Fourth International Workshop on Search and Exchange of e-le@rning Materials (SE@M'10)

Workshop Proceedings

27-28 September 2010, Barcelona, Spain

Table of Contents

Fourth International Workshop on Search and Exchange of e-le@rning Materials (SE@M'10)

Elena Shulman[1]

[1] European Schoolnet (EUN), Brussels, Belgium

Context and Objectives

The SE@M'10 workshop was held on September 27-28, 2010 in conjunction with the Fifth European Conference on Technology Enhanced Learning (EC-TEL'10) in Barcelona, Spain.

Over the last fifteen years, considerable effort has been spent on the development of standards and specifications in order to improve the interoperability of e-learning systems, repositories, and content. These efforts have led to significant improvements in the arena of technical interoperability enabling the emergence and expansion of successful federations and alliances of learning object repositories such as the LRE, GLOBE, etc. Building blocks for this success have been the creation, evolution and adoption of standards such as the IEEE Learning Object Metadata (LOM), the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), and the Simple Query Interface (SQI). More recent developments in this field include the definition of a Simple Publishing Interface (SPI) and the ongoing work within the IMS Global Learning Consortium on the Learning Object Discovery & Exchange (LODE) specification to facilitate the discovery and retrieval of learning objects stored across more than one collection. The development of best practices (for example by projects like ASPECT and ICOPER), enabling more efficient use of these standards and specifications, has also proven instrumental in the successes of current learning content retrieval infrastructures.

To further developments in this field, the main goal of this international workshop was to offer a forum where researchers and practitioners discussed theoretical aspects, open issues, and innovative approaches and shared the latest advances in the state of the art and practices for exchanging and describing learning content.

Program

Among the 12 papers submitted to the workshop, 11 (6 full papers and 5 short papers) were accepted by the SE@M'10 program committee. The papers were presented in the course of the two day workshop. Two keynote speakers opened and closed the workshop.

Papers explored innovative approaches and technical solutions for the management of learning object repositories, data models for efficiently organizing multiple metadata standards, content packaging, digital curation, keyword extraction, user generated metadata and a series of issues encountered in federated metadata management.

Organizers

David Massart and Elena Shulman, European Schoolnet (Belgium)

Program Committee

Vladimir Batagelj, University of Ljubljana, Slovenia
Jean-Noel Colin, University of Namur, Belgium
Ingo Dahn, University of Koblenz, Germany
Erik Duval, Katholieke Universiteit Leuven, Belgium
Joris Klerkx, Katholieke Universiteit Leuven, Belgium
Manuel Kolp, Universite catholique de Louvain, Belgium
Robert Kristoefl, BMUKK, Austria
Eugenijus Kurilovas, ITC, Centre of Information Technologies of Education, Lithuania
Tien-Dung Le, European Schoolnet (EUN), Belgium
Nikos Manouselis, Greek Research & Technology Network, Greece
Jon Mason, Queensland University of Technology, Australia
Nick Nicholas, Australian National Data Service (ANDS), Australia
Tomasz Orzechowski, AGH University of Science and Technology, Poland
Fredrik Paulsson, Umea University, Sweden
Jan Pawlowski, University of Jyväskylä, Finland
Alain Pirotte, Universite catholique de Louvain, Belgium
Daniel Rehak, ADL, USA
Griff Richards, Athabasca University, Canada
Bernd Simon, WU-Wien, Austria
Stefaan Ternier, Open Universiteit, The Netherlands
Frans Van Assche, Ariadne Foundation, Belgium
Riina Vuorikari, European Schoolnet, Belgium
Nigel Ward, University of Queensland, Australia
Tsuneo Yamada, Open University, Japan

Repository Services for Outcome-based Learning

Michael Totschnig[1], Michael Derntl[2], Israel Gutiérrez[3], Jad Najjar[1],
Roland Klemke[4], Joris Klerkx[5], Erik Duval[5], Franz Müller[1]

[1] Vienna University of Economics and Business
[2] University of Vienna, Faculty of Computer Science
[3] Universidad Carlos III de Madrid
[4] Open University of the Netherlands
[5] Katholieke Universiteit Leuven

`michael.totschnig@wu.ac.at`

Abstract. Despite the existence of numerous standards and specifications in technology-enhanced learning, there is a lack of interoperability of artifacts and services throughout the whole lifecycle of outcome-based education. In this paper we present the concept and prototypical implementation of an open architecture that aims to remedy these issues by providing a unified metadata and service layer for making key educational resources sharable, storable, findable, and interoperable. The reference model and its supporting technology architecture are tested by a family of prototypes implemented as extensions to or adaptations of existing mainstream systems like Moodle, .LRN, Elgg and Facebook.

Keywords: outcome-oriented education, learning object repository, middle layer

1   Introduction

Conceiving services that make learning resources usable for design and delivery of outcome-based learning is a challenge that current learning object repositories are not yet able to meet. On the one hand, learning designs are dissociated from the learning opportunities (course offerings) where they have been or will be put into context, and thus it is difficult for the users (both the learner and the teacher) of those repositories to understand how they can benefit from these resources in the most fruitful way. On the other hand, learning outcomes are not yet defined and linked in a systematic way to learning designs, and even if they were, information about the learning or teaching history of the user would be needed in a standardized format, if the system was supposed to make meaningful suggestions.

The *ICOPER Reference Model* (IRM) aims at providing a framework through which innovative learning processes that exploit rich linkages between teaching methods, learning designs, learner assessments, learning resources, learning outcome definitions, user profiles for achieved learning outcomes and for learning needs, and learning opportunities can be stored, shared and delivered through standardized services and data formats. The *Open ICOPER Content Space* (OICS) is conceived as the testbed for implementing the IRM. Figure 1 displays the key processes in designing

and delivering outcome-based education. From a pedagogical perspective, the alignment of learning outcomes with assessment methods and teaching methods is one of the core foundations of the Bologna Process [6]. Implementing this foundation, the IRM as a reference model and the OICS as its reference implementation represent a visionary yet realistic approach to fully (i.e. conceptually and technically) supporting key pedagogical processes. These include the definition of intended learning outcomes; the reuse and creation of teaching methods and learning designs; the selection of content; and the offering and delivery of concrete learning opportunities. Going beyond current pedagogical and institutional practice, this approach even enables management of learner assessments, including verification and certification of learners' achievement of intended learning outcomes; publishing of learning outcome profiles; and institutional exchange of achievement data (between and among universities and companies).

Fig. 1. Key processes in outcome-based education.

In this paper, we present the architecture of the services that the OICS offers to client applications. These services draw upon existing specifications for content aggregation, federated search and publication that have been validated in numerous projects and organizations such as MACE, MELT, GLOBE, EducaNext and ARIADNE. All of those have been integrated into a middle layer API that provides easy-to-use services for the support of outcome based learning and teaching scenarios.

After describing the conceptual and data models defined in the IRM, we explain how these models are made available through the OICS services, and how these services are consumed by the prototypical tools developed in the ICOPER project.

2 OICS Architecture

The OICS infrastructure is built as a composition of services developed and maintained at Vienna University of Economics and Business (WU) and K.U. Leuven (KUL). At both sides a learning object repository is used for storing metadata har-

vested from the ICOPER content providers. Metadata is synchronized between them through the Simple Publishing Interface (SPI) protocol

Content is fed into the OICS either through the OAI-PMH protocol or a publication service based on the SPI protocol[1]. The ARIADNE harvester [2] enables the management of OAI-PMH targets and is integrated with a set of services improving the management of the repositories and the quality of the metadata:

- The registry service provides a catalogue of up-to-date information about learning object repositories (LORs) and allows the harvester instance used for the OICS to retrieve information about the OAI-PMH endpoint.
- The validation service ensures that only metadata records are stored which comply with the ICOPER LOM Application Profile (see Section 3.1), which is based on IEEE Learning Object Metadata (LOM).
- The transformation service allows applying mappings between foreign vocabularies to ICOPER specific ones.
- The identifier service generates unique and persistent identifiers that are added to metadata records upon ingestion into the OICS. Resolution to multiple views of the resource is provided through a simple web service.
  These services are explained in more detail in [2].

2.1   OICS Middle Layer

The OICS middle layer tries to bundle the requirements of the IRM together in a coherent API that is accessible from a variety of systems and tools (see Figure 2). Its key focus is the integration of concepts and data related to the key processes in outcome-based education. The OICS middle layer provides services for search and retrieval of learning resources, for publication, for the management of users and groups and for the management of learning outcome profiles within these key processes.

Search and Retrieval. The search and retrieval service gives access to the OICS resources by providing specific access methods for the different types of objects (learning outcome definitions, teaching methods, learning designs and learner assessments). Three example bindings for this service have been implemented:

- The ATOM binding exposes all resources as ATOM feeds that can be filtered based on values in the LOM metadata.
- The JSON binding uses a REST interface and provides the results to the client tools in the JavaScript Object Notation data format. JSON is a lightweight data format heavily used by web developers due to its simplicity (e.g. native evaluation of results in JavaScript) compared to the traditional XML data format approaches, which often require cumbersome DOM-based processing.
- A PHP search script forwards PLQL [3] expressions to the SQI SOAP end point provided by the KUL repository.

Publication. The OICS implements the Sword/AtomPub binding of the SPI - specification, learning objects and metadata records can be published to collections. Extending the SPI specification, The OICS implementation allows updating and re-

---

[1] For metadata editing and demonstration purposes the OICS also provides a web UI.

trieving of parts of the metadata record thus making it very easy to query and manipulate individual metadata fields relevant in specific use cases.

Fig. 2. OICS middle layer

Learning Outcome Profiles. Users of the OICS can manage their learning outcome profile according to the Personal Achieved Learning Outcomes (PALO) data model (see Section 3.3) both through a web UI and the same publication service as used for publishing learning resources.

3   Conceptual and Data Models of the ICOPER Reference Model

The OICS as a repository managing shareable educational resources implements the following data models: (1) the ICOPER LOM Application Profile (AP) as a unified metadata layer above learning designs, teaching methods, assessment methods, learner assessments and learning content, (2) an IEEE RCD-based specification for representing learning outcome definitions (LOD) and (3) the Personal Achieved Learning Outcomes (PALO) specification [1] for learning outcome profiles.

3.1 ICOPER LOM Application Profile

A metadata schema is needed to describe and store resources in a referatory. As one of the most widely used, supported, and implemented standards, IEEE LOM was profiled to enable the description of learning design resources (teaching methods, learning designs, assessment methods, and learner assessments), but care was taken to make the same profile also applicable to other types of learning resources stored in the OICS. The resulting ICOPER LOM AP ensures that resources described using different standards and specifications like IMS Learning Design (LD), IMS Question and Test Interoperability (QTI), and so forth, become semantically interoperable.

As one central feature, the AP allows defining intended learning outcomes for all resources in the OICS; to enable this, the Educational category of LOM was extended with an element containing a link to a learning outcome definition. This simple extension enables several added-value use cases and scenarios, as indicated and prototypically demonstrated in Section 6.

3.2 Learning Outcome Definitions (LOD)

The Learning Outcome Definitions (LOD) data model defines a conceptual base schema for describing and sharing learning outcome definitions in the context of online and technology enhanced learning. The data model provides a way to capture the key characteristics of a learning outcome, independently of its use in any particular context or target group. This model should enable the storage and retrieval of learning outcomes across learning systems that deal with learning outcomes data.
This specification is based on, and is an application profile of, the IEEE Reusable Competency Definitions (RCD) standard. EEE RCD is the only widely accepted standard for describing competencies. It is a continuation (and replacement) of the early efforts on the development of IMS Reusable Definition of Competency or Educational Objective (RDCEO). This specification profiles IEEE RCD with one metadata element defining the type of the learning outcome and its associated value domain to capture whether a learning outcome refers to knowledge, skill or competence following the definitions of the European Qualification Framework [4]. Instances that conform to the LOD specification also conform to the IEEE RCD standard, which ensures interoperability between the OICS and other systems.

3.3 Personal Achieved Learning Outcomes (PALO)

The Personal Achieved Learning Outcomes (PALO) data model [1] is a simple schema proposed to capture information on knowledge, skills and competences achieved by a learner and relations between those outcomes. Information on the context where the learning outcomes are obtained or applied, evidence records and levels (e.g. proficiency level) associated to the outcomes are also part of this schema.

One of the main challenges of communities and systems that deal with learning outcome information is interoperability. Different communities and systems may use different data models to represent information on skills, knowledge or competence obtained by a person that is required for a job or a task. The PALO specification is a step towards a common model supporting the exchange of such data, to enhance inte-

roperability of personal learning outcome information between, for example, learning management systems, e-portfolios, social applications and recruitment systems.

This data model enables describing relations between learning outcomes of learners, in addition to contextual and evidence related information. The PALO schema should enable capturing the following:

- Relations between achieved learning outcomes, regardless of the taxonomies or ontologies they belong to;
- Contextual information on where the achieved learning outcome is obtained or applied;
- Information about all types of evidence and assessment that prove the achievement of a learning outcome;
- Information about levels and ranking of an achieved learning outcome, like proficiency level.

The PALO model has been proposed as EU specification for capturing data of personal achieved learning outcomes at CEN Workshop on Learning Technologies[2].

4 Linking Learning Outcomes, Teaching Methods, Learning Opportunities and Learning Designs

The following hypothetical scenario illustrates how university administration, faculty and learners can benefit from making the systems that manage teaching and learning processes interoperable through the adoption of the services described in the IRM and provided by the OICS.

1. At University X, a new curriculum for a course program is developed. Each course description is linked to learning outcome definitions (LOD) and suggested teaching methods (TM), both stored in the OICS.
2. Best practices for TMs have been elaborated by a consortium of universities, and for each of them a template has been elaborated that can be used by instructional designers.
3. An instructional designer creates a new learning design (LD) for the program. He retrieves the LODs mentioned in the curriculum from the LOD repository and searches the TM repository for a suitable template including assessment methods. He imports it into his authoring environment, adds resources and learner assessments retrieved from the OICS. Since learner assessments are linked to learning outcomes and assessment methods, he is able to retrieve the most relevant resources. The LD is made available for feedback in a restricted collection.
4. Additional links to LODs and TMs can be added by program management.
5. Once the LD has been approved by the program management, the LD is published.
6. It is imported into the institutional LMS, and automatically a learning opportunity is pushed to a registry service for learning opportunities.
7. Upon each completion of the learning opportunity, feedback from teachers and learners is collected and the metadata is enriched.

---

[2] See http://www.cen.eu/cen/Sectors/Sectors/ISSS/Activity/Pages/WSLT.aspx

8. Learners completing the learning opportunity, including the successful finishing of learner assessment (LA), have their learning outcome profile augmented with entries for each achieved learning outcome. The achieved learning outcomes are evidenced by assessment records (AR), results of the assessment process.
9. Prospective learners that use the OICS LOD repository for identifying learning outcomes that correspond to their learning needs, will be able to retrieve other learners that already have achieved these outcomes, and since their profile also links to units of learning and learning opportunities, relevant recommendations can be presented to the learner.

The different parts of this high level scenario have been implemented by partners of the ICOPER consortium in various prototypes that make use of the OICS repository services.. In the following, we describe for each of the processes in Figure 1 how they are realized in some of the prototypes.

4.1   Learning Outcomes - Definitions

Learning outcomes can be defined for a single course, taught by several teachers, or be standardized across universities or whole domains. Instead of describing learning outcomes from scratch every time a new course is created, instructional designers should be provided with a list of relevant learning outcome definitions that they can link to their courses. For example, IEEE and ACM Computer Science task force has defined the list of learning outcomes for all computer science courses. These learning outcomes have been imported into the OICS and can be reused across different course and universities. In this way, different universities use common learning outcome terms for describing what students would achieve by finishing a specific course.

The OICS also provides a simple-to-use web interface for creating and editing them. A browsing interface is currently implemented for the ICOPER public website. New definitions can also be created from within several authoring environments and LMS and are submitted to the OICS through the publication service.

4.2   Teaching Methods and Learning Design: OpenGLM

OpenGLM (short for Open Graphical Learning Modeler) is an open-source learning design editor, developed at Universität Wien. It is able to manipulate learning designs compliant with the IMS Learning Design specification. It was built to visually support the creation and reuse of teaching methods and learning designs. OpenGLM uses the search and retrieval services and the publishing service of the OICS Middle Layer; it uses those parts of the services that are related to retrieving, searching, editing, enriching and publishing learning designs (packaged as IMS LD units of learning), teaching methods and learning outcomes. OpenGLM is the key prototype supporting use cases in the "Teaching Methods and Learning Design" process in Figure 1.

4.3   Learning Content: OICS Roundtrip Authoring Re-use

Open University of the Netherlands and Humance AG have implemented an OICS roundtrip authoring re-use prototype, that brings together an asset management soft-

ware (MediaLibrary), an authoring environment (author42), and an LMS (OLAT). This prototype addresses the authoring and creation of units of learning with strong support for re-use of existing materials. It focuses on three main processes:

1. The collaborative collection and organization of media assets.
2. The collaborative creation of learning units based on these media assets.
3. The preparation for re-use of media assets and learning units.

The prototype has several connections to the OICS: through OAI-PMH targets individual media assets (MediaLibrary) as well as units of learning (author42, OLAT) are made available to the OICS infrastructure together with their LOM-based metadata. The OICS search services are integrated in author42 to enable seamless integration of existing media in the authoring process. More details on this prototype can be found in an accompanying paper [5].

4.4 Learning Opportunities

Knowledge Markets Consulting Ges.m.b.H. is implementing a prototype on top of its 2know2 platform, that allows to announce learning opportunities for courses that are linked to learning outcomes and teaching methods. The learning outcomes and teaching methods are directly stored in and retrieved from the OICS via the search, retrieval and publishing services of the OICS Middle Layer. New learning opportunities can be announced at the 2know2 platform with a news article and an RSS feed and can also be published in the repository for learning opportunities at the OICS.

4.5 Learner Assessment

University Carlos III of Madrid (UC3M) has developed a prototype integrated into the .LRN platform that demonstrates various use cases related to the learner assessment process:

(1) Outcome-based search of assessment resources. Once defined assessment method and learning outcomes of a course, the instructional designer should find appropriate assessment resources aligned to them, and this process is enabled by the .LRN module. The designer is able to search for assessment resources in the OICS by keyword, but he is also able to filter the results by the intended learning outcomes of the course (available in the learning outcome definitions repository) and by the assessment method he decided to use to orchestrate the resources.

The described use case makes use of the search service of the middle layer API of the OICS in order to search assessment resources from several repositories. This service also provides means for the filtering process by learning outcomes and/or assessment method.

(2) Sharing annotations about assessment resources. Once an activity involving a published assessment resource has finished, teachers can make use of the annotation system in order to provide information about the students' performance. Teachers could also collect students' feedback to annotate the assessment resource with. This information enriches the published resource for potential instructional designers willing to re-use it.

This sharing annotation system uses some services of the OICS via the middle layer API. Firstly, the publication service is used to publish the resources and their LOM metadata in the OICS. Secondly, the service that allows updating the metadata record of a resource is used to update it with annotation information.

4.6 Learning Outcomes – Achievements

UC3M's prototype also takes care of the publication of the achieved learning outcomes into a learner learning outcome profile on the OICS.

After the completion of the course, and therefore the assessment activities, it is time for teachers to officially close the course. The implemented application provides the teacher with an interface to facilitate this task. It shows a list of the students of the course and the assessment result (grade) of each assessment activity carried out in the course. There is also a final grade automatically calculated as the arithmetic mean of all the results, but the teacher can modify it taking into account other factors. The teacher can also provide some textual feedback for any student's final grade.

Once the teacher has finished this task, he can officially close the course. The results of this action is that the profiles of all the students passing the course (in this case it means a final grade of 5 or more because the prototype uses a grading scale 0 to 10, proper to the Spanish system) will be updated with the achievements of the course, that is, the learning outcomes achieved. These achievements are also evidenced by an assessment record, which is an official record corresponding to the final grade of the course and has the University as the assessing body.

This process uses the service of the OICS middle layer that allows updating the learner's profile with assessment records and achievements.

Other tools and environments have been integrated with OICS services, and are able to publish achievements into a learner's profile or display them:

- Umeå University, Sweden, has developed a Moodle block that imports learning designs into a course and exports learning outcomes for students enrolled in it.
- IMC, Germany, has extended its LMS Clix with functionality for the execution of outcome-based learning designs as well as the management of learners' PALO.
- eXact learning solutions (formerly Giunti Labs), Italy, has implemented access to OICS' learning outcome related features into its learn eXact enterprise LCMS.
- AGH – University of Science and Technology, Poland, has integrated the display of PALO profiles into Facebook.
- Tallinn University, Estonia, has integrated the e-portfolio environment, Elgg, with OICS services for searching learning resources and for publication of PALO data.

5 Conclusions and Future Directions

With the services we have designed and implemented for the OICS, we want to make learning resources available in the contexts where outcome related education takes place: LMS, personal learning environments, social networks. By providing richer linkages between learning needs, learning designs, teaching methods, learning outcomes and learning opportunities, teachers and learners will be able to make more

innovative use of available learning objects. The OICS is also designed to store information about user's experiences as annotations, and thus to create still more opportunities to discover relevant resources.

