Bueno-de-la-Fuente, G., Hernández-Pérez, T., Rodríguez-Mateos, D., Méndez-Rodríguez, E. y Martín-Galán, B. (2009). Study on the use of metadata for digital learning objects in university institutional repositories. En *Proceedings of the 13th International Conference on Electronic Publishing, ELPUB2009: Rethinking Electronic Publishing: Innovation in Communication Paradigms and Technologies, Milano, Italy*, pp. 587-594.

STUDY ON THE USE OF METADATA FOR DIGITAL LEARNING OBJECTS IN UNIVERSITY INSTITUTIONAL REPOSITORIES

Gema Bueno-de-la-Fuente[1]; Tony Hernández-Pérez[2]; David Rodríguez-Mateos[3]; Eva M. Méndez-Rodríguez[4]; Bonifacio Martín-Galán[5].

[1-5]Dept. Library & Info Science. University Carlos III of Madrid. c/Madrid 126, 28903 Madrid. Spain
e-mail: [1]gbueno@bib.uc3m.es, [2]tony@bib.uc3m.es, [3]pirio@bib.uc3m.es, [4]emendez@bib.uc3m.es, [5]bmartin@bib.uc3m.es

Abstract

This study analyzes the present use of metadata describing the educational resources that some universities include in the collections of their institutional repositories (IRs). The goal is to test the viability of implementing value-added services by offering educational resources from IRs in addition to those available from learning object repositories (LOR), based on their metadata.

We identify and analyze the different metadata models in a sample of university IRs, concentrating on: the use of one or multiple metadata schemas coexisting in the repository; the use of educational metadata schemas and application profiles such as IEEE LOM or DC-Ed; the possible extensions (qualifiers or any refinements) to DC-Simple; the specific metadata elements used to describe educational features (such as audience, type of educational material, learning objectives, etc.) and the values of the metadata elements, especially the use of specific vocabularies for elements of educational interest.

Keywords: Institutional Repositories; Learning Objects; Dublin Core Metadata; Metadata Schemas; OAI-PMH.

1. Introduction

Institutional repositories can house all kinds of content originating from the intellectual production of the institution concerned. Thus, in repositories of higher educational institutions, in addition to typical scientific products (articles, reports,

conference papers, etc.), any sort of resource 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, repository managers confront the difficulty of describing different types of resources that require specific meta-information to identify their particular characteristics.

Generally, Open access digital repositories have implemented the Open Archives Initiative - Protocol for Metadata Harvesting (OAI-PMH) as a mechanism to achieve interoperability in the exchange of metainformation. To do so, OAI-PMH Compliant repositories (or data providers) have to display their metadata records at least in the unqualified Dublin Core metadata schema (DC-Simple, Dublin Core Metadata Element Set, ISO 15836). Moreover, these data providers can describe their resources using any additional metadata schema, even exposing the records in these formats, as long as they are based on XML Schema, or their metadata elements map to those of DC and are displayed in oai_dc format.

This ability to use multiple metadata schemas would be the obvious approach to precisely describe different types of material (scientific, educational, administrative, etc.) with different metadata application schemas or profiles which allow a more accurate description. Thus, in a university IR it would be possible to use the metadata application profile SWAP (Scholarly Works Application Profile) for its collection of articles and preprints, together with ETD-MS for its collection of theses and dissertations, and IEEE LOM or the application profile for education DC-Ed for the educational material in the IR. This would make possible a selective harvesting of different kinds of materials and domain specific metadata formats in order to build value-added services. In the case of educational content, it could facilitate a joint service offering learning resources harvested from IR together with those from specific-purpose learning object repositories (LOR).

However, the success of these efforts highly depends on various aspects related to the quality of metadata records (e.g: the correct identification of resource types by means of metadata elements as dc:type, or the use of domain specific metadata schema); the options for selective harvesting (e.g.: creation of sets of different resource types); and the harvesters' functionalities (e.g: collection of metadata records in various schemas).

The aim of this study is to analyze how metadata for the description of digital learning materials are currently being used in different institutional repositories in higher education institutions 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 different to DC-S 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-S elements have been refined by means of element qualifiers.

2. Methodology

To undertake the metadata analysis we selected a sample of institutional repositories holding digital learning objects in their collections from the repository directory OpenDOAR, which allows us to select repository type (institutional) and type of content (learning objects). Of the 1128 repositories registered by OpenDOAR as of 22 April 2008, 124 fulfilled both conditions. Some selection filters were applied to this subset of IR: by language (excluding Asian and Cyrillic languages), by software (choosing only the most used options on a global level, DSpace, GNU EPrints, Fedora and Opus); and a natural filter related to the technical problems arising during the harvesting (several repositories could not be entirely harvested for various technical reasons), thus selecting only those repositories with 100% obtainable records. Taking all these criteria into account, the final sample was reduced to 47 repositories.

The harvesting was performed using the OAIHarvester2 tool, developed by OCLC. The tool was configured only to use "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 IR in a single XML file. The XML files were processed through two XSLT stylesheets to obtain an HTML document in tabular form. The tables were later transferred into spreadsheets, which are useful for doing quantitative studies and for reviewing the content of the records.

