types of audience: course attendees, learners, regional and provincial secretaries, and library users (*corsisti, discenti, segretari comunali e provinciali, utenti della biblioteca*).

6.5. Number of learning objects and their identification

Although in theory educational material could be considered a natural type of content for institutional repositories, the results show that in fact they are relatively scarce. Barely 3% of the harvested records (4492, to be precise), were identified as educational resources. Furthermore, only 2910 of them were found by the means of the `dc:type` element, while the rest were selected in subsequent examinations of the records and the repositories.

These learning objects are distributed very unequally among the 47 repositories of the sample. Only 9 repositories contained a considerable and obvious amount of these resources: whether they are LORs, like *Armida* at the University of Milan (Italy) and *TecMinho e-learning Repository* (Portugal), or whether they are institutional repositories with specific learning material communities, e.g. *Bromley University* (UK) or the *University of Barcelona* (Spain). However, more than half of the repositories (24) had an insignificant volume of learning objects, the average for these 24 being approximately 2%. In one of every three repositories studied (14) no learning objects were found as the definition stated here.

The `dc:format` values in learning objects records vary slightly from the general trend. The most common format is also pdf, present in 1451 objects (49.9%), but it is followed by other application formats as octet-stream (22.3%), msword (6.3%), or vnd.ms-powerpoint (1%), as well as digital images formats, with 7% of jpeg.

The localization of the learning objects in each repository was initially based on the analysis of the metadata records collected, observing the element `dc:type`. Various difficulties arose with this task, especially because the use of `dc:type` is frequent but not generalized in the institutional repositories studied (see Table 2), and because the values of the element `dc:type` are extremely heterogeneous. In some cases the inclusion of document type

values through elements like `dc:subject` or `dc:format` was detected.

The most widespread term for identifying educational materials is "Learning Object" (Table 4), but it is common to find different non-standard versions of equivalent terms in English *(educational material, teaching resource, teaching material or training material*) or in the language of the repository *(e.g. materiale didattico, objet d'apprentissage).* With these generic denominations for a learning object a total of 1072 records were found. Furthermore, a total collection of over 1800 records of diverse teaching support material or specific types of learning objects (*Examen, Dispensa, Guia Docente, Esercizi o Soluzione, Lectures,* amongst others) were obtained.

Table 4: Terms used to denote educational content in the institutional repositories analyzed.

Dc:type Value	No. of Records
Educational resource (general term)	1072
Learning Object	621
Interactive Resource	355
Materiale didattico	83
Training Material	7
Objet d'apprentissage / Learning Object	2
Teaching Resource	2
Educational material	1
Learning Material	1
Educational resource (specific term)	1838
Dispensa o Appunti	559
Vorlesungsverzeichnis	336
Inaugural Lecture	228
Dispensa	182
Programma o Bibliografia	119
Lectures	60
Estratti da libri o periodici	56
Farewell Lecture	49
Vorlesung	37
Seminar, speech or other presentation	35
Special lecture	34

Dc:type Value	No. of Records
Educational resource (specific term)	(cont.)
Esercizi o Soluzioni	32
Altri materiali	21
Guía docente	19
Examen	12
Materiali multimediali	12
Manuali e antologie	10
Programma	8
Manual	7
Lectures_Presentation	5
PublicLecture	5
CourseOutline	4
LectureNote	4
Syllabus	3
Training Guide	1
Educational materials selected by alternative methods to **dc:type** values	1582
Prova d'essame	1509
Lecture Note	32
Lecture Presentation	26
Learning Object	12
Presentation	3

Despite this, and due to the heterogeneity of material that can be used in a higher education learning context, particular care was taken to select only those types of objects which would allow

us to check that they actually are learning objects as in our definition. To perform the checking process we resorted to various methods, on the one hand, set analysis (`setName`, to be precise), obtained by consulting OAI ListSets; and on the other hand, direct observation of the repositories, using metadata elements queries, or browsing the indexes and structure.