In order to provide a good practice of enabling the design and delivery of outcome based learning, prototypes implemented in ICOPER project extended the functionality of existing systems that are already used by teachers and learners.

Currently, the implemented prototypes are being evaluated by target end users like learners, teachers and instructional designers. The goal of the evaluation is to determine:

- Ease of use, usefulness and completeness of implemented outcomes based functionalities;
- Added value and innovation of introduced functionalities to target users;
- The extent that such applications are being adopted by universities
- Interoperability level of data exchanged using the specifications adopted and implemented in this project.

References

1. Najjar, J., Derntl, M., Klobučar, T., Simon, B., Totschnig, M., Grant, S., Pawlowski, J.: A Data Model for Describing and Exchanging Personal Achieved Learning Outcomes (PALO), International Journal of IT Standards and Standardization Research (IJITSR), vol. 8, issue 2 (2010)
2. Klerkx, J., Vandeputte, B., Parra, G., Santos, J.L., Van Assche, F., Duval, E.: How to Share and Reuse Learning Resources: the ARIADNE Experience, European Conference on Technology Enhanced Learning (ECTEL) (2010)
3. Ternier, S., Massart, D., Campi, A., Guinea, S., Ceri, S., Duval, E.: Interoperability for Searching Learning Object Repositories. The ProLearn Query Language. In: D-Lib Magazine. vol. 14, nr. 1/2 (2008). http://www.dlib.org/dlib/january08/ceri/01ceri.html
4. EQF. The EQF for lifelong learning, Office for the publication of the EC, ISBN 978-92-79-0847-4 (2000)
5. Klemke, R., Schmitz, B.: Authoring for re-use in outcome-oriented learning scenarios. Se@m 2010, the 4th International Workshop on Search and Exchange of e-le@rning Materials, Barcelona, Spain (2010)
6. European Commission: The Bologna Declaration (2000). http://ec.europa.eu/education/policies/educ/bologna/bologna.pdf

The Ariadne Registry of LORs

Jose Luis Santos, Joris Klerkx, and Erik Duval

joseluis.santos, joris.klerkx, erik.duval
@cs.kuleuven.be
Dept. Computerwetenschappen
K.U.Leuven
Celestijnenlaan 200A
B-3000 Leuven
Belgium

Abstract. The ARIADNE registry is one of the core components in an architecture that promotes interoperability of networks of repositories that facilitates the access to the learning content and encouraging the share and reuse of digital content. This paper explains the development of the ARIADNE registry of learning object repositories (LORs) and the role played by standards and specifications.

1 Introduction

One of the problems of Technology Enhanced Learning (TEL) is that creating Learning Objects (LOs) is expensive and time-consuming [1]. Sharing is one of the possibilities to address this problem. When content creators search for materials to reuse, they typically do not care about where the resource is located, but want to find the best quality materials that satisfy their needs [2].

LOs are typically stored in Learning Objects Repositories (LORs). In ARIADNE, considerable effort has been spent on the development of standards and specifications for LORs [3], including IEEE LOM [4], SQI [5], SPI [6], OAI-PMH [7] and PLQL [8]. These allow effective share and reuse of LOs between different LORs and networks of LORs.

One of the problems for managing all these repositories within a network is the scalability. Currently, every repository is added manuallyin the harvester or in the federated search layer service. However, this time consuming process requires that one person configures the targets and its parameters such as the requirements (e.g. query languages for querying or metadata formats for harvesting).

In addition, some extra information about the content inside of the repositories can be useful for managing this architecture. For instance, if we know that one repository contains LOs focused on mathematics, and we are interested in them, we can select this target to be harvested. To enable the sharing and exchange of this information with other networks of repositories, the information needs to be structured and managed.

This paper focuses on how the ARIADNE implements the ARIADNE Registry that has been integrated in the GLOBE architecture [9] and how ARIADNE addresses the problems described above by:

1. using a reference implementation developed for the ASPECT project [10] based on the concept of content collection. This content collection contains relevant information about the content of the collections and technical information for accessing them;
2. creating a registry for managing this information;
3. using standards and specifications for increasing the interoperability within networks of the repositories;
4. creating a network of registries in order to exchange the information between them and to be able to access the LOs of other networks. In this way, all the modifications done in a network can be widespread through all the networks of the registries.

The paper is organized as follows: section 2 which shows a possible use case of the approach. Section 3 introduces an explanation of the Registry Data Model co-developed in ASPECT project. Section 4 explains the Registry Architecture. Some statistics and data are presented in section 5.

2 Use Case. ARIADNE Foundation integrates the ARIADNE registry in its infrastructure

This section focuses on a use case for the ARIADNE registry and discusses: (i) the integration of the ARIADNE Registry in existing architectures and (ii) increasing the collaboration between different institutions for exposing their LOs. Also, we introduce some technical details about the implementation. The final goal is to explain the use of the registry in a non formal way.

ARIADNE Foundation has several repositories where the LOs are described by Learning Object Metadata. They expose the metadata using SQI and OAI-PMH. However, they have noticed that the number of repositories is increasing and it's difficult to manage all the information from other content providers. These content providers describe their LOs using different specifications like LOM and Dublin Core(DC), on the other hand some content providers expose the LOs using OAI-PMH and/or SQI. Looking for a good solution, they decide to integrate the ARIADNE Registry to manage this information because:

1. The ARIADNE Registry allows the definition of LORs using IMS LODE. That is open in terms of using specifications.
2. The ARIADNE Registry exposes the information using SQI, RSS and OAI-PMH. SQI allows to query the registry. RSS alerts subscribers when a new target is added. OAI-PMH allows to harvest all information from the ARIADNE Registry. These three specifications allow to integrate their ARIADNE Federated Search Layer which queries different repositories to obtain different LOs from different repositories, the ARIADNE harvester which

harvests metadata from different repositories and to build a federation of registries which allows to collaborate with different institutions.

The benefits that they obtain from the integration of the ARIADNE Registry are:

1. They have centralized all the information from all the repositories where they harvest from. It saves time in terms of administration tasks.
2. They can implement other services on top of the registry to check the availability of the different services. This service allows that other services can check this information before trying to access the targets. This information optimizes the performance behavior of these services.
3. The integration of the registry allows that new targets added are widespread by different network of LORs which are federated with ARIADNE network.

3   ARIADNE Registry Data model

The ASPECT project [10] has co-developed an application profile of IMS LODE[11]. It is based on IMS Dublin Core Collections Application Profile specification and complemented with ISO 2146 and IEEE LOM [12]. This specification uses the concept of content collections. A content collection is defined as a group of digital content which is exposed to the world through some protocols based on standards or specifications.

ARIADNE has chosen this specification, because it is not restrictive in terms of use of specifications, and it increases the possibilities of interoperatibility between architectures. In addition, the model does not restrict how the content collections are created. This is an important issue, because the content providers can choose how they create them and can offer metadata information about the collections that they are interested in.

This model is represented by a schema that contains three main elements Content Collection, Metadata Collection and Protocol.

1. Content Collection contains information about access rights, authoring, title description, keywords, etc. This is information about the content itself.
2. Metadata Collection contains information about how the metadata is exposed. Here, the content provider has to define which specification are used to expose metadata. This part of the specification has an element for defining the specification called Protocol Implementation Description which is used for defining extra information like the query language supported by an SQI interface or sets supported by an OAI-PMH interface.
3. Protocol contains specific information about the specification used like the URL of the schema, namespace or the binding location.

Several examples can be found at the ARIADNE Registry site [13].

4 ARIADNE Registry Architecture

The implementation of the registry enables ARIADNE to build a federation of registries, to provide access to collection information using SQI and to publish new content collections using SPI.
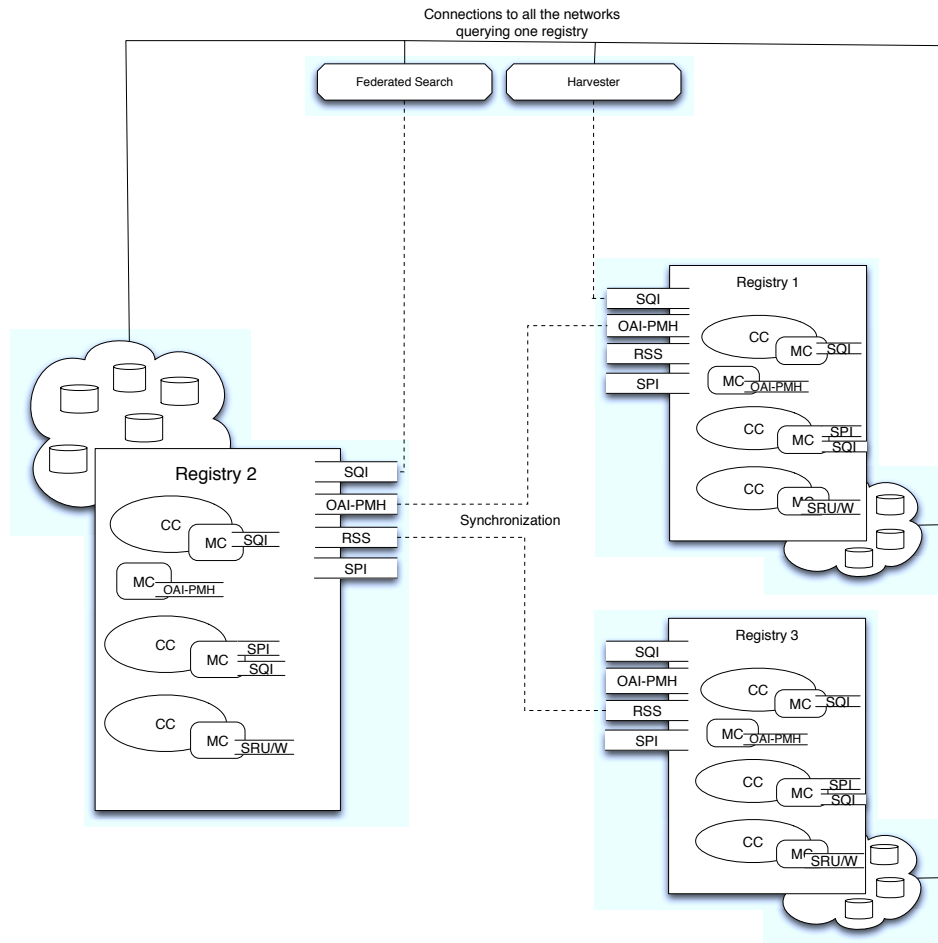

Fig. 1. ARIADNE architecture approach

A registry can contain metadata collections (MC in the figure 1) or content collections (CC). The latter contain metadata collections (MC).

The final goal of the registry is to create a network of networks of repositories, similar to current DNS functionality on the internet. All the registries

are synchronized so if a harvester or a federated search service queries a registry, they can access all the content collections or metadata collections in the different networks. This approach is important: it allows automated discovery, decreases time spent managing repository information, and allows automated widespread updates.

This architecture shows how the registry can be synchronized with other registries. The registry exposes its contents using OAI-PMH so that other registries can harvest its content. In addition, the registry exposes its updates also using the RSS 2.0 specification, so that registries can be synchronized also with RSS feed readers. This RSS system can be used as a notification system, as it contains all the targets published.

The registry uses SQI for querying because it is neutral in terms of query language or results formats. Consequently, all SQI clients can query the registry. The registry supports different query languages like PLQL, VSQL and Lucene Query Language.

Finally, the registry implements an SPI interface for publishing content collections. For instance, when other content collections are harvested, they are inserted using SPI to publish content by reference, because the metadata in this case defines a content collection already published elsewhere.
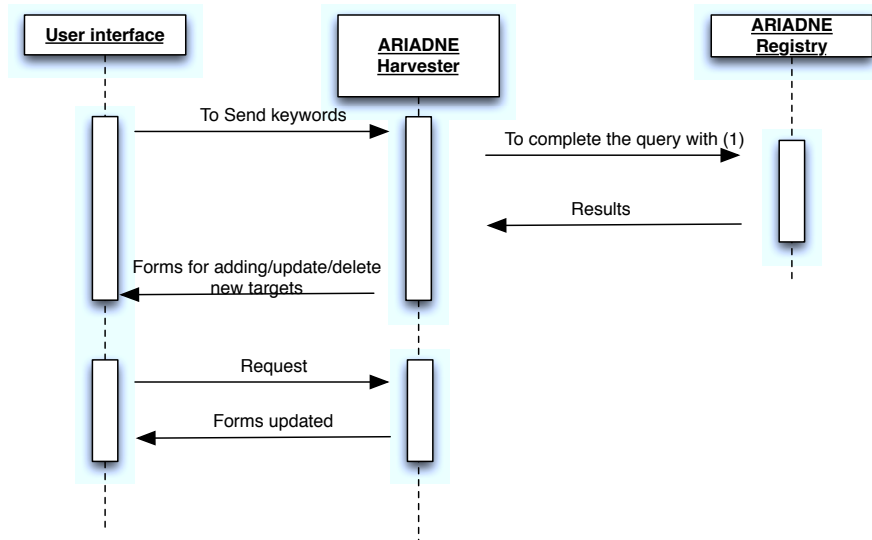
One of the main advantages of this implementation is that ARIADNE Registry is based on the ARIADNE Repository. This software is a flexible implementation that allows different models of metadata. For instance, it supports LOM, ILOX, Dublin Core and this paper explains how it supports IMS LODE Registry specification. The idea behind this software is that we can index all kind of xml document. However, the final goal is that we can built combined queries, for instance, using PLQL, based on the content of different tags.

Trying to validate this architecture, we have integrated the ARIADNE harvester with the registry. The following sequencing diagram 2 shows how a system administrator can configure the harvester using the current implementation where the system administration can query the registry and add the targets that he/she is interested in.

5   Related Work

One of the main goals of ARIADNE is to encourage sharing and reusing LOs. For this purpose, we have implemented a registry using the concept of content collection, integrating some technical information in the data model and allowing the federation of registries. Table 1 shows the similarities between existing approaches and the ARIADNE approach.

Domain Name System (DNS)[14] has conceptual similarities with the approach explained in this paper. While DNS translates a name identifier to a number identifier, the registry translate abstract information contained by the content collection to a technical information contained in the metadata collection. In addition, the federation of DNS allows the propagation between them of the new sites. It is a similar approach that ARIADNE Registry implements with

(1) metadataCollection.target.targetdescription.protocolIdentifier.entry = "oai-pmh-v2"

Fig. 2. Interaction diagram - Configuring the harvester

Table 1. Similarites with an existing approaches (Non-exisiting, Weak and Strong)

$Different\ approaches$	$Content\ collection$	$Extra\ technical\ information$	$Federation$
DNS	Non-exisiting	Non-exisiting	Strong
$MWSDI$	Weak	Strong	Weak
$IESR/OCKHAM$	Strong	Weak	Weak
$CORDRA$	Strong	Weak	Weak

the federation of the registries (the new content collections should be propagated among them)

METEOR's Web Service Discovery Infrastructure (MWSDI) uses UDDI specification for defining the interfaces (web services) and it defines an ontology for providing support to the infrastructure using OWL. MWSDI is based on the need to decentralize an architecture for several reasons such as geographical location, nature of registered services, business functionality, technical specifications and so on [15][16]. For this reason, it contains an ontology with details of the domain, Registries, Ontologies and Registries federation and network of relationship among them. The main difference is that MWSDI doesn't contain a content collection approach and the ARIADNE registry doesn't implement an hierarchical architecture of registries.

JISC Information Environment Service Registry (IESR)[17] and OCKHAM NSF / NSDL Registry [18] are developed under Global initiatives Registries [18]. They introduce the concept of content collection approach, while IESR uses Research Support Libraries Programme Collection Description schema (RLSPCD) and Dublin Core, which allows the definition of some technical details of the services which provide access to the collections[17], OCKHAM NSF/NSDL uses Dublin Core. They consider the definition of how the repositories expose the metadata considering protocols like Z39.50, Web Service SOAP, SRW and OAI-PMH. However, they don't provide extra information in their implementations such as sets supported by OAI-PMH or query languages. The registries expose their metadata through a SRU client [19, 16].

CORDRA / ADL registry (ADL-R) merges the concepts or MWSDI and IESR/OCKHAM, introduces a hierarchical structure of registries and the concept of content collection based on repositories which accept LOM as a metadata of LOs [20] .

At the end, we can conclude that Ariadne Registry incorporate the best characteristics from other approaches:

1. Conceptually, we implement similar approach to DNS widespread functionality.
2. IMS LODE Registry allows the definition of the services like UDDI specification concept do.
3. IMS LODE Registry allows the content collection approach like the above approaches mentioned do.
4. ARIADNE Registry allows the federation of the registries.

6   Statistics

In this section, we include some information about the metadata collections inserted and how they expose the metadata. It includes the number of records harvested using the OAI-PMH protocol. However, SQI or SRU/W don't provide a way to get all the possible results, thus the SQI and SRU/W targets do not contain information in date and records harvested field (table 2).

Table 2. Records harvested and protocols

id	date	records	SQI	SRU	OAI	SPI
dum_rvp_cz	09 Apr 2010	2323	No	No	Yes	No
atlas_fri_uni − lj_s	12 Apr 2010	4400	No	No	Yes	No
openlearn_open_ac_uk	12 Apr 2010	535	No	No	Yes	No
sodis_de	12 Apr 2010	2277	No	No	Yes	No
sdt_sulinet_hu	12 Apr 2010	18864	No	No	Yes	No
ait_opetaja_ee	12 Apr 2010	4050	No	No	Yes	No
193_43_17_27	12 Apr 2010	19309	No	No	Yes	No
193_93_132_187	12 Apr 2010	3541	No	No	Yes	No
www_yteach_com	12 Apr 2010	10	No	No	Yes	No
aspect	12 Apr 2010	124933	Yes	No	Yes	Yes
aspect_cup_cam_ac_uk	12 Apr 2010	8839	No	No	Yes	No
melt_contento_se	12 Apr 2010	3014	No	No	Yes	No
spindeln_iml_umu_se	12 Apr 2010	6699	No	No	Yes	No
lom_emokykla_lt_u_1	12 Apr 2010	1101	No	No	Yes	No
195_82_131_106	12 Apr 2010	196	No	No	Yes	No
aplitic_xtec_cat	12 Apr 2010	9876	No	No	Yes	No
www_klascement_net	12 Apr 2010	17739	No	No	Yes	No
www_portaldasescolas_pt	12 Apr 2010	1554	No	No	Yes	No
fenix_isftic_mepsyd_es	12 Apr 2010	26666	No	No	Yes	No
Mace_0	13 Apr 2010	180461	No	No	Yes	No
Globe	13 Apr 2010	385739	Yes	No	Yes	No
caad_asro_kuleuven_be	3 May 2010	450	No	No	Yes	No
caad_asro_kuleuven_be_0	3 May 2010	65	No	No	Yes	No
caad_asro_kuleuven_be_1	3 May 2010	171	No	No	Yes	No
caad_asro_kuleuven_be_2	3 May 2010	890	No	No	Yes	No
caad_asro_kuleuven_be_3	3 May 2010	2773	No	No	Yes	No
caad_asro_kuleuven_be_4	3 May 2010	50	No	No	Yes	No
caad_asro_kuleuven_be_5	3 May 2010	2106	No	No	Yes	No
cumincad_scix_net	3 May 2010	9636	No	No	Yes	No
ariadne_members	3 May 2010	3290	Yes	No	Yes	No
ariadne_partners	3 May 2010	71600	Yes	No	Yes	No
icoper	3 May 2010	16454	Yes	No	Yes	No
eun	3 May 2010	187722	Yes	No	Yes	No
ourj_code	3 May 2010	297	Yes	No	Yes	No
iskme	3 May 2010	30907	No	No	Yes	No
oai_rails_it_com	3 May 2010	211	No	No	Yes	No
oai_rails_it_com_0	3 May 2010	2223	No	No	Yes	No
oai_rails_it_com_1	3 May 2010	29	No	No	Yes	No
kocw	Not applicable	Not applicable	Yes	No	No	No
merlot	Not applicable	Not applicable	Yes	No	No	No
educanext	Not applicable	Not applicable	Yes	No	No	No
lornet	Not applicable	Not applicable	Yes	No	No	No
espol	Not applicable	Not applicable	Yes	No	No	No
agrega	Not applicable	Not applicable	Yes	No	No	No
nime	Not applicable	Not applicable	Yes	No	No	No
fedora	Not applicable	Not applicable	No	Yes	No	No
Total		1151000	14	1	38	1

Table 2 shows that the registry contains fourteen repositories which support SQI, one repository which supports SRU/W, twenty-five which support OAI-PMH and one which supports SPI. As the ARIADNE harvester relies on the registry, we know how many objects are exposed through OAI-PMH. The information is shown in the table.