Together with the harvesting of metadata, some other data collection methods were performed, such as the direct observation of metadata records and content browsing, and the verification of the multiple metadata formats, if used, of each IR.

3. Results and discussion

In accordance with the methodology explained above, 47 repositories from 18 countries and 141,883 metadata records were studied. The main language in 75% of them was English. By software used, 72% used Dspace, 17% E-Prints, 9% Opus and 2% Fedora.

Element	No. records using element	% Records using element	No. IR using element	% IR 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%

Table 1. Usage of DC-S metadata elements in the repositories and records analyzed.

Based on these data (summarized in Table 1) it is possible to define three levels of usage for the DC-S metadata elements in the IRs harvested, following a distribution similar to that found by other DC-S metadata use studies [1]:

- Generalized usage: elements used in 98-100% cases (dc.date, dc.title and dc.identifier);
- Frequent usage: those used in 65-90% of the records (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 (dc.contributor, dc.relation, dc.rights, dc.source and dc.coverage).

Six non oai_dc elements were found, which were used in just one single repository (except dc.audience, found in two). In total, these elements appear in approximately 10,000 records. The usage of these elements is representative within the repository that has added them but not particularly significant with respect to the total records harvested (141,883).

The discovery of 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, and because the values of the element dc.type are extremely heterogeneous, even for learning resources (*Learning Object; Interactive Resource; Materiale didattico; Training Material; Objet d'apprentissage; Teaching Resource; Educational material; Learning Material; Dispensa o Appunti; Vorlesungsverzeichnis; Lectures; Farewell Lecture; Vorlesung; Seminar, speech or other presentation; Special lecture*, etc.)

These learning materials are distributed very unequally among the 47 repositories of the sample. Only 9 repositories contained a considerable and obvious amount of LOs: whether they are learning object repositories (LORs) like Armida, at the University of Milan (Italy), or the TecMinho e-learning Repository (Portugal); or whether they are institutional repositories with specific teaching material communities, e.g. Bromley University (UK) or the University of Barcelona (Spain). More than half of the repositories (24) had an insignificant volume of LOs, the average for these 24 being approximately 2%. Finally, in one out of every three repositories studied (14) no learning objects were identified apart from the students' studies, works and theses.

4. Conclusions

Several interesting conclusions have been drawn from this study, with the following most relevant ones:

- The inclusion of digital learning objects in university IRs has not been particularly widespread. Scarcely 4,492 learning objects (about 3% of the total harvested records) were identified, of which only 2,910 could properly be called dc.type educational material, while the rest were found in subsequent examinations of the records and the repositories. Despite the data reflected in the directory OpenDOAR, 1/3 of the repositories studied did not have learning objects in their collections;
- The discovery of these learning objects and consequently the building of value-added services based on this material are far from easy. Despite the potential of

digital repositories and OAI-PMH for collecting metadata, various limitations make selective retrieval difficult. These limitations are connected with the software used to create the repository, the application of the OAI protocol, and the quality of metadata records;

- The harvesting of OAI metadata was one of the main methods used for data collection, which presented a major technological challenge: the inadequate level of protocol compliance on the part of the metadata providers. Throughout our research we encountered some of the more common problems [2]: incomplete retrieval, invalid or malformed XML documents, etc. which made it necessary to manually check the metadata obtained;
- The single use of DC-S has been found inadequate for the great heterogeneity of content that an IR may hold, for example the learning objects analyzed in this study. This corroborates the need to use metadata schemas that provide more detail about the resources than oai_dc elements;
- The great majority of the IRs analyzed (37 cases, 79%), have only implemented DC Simple and only displayed their records in oai_dc. Only 10 repositories (21%) use more than one metadata model, offering in total 14 schemas other than oai_dc and none of them are educational metadata schemas. Quite different is the case of theses and dissertations specific metadata models: together with oai_etdms (used in repositories worldwide), several schemas are used locally (uketd_ms, uppsok or XMetaDiss) for describing the specific characteristics of this type of digital object;
- The use of controlled vocabularies to assign values to the dc.type element that could ease the identification of learning objects is still underdeveloped and not fully standardized. Despite the existence of content schemes for this element, like DCMI Type, and the subtype draft for DCMI Type, EPRINTS Type, or vocabularies of educational resource types, like LearningResourceType from IEEE LOM, ResourceType from RDN/LTSN, and even the NSDL Learning Resource Type Vocabulary, 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 their collections hold.

Notes and references

[1] WARD, J. 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, p. 315-317. Available at: http://portal.acm.org/citation.cfm?id=827140.827196 (March 2009).

[2] CHUMBE, S. et al. Overcoming the obstacles of harvesting and searching dig-

ital repositories from federated searching toolkits, and embedding them in VLEs. In *Proceedings of the 2nd International Conference on Computer Science and Information Systems*, 27-30 July 2006, Athens, Greece. Available at: http://eprints.rclis.org/6394/ (March 2009)

June 2009
Printed on demand
by *"Nuova Cultura"*
www.nuovacultura.it

Book orders: ordini@nuovacultura.it