Selection through `setName` was not very successful as its use is far from standardized throughout the different institutional repositories. This heterogeneity, in most cases, is conditioned by the software used. In DSpace repositories, apart from the sets automatically generated by its system of organization in communities and collections, the creation of new specific sets is almost non-existent. However, EPrints and OPUS repositories usually offer sets by document or resource type. In the latter, `pubtype` sets allowed us to select some additional learning objects non found by `dc:type`, as in the following `setName` values: `pubtype:26`, corresponding to 'LearningObject', `pubtype:25` to 'LectureNote', and `pubtype:97` to 'CourseMaterial'.

Additionally, we performed queries by `dc:type` element, but in some DSpace repositories this function was not activated. Moreover, most of the EPrints and OPUS repositories offered a predefined list of resources types available in the advanced search form, allowing us to select those related to learning objects.

Lastly, we resorted to assisted browsing through the indexes offered by each repository, in those cases where a *type of resource* index existed. Again, this functionality is generalized in the case of repositories based on EPrints (although sometimes this is not displayed), OPUS and Fedora, but is less usual in DSpace repositories.

Finally, we also attempted to identify communities or collections of learning objects by direct browsing through the repository organisation system, with negligible results.

6.6. Metadata schemas and encoding options to display educational content

Obviously, all the repositories analyzed use the DC-S metadata schema with the OAI-PMH protocol (`oai_dc`). In addition to this, the great majority of the institutional repositories analyzed (37 cases, 79%), have only implemented DC-Simple and only display their records in `oai_dc.` Thus, only 21% of cases (10 repositories) use more than one metadata model, ranging from one to five, which in

total means 14 schemas other than `oai_dc`.

The most commonly used schemas are: DIDL**,** MPEG21 (Digital Item Declaration Language); Epicur**,** a schema that allows allocation of a permanent identifier or German URN, provided by the German National Library; and `oai_etdms`, for electronic theses and dissertations (ETDs). As well as `oai_etdms`, several of the schemas added (`uketd_ms`**,** `uppsok` or `XMetaDiss`) are designed to provide a set of metadata elements for ETDs.

We observed absolutely no educational metadata schemas displayed, although we did find one very interesting example, the *Repositorio de Material Educativo*, of the *Universidad Técnica Particular de Loja*, Ecuador [28], which has adapted the DC Namespace of the DSpace repository so that the labels correspond to those of the educational metadata standard IEEE LOM (see record below).

```
<OAI-PMH xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/
http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
<responseDate>2008-05-20T07:58:05Z</responseDate>
<request metadataPrefix="oai_dc" verb="ListRecords">http://eva.utpl.edu.ec:8080/dspace-
oai/request</request>
<ListRecords>
<record>
<header><identifier>oai:eva.utpl.edu.ec:123456789/926</identifier>
<datestamp>2007-11-28T02:09:54Z</datestamp>
<setSpec>hdl_123456789_1182</setSpec>
</header>
<metadata><oai_dc.dcxsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/
http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
<dc:general>103</dc:general>
<dc:general>Concepto de Doctrina Social de la Iglesia</dc:general>
<dc:general>es</dc:general>
<dc:general>Comprenda el concepto de doctrina social de la iglesia.</dc:general>
<dc:general>definición de doctrina social</dc:general>
<dc:general>doctrina social de la iglesia</dc:general>
<dc:general>religión</dc:general>
<dc:general>doctrinas</dc:general>
<dc:general>iglesia</dc:general>
<dc:general>ciencias humanas y religiosas</dc:general>
<dc:lifecycle>autor</dc:lifecycle>
<dc:lifecycle>UTPL</dc:lifecycle>
<dc:lifecycle>2007-11-28T02:09:53Z</dc:lifecycle>
<dc:technical>http://eva.utpl.edu.ec/dspace/handle/123456789/926</dc:technical>
<dc:educational>Animacion</dc:educational>
<dc:educational>Objeto de Aprendizaje</dc:educational>
<dc:date>2007-11-28T02:09:53Z</dc:date>
</oai_dc.dc>
</metadata>
</record>
```

Even though it was impossible to undertake the retrieval of this repository (as it did not fall within the scope of this study), we did discover how the elements of the main IEEE LOM categories were rendered, based on the XML syntax of Dublin Core labels: `dc:general`, `dc:lifecycle`, `dc:educational` or `dc:technical`.