The result is that we expose 1151000 metadata instances through the ARIADNE registry so far, although, we have to remark that some of them are duplicated, because some targets harvest other targets. But it means that we can expose a large amount of metadata instances specifying the protocol to access to them. All the harvesters which implement OAI-PMH can connect easily to the registry, using an SQI client for finding the targets and harvesting all the metadata from all the repositories described in the registry.

7 Conclusions and future work

We have developed a LOR registry, which enables us to set up a federation of registries. Different federations of repositories can easily obtain information from other federations through the registries. In this context, we will deploy the ARIADNE Registry in the ASPECT project, and we will evaluate the pros and cons of the federation.

In the scope of this work, we have considered some of the main specifications about querying, harvesting and publishing like SQI, OAI-PMH and SPI. We have spent some efforts to define extra information for the SRU/W. However, we have to validate this specification. Also, we would like to consider more specifications such as OKI OSID.

Now, we have integrated the ARIADNE harvester with the registry, in addition, we have to work on the Federate Search Layer to allow querying based on the content of the registries.

References

1. Ochoa, X., Duval, E.: Measuring learning object reuse. In: EC-TEL '08: Proceedings of the 3rd European conference on Technology Enhanced Learning, Berlin, Heidelberg, Springer-Verlag (2008) 322–325
2. Ternier, S., Bosman, B., Duval, E., Metzger, L., Halm, M., Thorne, S., Kahn, J.: Connecting oki and sqi: One small piece of code, a giant leap for reusing learning objects. Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications (2006) 825–831

3. Duval, E., Forte, E., Cardinaels, K., Verhoeven, B., Van Durm, R., Hendrikx, K., Forte, M.W., Ebel, N., Macowicz, M., Warkentyne, K., Haenni, F.: The ariadne knowledge pool system. Commun. ACM **44**(5) (2001) 72–78

4. Neven, F., Duval, E.: Reusable learning objects: a survey of lom-based repositories. In: MULTIMEDIA '02: Proceedings of the tenth ACM international conference on Multimedia, New York, NY, USA, ACM (2002) 291–294

5. Bernd Simon, Daniel Olmedilla, N.S.e., ed.: A Simple Query Interface for Interoperable Learning Repositories. Volume 11-18., Chiba, Japan (2005)

6. Ternier, S., Massart, D., Van Assche, F., Smith, N., Simon, B., Duval, E.: A simple publishing interface for learning object repositories. Proc. World Conf. Educational Multimedia, Hypermedia, and Telecommunications, Assoc. for the Advancement of Computing in Education (2008) 1840–1845

7. Lagoze, C., Lagoze, C., Van de Sompel, H.: The open archives initiative: Building a low-barrier interoperability framework. IN PROCEEDINGS OF THE ACM/IEEE JOINT CONFERENCE ON DIGITAL LIBRARIES (2001) 54–62

8. Ternier, S., Massart, D., Campi, A., Guinea, S., Ceri, S., Duval, E.: Interoperability for searching learning object repositories: The prolearn query language. D-Lib Magazine **14**(1-2) (2008)

9. Klerkx, J., Vandeputte, B., Parra, G., Van Assche, F., Duval, E.: How to share and reuse learning resources: the ariadne experience. In: Proceedings ECTEL Conference 2010. Accepted for publication. (2010)

10. project, A.: http://www.aspect-project.org/ (2010)

11. LODE, I.: http://www.imsglobal.org/lode/spec/imslodev1p0bd.html (2010)

12. Massart, D., Smith, N., Tice, R.: Design of data model and architecture for a registry of learning object repositories and application profiles. deliverable 2.2 from aspect project. ASPECT PROJECT (2009)

13. Example Registry Instance, I.: http://ariadne.cs.kuleuven.be/ariadne-registry/search/showmetadata.jsp?query=metadatacollection.identifier.entry

14. Mockapetris, P., Dunlap, K.J.: Development of the domain name system. SIG-COMM Comput. Commun. Rev. **18**(4) (1988) 123–133

15. Sivashanmugam, K., Verma, K., Sheth, A.: Discovery of web services in a federated registry environment. In: ICWS '04: Proceedings of the IEEE International Conference on Web Services, Washington, DC, USA, IEEE Computer Society (2004) 270

16. Xiang, X., Morgan, E.: Exploiting 'light-weight' protocols and open source tools to implement digital library collections and services. D-Lib Magazine **12**(10) (2005)

17. Apps, A.: A registry of collections and their services: from metadata to implementation. In: DCMI '04: Proceedings of the 2004 international conference on Dublin Core and metadata applications, Dublin Core Metadata Initiative (2004) 1–10

18. Registries, G.I.: http://www.globalregistries.org/ (2010)

19. Morgan, E., Frumkin, J., Fox, E.: The ockham initiative - building component-based digital library services and collections. D-Lib Magazine **10**(11) (2004) 51–59

20. Jerez, H., Manepalli, M., Blanchi, C., Lannom, L.W.: Adl-r: The first instance of a cordra registry. D-Lib Magazine **12**(2) (2006)

Where is the user? Filtering Bots from the Edurep Query Logs

Wim Muskee

Kennisnet Foundation
Paletsingel 32
2718 NT Zoetermeer, NL
`w.muskee@kennisnet.nl`

Abstract. Edurep indexes learning object metadata from several repositories, offering a webservice interface on which portals can build their own search implementation.

At Edurep query log level, no obvious distinction can be made between human users and webcrawlers visiting these portal sites. This makes it impossible to gather any meaningful data on user search behaviour.

Four query types, distinguished from the six largest portals' websites were related to one month of query logs. For two query types a distinction between human and automatic generated traffic could be found. However, these results can only be used to advise connected portals on their interface implementations. More research is needed to actually perform any reliable filtering.

Keywords: webservice, crawler detection, log analysis

1 Introduction

Edurep is a Dutch learning object search engine, indexing harvested learning object metadata from more than 50 different repositories. Search portal developers can interface with the search engine using the Edurep webservice, available through the SRU/SRW protocol (Figure 1).

Although operational for some years [11], the operators gained access to the search query logs only recently (december 2009). Through analysis of these logs and webserver logs of one portal, the operators discovered that a significant amount of queries came from various search engine bots[1]. Among several harm-
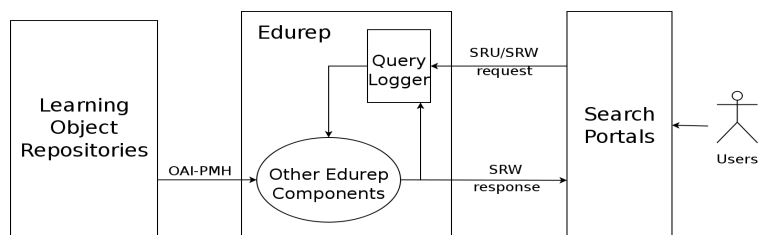

Fig. 1. A simplified diagram of Edurep in its context.

ful aspects, Edurep is affected by two in particular. First, and obviously, webcrawlers generate extra traffic, possibly limiting performance for human users. Secondly, webcrawlers generate automated traffic, making it harder for the operators to infer meaningful human interaction results from the Edurep query logs.

Most of these search engine bots can be identified at search portal level based on their HTTP request User-agent string or IP adress [12,9]. However, this information is no longer available when the request reaches Edurep.

This problem is not typical for Edurep, but applies to any webservice which allows connections from a third-party search interface. Examples of these in the learning object context include the LRE [3], MACE [14] and the Spider project [4], all of them available through the SQI protocol [7].

With Edurep as context, this paper aims to explore methods to make a distinction between automated and human queries in webservice query logs. To this end, four query types were distinguished from several search portal web interfaces. The SRU representations for each query were used to filter the logs for a specific query type and analyze it more closely. The paper ends with a discussion of the results.

---

[1] A type of webcrawler; a program which gathers information from the internet by recursively following found hyperlinks.

2 Modeling Automated Queries

Because webcrawlers only follow hyperlinks, automated searches are caused by the presence of hyperlinks which cause an Edurep search query. An analysis of the portals' search interfaces is necessary to combine hyperlinks with logged SRU queries.

2.1 Portal Search Interfaces

Looking at the search interfaces of the six largest portals (consisting of 97% of query total), four types of hyperlinks were distinguished.

- *search links:* Issuing a search to retrieve a first page resultset.
- *pagination links:* Issuing a search to retrieve another resultset page.
- *result links:* Issuing a search to retrieve a specific record.
- *facet links:* Issuing a search to retrieve the amount of records for that facet.

Typically, the portals retrieved either 5 or 10 results after a search query. The number of navigation links ranged from 5 to 20, always including a next and/or previous link and sometimes including links to the first and/or last page. A few included result and facet links.

Only one portal (C) performed a search on page arrival. The resulting page included all link types. All the portals' queries were represented as a url in the browser navigation bar, meaning they can be pasted easily on other webpages for others to click on, including bots. When searching for the portals' url query prefixes on Google, indeed some results were found. Also corresponding queries were discovered in the query logs.

2.2 SRU/SRW

Edurep can be queried using the *searchRetrieve* operation of the SRU/SRW protocol [13]. Among several supported request parameters [10], the *startRecord* parameter determines which record of the resultset is displayed first. When omitted, it defaults to 1. The *maximumRecords* parameter sets the number of records each resultset contains. Edurep's default is 10.

A search query typically has no *startRecord* value at all or a value of 1. Also, to present a reasonable amount of results, the *maximumRecords* value is set to 5 or higher, or left out to return 10. Pagination queries have a *startRecord* value higher than 1.

In a result query, the *startRecord* value is omitted or 1. Since a result of 1 is expected, the value for *maximumRecords* does not need to be 1. However, because a specific record is requested, part of the *query* value is characteristic. In Edurep, a specific record can be requested by filtering on lom.general.identifier or lom.general.catalogentry, the LOM identifier, or meta.upload.id, Edurep's internal unique identifier.

Facet queries can be performed inside a search query by adding Edurep's *x-term-drilldown* parameter to the SRU query. In addition to the search results,

a count drilldown for each facet of the requested field is retrieved. Because this function is not supported for all LOM fields, separate facet queries can also be executed. These have a *startRecord* value of 1 or none at all. Also, the value for *maximumRecords* is 0 or 1 [2].

3 Dataset

The logs of January 2010 were used as dataset and the analysis is done in R [5]. Each log entry consisted of the portal's ip adress, the timestamp when a search query entered the system (UTC), the size of the response data in kilobytes, the processing time in seconds, the entrypoint of a query on the server indicating the used protocol (SRU or SRW), and the SRU search query.
Five variables from each query were used. The IP adress, *startRecord* and *maximumRecords* values were used unprocessed. The *query* argument was used as a whole, assuming each portal constructed their queries in the same way and query uniqueness was not compared across portals. An identifier boolean was set to 1 if a result link was detected.

4 Results

Concerning *search queries*, the distinction between human and automatic induced queries can be made based on the occurence of the queries. Automatic induced queries will appear more often in relation to human generated ones. While Portal C's startup page query appeared more than 6 times than any of its other queries, a good threshold could not be determined.
Assuming most users will never click past the second page of search results [1], *facet queries* with a *startRecord* value over 200 will probably be auto-generated (PAG1). A more elegant method for determing automatic facet queries is to scan the logs for pagination ranges. A range was crudely defined as a set of SRU queries (min. 10) with equal *query* values, a *startRecord* difference of *maximumRecords* and a maximum *startRecord* value higher than 200 (PAG2).
Based on occurence of *result queries*, no clear evidence for automatic querying was found in the logs. This was attributed to the dynamic nature of Edurep's content, with changing resultsets, different results will be queried.
After plotting the unique *facet queries* of Portal C (Figure 2), the small layer of queries below the top coincided with the facet queries executed on entering the search page. Observing that 10 of the 12 sub-top queries were executed about 2330 times, it was assumed they were caused by automatic querying. From the queries of these types, that amount could be subtracted, leaving their human induced occurences (FACET). Following from this assumption, at least the same amount of automatic hits were generated by Portal C's startup search query, and could thus also be subtracted.

---

[2] Technically, by setting this value to 0, the same total can be retrieved, but since the usage of 1 had been observed, it was included
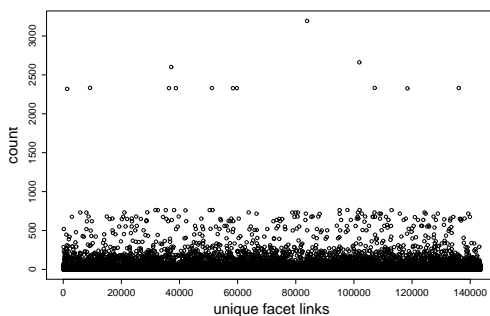
Fig. 2. Unique facet link queries plotted against occurences.

	total	PAG1	PAG2	FACET
Portal A	41690	-15237	-13355	
Portal B	126340	-105026	-89710	
Portal C	1293902	-15255	-15654	-30290
Portal D	48841	-47	-62	
Portal E	232341	-1778	-1815	
Portal F	82527	-406	-205	
total	1825641	-137749	-120801	-30290

Table 1. Subtractions of filter method implementations..

The subtractions from each filtering method are displayed next to to each portal's total amount of queries in table 1.

5 Discussion

Considerable automatic induced querying was observed. In terms of bandwidth the found ranges from *PAG2* alone caused 13,3 Gb of traffic, $26,5\%$ of the total A-F amount. Concerning the amount of queries, *PAG2* and *FACET* accounted for $8,4\%$ of the total A-F amount of queries.

However, assumptions were made and the used filter methods are still rudimentary and incomplete. In using *PAG2* for instance, tails or heads of the ranges may lie outside the used dataset. Also, the dataset probably contains heads or tails of ranges from other months. This is even more true when considering the pagination queries don't need to appear on the timeline in the same order as they appear on the page [2]. Secondly, first- and lastpage pagination queries were not considered in *PAG2*.

The immediate findings of this study make it possible to tailor our advise for portals. One aspect of this is related to blocking crawlers at the portal by implementing the Robots Exclusion Standard [6]. Use of this standard could also be enforced through Edurep's user level agreement. As an unintended side effect, automated usage amplified some examples of inefficient quering on Edurep. Another aspect of the advise should include information on how to interface with Edurep better.

Use of various scripts to parse and filter the log files proved very useful during the conduct of this study. Automating the used scripts will allow the administrators to detect undesirable behaviour in an earlier stage and act on it sooner, leaving Edurep free to be used by actual users.

Future research should improve on several aspects. First of all, more months of logging need to be used to combine and compare with current results. Secondly, the SRU *query* values need to be parsed fully to allow more accurate filtering options and to compare queries across portals. Last is the usage of the portal website. Parameters like the size and format of the pagination links, and the types of search, result or facet links on the page could prove useful in implementing better automatic detection methods.

A more long term product change would be to also request the end user's orginal User agent string in the query to Edurep. Also requesting the original IP adress could lead to privacy concerns. Since lots of crawler User agent strings are publicly available [8], this information could greatly enhance our filtering efforts.

An new Edurep component could be introduced, making it possible to block requests before they are processed by the system. However, at this point it is unclear if such an extra check on all requests outweighs the benefits of not having to process the blocked requests. For now, such a filtering component will have to implemented before the logs are processed by our business level reporting tool.

While the ideas in this paper could be used in similar architectures, the actual scripts cannot because they are made for SRU and Edurep's query log format. With more standardization in repository query languages (like SQI), corresponding logging standards can be thought of, making sure developed analysis tools benefit many and query logs can be shared easily.

Filtering automatic queries is after all needed to look more closely at the human ones. The focus of interest is teacher search behaviour, not only on Edurep but beyond our borders.

References

1. Aula, A.: Studying user strategies and characteristics for developing web search interfaces. Dissertations in Interactive Technology 3 (December 2005)
2. Dikaiakos, M.D., Stassopoulou, A., Papageorgiou, L.: An investigation of web crawler behavior: characterization and metrics. Computer Communications 28(8), 880–897 (2005)

3. Massart, D.: Towards a pan-european learning resource exchange infrastructure. Lecture Notes in Computer Science 5831/2009, 121–132 (2009)
4. Paulsson, F.: Connecting learning object repositories: Strategies, technologies and issues. Internet and Web Applications and Services, International Conference on 0, 583–589 (2009)
5. R Development Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2009)
6. robotstxt.org: The web robots page. Retrieved August, 3 2010, from http://www.robotstxt.org. (2007)
7. Simon, B., Massart, D., van Assche, F., Ternier, S., Duval, E., Brantner, S., Olmedilla, D., Miklós, Z.: A simple query interface for interoperable learning repositories. In: Proceedings of the 1st Workshop On Interoperability of Web-Based Educational Systems. pp. 11–18 (2005), http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.67.7745
8. Staeding, A.: List of user-agents (spiders, robots, browser). Stichting Kennisnet. Edurep wiki. Retrieved August 5, 2010, from http://www.user-agents.org
9. Stassopoulou, A., Dikaiakos, M.: Web robot detection: A probabilistic reasoning approach. Computer Networks 53(3), 265–278 (February 2009)
10. Stichting Kennisnet: Edurep wiki. Retrieved June, 3 2010, from http://edurep.wiki.kennisnet.nl
11. Stichting Kennisnet ICT op School: De educatieve contentketen: leertechnologische afspraken voor de toekomst. Retrieved May, 2 2007, from http://contentketen.kennisnet.nl/attachments/990312/De_Educatieve_contentketen_-_Leertechnologische_afspraken_voor_de_toekomst.pdf (December 2006)
12. Tan, P.N., Kumar, V.: Discovery of web robot sessions based on their navigational patterns. Data Mining and Knowledge Discovery 6(1), 9–35 (January 2002)
13. The Library of Congress: Sru: Search/retrieval via url. Stichting Kennisnet. Edurep wiki. Retrieved August 5, 2010, from http://www.loc.gov/standards/sru/., http://www.loc.gov/standards/sru/
14. Wolpers, M., Memmel, M., Klerkx, J., Parra, G., Vandeputte, B., Duval, E., Schirru, R., Niemann, K.: Bridging repositories to form the mace experience. New Review of Information Networking 14(2), 102–116 (2008)

Handling Multiple Metadata Streams Regarding Digital Learning Material

Jasper Roes[1], Jeroen van Vuuren[2], Nico Verbeij[2] and Henk Nijstad[3],

[1] TNO Information- and Communication Technology, Colosseum 27, 7521 PV, Enschede
[2] Verdonck, Klooster & Associates, Baron de Coubertinlaan 1, 2719 EN, Zoetermeer
[3] Stichting Kennisnet, Paletsingel 32, 2718 NT, Zoetermeer
Jasper.Roes@tno.nl, Jeroen.vanVuuren@vka.nl, Nico.Verbeij@vka.nl,
H.Nijstad@kennisnet.nl

Abstract. This paper presents the outcome of a study performed in the Netherlands on handling multiple metadata streams regarding digital learning material. The paper describes the present metadata architecture in the Netherlands, the present suppliers and users of metadata and digital learning materials. It furthermore describes the roles of suppliers and users that were defined. Based on the needs of the present and new roles, and based on the present architecture the necessary changes to the metadata architecture are described to make it more future-proof and to ensure it can handle multiple metadata streams.

Keywords: Metadata, architecture, learning materials, interoperability, reusability, edurep, eck, educational content chain

1 Introduction

In the last couple of years, the use of metadata to be able to search and find appropriate learning materials has increased drastically. With this increase, more organizations are interested in producing digital learning materials. To be able to find appropriate learning materials, metadata records need to be provided. In the past, most digital learning materials were created by publishers, who also provided metadata records for their materials. Now that not only publishers are creating digital learning materials, but also schools, individuals and many more, the metadata records for one piece of learning material are not provided by one group anymore. These new groups of people and organizations providing metadata create new streams of metadata records. Metadata streams are metadata records flowing from the creator of the record, to the central database where the records are used to find learning materials. These new streams were not foreseen when developing the present metadata architecture[1]. Instead of handling only one metadata stream, the educational content chain now has to handle many streams, all with different characteristics, and often providing additional metadata information about the same piece of learning material. This leads to difficul-

---

[1] A definition of how the systems handling metadata interoperate.

ties in the educational content chain[2], as the present architecture is based on the principle of only one metadata record for each piece of learning material.

Based on these developments in the educational sector, TNO [1] and VKA [2] were asked by Kennisnet [3] to perform a study on this subject, and to provide Kennisnet with recommendations on changes to the present educational content chain making it able to handle multiple metadata streams and making the educational content chain more future-oriented.

Kennisnet is the foremost public knowledge centre regarding information and communication technologies (ICT) and education in the Netherlands. TNO is an applied scientific research company that applies scientific knowledge with the aim of strengthening the innovative power of industry and government. Verdonck, Klooster & Associates (VKA) is an independent consultancy company with extensive experience at the intersection of strategy, implementation and ICT in the public sector.

2 Present Educational Content Chain

Constant developments in society, and the need of the Netherlands to extend their competitive knowledge industry, lead to a need for changes in the Dutch educational sector. A major development in the educational sector is the need to educate people during their entire life; education does not stop anymore after obtaining a degree. The digital era provides huge possibilities to support lifelong learning, and furthermore the present generation of students has high expectations about the digital possibilities in education.

These developments lead to more flexible education, and therefore ask for new educational material: digital material. To ensure a good user experience, and to ensure interoperability and reusability of digital material there is a need for agreements. Kennisnet therefore developed an educational content chain to make agreements about the use of digital learning materials and to ensure interoperability, reusability and usability [4].The educational content chain in the Netherlands consists of five major steps:

1. Developing (creating digital learning material)
2. Making available (placing material on the web and providing metadata)
3. Finding (searching in the central database)
4. Arranging (creating new material based on existing materials)
5. Using (using the material during courses)

These steps are often represented as a circle as the use of digital materials often leads to the development of new, or renewed, digital material.

---

[2] The educational content chain consists of all users and suppliers that create, distribute, make available or use digital learning materials. This can for instance be publishers, schools, teachers, students, etc.

3 Present Content Chain Issues

The present content chain has implicitly been developed with only one main stream of provisioning *(authorative [5])* metadata information: the stream that from the suppliers of the learning material to the central database where all metadata is collected and which the users can use question to find learning material. As long as this assumption is valid, for each piece of learning material only one metadata record is provided and users searching the database will only find each piece of learning material once. Examples of these *single (authorative [5]) metadata-stream* central search platforms are Edurep [6] in the Netherlands, the LRE [7] and the Spider project [8].

Recent developments though, show more and more sources providing complete metadata records, or supplying additional metadata to the original record. These new records lead to more than one metadata record in the central database for most of the learning materials. The availability of multiple metadata records for one piece of learning material in the present content chain makes it difficult for users to locate useable learning materials as the search results show one piece of learning material multiple times (once for each metadata record).

Furthermore these multiple records also have another effect: users cannot search in the richest metadata, resulting in not finding useable learning materials. Table 1 shows two metadata records for the same piece of learning material. If a user searches for 'Geography' as course, he will find the material. If he searches for 'Layer of the earth' he will also find the material. When a user searches for the combination of both keywords the present content chain will not return the material as the combination of the key words are not available in one metadata record.

Table 1. Multiple metadata records from different suppliers for one piece of learning material.

Supplier	Course	Key word
A	Geography	-
B	-	Layer of the earth

Based on the information available, the study [9] concluded that the two main issues of the present content chain are:
1. The present chain is incapable of handling multiple metadata streams
2. The present chain cannot link multiple metadata records that describe one piece of learning material

4 Roles of Users and Suppliers

The study showed that it is difficult to the characterize parties (people or organizations supplying learning materials and metadata). There is no such thing as a general publisher or a general user. Each party has different requirements and performs different roles, as has been argued in [10]. In this study six roles are defined [9]:
1. Suppliers: the expert creator/publisher of the learning material.
2. Prosumers: a consumer that also produces learning material.

3. Scouts: scouts locate learning material on the internet.
4. Enrichers: enrichers provide additional metadata for a specific target group.
5. Reviewers: reviewers review learning material for a specific target group
6. Users: a user that supplies metadata and/or reviews learning material for a specific target group

The needs of the roles are: suppliers, prosumers, and scouts want to provide metadata to make the learning material findable. Enrichers, reviewers and users want to provide additional metadata to make the material more findable for specific target groups.

5 Desired Educational Content Chain

Based on the defined roles and their needs, a new architecture for the educational content chain was defined. In this architecture every role can provide the information they define or have present. Figure 2 presents the old and new architecture. The new architecture includes everything that is in the figure; the old architecture has the same structure, but only consists of the roles and records that are marked in a grey. On the left side of the figure the different roles are represented. The different roles are divided into three groups to give the users of the search engine the possibility to make a distinction while searching. Because of the distinction in three groups, users have the possibility to search only in metadata records provided by the source of the learning material. Each group has its own contact point at the central search platform (Edurep [6] in the Netherlands) to make the distinction described possible. On the right side of the figure the users that search for materials are represented.
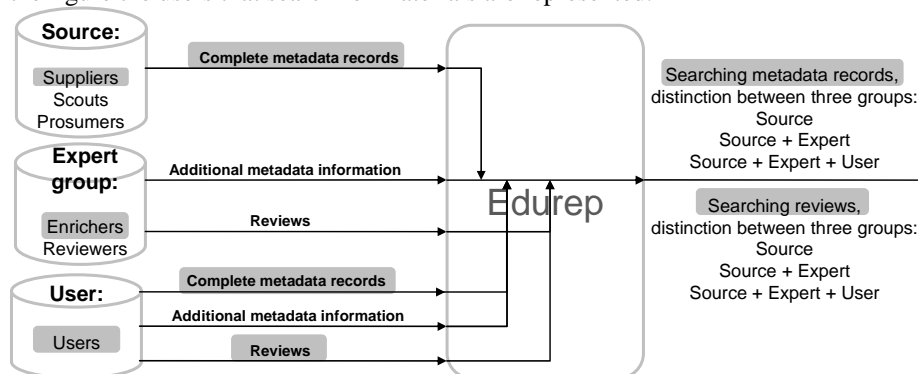

Fig. 1. Old and new architecture of the Dutch educational content chain

6 Recommendations

To be able to create the new architecture as described in section 4, it is necessary to make some changes to the present educational content chain. The most important rec-

ommendation that was defined is: "Ensure that each piece of learning material has a unique identifier, and ensure that the user can search in the richest metadata information (a combination of all available metadata information) available" [9].

By ensuring that each piece of learning material has a unique identifier, it is possible to reference different metadata records that apply to the same piece of learning material. If these references can be made by the central database, it is possible to search in the richest metadata possible. It is furthermore possible to combine multiple pieces of metadata to create the best set of metadata for one piece of learning material. To ensure unique identifiers for each piece of learning material, agreements have to be made on how to generate unique identifiers. The suggestion made in the study is to use a formal URN (Uniform Resource Name) [11] as the unique identifier for the learning material. Kennisnet could apply for a formal URN to be used in the Netherlands, and could supply each provider of learning materials with its own unique number. The combination of the URN, the unique code for the provider and a unique number to be defined by the provider of the piece of learning material ensures that each piece of learning material has a unique identifier.

Furthermore, two other recommendations have been defined based on the study: "Ensure that the educational content chain is managed" & "Provide a service to add additional metadata". The first recommendation is important since the management of the content chain is currently not explicitly defined. The second recommendation is important to ensure that all roles defined have the possibility to provide metadata. Most roles already have some kind of tooling available, but f.e. for enhancers and reviewers, no tooling is available.

References

1. TNO, http://www.tno.nl
2. VKA, http://www.vka.nl
3. Kennisnet, http://www.kennisnet.nl
4. Kennisnet Educational Content Chain, http://contentketen.kennisnet.nl
5. Recker, M. & Wiley, D.A.: A non-authoritative educational metadata ontology for filtering and recommending learning objects. In: Journal of Interactive Learning Environments, pp. 255-271, Taylor & Francis, London (2001)
6. Edurep, http://edurep.kennisnet.nl
7. Massart, D.: Towards a pan-european learning resource exchange infrastructure. Lecture Notes in Computer Science 5831/2009, pp. 121-132, Springer-Verlag, Berlin (2009)
8. Paulsson, F.: Connecting Learning Object Repositories: Strategies, Technologies and Issues. In: 2009 Fourth International Conference on Internet and Web Applications and Services, pp. 583-589, IEEE Press, New York (2009)
9. Roes, J., Verbeij, N., van Vuuren, J.: Adviesrapportage metadatastromen. Kennisnet, Zoetermeer (2010)
10. Manouselis, N., Sampson, D.: Learning Resources Brokerage Systems: An Agent-Based Virtual Market Model. In: Third IEEE International Conference on Advanced Learning Technologies, pp. 424, IEEE Press, New York (2003)
11. IANA URN namespaces, http://www.iana.org/assignments/urn-namespaces/

Managing Broken URLs in Federated Metadata⋆

Tien-Dung Le and Elena Shulman

European Schoolnet, Rue de Trèves 61, B-1040 Brussels, Belgium.
{dung.le,elena.shulman}@eun.org

Abstract. In a large federation of learning object repositories, learning object locations or learning object URLs in metadata can become out of date when learning objects are moved or deleted. Regular checking of learning object URLs is essential. However, fully checking across a large federation imposes an unsustainable burden on time and resources while negatively impacting networks and repositories. This paper describes a broken URL handling system with a heuristic model that can provide a sustainable solution for federation service managers while enhancing communication and collaboration among federation stakeholders.

Key words: Broken URLs, federated metadata, heuristic model, learning resource exchange

1 Introduction

The Learning Resource Exchange (LRE) is a service that allows European teachers to get access to digital educational content from many different countries and providers (Figure 1). Content providers produce metadata, i.e., machine-readable descriptions of the educational content they want to make available to teachers within and beyond their national learning object repository systems. The LRE provides unified access to Learning Objects (LOs) stored in these different repositories. Each content provider exposes their metadata so that it can be easily accessed by the LRE. The LRE collects metadata from the different content providers and compiles them to produce a digital catalog of learning resources that can be consulted by teachers using the LRE [4]. Along with other information relevant to pedagogical contexts, metadata contains the locations (i.e., Uniform Resource Locators - URLs [1]) where resources can be obtained. Typically, users of the LRE catalog can obtain learning resource of potential interest by following the URL provided with the resource description. If this URL does not point to the expected resource, it is said to be "broken". Although broken URLs can sometimes be the products of transient network or server problems, in most of the cases they are caused by outdated metadata.
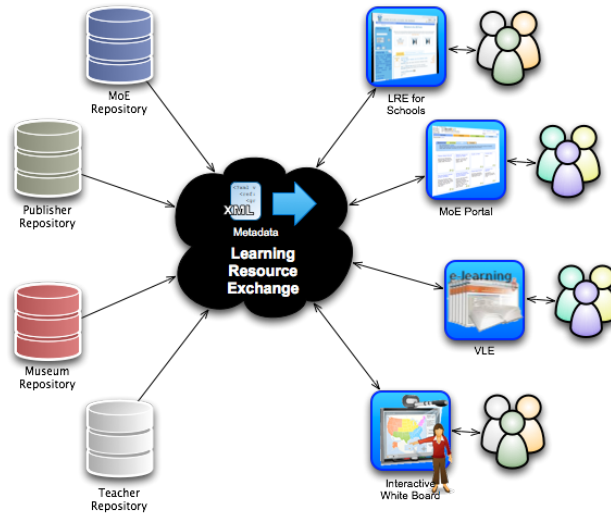
Fig. 1. Learning Resource Exchange.

This paper presents the results of our efforts to develop a solution that will effectively detect broken URLs, automate a system for communicating with content providers and allow greater flexibility for LRE service managers in resolving issues that can negatively impact user experience with the LRE.

Section 2 presents the background to the problem. Section 3 presents the architecture of the broken URLs handling system and details our proposed heuristic algorithm to automatically detect the broken URLs. Section 4 details the guidelines developed to facilitate effective and expeditious communication with content providers and procedures for LRE managers and system responses depending on the status of broken URLs once reported. Section 5 is a discussion including comments on some elements of comparison with related work.

2   Background

Broken URLs are problematic on several levels for a service that manages a large federation of repositories. Unsurprisingly, broken URLs significantly impact user satisfaction with the system. Teachers have reported a loss of confidence in the LREs catalog and perceived quality of the LRE more broadly [6]. From the users point of view, broken URLs are also a source of frustration when teachers attempt to access resources they have previously placed in a favorites folder or when they recommend to colleagues resources that have become suddenly unavailable. As depicted in Figure 1, the LRE federates LO repositories from various origins. LRE content is provided by ministries of education (MoE), commercial and non-profit content providers (Publisher), and cultural heritage organizations (Museums). Some of these repositories might store user-generated

content (Teachers). Because the LRE does not host the objects or control access to objects directly, the LRE must rely on the timely collaboration of content providers depicted in Figure 1 to update their metadata whenever object locations have changed.

Experience has shown that broken URLs can become a significant problem. Part of the solution we propose is based on a heuristic algorithm developed and tested for this purpose. This paper also illustrates the way we have linked a solution for broken URLs detection to support communication and collaboration with content providers.

The LRE has a number of techniques in place intended to ameliorate the issue with metadata quality but none has provided a sustainable solution in a system reliant on coordinated actions in a growing federation. First, to keep the LRE catalog up-to-date, metadata is collected on regular basis (e.g., everyday). This technique ensures that updates to the metadata on the part of content providers are reflected in the main catalog. However, this requires content providers to maintain the accuracy of their own metadata. We have found that this technique has its own set of drawbacks because some providers do not use the metadata they produce for the LRE. Another technique relies on users' scrutiny. Teachers who find broken URLs are invited to report them.

Reported URLs are checked by the system and, if the problem is confirmed, the corresponding entries are removed from the catalog and their content providers receive a request to correct the problem. The main drawback of this technique comes from the fact that it is a highly individualized solution impacting the quality of a few records at a time. Moreover, this technique is fully reliant on the goodwill of the catalog users and does not ultimately shield them from unsatisfactory search and retrieval experience culminating in the discovery of broken URLs.

A third technique consists of systematically checking all the URLs of the catalog on a regular basis. Unfortunately, this solution does not scale. As the number of catalog entries grows, the time necessary to check the catalog fully can be prohibitively extensive in duration. For example, in the LRE a full check of more than 200,000 URLs takes more than two days. To make matters worse, systematically checking all the learning resource URLs available on a system can sometime be considered as "unfriendly" by their administrators.

3   Broken URL Handling System

While allowing the LRE to efficiently detect and make decisions to remove catalog entries with broken URLs from the search, the proposed solution is ultimately intended to support and automate more effective communication with content providers and to facilitate their collaboration in expeditiously updating their metadata. The handling system (Figure 2) and guidelines described in this article accomplish several objectives. They demonstrate the feasibility of a heuristic model in efficiently detecting broken URLs and propose a mechanism to trigger ameliorative actions from providers. They illustrate our efforts at establishing

clear guidelines to support decision making and communication with providers for LRE service managers responsible for metadata quality and user satisfaction. A component to detect and hide broken URLs is only part of the solution. The proposed solution also takes into account and avoids broken URL detection techniques that can appear to be denial of service attacks on partner systems.
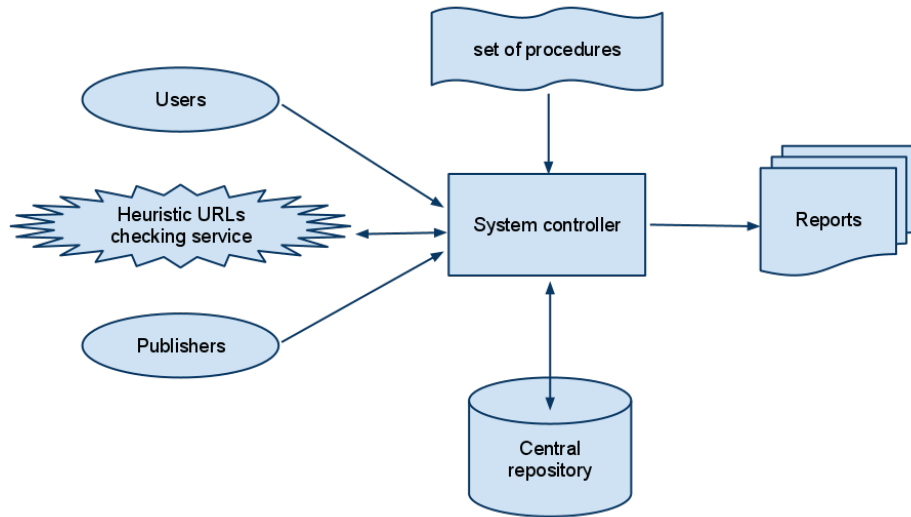

Fig. 2. Broken URLs Handling System.

There are several steps in the detection and handling of broken URLs. The first stage was a full system check for broken URLs to determine the scale of the problem and to create a baseline to test the proposed algorithm. As we discussed, while this kind of check can produce useful data on URLs, it is not feasible to use this technique on a regular basis for reasons already mentioned. The second stage is our proposed long term solution to detect and trigger corrections for bad URLs based on a check of smaller subsets and provide better automated channels for communicating with content providers. Techniques for determining the size and location of the sets to be checked are described below.

As mentioned above, it is necessary to put in place a heuristic checking algorithm to regularly check URLs of the catalog but not to "attack" the systems hosting LOs. Therefore only a subset of URLs should be checked at a time. Our analysis of the LO location distribution shows the URL selection should be based on the LO domain name system (or domain for short). For each domain, only a subset of its URLs should be checked. The first step is to determine which domains should be targeted for checks. To determine if the system stops or continues checking URLs in a domain, a sampling plan is applied after all selected URLs are checked.

3.1 Learning Object Location Distribution

Learning objects in the LRE federation are hosted in one or several domains. Typically, a URL or a LO location refers only to one domain while a domain normally hosts more than one LO. We analyzed the relationship between LOs and domains based on 2 aspects. First we looked at the domain distribution in the catalog, which indicates how many domains have a certain number of LOs. Second we analyzed the distribution of LOs to determine how LOs are distributed among domains.

In June 2010, there were 842 domains in the LRE. Figure 3 shows the domain distribution and Figure 4 shows the LO distribution. A significantly large percentage of domains (60%) hosted just a single learning object. Even more important for the development of our solution was the finding that only 4% of domains hosted more than 1000 LOs and 70% of LOs are located on domains presently hosting over 10,000 LOs. The fact that a small number of domains are hosting 96% of all LOs was instrumental in building a heuristic checking algorithm.
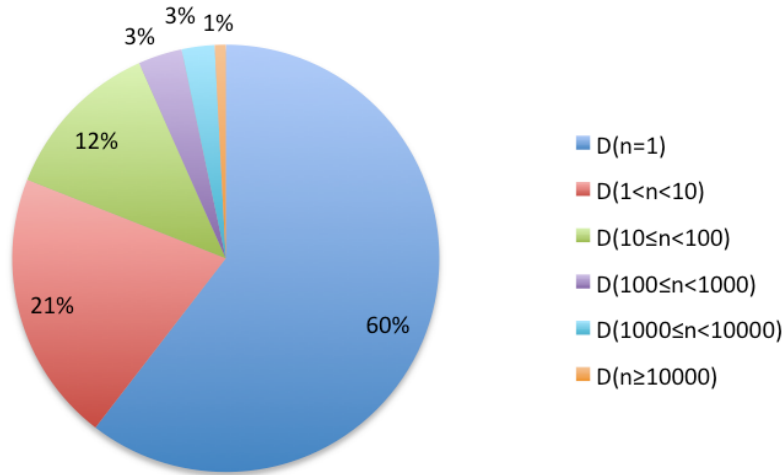

Fig. 3. Domain distribution in the LRE.

3.2 Heuristic Checking Algorithm

Based on the analysis of LO distributions, we propose to check each domain separately. For each domain, two checking steps are applied. First, the system checks all reported-broken URLs, which are reported as broken URLs in the last check, and calculates the confidence of the system on the domain. The second step in the check applies a sampling plan to check all previously good URLs. The sampling plan (Figure 5) is a simplified solution of the sequential sampling plan [7].
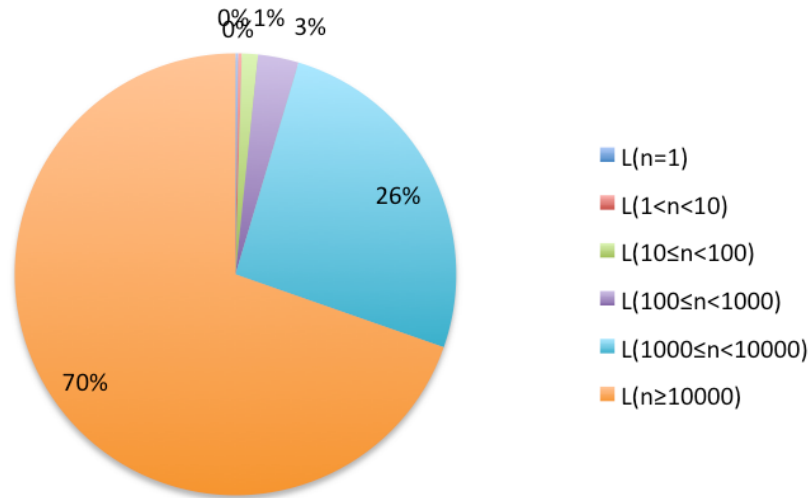
Fig. 4. LO location distribution in the LRE.

Sampling Plan

- Single sampling plan
  One sample of URLs on a domain, such as 10% of URLs, is selected at random. In case there are more than p_2 of selected URLs, such as 90%, are good, all unchecked URLs are considered as good URLs. Otherwise, the whole URLs on the domain are checked.
- Sequential sampling plan
  A set of URLs in a domain, for instance a set of 100 URLs, is selected at a time and after inspection a decision is made to accept (stop checking) or reject (check all URLs) in the domain or select another set in the same domain (continue sampling). In general, this sequential sampling allows for quick decisions, especially when frequency of broken URLs in a domain is particularly high or particularly low.
- Sampling plan in the algorithm
  Because we recognize the difficulties in calculating the slope of accept and reject lines, the checking algorithm uses sequential sampling plan with a simplification - the slope is zero (Figure 5).