Specific metadata schemas

One of our objectives was to determine whether any institutional repositories had adopted their own metadata model specifically for describing their educational resources. During the direct observation phase, the only example found was the teaching material repository of *AMS Campus,* University of Bologna (Italy). Thanks to the configuration functionalities offered by the GNU EPrints system, the internal metadata schema has been personalized by adding two specific categories of educational interest: education (*insegnamenti*) and teacher (*docente*). Both categories comprise different metadata elements that enable a good level of detail in the description of the institution, qualification, and subject for which the item has been used, and for the identification of the author or person responsible for the material. However, it does not offer specific elements to characterize the object itself, such as learning object type, level of difficulty, learning objectives, learning time, etc.

In addition to *AMS Campus*, nine other repositories using different metadata elements and/or added qualifiers to `oai_dc` elements were identified. Apart from the qualifiers which DSpace includes [29], based on the Dublin Core Library Application Profile (DC-Lib) [30], (e.g. `dc:contributor:author`), most of the refinements added are used to record the institutional origin of the author or contributor (university and/or department); to indicate the type of document and subject or discipline; and even to refer to the title or date of the course for which the object was used. In the case of the Repository of the University of Alicante (Spain), *RUA*, the DSpace metadata input form was simply adapted to their teaching community, with labels like department, subject and subject code, studies in which used, or knowledge area, but the elements remain DCMES or DC-S and are exposed as such for the harvesters (`oai_dc`).

In any case, it cannot be claimed that in the sample studied one or several specific metadata application profiles are used for learning objects. Although in some repositories, such as *Armida@UniMI*, at the University of Milan, the High Schools Local Public Administration (*Scuola Superiore Pubblica Amministrazione Locale)* repository, or the University of Guelph repository in Canada, educational elements or qualifiers have been added (`dc:CourseTitle`, `dc:CourseDate`, `dc:contributor:reader`, `dc:description:teacher`, as well as

dc:audience). This was not exclusively but together with other elements of a different nature and purpose, configuring a tailor-made model for each institution.

The case of doctoral theses and dissertations is different, as, in addition to the repositories that use a metadata schema for ETDs directly (oai_etdms, uketd_ms, xMetaDiss, uppsok), some archives have added specific elements for the description of this kind of material, adapting the oai_dc schema themselves. The Academic Repository of the University of Chile, *Kluedo* at the University of Kässel (Germany), the Repository of the University of Göteborg (Sweden), and the Swedish School of Economics and Business Administration, Hanken (Finland), are some examples.

DC-Simple elements for allocating educational information

In addition to the elements and qualifiers that it has been possible to add to oai_dc, it was found that several of the 15 DC-S elements were being used to assign educational information. In principle, the repositories studied were those with a sufficient volume of LOs to justify the adaptation of the metadata model, i.e. repositories of educational objects and institutional repositories with a significant volume of educational objects. Of the nine repositories which comprised this group, only three - *OpenstarTS*, of the Athenaeum of the *Universitá degli Studi di Trieste* (Italy), Bromley University (UK) and the University of Parma (Italy) - do not use any metadata mechanism (neither DC-S elements to assign educational information, nor the inclusion of qualifiers for the same purpose). The other six repositories have taken advantage of elements like dc:relation, dc:description, dc:contributor, dc:identifier, dc:subject and dc:type, with various qualifiers, to record many different aspects of educational interest. The commonly described characteristics are: firstly, subjects, courses, and qualifications; secondly, the knowledge area, department or institution; and to a lesser level, the type of learning object.

In general, we observed that those repositories with a significant volume of teaching material needed new elements different from oai_dc, which would allow them to allocate this specific information of their educational resources.

Use of vocabulary encoding schemes in the dc:type element

With regard to the content of the metadata elements, it is particularly interesting to determine which vocabularies the institutional repositories analyzed are using to codify their content. Specifically, the dc:type element was analyzed, looking for the use of different vocabularies for types of learning objects.

Only in one case an educational specific vocabulary was used, at the AMS Campus, University of Bologna, with seven different types of LOs: *Altri materiali, Dispensa o Appunti, Esercizi o Soluzioni, Estratti da libri o periodici, Manuali e antologie, Materiali multimediali,* and *Programma o Bibliografia*.

In all the other cases, the values are not used exclusively to refer to educational material, and in many cases, no standardized schema apart from DCMI Type Vocabulary [25] are even adhered to strictly, being very common to use values from different vocabularies, and own-created values mixed with controlled terms from one or more vocabularies. Moreover, as explained in 6.4, the values that designate learning objects are not standardized (see Table 4). This makes achieving semantic interoperability on an international level far less likely.

7. CONCLUSIONS

The quantitative and qualitative analysis of learning objects in open institutional repositories and of the metadata models used to describe them has enabled us to draw a set of conclusions and in some cases to ratify the claims of other works or our own hypotheses.

- Until now, the inclusion of digital learning objects in institutional university repositories has not been particularly widespread. Except in some specific institutional repositories with a clear orientation towards educational resources (like the *Diposit Digital*, University of Barcelona, *ARMIDA@UniMI* or the repository of the SSPAL), the majority do not include or do not sufficiently identify the existence of this type of digital object. Despite the data reflected in the directory OpenDOAR, 1/3 of the repositories studied did not have learning objects (as in our definition) in their collections.

- The localization of these learning objects and consequently the building of value-added services based on this material are far from easy. Despite the potentialities of digital repositories and OAI-PMH for collecting metadata, various limitations make selective retrieval difficult. Some of these limitations, as suggested in this paper, are connected with the limitations of the very software with which the repository was created, with the application of the OAI protocol itself, and with the quality of description with metadata.
- The harvesting of OAI metadata was one of the main methods used to collect data for this study, which obliged us to confront one of the technological challenges of the protocol: the inadequate level of compliance on the part of some data providers. Throughout our research we encountered some common problems [18]: incomplete retrieval, invalid or malformed XML documents, etc. which made it necessary to check the metadata obtained.
- The data providers do not apply the `oai_dc` metadata format strictly, which creates problems with the quality of harvestable OAI metadata. Some bad practices detected include the use of inappropriate elements to present information, or the lack of complete data when the record belongs to a local collection of documents.
- There is great diversity in resource description practices and in the use and interpretation of DC-Simple metadata elements [22], despite the existence of initiatives to minimize this phenomenon [31]. Unqualified DC or DC Simple is a potential source of interpretation problems, given that the OAI data providers have total freedom to enter anything they like in fields like author, publisher or abstract. In addition to the important internal functions of metadata in describing, organizing and storing the resources of an institution, if the OAI-PMH protocol is applied as an interoperability mechanism, its application will have an important effect on the harvesting of metadata and in subsequent services offered.
- According to the results obtained, DC-S proves to be inadequate for the great heterogeneity of content an institutional repository may hold, and corroborates the need to use metadata schemas that provide more detail about specific domain resources. However, the internal use of formats other than `oai_dc` is not common with regard to the description of educational material.

- In some repositories based on DSpace, educational resource-specific metadata elements and qualifiers had been added, but very few. On the contrary, added elements or specific formats for the description of ETDs (e.g. `etd_ms`, `uketd_dc`) are quite common.
- The use of standardized vocabularies and other encoding schemes that guarantee the consistency and quality of the records of a single repository is an underdeveloped aspect of higher educational institutional repositories. In the case of the standardization of object types, this lack of vocabulary makes it difficult to create value-added services by means of filtering material by typology, as well as to research studies like this one.
- Despite the existence of content schemes for `dc:type`, like DCMI Type [25], and the subtype draft for DCMI Type [32], EPRINTS Type [26], or vocabularies of educational resource types [33], like LearningResourceType from IEEE LOM [7], ResourceType from RDN/LTSN [34], and even the NSDL Learning Resource Type Vocabulary [35], they are not used consistently. The different repositories adapt them to suit their own needs, even using their own values to designate the different document typologies in their collections.
- With regard to the checking methods used to correctly identify the learning objects, apart from the retrieval of OAI metadata, various conclusions have also been reached. Firstly, significant heterogeneity was observed in relation with the organization of the repositories, this being an aspect with no standardization. In general, very broad and complex organizational systems and even obscure collection titles were found. This makes the identification of collections of learning items much more difficult. Secondly, the use of sets and indexes is highly conditioned by the functionalities of the different repository tools, and they tend to mirror the same heterogeneity and variability in the generation of collections. Thus, a concept that should have facilitated the selective harvesting of metadata records has ended up by making it more difficult.