Sampling Plan's Parameters

- Pre-defined threshold p_1 presents the low confidence of the checking system on the domain. $p = p_1$ is the reject line in the sampling plan.
- Two pre-defined thresholds p_2^- and p_2^+ present the low-value confidence and the high-value confidence of the checking system on the domain .
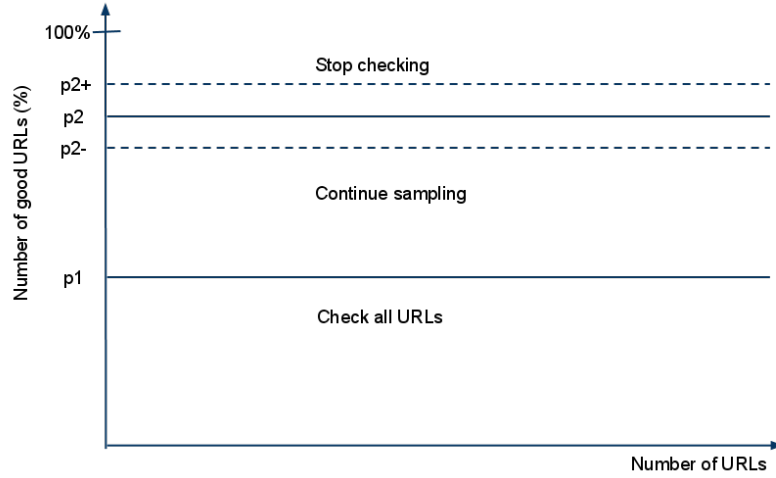
Fig. 5. Sampling plan.

- N_{pb} is the number of broken URLs in the previous check, N_{psb} is the number of stay-broken URLs in the current check, the rate r is calculated as follow

$$r = N_{psb}/N_{pb} \tag{1}$$

- A threshold p_2 presents the confidence of the checking system on the domain. p_2 is calculated as follow

$$p_2 = p_2^- + r * (p_2^+ - p_2^-) \tag{2}$$

$p = p_2$ is the accept line in the sampling plan.

Sampling Plan Procedure

- A group G of URLs is selected at a time for checking. Assume that N_c is the total number of checked URLs in the current check, N_g is the total number of good URLs, p is the percentage of good URLs calculated by

$$p = N_g/N_c \tag{3}$$

- If $p < p_1$ then the system checks all URLs of the set. If $p_2 \leq p$ then the system assumes that all the other URLs on the domain are good and stops checking. Otherwise, the system continues sampling and checking (Figure 5). Algorithm 1 shows the detail of the checking technique.

Algorithm 1 Heuristic checking algorithm

1: **procedure** CHECK
2:     **for all** domain $\in$ domains **do**
3:         **Re-check all broken URLs on domain**
4:         **Calculate p_2 using equation 2**
5:         **sampling = true**
6:         **while (sampling and (domain has unchecked URLs)) do**
7:             **Take a group of URLs for checking**
8:             **Check this group**
9:             **Update p using equation 3**
10:            **if $p < p_1$ then**
11:                **Check all other URLs**
12:                **sampling = false**
13:            **else**
14:                **if $p_2 \leq p$ then**
15:                    **sampling = false** $\triangleright$ assume that all other URLs are good
16:                **end if**
17:            **end if**
18:         **end while**
19:     **end for**
20: **end procedure**

3.3 Experiment

To compare our algorithm with the full system check, we first executed a full system check to detect all broken URLs in the LRE, then we executed the heuristic checking algorithm 3 times in simulation mode with the findings from the full system checking results for comparison.

In June 2010, the total number of URLs in the LRE was 236763, among them 45711 (19,31%) URLs were broken[1]. The first run was executed with an assumption that all URLs are good, while the second run and the third run used knowledge from the previous run. The following parameters were used: $G = 100, p_1 = 50\%, p_2^- = 90\%, p_2^+ = 95\%$.

The first run only checked 17,36% of URLs but was able to detect 73,48% of broken URLs. The second and the third checks detected 94,92% and 98,67% of broken URLs. Overall, after 3 runs, the heuristic check algorithm only checked 36,78% of all URLs on average but detected 98,67% of broken URLs.

4 Procedures to Correct Broken URLs

Given that the LRE federates LO repositories from various origins, effective handling broken URLs depends simultaneously on the heuristic checking algorithm that can be deployed to test discrete domains and automation of communication channels with content providers. Therefore, we established a schedule for check

---

[1] The experiment took place on June 23rd, 2010.

Table 1. Results

Run	Selected URLs (in number and in percentage of the total number of URLs)	Broken URLs (in number and in percentage of the total number of broken URLs)	Rate (number of broken URLs /number of selected URLs)
1	41102 = 17,36%	33587 = 73,48%	81,72%
2	105126 = 44,40%	43388 = 94,92%	41,27%
3	115005 = 48,57%	45101 = 98,67%	39,22%
Overall	710289 = 36,78% (in average)	45101 = 98,67% (max)	54,07% (in average)

sequences and conditions under which LRE service managers could act on the findings of the checking algorithm, initiate the removal of records from search results, set in motion further checks and determine when it was appropriate to return records to the search. Making these conditions explicit allows for better management of metadata quality in the LRE, shields users from negative experiences during search and retrieval and opens opportunities to collaborate with content providers in improving the quality of their own metadata.

The first step in this process begins when a broken link is first found and the LRE system generates an automated notification for providers. System controller coordinates all other components and generates reports. These machine readable reports are set to trigger a fix and can be exposed to providers using RSS, ATOM feeds or downloadable excel sheets depending on the preferences of the providers. After seven days, the LRE system is set to recheck the problem URLs. If the problem persists and there has been no other response from the provider, an automated email message is generated for the LRE system manager notifying them of URL's status, providing relevant details about the record(s) in question and contact information for the provider. Using this data, the LRE manager sends a personal email to the repository manager. The repository manager will be notified that the problem has not been corrected despite an automated report. The guidelines stipulate that after three weeks records with broken URLs will be removed from the search results. At this point we hope to begin a dialog with the repository manager, determine the source of the problem, offer advice or other assistance if needed and provide feedback more generally to repository administrators. In order to mitigate the detrimental effects of broken URL on user experience, we have instituted a time limit for how long records with broken URLs having been identified and reported to providers can remain in the LRE search. One month after the initial discovery of the broken URL, the system will initiate another check of the domain using the algorithm described above. If the problem has not been corrected, records with broken URLs will be removed from search results. This action is accompanied by an automated notification sent to the provider listing items that are no longer in the search results and the reason for their removal. Once records have been removed from the search, a

Fibonacci sequence is implemented to determine if and when the problem URLs are corrected by providers. When the check sequence indicates that the broken URL has been corrected, the records will be returned to the search results. An automated notification to providers, both machine readable and as an email, will be generated to alert the providers of their record(s)' change in status. This ability to return the records to the search takes into account scenarios in which providers are not able to respond expeditiously to reports of broken URLs without unduly impacting the quality of LREs services.

The guidelines described above take into account the need to provide flexibility of responses for LRE service managers and opportunities to improve collaboration between system administrators within the federation.

5 Discussion and Related Works

Although, recognition of the problems and attempts to grapple with the challenges of broken URLs or "link rot" are not new, most of existing approaches to automatically detect broken URLs are based on the relationship between resources or pages [2],[8], [12] which is not applicable in a LO federation where there is almost no link between two learning objects locations.

There are also some methods such as [9], [11], [3] to assure that the links are always available. These methods could be divided into two categories. The first approach involves creating copies and keeping resources in a local repository [9], [11]. This approach assumes that the LOs will not be modified or requires a check of the links on the fly to determine if and when the system needs to switch the links.

The second approach uses a persistent identifier service to resolve the URLs [3]. However, this approach only works well if such service is available for content providers sites, which is not the case in the LRE.

The proposed mechanism to manage broken URLs allows LRE managers to control the quality of LO metadata in a federation. It combines different ways to detect broken URLs and report problems to the content providers. Because the heuristic checking algorithm is able to detect broken URLs without blocking network communication it can be implemented to run in a sequence and schedule intended to provide detection, reports and for corrective actions either on the part of providers or, if necessary, on the part of LRE managers.

6 Conclusion

In this paper, we have presented a framework with a heuristic checking algorithm that allows for the correction of broken URLs in federated metadata. Future work will focus on further enhancements of this framework to detect broken URLs when metadata arrives in the LRE. This future work will involve the development of a filter to remove broken URLs from metadata collected via the Open Archives Initiative Protocol for Metadata Harvesting OAI-PMH [5] or the Simple Publishing Interface SPI [10].

References

1. Berners-Lee, T., Masinter, L., McCahill, M.: Uniform Resource Locators (URL) (RFC 1738). Network Working Group (1994)
2. Ingham, D., Caughey, S., Little, M.: Fixing the "Broken-Link" Problem: The W3Objects Approach. In: Computer Networks and ISDN Systems, 28, pp. 1255-1268 (1996)
3. Kahn, R., Wilensky, R.: A Framework for Distributed Digital Object Services. In: International Journal on Digital Libraries, 6, pp.115-123 (2006)
4. Massart, D.: Towards a Pan-European Learning Resource Exchange Infrastructure. In: Feldman, Y., Kraft, D., Kuflik, T. (eds.) NGITS'2009, LNCS, vol. 5831, pp. 121-132. Springer, Haifa, Israel (2009)
5. Lagoze, C., Sompel, H. V.: The Open Archives Initiative Protocol for Metadata Harvesting http://www.openarchives.org/OAI/openarchivesprotocol.htm (2001)
6. MELT: Final Evaluation Report, D 7.3. http://info.melt-project.eu/shared/data/melt/MELT_D7_3_Final_Evaluation_Report.pdf (2009)
7. NIST/SEMATECH: e-Handbook of Statistical Methods. http://www.itl.nist.gov/div898/handbook (2010)
8. Popitsch, N., Haslhofer, B.: DSNotify: Handling Broken Links in the Web of Data. In: Proceedings of the 19th international WWW conference (WWW2010), NC, USA (2010)
9. Reich, V., Rosenthal, D.: LOCKSS: A Permanent Web Publishing and Access System. D-Lib Magazine, vol. 7 (2001)
10. Ternier, S.: Interim Report: The Simple Publishing Interface Specification. http://ariadne.cs.kuleuven.be/lomi/images/b/ba/CEN_SPI_interim_report.pdf (2009)
11. Veiga, L., Ferreira, P.: RepWeb: Replicated Web With Referential Integrity. In: Proceedings of the 2003 ACM Symposium on Applied Computing, pp. 1206-1211, New York, NY, USA (2003)
12. Wiley, G., Thomas, W.: Improving OpenURL Metadata. Serials Librarian, vol. 56, pp. 282-286 (2009)

Preliminary Discussion on a Digital Curation Framework for Learning Repositories

Nikos Palavitsinis[1;2], Nikos Manouselis[1], Salvador Sanchez-Alonso[2]

[1]Greek Research & Technology Network (GRNET), Greece

{palavitsinis; nikosm}@grnet.gr;

[2]University of Alcala de Henares (UAH), Spain

{palavitsinis; salvador.sanchez}@uah.es

Abstract. Learning Object Repositories have met significant development during the last few years. Researchers have extensively discussed the concept of learning objects and their accompanying metadata. Metadata in specific, were introduced and developed by the community of librarians for cataloguing purposes. From the same background, digital curation has emerged as a field of research directly linked to the needs of preserving large datasets over time and platforms. This paper links digital curation with learning objects and mostly discusses the process of digital curation whereas at the same time, attempts to identify possible research directions for digital curation in LORs.

Keywords: learning resources, curation, metadata, lifecycle

1 Introduction

Learning Object Repositories (LORs) are databases used for storing and/or enabling the interoperability of Learning Objects (LOs) as defined by McGreal [13]. Because not all repositories store the actual object files, a key function of repositories is to identify the storage location of the objects and provide an indexing system that enables the efficient search and discovery of the objects [16]. A growing body of learning repositories is making digital learning resources available to the user searching for educational content on various topics, through learning repositories (i.e. MERLOT, MIT's OpenCourseWare, ARIADNE, LRE for schools, Organic.Edunet, MACE Project).

This vast amount of objects calls for specific actions to maintain them over their lifecycle and make them available for current and future generations [2]. Addressing this issue, the term "Digital Curation", which implies a transfer of existing curatorial approaches from analogue resources to their digital counterparts, was introduced at the "Digital Curation: digital archives, libraries and e-science seminar" in 2001.

Beagrie [2] defines "Curation" *as the actions needed to maintain digital research data and other digital materials over their entire life-cycle and over time for current and future generations of users.* Pennock [14] quotes that *Digital Curation is about maintaining and adding value to a trusted body of digital information for both current and future use*, adding the aspect of added value in the process of Digital Curation.

In a comprehensive paper, Campbell [3] identified some of the issues that need to be taken into account when considering the Digital Curation of both learning objects and the metadata that describes them. Specifically, Campbell concluded by pointing out that it is certain that the use of digital objects to facilitate teaching and learning will continue to grow and that the metadata requirements of the communities of practice that use these resources will become increasingly complex.

Building on this notion, this paper emphasizes on the importance of Digital Curation on Learning Object Repositories (LORs) and suggests the use of the DCC Curation Lifecycle Model [10] in Learning Repositories.

To this direction, the first section provides definitions to set a common level of understanding on the basic concepts used in this paper while the second section presents in detail the Digital Curation process through existing work and discusses the notion of applying such techniques to Learning Object Repositories. Finally, the last section identifies limitations of the specific research and suggests future directions of research.

2   Background

The term Digital Curation implies not only the preservation of digital resources or the maintenance of a collection to keep it accessible but it also includes some degree of added value and knowledge [2][8]. Curation actions can be carried out on a broad range of scientific data and resources in multiple disciplines, ranging from arts & humanities to life sciences, physical sciences, medical sciences etc [1] to all stages of the digital resources' lifecycle [10]

In general, digital learning resources are significantly different when compared to digital datasets that are used in sciences such as physics, astronomy, biology, etc. This difference also stems from the specific nature of learning object metadata as IEEE [11] states these should take into account *"the diversity of cultural and lingual contexts in which the learning objects and their metadata will be exploited"*. Supporting this, McGreal [12] stated that *LOs can be defined as any reusable digital resource that is encapsulated in a lesson or assemblage of lessons grouped in units, modules, courses, and even programmes.* Polsani [15] defined reusable learning objects (RLO) as *independent and self-standing units of learning content predisposed to reuse in multiple instructional contexts.* These definitions emphasize on the educational uses of learning resources, already providing some evidence as regards their unique characteristics.

3 Digital Curation Framework for Learning Object Repositories

The DCC (Digital Curation Centre of the University of Edinburgh) Curation Lifecycle model (Fig. 1) is an existing curation model that can be used for curation actions as it is generic enough to be applied to different contexts and serve different communities. Additionally, the authors feel that the generic nature of DCC Curation Lifecycle Model can serve as a basis for elaborating on some initial thoughts on curation issues in LORs, providing the ability to apply them to similar models dealing with digital curation.

The DCC Curation Lifecycle Model (DCC-CLM), provides a graphical high-level overview of the stages required for successful curation and preservation of data from initial conceptualization or receipt. The authors feel that each stage of the DCC-CLM should be carefully examined, always taking into account the specificity of LORs and LOs to identify challenges and issues that may arise for digital curation in the case of LORs. Examining the model in such a way, could possibly indicate whether or not curation actions will be needed for educational resources in contrast to the scientific data already being largely curated.
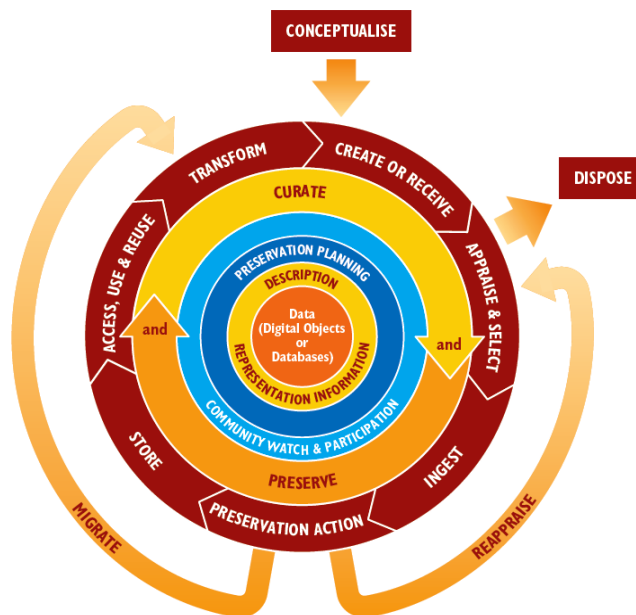

Fig. 1. DCC Curation Lifecycle Model as defined by Higgins (2008)

Lastly, even if, the model presented may not be one hundred percent applicable to the case of learning resources and learning repositories, but may as well present specific challenges for applying parts of the model in LORs. An important research question has to do with the degree to which existing metadata standards and specifications,

48

store preservation data for the learning resources. Because if specific preservation actions are proven to be relevant for learning resources as well, then for sure, existing standards should be capable of storing such data. This and similar issues will be further analyzed in future, more extensive studies.

4 Conclusions and Limitations of the Study

The present paper opened a discussion on whether or not Digital Curation can take place in the context of LORs. Overall, this paper attempted to build upon relevant studies on digital curation issues for educational metadata. By doing so, the authors attempt to open a discussion on whether or not, digital curation or some key processes it involves are relevant for Learning Object Repositories.

The first limitation of this paper lies within Digital Curation itself, as it is yet an emerging field with many different contributions from a great number of scientists that make it even more difficult to define concepts and theories. Another important limitation of this paper lies in its theoretical nature. This fact is mainly attributed to the need for an initial discussion, even on a theoretical level, on some potential research directions which will be documented in follow-up papers and examined in depth through case studies on existing LORs.

Future research that will extend the initial findings of this paper will focus on specific steps of the DCC Curation Lifecycle Model, reviewing existing literature from the curation experts as well as the learning repository ones, trying to also quantitatively prove that digital curation is significant as a context-specific curation. Once this is proven, future directions will include composing a DCC Curation Lifecycle Model for Learning Object Repositories.

Acknowledgements

References

1. Ball, A.: Review of the State of the Art of the Digital Curation of Research Data. Project Report. Bath: University of Bath, (ERIM Project Document erim1rep091103ab12). Retrieved on June 20, 2010 from: http://opus.bath.ac.uk/19022/2/erim1rep091103ab12.pdf (2010)
2. Beagrie, N.: Digital Curation for science, digital libraries, and individuals.International Journal of Digital Curation, 1(1). Retrieved June 22, 2010, from: http://ijdc.net/index.php/ijdc/article/viewFile/6/2 (2006)

3. Campbell, L.: Learning Object Metadata (LOM), DCC Digital Curation Manual, S.Ross, M.Day (eds), Retrieved June 20, 2010, from http://www.dcc.ac.uk/resource/curation-manual/chapters/learning-object-metadata (2007)

4. Carrier, S.: The Dryad Repository Application Profile: Process, Development, and Refinement, DOI: http://hdl.handle.net/1901/534 (2008)

5. Currier, S., Barton, J., O'Beirne, R., & Ryan, B.: Quality assurance for digital learning object repositories: issues for the metadata creation process. ALT-J, 12(1), pp. 5-20 (2004)

6. Chen, Y., Chen, S., & Lin, S.: A metadata lifecycle model for digital libraries:methodology and application for an evidence-based approach to libraryresearch. In *Documents in Information Science, Working Papers Series*.Retrieved June 18, 2008, from http://www.ifla.org/IV/ifla69/papers/141e-Chen_Chen_Lin.pdf (2003)

7. Collis, B. & Strijker, A.: Technology and human issues in reusing learning objects, Journal of Interactive Media in Education, vol. 4, pp. 1–32, (2004)

8. Conyers, A., Dalton, P.: Preservation of e-Learning Materials: an attitudinal study, Birmingham City University. Retrieved on June 23 from: http://www.jisc.ac.uk/media/documents/programmes/preservation/elopres.pdf (2008)

9. Greenberg, J.: Theoretical Considerations of Lifecycle Modeling: An Analysis of the Dryad Repository Demonstrating Automatic Metadata Propagation, Inheritance, and Value System Adoption, Cataloging & Classification Quarterly, 47(3), pp. 380 – 402 (2009)

10. Higgins, S.: The DCC Curation Lifecycle Model, International Journal of Digital Curation, 3(1), http://www.ijdc.net/index.php/ijdc/article/view/69/69 (2008)

11. IEEE: Draft Standard for Learning Object Metadata. Retrieved on June 22, 2010 from http://ltsc.ieee.org/wg12/index.html (2001)

12. McGreal, R.,: Learning objects: A practical definition. Retrieved June 15, 2010 from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.118.9253&rep=rep1&type=pdf#page=25 (2004)

13. McGreal, R.: A typology of learning object repositories, [pre-print]. Retrieved June 22, 2010 from http://hdl.handle.net/2149/1078 (2007)

14. Pennock, M.: Digital Curation: a life-cycle approach to managing and preserving usable digital information. Library and Archives Journal, 1. Retrieved: June 22, 2010, at http://www.ukoln.ac.uk/ukoln/staff/m.pennock/publications/docs/lib-rch_curation.pdf (2007)

15. Polsani, P.R.: Use and Abuse of Reusable Learning Objects. Journal of Digital Information, 3(4) (2003)

16. Richards, G., McGreal, R., Hatala, M., & Friesen, N.: Learning object repository technologies: Portals for on-line objects for learning. Journal of Distance Education, 17(3), pp. 67-79 (2002)

Automatic Keywords Extraction –
a Basis for Content Recommendation

Ivana Bosnić[1], Katrien Verbert[2], Erik Duval[2]

[1] Faculty of Electrical Engineering and Computing, University of Zagreb,
Unska 3, HR-10000 Zagreb, Croatia
[2] Dept. Computerwetenschappen, Katholieke Universiteit Leuven,
Celestijnenlaan 200A, B-3001 Leuven, Belgium
ivana.bosnic@fer.hr, {katrien.verbert, erik.duval}@cs.kuleuven.be

Abstract. This paper describes a use case for an application that recommends learning objects for reuse and is integrated in the authoring environment. The recommendations are based on the automatic detection of content being authored and the context in which this resource is authored or used. The focus of the paper is automatic keyword extraction, evaluated as a starting point for content analysis. The evaluations explore whether automatic keyword extraction from content being authored is a sound basis for recommending relevant learning objects. The results show that automatically extracted keywords are suitable for this purpose, if some observed issues are appropriately addressed.

Keywords: content, reuse, recommendations, keywords, keyword extraction

1 Introduction

Content reuse today – although somewhat increased by new technologies and interfaces to aggregate and remix the content – is still not straightforward for mainstream authors of educational content. Barriers limiting content reuse include the immaturity or absence of support for discovering and reusing learning content in authoring tools and difficulties associated with combining and referencing reused learning materials [1]. The goal of our research is to analyze the reuse potential of learning objects and to support their discovery, recommendation and reuse within available authoring tools. Recommendation is based on both the content being authored and the context in which the content is authored or used. This paper analyzes whether the results of automatic keyword extraction from the content being authored can be a basis for recommending resources relevant to the author. These keywords are generated based on both the on-the-fly analysis of content the author is editing, and context data that is available in an authoring or learning environment. Our research, presented in this paper, focuses primarily on the results of keyword extraction analysis, and on describing the process of content reuse which is based on this topic analysis and integrated in the authoring environments.

The paper is organized as follows: The application use case is presented in section 2. Automatic keyword extraction services are presented in section 3. Section 4 describes the comparison between two keyword generation services, while section 5

describes the keyword evaluations in the application prototype. The paper wraps up with conclusions and future work in section 6.

2 Application Use Case

The application purpose is to help authors of educational content, by:

- recommending relevant content during authoring, without manual searching by the author;
- enabling easier content reuse and remix, particularly of small fragments, by referencing or using advanced copy-paste functionalities;
- integrating these functionalities in the authoring or learning environments through extensions of applications such as wikis, blogs, or presentation software.

One of the application use cases can be described with the following steps:

1. The user authors the content in his authoring environment (e.g. Wiki);
2. The application collects the content being authored, together with context data available (e.g. age range, difficulty level) and proposes the recommendations;
3. The user views the recommendations to decide whether they are relevant to him;
4. If the content is useful for either copying partly or just for getting ideas, then the user chooses to reference this content. The reference is automatically inserted in the content being authored, in the appropriate format (e.g. WikiMarkup, HTML markup or plain text);
5. As the user continues to edit the content, the changes are incorporated and new recommendations are presented.

In order to discover the resources, the application, integrated in the authoring environment, analyzes the content being authored. An automatic keyword extraction service extracts keywords from the text. Additional context is obtained from the authoring or learning environment (the purpose of the course, the preferred format of resources to be reused, etc.). Together with the keywords, this context data is used to search and retrieve relevant resources from content providers, including large learning object repository networks and social bookmarking websites.

3 Keyword-Based Content Discovery

The usual way of querying content providers is by using keywords as search terms. In the case of repositories containing learning object metadata, search terms can be used to query fields such as title, description or keywords and further refined by using additional metadata fields that capture the context in which the learning content is used. In this section, automatic keyword extraction services that can be used as a basis for generating search terms are presented.

Keyword extraction services can be divided in two groups, based on the usage of algorithms for constructing the semantic context:

- **term extraction services** – this group of services extracts the keywords from a text. Examples include Yahoo Term Extraction Web Service [2] and Fivefilters [3].

- **semantic entity extraction services** - this group of services not only extracts the keywords, but also detects the concepts related to the text, which are not present in the text itself. These services often have semantic linking features, i.e. they include additional encyclopedia links, images, articles, etc. Examples of such services are Zemanta [4], OpenCalais [5], Evri [6] and AlchemyAPI [7].

Most services provide interfaces for online use, mainly REST or SOAP. The usual result outputs are represented in RDF, XML, JSON or plain text. The services mostly use keyword classification schemes, such as the DBpedia ontology [8], Wordnet [9] or dmoz Open Directory Project [10]. Some services have their own entity databases.

Several comparisons of keyword extractors and semantic APIs exist. Zemanta and OpenCalais are recommended in [11], AlchemyAPI and Evri in [12], while [13] focuses on the characteristics of services for semantic tagging, without specific recommendations. Services from both groups were evaluated to compare and contrast their efficiency and potential use within our application:

- **Yahoo Term Extraction Web service** (Yahoo in the following text) is a popular keyword extractor with a RESTful interface, which returns up to 20 keywords that are found in the text. The keywords are not ranked internally. This service is successfully used in automatic metadata generation frameworks like SAmgI [14]. As SAmgI generates metadata for a subset of objects in the GLOBE network of repositories [15] that is used in our research, this was an additional reason to evaluate it for our purpose.

- **Zemanta** is a semantic entity extraction service with both RESTful and JavaScript interface. It returns up to 8 ranked keywords. Additionally, it recommends images, links to ~20 Web sites (Wikipedia, Youtube, IMDB, etc.) and blog/news articles from ~10000 sites. Optionally, Zemanta provides the keywords according to the *dmoz* keyword classification. Moreover, its extraction process can be influenced by emphasizing selected words.

The following section describes the comparison of these two services and the evaluation of their potential for automatic content discovery. In this evaluation, Zemanta and Yahoo were used to extract the keywords from several already existing presentations. These keywords were graded by users. In addition, the users were asked to manually provide keywords for the presentations and the keywords extracted by Yahoo and Zemanta were compared with these, user-generated keywords.

4 Evaluation of Keyword Extraction Services

4.1 Evaluation Methodology

The goals of this evaluation were to test the keyword extraction services with the examples of existing educational content, to compare the keywords extracted by Zemanta and Yahoo, and also to compare those to the user-generated keywords.

In the evaluation, 9 presentations were used – 3 for each topic (open source, databases and gravity force), different in their characteristics, which is expected to influence the quality of extracted keywords. A topic of open source mostly uses

general words, descriptions and a smaller number of specific terms; a topic of databases is a more specific one, while an explanation of a gravity force contains formulas and lots of specific physics-related terms.

The presentations were gathered from Google's first page result on queries for "what is open source", "what is database" and "what is gravity", with file type filtering for Microsoft PowerPoint presentations. The excerpts chosen were text-only contents of 3 adjacent slides of each presentation, to better describe the context. Some slides had examples from other fields to help illustrate the concepts. Some texts were written as sentences, while others had only a few words per bullet. An assumption is made that the extraction services will have less success with shorter texts, partial sentences and the examples from different fields. However, these are often found in presentations, thus it should be tested whether keyword extraction gives satisfying results in those cases, too.

Six users were involved in the evaluation, which consisted of two parts:
1. The users were asked to read 9 text excerpts, and write the queries which they would use in search engines. They could type as many queries as they wanted.
2. For each of the 9 presentations, the users were presented with 8 keywords from Zemanta and the first 10 keywords from Yahoo. They were asked to grade the relevancy of each keyword, which, of course, could consist of one or more words.

4.2 Automatically Extracted Keywords

Two keyword extraction services were compared by the following criteria:

User keyword relevancy grading. Fig. 1 shows the average of relevancy grades per presentation. Zemanta is graded higher in 7 of 9 presentations.

If the same average is calculated for 3 **presentation topics**, it shows that the keywords from both services are graded higher as the topic specificity increases (Fig. 2). In all three topics, users have graded the keywords from Zemanta higher.
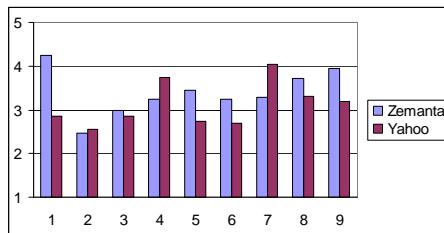

Fig. 1. The average of keyword relevancy grading per presentation. For each of the 9 presentations (X-axis), the users were grading the relevancy of 8 keywords from Zemanta and 10 keywords from Yahoo, with grades 1-5 (5 being the most relevant). The average of grades is calculated for two services separately (Y-axis). The grades for the same keywords were equally distributed among users.
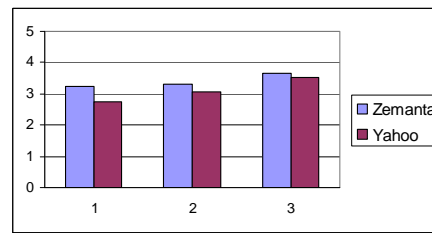
Fig. 2. The average of keywords relevancy grading (Y-axis) per presentation topic (1 – open source, 2 – databases, 3 – gravity) on the X-axis.

Fig. 3 shows the average of user grading for the keywords for each of the 8 Zemanta ranks. In general, the grading tends to drop as Zemanta ranking lowers, which justifies the decision to make queries by combining the highest Zemanta ranked keywords. Yahoo provides the keywords in order of appearance in the text, without any ranking mechanism, so this service could not be evaluated in this way.
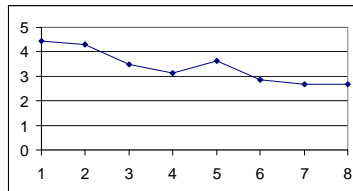

Fig. 3. The average user grading of keywords per particular Zemanta rank. The X-axis presents 8 Zemanta internal ranks. The Y-axis presents the average of user grades for the keywords in each Zemanta rank. In this diagram, the keywords from all 9 presentations were included.

4.3 User-Generated Keywords

To see how different the user keywords are from automatically extracted ones, the comparison of these two sets was made. This comparison is used to analyze how different are the results provided by keyword generation services from the user-proposed search queries - keywords. Only the keywords shared by at least two users were included, to provide more comprehensive and relevant results.

Two comparisons were made:

- exact match – checking whether the exact user-generated keyword was included in the list of extracted keywords. The difference in singular/plural form of nouns was counted as exact match, as most indexing services used can internally match these.
- similar match – checking whether a similar user-generated keyword was in the list of automatically extracted ones. The keywords as subsets of other keywords are considered similar (e.g. keyword "open source" is similar to "open source definition"), as well as the ones which could be matched with grammatical or syntax changes (e.g. keyword „gravity law" is similar to „law of gravity").

Fig. 4 shows the number of common user-generated keywords and the number of matches with automatically-generated keywords. The results show that the more important keywords – the ones which are common to more users – have a higher match rate. This is especially visible if similar matches are considered, which is an argument for use of advanced methods to find the keywords similar to automatically generated ones.
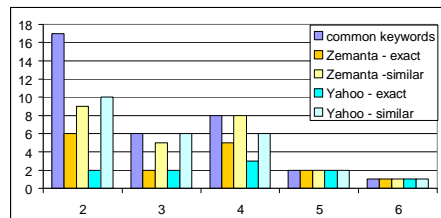
Fig. 4. The number of exact and similar matches between user-generated and automatically extracted keywords, in comparison to common keywords – the ones proposed by more than 2 users (Y-axis). The keywords are distributed by the number of users which proposed this keyword, shown in X-axis. With the exact match, Zemanta matches more words than Yahoo in 2 sets and in 3 is equal to Yahoo. With similar match, Yahoo matches more words than Zemanta in 2 sets, less words in 1 set and in 2 is equal to Zemanta.

The following section describes the initial keyword evaluations carried out in the application prototype environment, where the keywords had to be extracted during the presentation authoring. This approach poses additional challenges in text preparation and automatic keywords extraction, which are described in the following text. In these evaluations, the Zemanta extraction service was used.

5 Keyword Evaluations in the Authoring Environment

5.1 Evaluation Methodology

Two keyword evaluations were carried out. The overall goal of these evaluations was to determine whether automatic keyword extraction from content being authored is a sound basis for recommending relevant learning objects to the author. More specifically, the relevancy and ranking of the extracted keywords were evaluated. The evaluations were done as a part of an overall evaluation according to the *discount usability engineering* principles [16]. Therefore, it should be noted that these are not the results of thorough evaluations, rather of basic, initial user tests.

The users were asked to create an informative presentation about a programming topic familiar to them. The time was limited to 15 minutes. Specifically, the users were given an empty presentation template in the MediaWiki service, enhanced by the WikiPres extension – a MediaWiki plugin for collaborative presentation authoring using WikiMarkup [17]. They were advised to make use of the recommendation application, and to properly attribute reused resources.

Once the presentation was finished, the users chose one of the more content rich slides they authored (not the title or introduction slide). They were presented with 8 keywords generated for that slide and asked to rank the 5 keywords they considered the most relevant. Fig. 5 presents the relation of the user ranking and Zemanta ranking. Fig. 6 shows the averages of user rankings for keywords in the same Zemanta rank.

5.2 Evaluation 1

Four users ranked the keywords extracted and ranked by Zemanta. Of course, the generated keywords were different for each user: the user ranking is compared with that of Zemanta.
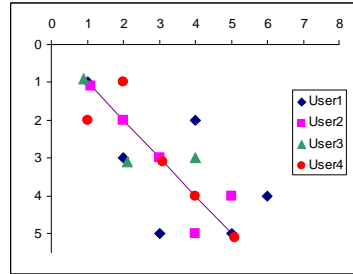
Fig. 5. The relation between the user and Zemanta ranking. The X-axis presents Zemanta ranks, from 1-8 (1 being the highest-ranked). The Y-axis presents user ranks from 1-5 (1 being the highest-ranked). The ranking itself is marked with a dot of a different type for each user. Ideally, the user and internal rankings would be identical, with all the dots on a diagonal line. Here, the dots are dispersed, but still near the diagonal line. The majority of dots are placed in the first five columns (Zemanta rank 1-5): this shows that users and Zemanta largely agree on what are the 5 most relevant keywords.
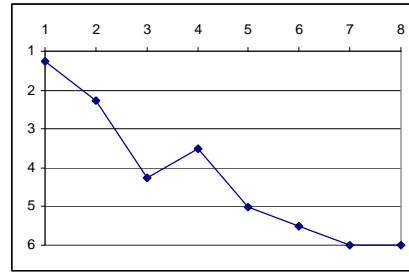
Fig. 6. The average user ranking. The X-axis presents Zemanta internal ranks. The Y-axis presents the average of user rankings for all keywords in a particular Zemanta rank. For instance, the highest-ranked keywords by Zemanta got 1, 1, 1 and 2 as user ranks, which gives an average of 1.25 out of 5. The diagram shows that the user ranking lowers together with Zemanta ranking; the keywords with the lowest Zemanta rankings are not among the most relevant to the users. For this calculation, the keywords not being among the 5 most relevant were given the rank 6.

Lessons learned. The interpretation of evaluation results shows that users mostly agree with Zemanta ranking, which is important for our purpose. Looking into the example of extracted keywords, it can be seen that there are also some irrelevant keywords. In addition, during the evaluation, the following issues were observed:

- **Content *cold start***. At the beginning of authoring, a number of words should be present for satisfactory results. Otherwise, irrelevant initial keywords are extracted.
- **Semantic relation of words**. Typically, users would test the application by typing a few words to start with, without making any sentence structure or phrases. As Zemanta tries to extract semantic relations from phrases, a text where the words do not make at least a phrase poses a problem for keyword extraction. The influence of this style of writing on keyword extraction should be further evaluated.
- **Unnecessary text markup**. The content submitted to the keyword extraction service contained XML tags, which were internally defining the layout. These were not removed automatically, and thus influenced the keyword extraction.
- **Ambiguity**. For small-size texts, keyword generation was sometimes biased by particular meanings of phrases, as the phrase context could not be determined.

Implementation modifications. Several modifications related to keyword extraction were implemented after the first evaluation:

- **Including the content from previous slides.** To address the cold start issue which occurs when a new slide is started, the content from two previous slides has been included in the keyword extraction, to provide a larger context. As even the completed slides can have a small number of words, this can be very useful. However, a problem can occur if there is a major topic change in adjacent slides.

- **Title emphasis.** To help solving semantic problems, the slide title was marked as *emphasized*, which is an additional Zemanta option to focus the extraction on particular words. Depending on the writing style of the author, this can improve the keyword extraction, but it can also degrade it (e.g. slide title "History", as the history of a technology, could bias the generator towards general human history).
- **Text cleaning.** The text submitted to the keyword extraction service was additionally cleansed of XML tags, as it was not done by Zemanta automatically.

5.3 Evaluation 2

The goal of the second evaluation was to analyze the influence of different text scenarios in presentation authoring: including an example, changing the sub-topic of the presentation and writing about a more general topic.

Four users were involved in the evaluation. The process was the same as in the first evaluation: authoring the introductory slides on a topic in the computer science field. To analyze the text scenarios, one user was asked to include a real-world example, while a second user was asked to focus on a specific subtopic in some slides. The third user was writing about a more general topic ("open source"). The fourth user was writing a presentation without a specific scenario. It was expected that the different text scenarios and one more general topic would lower the similarity between the user and Zemanta keyword ranking.

Fig. 7 and Fig. 8 present the evaluation results in the same way as the diagrams in the first initial evaluation. Fig. 7 shows the relation of the user ranking and Zemanta ranking. Fig. 8 shows the averages of user rankings for keywords in the same Zemanta rank. The highest-ranked keyword is ranked on average with 1.75, and the user relevancy ranking average drops as Zemanta ranking lowers, to an average of 5.5, for the fifth keyword.
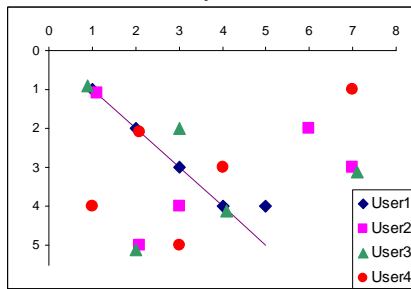

Fig. 7. The relation between user and internal ranking. The X-axis presents Zemanta internal ranks, from 1-8 (1 being the highest-ranked). The Y-axis presents user ranks from 1-5 (1 being the highest-ranked). The actual ranking is marked with a dot of a different type for each user.
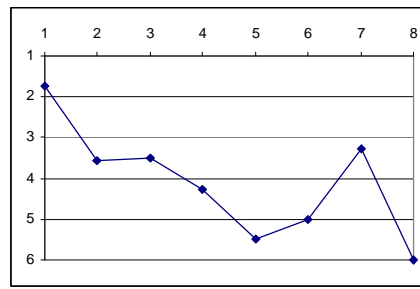
Fig. 8. The average user ranking. The X-axis presents Zemanta internal ranks. The Y-axis presents the average of user rankings for the keywords in a particular Zemanta rank. For this calculation, the keywords not being among the 5 most relevant were given the rank 6.

Some keywords most relevant to users occur in the lower Zemanta ranks (6-8):

- an example from banking for database systems was included, which caused the keywords related to the example (e.g. "bank") to be extracted (User 2);
- in the presentation about a less specific topic ("open source"), a keyword which was relevant to the user was in the lower Zemanta ranking (User 3);
- in the presentation about HTML, the user was creating a slide specifically for dynamic HTML. As the previous slides were about HTML in general, the keywords were more related to HTML. The most important keyword – "*dynamic HTML*" – was ranked seventh by Zemanta (User 4).

One way to solve these problems is providing a larger context, from the content itself (additional slides) or from the external environment. Another solution is to give users the option not to include the context of previous slides (useful for changing topics) and not to emphasize the slide titles (useful for misleading titles), but this could reduce the application usability as the user needs to manually select these options. Detecting the change of topics can be done based on the slide layout changes, as some authors divide the presentations in subtopics with slides of a particular layout, or by heuristics based on the topic changes per each slide or per slide sets.