REFERENCES

[1] *Open Archives Initiative*. 2008. http://www.openarchives.org

[2] Dublin Core Metadata Initiative. 2008. *Dublin Core Metadata Element Set, ISO 15836*. http://dublincore.org/documents/dces/

[3] Open Archives Initiative. 2002. *Metadata schema for unqualified Dublin Core, `oai_dc` format*. http://www.openarchives.org/OAI/2.0/oai_dc.xsd

[4] Digital Library Production Service, University of Michigan. 2008. *OAIster*. http://www.oaister.org

[5] JISC Digital Repository Wiki. 2008. *SWAP (Scholarly Works Application Profile)*. http://www.ukoln.ac.uk/repositories/digirep/index/Eprints_Application_Profile

[6] Atkins, Anthony, Edward Fox, Robert France, and Hussein Suleiman. 2008. *ETD-MS: an Interoperability Metadata Standard for Electronic Theses and Dissertations*, version 1.00, revision 2. http://www.ndltd.org/standards/metadata/etd-ms-v1.00-rev2.html/

[7] IEEE LTSC committee, WG12. 2002. *IEEE Standard for Learning Object Metadata*. http://ltsc.ieee.org/wg12/par1484-12-1.html

[8] Dublin Core Education Application Profile Task Group. 2008. *DC-Education Application Profile*. http://dublincore.org/educationwiki/DC_2dEducation_20Application_20Profile

[9] Ward, Jewel. 2003. A quantitative analysis of unqualified Dublin Core Metadata Element Set usage within data providers registered with the Open Archives Initiative. In *Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital libraries*, 27-31 May 2003. 315-317. http://portal.acm.org/citation.cfm?id=827140.827196

[10] Dushay, Naomi, and Diane Hillmann. 2003. Analyzing metadata for effective use and re-use. Paper presented at *DC-2003 Dublin Core Conference: Supporting Communities of Discourse and Practice. Metadata Research and Applications*. 28 Sep - 02 Oct, Seattle, USA. http://www.siderean.com/dc2003/501_Paper24.pdf

[11] Efron, Miles. 2007. Metadata Use in OAI-Compliant Institutional Repositories. *Journal of Digital Information*, 8 (2). http://journals.tdl.org/jodi/article/view/196/169

[12] Shreeves, Sarah L., Joanne S. Kaczmarek, and Timothy W. Cole. 2003. Harvesting cultural heritage metadata using the OAI Protocol. *Library Hi Tech*, 21 (2), 159-169.

[13] Najjar, Jehad, Stefaan Ternier, and Erik Duval. 2003. The actual use of metadata in ARIADNE: an empirical analysis. In *Proceedings of the ARIADNE 3rd International Conference*, October 2003, (Duval,

E., ed.), 1-6. http://www.cs.kuleuven.ac.be/~stefaan/papers/ActualUseOfMetadata.pdf

[14] National Science Digital Library (NSDL). 2008. http://nsdl.org/

[15] University of Illinois at Urbana-Champaign. 2003. *Open Archives Initiative Metadata Harvesting Project*. http://oai.grainger.uiuc.edu/projectinfo.htm

[16] Barton, J., S. Currier, and Hey, J. M. N. 2003. Building quality assurance into metadata creation: an analysis based on the learning objects and e-prints communities of practice. In *DC-2003 Dublin Core Conference: Support Communities of Discourse and Practice - Metadata Research and Applications*, 28 Sep - 02 Oct, Seattle, USA. http://eprints.cdlr.strath.ac.uk/2338/01/Barton_201_paper60.pdf