5.4 Lessons Learned

The majority of best-ranked keywords in these two evaluations were in the first 5 of the keywords suggested by Zemanta. Due to the specifics of the scenarios, some keywords which users chose as most relevant were in the lower Zemanta ranks.

The users were creating presentation texts for evaluation purposes, not for real presentations. Therefore, some presentations contained very few words, which were not semantically connected. Although some authors prefer to create presentations without many words, the majority of authors still write at least a set of phrases on the slides, which is necessary for obtaining the relevant terms from keyword extraction services.

6 Conclusions and Future Work

The evaluations performed confirm Zemanta as a sound basis for the intended purpose, based on the results and available features such as proposing the keywords - mostly abstractions - which are not present in the text, emphasizing the words to influence the extraction and internal ranking. The five highest-ranked keywords extracted by Zemanta will be used, as the users graded these keywords on average with more than grade 3 (the average of grades 1-5).

Future improvements of keyword extraction include the use of keyword classification schemes to detect similar terms and exploring folksonomies as an additional way to find tags that are often used together. To address the problems observed in various text scenarios, two options will be implemented if the user wants to adapt the keyword list: removing a keyword from the list and simple user rating. If rating is used, Zemanta ranking will be combined with the user rating to form a more relevant keywords list.

Several questions remain: Will extracted keywords be found in metadata entries? Do more relevant keywords in the queries produce more relevant recommendations? What can be done not to omit the relevant content, while using this approach? These questions are certainly important and should be investigated.

Besides the keywords, other research segments not discussed in this paper, such as including context information from the environment, will influence the quality of final recommendations. Therefore, further research will focus on usability of content reuse workflows, extraction of context from the authoring environments or learning management systems and mapping such context to learning object metadata. The proposed solutions will be evaluated using the developed prototype application.

References

1. Wirski, R., Brownfield, G., Oliver, R.: Exploring SCORM and the national flexible learning toolboxes. Proceedings of the 21st ASCILITE Conference, Perth. (2004).
2. Term Extraction Web Service - YDN,
   http://developer.yahoo.com/search/content/V1/termExtraction.html.
3. term extraction | fivefilters.org, http://fivefilters.org/term-extraction/.
4. Blog Smarter | Zemanta Ltd., http://www.zemanta.com.
5. Home | OpenCalais, http://www.opencalais.com/.
6. Developer Portal - News - Evri, http://www.evri.com/developer.
7. AlchemyAPI - Transforming Text Into Knowledge, http://www.alchemyapi.com/.
8. Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., Hellmann, S.: DBpedia - A crystallization point for the Web of Data. Web Semantics: Science, Services and Agents on the World Wide Web. 7, 154-165 (2009).
9. Fellbaum, C., others: WordNet: An electronic lexical database. MIT press Cambridge, MA (1998).
10. ODP - Open Directory Project, http://www.dmoz.org/.
11. Entity Extraction & Content API Evaluation « ViewChange Development Blog, http://blog.viewchange.org/2010/05/entity-extraction-content-api-evaluation/.
12. Puzzlepieces – Comparing NLP APIs for Entity Extraction,
    http://faganm.com/blog/2010/01/02/1009/.
13. Dotsika, F.: Semantic APIs: Scaling up towards the Semantic Web. International Journal of Information Management. 30, 335-342 (2010).
14. Meire, M., Ochoa, X., Duval, E.: Samgi: Automatic metadata generation v2. 0. Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications. p. 1195–1204 (2007).
15. GLOBE | Connecting the World and Unlocking the Deep Web, http://globe-info.org/.
16. Nielsen, J.: Usability engineering at a discount. Proceedings of the third international conference on human-computer interaction on Designing and using human-computer interfaces and knowledge based systems (2nd ed.). (1989).
17. Bosnić, I., Pošćić, A., Ačkar, I., Žibrat, Z., Žagar, M.: Online Collaborative Presentations. Proceedings of the 32nd International Conference on Information Technology Interfaces - ITI 2010. pp. 1-6 , Cavtat/Dubrovnik, Croatia (2010).

Conversion of the YDP Learning Content to Common Cartridge Package

Jaroslaw Dziedzic

Young Digital Planet S.A. Educational ePublishing, Content and Technology,
Ul. Slowackiego 175, 80-298 Gdansk, Poland
jaroslaw.dziedzic@ydp.com.pl
http://www.ydp.eu

Abstract. This paper describes experiences earned during the development of a software application that could convert some YDP specific learning content into Common Cartridge.

1 Introduction

Young Digital Planet has almost twenty years of experience in digital learning content creation. We have created several generations of applications and content. The experience and ideas have accumulated over years, resulting in broad range of top-quality educational content, advanced applications and tools, as well as effective and flexible data formats for storing and processing the digital content.

We observe emerging standards in the field of digital learning content data formats with interest, and we keep evaluating them and assessing their usefulness for the packaging and distribution of our content. The ASPECT project gives us the opportunity to take a look at the new Common Cartridge emerging standard, analyse it thoroughly and decide, if we can implement it in our products, and how to do it.

It would not be possible to reach a deep understanding of any new concept just by reading the documentation. Therefore, in order to gain some knowledge and experience, we decided to produce a tool that will try to convert our existing content to the CC format. Our goal was to have our content with all, or most of its features, packed into the Common Cartridge and running in the CC player.

2 YDP Content Concept vs. Common Cartridge Concept

Right from the start of reading the CC specification it became obvious, that the CC concept differs in many ways from our existing content. These differences could break the project just at the beginning, but fortunately we found some methods of changing our content to suit the Common Cartridge way. Let's look first at the key concepts for YDP content and CC.

2.1 YDP Content

Our content is usually organised in lessons that consist of several pages. Each page contains some educational content (text, pictures, slideshows, video, sound clips, interactive animations etc.) and exercises related to this content. The student learns not only by reading/viewing the static or interactive content, but also by solving exercises and observing the feedback from the computer. The process is interactive, exercises are tightly integrated with the learning content and it is impossible to separate them from the material presented on the page. The scoring of all the exercises in the lesson is usually presented and summed up on the last page of the lesson, and the student has the possibility to return back to any exercise, read the lesson again, sort out his answers and see the result.

We use several different kinds of interactive exercises in addition to typical multiple choice and fill-in-blanks. Pages can have sub-pages (using tab-sheet controls), buttons that launch some actions or just open popup pages with additional content etc. Everything is laid out by our graphic designers to form harmonious and attractive environment for learning content and exercises.

Our entire content is presented to the student by means of a dedicated player written in Adobe Flash. We have no concept of static content displayed outside of the player − for instance as a PDF or HTML page. Everything is played by our flash application. The player must always accompany the XML and data files.

2.2 Common Cartridge

As we understand from the specification, the Common Cartridge is based on the concept of the complete separation of a learning content and exercises. The only reason the exercises are included − is to test the student's knowledge, acquired while reading/viewing the static resources. Exercises cannot be mixed with the learning content and do not participate in student's interactive learning experience. They simply serve as an assessment tool or they form a pool of exercise resources (Object Banks) aimed for a teacher, who wants to use them in his custom-made lessons and examinations.

While there is really great freedom about the way the static content can be built (HTML, PDF, flash objects etc.), this is not the case with interactive exercises. They can be described only by means of the old QTI 1.2.1 data format (constrained even further by the CC profile), and played back by the player integrated with the server software. There is no place for the dedicated player here, so the content producer has really limited possibilities with regard to choice of exercises and the way they are presented to the student. In fact − even though the CC specification claims that it can handle 6 exercise types − there are just two distinct exercise types to choose: multiple choice (in 3 simple variants) and fill-in-blanks. And even these two types have important limitations that allow for their use in really simple cases only.

On the other side − the lack of a dedicated player can be an advantage for some content producers that are interested in simplicity. Creating only the

authoring software for QTI 1.2.1 could be − in theory − simpler than creating both authoring software and a dedicated player. It would be true − if the QTI 1.2.1 was straightforward and easy to implement. Unfortunately − it is not the case, and I will explain it in details later in this document.

One advantage of a CC approach without the dedicated player can be easily seen. The CC package doesn't have to care about communication with the LMS server, there is no need to think about the exercise status, scoring etc. − because all these issues are handled by the standard player integrated with the server. The content producer simply has to describe exercises using some standard data format. It is a really great idea. But the choice of QTI 1.2.1 as the data format for interactive exercise representation, and restricting this format in the CC profile, makes it difficult to create anything except some really simple content.

2.3 Key Differences

The main difference between the YDP and CC content is the existence of the dedicated player for the YDP content. It doesn't exist in Common Cartridge and all interactive content for CC must be described by means of QTI 1.2.1 files.

Another important difference is the complete separation of learning and examination in CC − while in YDP content most exercises are alternated with static content on lesson pages, and they together form an interactive learning experience. Obviously, we also make use of the "examination" concept (just like in CC) and the series of exercises that form an assignment also exists in our content − but they are not dominant. We believe, that solving exercises is a part of a learning process and exercises should not be limited to examine the students only.

The really limited number of exercise types available for CC is another problem for implementing a conversion of our content. It is not possible to use connection exercises, word jigsaw, sentence jigsaw, element grouping, text item identification, crosswords, marking elements, colouring pictures, solving mathematical equations and many more It makes the straight conversion impossible.

And last, but not least − the visual appearance. We put much effort in graphic design of our lessons. Pages are beautifully laid out by graphic designers to satisfy both the aesthetics and the ergonomics. As this would not be a problem to obtain the same level of aesthetics on the Common Cartridge static pages − it is impossible to achieve it with assignments. Using the old QTI 1.2.1 there is no way to express the layout and graphical appearance of the page. We must reconcile ourselves to simple text exercises laid out one per page. We are aware of the fact that some aspects of visual appearance can be defined using "text/html" instead of plain text. But this potentially powerful feature is not well explained in the specification − so we don't know, what is the intended scope of the HTML that can be used in QTI <mattext> elements. It would be rather unrealistic to believe, that one can put a complicated HTML there − with pictures, tables or JavaScript for example. So we believe, that "text/html" means just some text attributes − like bold, italic or font colour and size. While it is useful to have

such possibilities, they would not allow us to design the layout and create a visually stunning page.

2.4   The Solution

As described before, it is impossible to make a straight conversion of our content into CC, because of the limited number of exercise types available, and because of the strict separation of lessons and exercises in CC. However, we've invented a workaround for this problem.

The CC specification allows us to put any "web content" into the package, so we can also include our player − which is, in fact, just a big SWF file that can be run in a flash plug-in in a HTML browser. The player will not be able to communicate with the LMS server (because it is not the CC way of doing things), but it will be able to play our content without any problems, with all kinds of exercises and multimedia files, with attractive appearance and with the pages layout that we designed. But from the CC server's point of view − it will be just the static content and the exercise results and state will not be reported to the server. From the student's perspective, however, lessons with educational content alternated with exercises still form a very valuable learning experience: interactive exercises can check student's answers, show hints, display feedbacks and show results. The results and exercise state are transient and will not survive when the student closes his session with the system, but they assist the student during the course of learning.

And then it comes to the examination. We have to stick with the CC convention and try to represent some exercises in QTI 1.2.1. We decided that for each lesson we will prepare some additional pages with simple exercises that can be converted to QTI, and they will form an examination after the lesson.

So each of our packages will consist of an interactive lesson with exercises that do not report results to the LMS, and the assignment with some exercises related to that lesson − prepared in QTI 1.2.1.

3   Implementation of the YDP → CC Converter

We have designed and implemented the application that takes a number of lessons from our existing content and includes them in the Common Cartridge package. The application is written in Java, and it has the user interface that makes it possible to select lessons for conversion and specify the conversion type. The lesson data can be interpreted in two different ways, depending on the type selection made. It can be either "lesson" or "test".

3.1   Lesson Conversion

If a "lesson" conversion type is selected, all the data files (various XML and media files) are copied into the package directory structure and appropriate entries for each file are added in imsmanifest.xml file. The directory containing

all the files looks almost identical, like in our original content; the location of files in relation to each other is preserved. The only change to the original content was the necessity to change filenames of most xml files − I will explain it in details later.

For each directory that contains the data for a single lesson, a HTML file is generated during the conversion. This HTML contains a Flash Player applet, that loads our dedicated player (in SWF file), and this player loads end executes the lesson content described in XML files.

The dedicated player is also copied to the package into a separate directory, and is referenced from each generated HTML file. All the files described above are registered in imsmanifest.xml file as a "webcontent" data type. From the server's point of view, they just form a HTML page with the embedded SWF file.

Everything should run on the CC server as a static content − i.e. without the possibility to report the exercise results to the server, and without the persistent storage of exercise state. However, the content in the cartridge should look and behave identically to our existing applications.

We validated generated packages using the Cartridge Validator application (Common Cartridge Test Tool). This application proved really useful and helped to find some bugs in the package structure. Finally, we had the validated package that truly conformed to the CC specification.

The final step for static content conversion was to run it in some reference Common Cartridge player and verify that it works just like expected. The problem was the apparent lack of the reference CC player on the IMS Global site. We couldn't find this important resource, so the only option was to use the player from Icodeon for testing purposes. Thankfully, we had no difficulties getting the access to Icodeon site to try their player.

We had only one problem during this stage of implementation. The package generated just like described above didn't seem to work It seemed strange, because we expected that the package that validates correctly will, at least, show some signs of working But we had just a grey rectangle. After some debugging we found that most of our XML files looks like they did not exist on the server (we got error 404 for them), while some others XML files and all media files (JPG, SWF, MP3 etc) were okay. All the missing files had something in common: they were XML files, but their filename extension was different: .page or .flo for example. The possible explanation of the problem could be the MIME-type settings on the server (ignoring all unknown file types), or the package import algorithm, that rejects unknown file types. Anyway − we had to rename all the files to XML. It was not that easy, because there was the need to find all internal references to renamed files in XML and change them. Of course, we finally succeeded and we've got our content perfectly running on the Common Cartridge server!!!

3.2 Exercise Conversion

If a "test" conversion type is selected, the YDP content is treated in a different way. It has to be converted to QTI file. Since this conversion is impossible for our stock content, we use the content created especially for this purpose. It contains only two exercise types (single/multiple choice and fill-in-blanks), and text modules for presentation of questions. We tried to get the most out of the QTI specification and used advanced features, like feedbacks. However, the results were mixed, some features that we wanted to implement were acceptable in the CC specification, some were not. The specification is not clear in many places, and it was an additional difficulty. We were also surprised with the fact, that even these simple exercise types allowed in the "CC profile" do not always work as expected: we couldn't convert our fill-in-blanks exercises to CC, because they usually (in fact − almost always) contain many blanks in a text. And it looks that "CC profile" restriction to the QTI doesn't allow us to have more than one gap in a single exercise, and this single gap must be at the beginning or at the end of a sentence...

But the main problem was the exercise scoring method in QTI 1.2.1, which is overly complicated and difficult to implement. Wouldn't it be better to select some simple and standard scoring methods for the exercise, instead of this big and complicated <resprocessing> section for each item? Anyway − the problem with <resprocessing> is much more serious: despite its complexity, its abilities are really small. This is because of the CC Profile restrictions that force to use the SCORE variable only in "all or nothing" manner. It means that if we have the multiple choice exercise with ten checkboxes, five of them being correct answers − than we have no way to express that we want 20 points for each correct answer. All we can do is to set 100 points if all five correct checkboxes were selected. In case of any error, we have to set 0 points for the entire exercise. This single deficiency makes the whole CC system useless for the representation of any high quality content. Really, only very simple exercises, like choice with a single correct answer (with radio buttons instead of checkboxes), or fill-in-blanks with a single gap, can be correctly scored.

The YDP converter creates QTI files that can be positively verified by the Common Cartridge Validator. However, when we run them in the CC player, which is available for us, only fragments of exercises appear and they do not act as expected. Perhaps it is because the Icodeon player is also under development and maybe we just run into some not implemented features, or just bugs. But at this stage of development we really need a rock-solid reference player to verify our implementation in practice.

3.3 Problems and Suggested Solutions

The most difficult part of this project was to study the specification. The Common Cartridge Specification document [1] is not precise, not coherent and even sometimes internally contradictory. Some simple, obvious topics are explained

with such a great amount of detail that makes them almost unreadable (chapters 3.3 and 3.4 about file location in a package and references between them, for instance) and difficult to understand. On the other hand, some topics that should be explained with more details are too brief. This is the case with the QTI description (chapter 4.9). The diagrams for QTI are helpful, but they are contradictory to the description sometimes, especially when it comes to differences between QTI 1.2.1 and QTI Common Cartridge profile. We couldn't find definite answers for many questions − for instance whether one can use pictures or video in exercises. The diagram in paragraph 4.9.8 of the specification [1] suggests that it is possible (see MaterialSelection box), the description in paragraph 4.9.1.4 and 4.9.2 says − that only text is allowed. Such inconsistencies are frequent in this document. Writing a new version of the "CC profile" specification, similar in quality to the QTI 1.2 specification document [2] (which is, in contrast, very clear, precise and easy to read) would be very much appreciated by developers.

The Validator tool proved to be really very useful. However − it didn't find a problem with XML files that have the filename extension different than .xml. We suggest updating this tool, including also filename extension validation.

There is an urgent need for the official reference Common Cartridge player, which would be available to any developer that is interested in CC package creation. We just can't convert anything to CC if we can't see the proof that everything is working correctly. The specification, that doesn't give clear answers for many questions, makes the problem even bigger. Many issues just have to be tested in a reliable player. The reference CC player is on the top of my wish list today.

4 Conclusions

Despite the obvious limitations of the Common Cartridge, we want to continue our efforts to familiarise ourselves with this standard and prepare some software tools, that would allow the conversion. However, Common Cartridge has too many limitations at the moment to be considered our main format for packaging the content. But we acknowledge, that it has not been designed for such a task. It was designed to make possible the creation of small exchangeable items, learning objects that can form building blocks for a teacher, who wants to create her own course. We would be happy to be able to create such small building blocks in the future, based on our existing content. But for this application to be successful, the CC standard should evolve and allow for more interesting, more complicated content.

We would appreciate some improvements and changes − the replacement of the "QTI 1.2.1 CC Profile" with something better would be our first postulate. One possible solution could be a move to QTI 2.0 that solves many problems of the old version (it has quite broad choice of exercise types, more logical and concise syntax, support for HTML-like layout control and CSS styles). Or, as a minimum, the new version of the CC profile still based on QTI 1.2.1, but with less restrictions (especially in <material> sections and in the <resprocessing>

area). We are open to discuss these issues and exchange ideas to make this standard better and really useful for the software industry.

References

[1] IMS Common Cartridge Profile, Version 1.0 Final Specification
    (`http://www.imsglobal.org/cc/ccv1p0/imscc_profilev1p0.html`)
[2] IMS Question & Test Interoperability, Final Specification Version 1.2
    (`http://www.imsglobal.org/question/index.html`)

Authoring for re-use in outcome-oriented learning scenarios

Roland Klemke[1], Birgit Schmitz[2]

[1] Centre for Learning Sciences and Technologies
Open University of The Netherlands
Valkenburgerweg 177, 6419 AT Heerlen, Netherlands
roland.klemke@ou.nl

[2] Humance AG
Goebenstraße 10-12, 50672 Köln, Germany
bsc@humance.de

Abstract. Content production processes currently experience a shift in focus. Due to the growing trend of highly individualized learning scenarios they have to face and to combine a multitude of different standards. The European eContent+ network of Excellence ICOPER researches possible strategies and implementations to deal with this new situation. In its course the project develops a comprehensive set of prototypes that use, evaluate and propose extensions to a large number of currently relevant standards such as IEEE RCD, SCORM or OAI-PMH. This paper describes a collaborative, re-use based authoring approach that was realized with one of these prototypes.

Keywords: authoring, re-use, interoperability, outcome-oriented learning

1 Introduction

Today's technology enhanced learning scenarios focus on learning outcome oriented delivery of learning processes, contents, and services. Rather than pre-defining static curricula individual learning processes are enabled: learner profiles indicate individual gaps, learning outcomes describe the required skills, recommended learning materials help to close these gaps, and assessments indicate successful mastery which is reflected back into learner's profiles. This way, the learner gains a large factor of *ownership of learning* [9].

In such complex learning situations interoperability issues become important: standards to describe learning outcomes, learner profiles, assessment items and learning materials are needed. Technical interoperability between different components of an overall learning delivery toolset is required in order to ensure seamless learning processes. Web-service based approaches help to simplify technical interoperability [12].

The learning outcome orientation also changes the way learning contents are produced and organized. The traditional "one-size-fits-all" approaches deliver the

same content to a large target audience. The production of learning content at high cost is therefore justified. Outcome-oriented content production processes however, deliver highly individualized content and have to cope with a large diversity of interoperability standards due to the multitude of sources they are based on (different learning objects, learning outcomes, learning designs and learner profiles). The need for re-use is therefore essential.