[17] Tennant, Roy. 2004. *Bitter Harvest: Problems & Suggested Solutions for OAI-PMH Data & Service Providers*. California Digital Library, OAI Harvesting Infrastructure Project. http://www.cdlib.org/inside/projects/harvesting/bitter_harvest.html

[18] Chumbe, Santiago, et al. 2006. Overcoming the obstacles of harvesting and searching digital repositories from federated searching toolkits, and embedding them in VLEs. In *Proceedings 2nd International Conference on Computer Science and Information Systems*, 27-30 July, Athens, Greece. http://eprints.rclis.org/6394/

[19] University of Nottingham. 2008. *OpenDOAR (Directory of Open Access Repositories)*. http://www.opendoar.org

[20] OCLC. 2008. *OAIHarvester2 Open Source Software (OSS) project.* http://www.oclc.org/research/software/oai/harvester2.htm

[21] Osborne, Shaun. 2004. *OAI Harvester of the 'Harvesting the Fitzwilliam' (HTF) Project*. http://www.fitzmuseum.cam.ac.uk/projects/htf/docs/oaih.php.txt

[22] DCMI. 2008. *DC /terms/ namespace*. http://dublincore.org/documents/dcmi-ter

[23] National Library of Sweden / LIBRIS Department. 2008. *Uppsök project*. http://uppsok.libris.kb.se/sru/uppsok

[24] Linde, Peter, Carin Björklund, and Aina Svensson. 2005. Putting a National Portal for Undergraduate Theses into Production. In *Proceedings ELPUB2005 Conference on Electronic Publishing,* Kath. June 2005, Leuven, Belgium. http://elpub.scix.net/data/works/att/155elpub2005.content.pdf

[25] DCMI. 2008. *DCMI Type Vocabulary*: http://dublincore.org/documents/dcmi-type-vocabulary/

[26] Repositories Research Team Wiki (DigiRep), JISC Digital Repositories Programme. 2008. *EPRINTS Type Vocabulary*.

http://www.ukoln.ac.uk/repositories/digirep/index/Eprints_Type_Vocabulary_Encoding_Scheme

[27] Freed, N., J. Klensin, and J. Postel. 1996. *Multipurpose Internet Mail Extensions (MIME) Part Two: Media Types, RFC 2046*. http://www.rfc-editor.org/rfc/rfc2046.txt

[28] Universidad Técnica Particular de Loja (UTPL), Ecuador. 2008. *Repositorio de Material Educativo (RME)*. http://eva.utpl.edu.ec:8080/dspace/

[29] DSpace Foundation. 2008. *Metadata.* http://www.dspace.org/index.php?option=com_content&task=view&id=141

[30] DCMI. 2004. Dublin Core Library Application Profile (DC-Lib): http://dublincore.org/documents/library-application-profile/

[31] Repositories Research Team Wiki (DigiRep), JISC Digital Repositories Programme. 2008. *Issues with current use of simple DC:* http://www.ukoln.ac.uk/repositories/digirep/index/Issues_with_current_use_of_simple_DC

[32] IFLA. 2000. *DCT2: Dublin Core Type Vocabulary: Subtypes Working Draft*. http://www.ifla.org/udt/dc8/subtypes.htm

[33] Currier, Sarah, Lorna M. Campbell, and Helen Beetham. 2005. *JISC Pedagogical Vocabularies Project. Report 1, Pedagogical Vocabularies Review*. Final Draft, 23rd December 2005. http://www.jisc.ac.uk/uploaded_documents/PedVocab_VocabsReport_v0p11.doc

[34] Barker, Phil, et al. 2007. *RDN/LTSN Resource Type vocabulary*, Version 1.0 http://www.intute.ac.uk/publications/rdn-ltsn/types/

[35] NSDL Registry. 2006. *NSDL Learning Resource Type Vocabulary*. http://metadataregistry.org/concept/list/page/1/vocabulary_id/11.html