The European ICOPER project [8] analyses and discusses state-of-the art implementations of current standards as a base for the development of a comprehensive set of prototypes that support individual learning, teaching and authoring. In the course of ICOPER, the Open ICOPER Content Space (OICS) was developed, which combines learning object metadata repositories, learning outcome repositories, learning design repositories and learner profile repositories. The OICS offers a service interface, which allows to retrieve from and publish into the OICS [10]. Some of the standards the OICS works with, comprise:

- *IEEE Reusable Competency Definitions (RCD)* defines a data model for describing, referencing and sharing competency definitions. The *ICOPER Learning Outcome Definition (LOD)* is an application profile based on RCD that can be used to create *Personal Achieved Learning Outcome (PALO)* profiles [7].
- *Sharable Content Object Reference Model (SCORM)*, a standard to describe structure and behavior of content and components [1].
- *Learning Object Metadata (LOM)*, an IEEE standard to describe metadata for learning objects in a standardized way [5].
- *Open Archive Initiative's Protocol for Metadata Harvesting* (OAI-PMH), a protocol specifying the harvesting of metadata for learning objects residing in repositories [6].

In the course of the ICOPER project, a collaborative, re-use based authoring prototype was developed, that is based on the OICS.
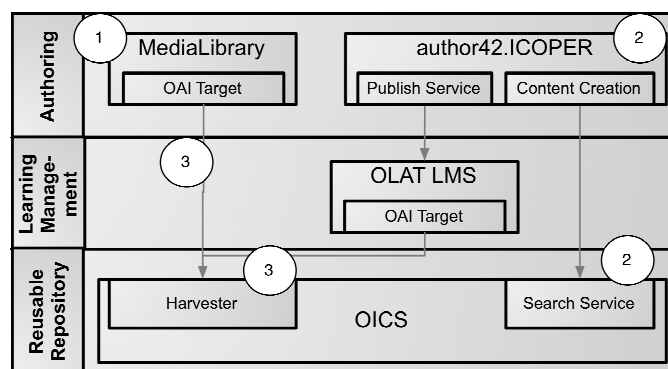

Figure 1. Architecture of the ICOPER authoring for re-use prototype

The prototype focuses on three main processes:

1. **The collaborative collection and organization of media assets**.
   Media assets comprise individual content elements such as texts, pictures, videos, and audios, which form the basis of all content productions.
2. **The collaborative creation of learning units based**.
   Learning units are navigable and interactive learning contents built out of individual media assets.
3. **Preparation for re-use**.
   The previous two processes are supported by a background harvesting process which updates the metadata repository of the OICS in order to make updated contents searchable and retrievable.

2 Collaborative collection of multi-media assets

A common problem to all collaborative, re-use based production processes is the retrieval, organization and management of media assets. Especially, when production processes have to cope with heterogeneous target groups (e.g. different languages, support for disabled people) or dynamic topic domains (with many subsequent content versions) media asset management is a complex matter.

Within the ICOPER project we therefore developed the MediaLibrary prototype. This online tool offers the opportunity to share media within a community. As opposed to other publicly available platforms (such as Flickr or YouTube), the MediaLibrary is designed to support media production processes, content-re-use, complex media models (with different versions, variants, languages, media formats) and corresponding metadata. The MediaLibrary is connected to the OICS via an OAI-PMH interface through which media assets within the MediaLibrary can be searched, browsed, and re-used. This way, media asset collections become part of larger learning content and metadata repositories.

3 Collaborative creation of learning units

The creation of learning units involves different tasks comprising the development of didactical concepts, storyboards and the actual content production which can be supported by content authoring environments. In our authoring prototype we use a customized version of author42 (named author42.ICOPER) that is a web-based authoring environment with extensible interfaces [2] which enables teams of authors to collaboratively create learning contents.

author42.ICOPER is integrated with the OICS: as a result the whole repository (including contents from the MediaLibrary and other sources) can be searched directly from the content production environment and retrieved results can be seamlessly integrated in the current production process.

To enable this integration, author42.ICOPER was extended with a customized search interface that can be launched from within any content production step. The search interface allows searching the OICS using combinations of different metadata and keyword fields. Search results can be selected and integrated into the current content page just as the users own content would be: page layout, element sizing and positioning functionalities can be used to fit the search results into the content page.

Through a web-service-based publishing interface, the content created can be published from author42.ICOPER into different repositories. In our prototype, we chose the open source learning management system OLAT [4] which we extended with an OAI-PMH target to connect it to the OICS again. This way, the content is available for further re-use.

4 Preparation for Re-use

Both prototypes, the MediaLibrary and the author42.ICOPER, offer OAI-PMH targets to access their metadata and contents from external repositories. The OICS contains an OAI-PMH compliant harvesting module [11] that is capable of accessing these targets and retrieving the metadata accordingly. That way, the OICS maintains a searchable repository of metadata that refers to the original contents.

The search function of the OICS can be used in two different ways: (a) directly, through the OICS' own search interface, or (b) integrated into other applications using the OICS' web-service interface. The latter being the case in the author42.ICOPER integration of the OICS, which enables search results to be automatically and seamlessly embedded into the content production process.

5 Conclusion and Outlook

We have presented a prototype to support collaborative, re-use based authoring for modern, outcome-oriented learning approaches. In a first evaluation of the abovementioned prototypes and their interoperability with the Open ICOPER Content Space (OICS), we performed an evaluation workshop at the JTEL SummerSchool 2010 in Ohrid, Macedonia. Ten participants with different backgrounds covering teachers, researchers, and students from computer science, technology enhanced learning and other fields took part in the evaluation.

In a first evaluation step, participants were asked to organize and metatag MediaAssets according to a selected LOD in the MediaLibrary prototype. These assets were re-used in the online authoring system author42.ICOPER to produce a unit of learning. Finally, participants published the results to enable further re-use via the OICS. Despite some technical problems that arose due to the prototypical status of the tools in use, participants could effectively solve their tasks and rated the toolset to be highly relevant.

In the meantime and with the prototype still under way, a new approach to connect repositories and authoring tools undergoes standardization efforts: the Simple Publishing Interface (SPI). This draft standard [3] especially focuses on the

integration of publishing tools (like the authoring tools used in our prototype) and repositories (like the OICS). We are planning to use SPI for the next version of our prototype to support direct collaboration and immediate re-use of results.

References

1. ADL SCORM (2004). Advanced Distributed Learning Sharable Content Object Reference Model, 2nd ed., Overview, 22 July, Alexandria, VA.
2. Cristea, A. (2006). Interfacing adaptive solutions with corporate training systems. PROLEARN Deliverable 1.6, retrieved online 19.06.2010 at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.99.3240&rep=rep1&type=pdf.
3. CWA 16097 (2010). The Simple Publishing Interface (SPI) Specification. February 2010. Available at: ftp://ftp.cen.eu/CEN/Sectors/TCandWorkshops/Workshops/CWA16097.pdf.
4. Fisler, J. & Schneider, F. (2008). Creating, Handling And Implementing E-Learning Courses and Content Using the Open Source Tools OLAT and eLML at the University of Zurich. ISPRS Conference 2008 -TS ThS-16: New Approaches and Tools for Education and Capacity Building International Society for Photogrammetry and Remote Sensing (ISPRS), Beijing, China, July 3-11 2008.
5. IEEE LTSC (2002). Final Draft Standard for Learning Object Metadata. Retrieved 01/15/10 at http://ltsc.ieee.org/wg12/files/LOM_1484_12_1_v1_Final_Draft.pdf.
6. Lagoze, C. and Van de Sompel, H. (2001) The Open Archives Initiative: Building a Low-Barrier Interoperability Framework. Proc. 1st ACM/IEEE-CS Joint Conf. Digi- tal Libraries, ACM Press, pp. 54–62.
7. Najjar, J., Derntl, M., Klobucar, T., Simon, B., Totschnig, M., Grant, S., & Pawlowski, J. (to appear 2010). A data model for describing and exchanging Personal Achieved Learning Outcomes (PALO). International Journal of IT Standards and Standardization Research.
8. Najjar, J., Simon, B. (2009). Learning Outcome Based Higher Education: iCoper Use Cases, IEEE International Conference on Advanced Learning Technologies (ICALT), workshop on European Workshop On E-Learning Standards Best Practice, Riga - Latvia, July 14 - 18, 2009.
9. Savery, J. R., & Duffy, T. M. (1995): Problem based learning: An instructional model and its constructivist framework. EDUCATIONAL TECHNOLOGY-SADDLE BROOK NJ-, 35, 31–31.
10. Totschnig, M., Derntl, M., Gutiérrez, I., Najjar, J., Klemke, R., Klerkx, J., Müller, F. (2010). Repository services for outcome oriented learning. 4th international workshop on Search and Exchange of e-le@rning Materials (SE@M), Barcelona, 27-28 September 2010.
11. Totschnig, M., Klerkx, J. Klobučar, T., Law, E., Simon, B., Ternier, S. (2009). D1.1 Open ICOPER Content Space Implementation of 1st Generation of Open ICOPER Content Space including Integration Mini Case Studies. ICOPER project deliverable. Retrieved Online at 28.06.2010 from http://www.educanext.org/dotlrn/clubs/icoper/new-lors/Deliverables/Deliverables_-_Submitted/D1.1/D1.1-final.pdf
12. Vossen, G. and Westerkamp, P. (2008). 'Why service-orientation could make e-learning standards obsolete', Int. J. Technology Enhanced Learning, Vol. 1, Nos. 1/2, pp. 85–97.

Exploring Quality Issues in the Use of LOs:
To Tag or Not to Tag?

Rune Hjelsvold[1], Jingjing Fan[1], Yngve Nordkvelle[2], and Kjell Are Refsvik[1]

[1] Gjøvik University College, Gjøvik, Norway
[2] Lillehammer University College, Lillehammer, Norway

Abstract. In this paper we describe a study where students tagged learning objects created by their professors. The study shows that the student tags extend the professors' view of the contents of the learning objects and add a wider context for interpreting the content of the learning objects. During interviews conducted at the end of the study, the professors reported that the students' tags represented a form of feedback that would help them recognize discrepancies between the learning objects' intended purpose and the perceived purpose.

1   Introduction

According to Kay and Knaack [4], the majority of researchers have emphasized technological issues such as "accessibility, adaptability, the effective use of metadata, reusability and standardization" when defining the term *Learning Object* (abbreviated LO in this paper). Kay and Knaack, however, define learning objects as tools that should enhance, amplify, and guide the cognitive processes of the learners [4]. The aim of our study is partly to investigate whether student-generated tags may be considered useful as content-descriptive metadata and partly whether the student tagging process itself will have a positive effect on the pedagogical value of the use of learning objects.

1.1   Tags as Metadata

Researchers currently have different and disagreeing views on the quality of user-provided tags and folksonomies (i.e., a vocabulary that has emerged organically as a result of the end-user tagging activities). Guy and Tonkin [2], for instance, describe tag sets as uncontrolled and chaotic and not very well suited for supporting searching. In their study, Sen et al [6] found that only 21% of user-provided tags were considered worthy of general display by other users.

Al-Khalifa and Davis [1], on the other hand, found that folksonomy tags were better in terms of search and contextual coverage than the metadata created by the human expert. Vuorikari et al. [7] found that tags can enrich and add value to controlled vocabularies. In our study, we will investigate this further.

1.2 Learning Objects in a Pedagogical Context

The idea that what the teachers intend to communicate is received by students is much criticized. In radical constructivism, it is claimed that it is less than likely that the received message is congruent with the sent (Qvortrup [5]). The didactical operations performed by the teachers are produced in order to increase the chances for the student to achieve a higher degree of understanding. On the other hand, students' interpretations and misinterpretations are both regarded as fruitful processes. Hansen and Brostrøm [3] argue that professors need to develop their skills as teachers in tagging their LOs. The discrepancies between the outcomes of the tagging performed by the professor and their students are indicative of the degree of overlap between teachers' intentions and students' reception of the same message.

2 The Student Tagging Study

In this project we studied two groups of master students at Gjøvik University College. Each group consisted of approximately ten master students. Group 1 was a group of first year students on the Master of Media Technology program participating in a course on media data coding and compression. Group 2 was a group of second year students on the same program who had previously participated in a course on semantic web. The courses were taught by two different professors – one for each course.

Both courses were offered as blended learning in which regular lectures were recorded. The recorded lectures along with lecture notes in PDF were used to produce LOs stored in an LMS. The professors assigned keywords to each learning object as content-descriptive metadata.

The project was divided in two main parts. Firstly, the students used a web-based application to tag LOs blindly (i.e., students did not get to see other students' tags). Secondly, semi-structured interviews were conducted in which the two professors and some of the students were interviewed individually.

The key characteristics of the two groups and the generated metadata can be summarized as follows: The individual student in group 1 generated 3.5 tags per LO on average, while the average in group two was 3.3. At the same time, the professor teaching group 1 assigned 14.6 per LO on average while the professor for group 2 assigned only 6.9 keywords per LO on average. On average, 1.3 tags generated by the individual student in group 1 also appeared as keywords assigned by the professor. The corresponding value for group 2 was 1.35. The overlap between the sets of tags generated by the students and the sets of keywords assigned by the professor per LO is further illustrated in Fig. 1. The upper part of the bars shown in brighter colors exposes the number of keywords assigned by the professors that did not appear in the sets of tags generated by the groups of students. The lower part of the bars shown in darker colors exposes the number of the number of tags generated by the groups of students that did not appear in the sets of keywords assigned by the professors. The mid-

dle part shown in the darkest color illustrates the amount of overlap between student-generated tags and professor-assigned keywords.
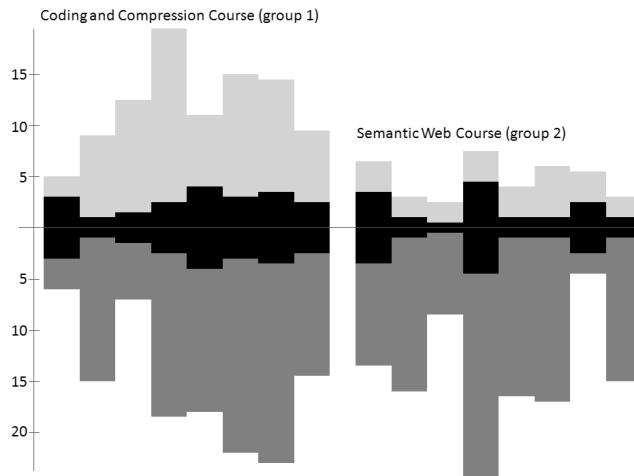

Fig. 1. Number of common terms (darkest color), compared to number of professor keywords only (lightest color) and to total number of student tags (bottom part) for the two courses

Finally, Figure 2 illustrates the level of agreement among students with regards to the tags. The ten most popular tags are enumerated along the horizontal axis (tag 1 being the most popular for the given LO). As can be seen on the figure, student tags mostly differ from the keywords chosen by the professors. Our study therefore indicates that the students' interpretation of content of the LO is different from the professors'. It thus seems like student-generated tags would be useful as a complementary type of metadata to professor-assigned keywords.

3 Value of Student Tagging: Beyond Content Descriptive Metadata

The two professors involved in the study and six of the students from group 1 were interviewed at the end of the study. The purpose of the interviews was to investigate whether student tagging added value beyond producing content descriptive metadata.

The students all agreed that the keywords provided by the professors were useful for them in interpreting the content of the LO. They even requested that student tagging of learning object should be introduced in all the courses they were signed up for. They found the keywords especially helpful in finding what the key aspects of the LO were.
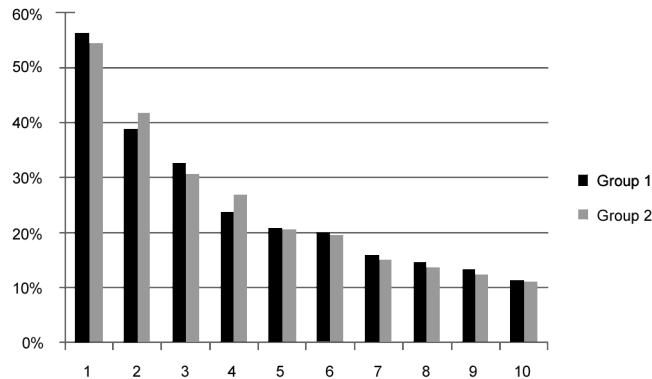
Fig. 2. Level of agreement among students for the most popular tags

A majority of the students also considered it useful to be able to view fellow students' tags. They also thought it would be useful if they had the opportunity to view students' tags from earlier years of the course – even though no such tags existed in the demo system.

Both professors agreed that student-defined tags were quite similar to their own keywords although they were not the same. Most importantly, however, the participating professors emphasized that the tags represented a sort of feedback that would help them understand how well the students were able to grasp the contents of the LO.

The two participating professors both stated that the feedback from the students would have a real impact. In some cases, they found that the students had chosen tags that they would like to include as one of the professor-assigned keywords in the future. In other cases, the feedback indicated that the students had missed some important messages or misinterpreted the LO. In these cases, the professors said they would either modify the original lecture and accompanying LO, or they would repeat these issues again in future lectures to help the students achieve a correct and deeper understanding.

One striking observation is that none of the professors considered removing even one single keyword from any of the LOs even if the students did not use it as one of their tags. The professors saw the student-generated tags as a possible extension to the set of keywords they assigned – not as a potential replacement.

Although the study of student tagging of LOs and the follow-up interviews were rather small, they indicate that content-descriptive metadata in the form of professor-assigned keywords *and* student-generated tags add value beyond supporting searchability and reusability. The students consider professor-assigned keywords as a help in interpreting the contents of the LOs and the professors see student-generated tags as valuable feedback from the students both when it comes to the quality of the LO and the quality of the accompanying set of metadata.

4 Conclusion and Further Work

Our study was rather small, involving some 20 students, two professors, and 16 LOs. Still, the study gives some interesting indications on the usefulness of student tagging and interesting paths for further work.

Our study shows that there is some overlap between the tags that students create for LOs and the keywords the professors assign to the same LOs. There is, however, also a significant difference between the two. This difference may provide a wider context for interpreting the content and context of the LO. Our study also shows that student tags may be utilized in the quality management of the LOs. The professors could recognize discrepancies and patterns of differences between the student-generated tags and the original set of professor-generated keywords that might call for quick or dramatic alterations of LOs. As a means for quality improvement this dimension goes far beyond the quest for retrieval and reusability question to include issues of feedback, evaluation and might improve the level of collective engagement and learning retrieval.

This project is part of an ongoing process in developing a Learning Object Repository at Gjøvik University College that will provide student tagging capabilities.

References

1. Al-Khalifa, H.S., Davis, H.C.: Replacing the Monolithic LOM: A Folksonomic Approach. In Proceedings of the Seventh IEEE International Conference on Advanced Learning Technologies (ICALT 2007), 665-669 (2007)
2. Guy, M., Tonkin, E.: Folksonomies: Tidying up tags? D-Lib Magazine, doi:10.1045/dlib.magazine (2006)
3. Hansen, P.S., Brostrøm, S.: Tagging the didactic functionality of learning objects. World Conference on Educational- Multimedia, Hypermedia & Telecommunications Chesapeake, Va. : Association for the Advancement of Computing in Education (2002)
4. Kay, R., Knaack, L.: Evaluating the learning in learning objects. Open Learning, doi:10.1080/02680510601100135 (2007)
5. Qvortrup, L.: Society's Educational System - An introduction to Niklas Luhmann's pedagogical theory. Seminar.net, 1:1 (2005)
6. Sen, S., Harper, F.M., LaPitz, A., Riedl, J.: The Quest for Quality Tags. In Proceedings of GROUP'07, 361-370 (2007)
7. Vuorikari, R., Sillaots, M., Panzavolta, S., Koper, R.: Are Tags from Mars and Descriptors from Venus? A Study on the Ecology of Educational Resource Metadata. In Proceedings of the International Conference on Web-based Learning (ICWL 2009), 400-409 (2009)

Educational Resources Packaging Standards SCORM and IMS Common Cartridge – The Users Point of View

[1]Kati Clements, [2]Àgueda Gras-Velázquez, [1]Jan M. Pawlowski1

[1]Global Information Systems, University of Jyväskylä, 40014 Jyväskylä Finland,
[2]European Schoolnet, Rue de Lalaing 24, 1040 Brussels, Belgium
kati.clements@jyu.fi, agueda.gras@eun.org, jan.pawlowski@jyu.fi

Abstract. This paper reports on an exploratory study analyzing the Educational Resources' packaging standards SCORM and IMS Common Cartridge's (CC) regarding interoperability from the point of view of key users: teachers. The two specifications SCORM and CC to package Educational Resources have been developed to help the users to re-use Learning Objects from Learning Object Repositories (LORs) in Learning Management Systems (LMS) of schools. In our study, we found that teachers find packaging solutions highly useful, especially the interoperability between LORs and LMSs. Teachers also appreciated that they can modify packaged content after it has been uploaded to the LMS. The teachers also strongly appreciated the additional functionalities of CC packages while teaching courses online or giving home work/extra assignments to their students.

Keywords: Educational Resources, content packaging standards, Learning Object Repositories, SCORM, IMS Common Cartridge, teachers

1 Introduction

This paper shows the views of teachers on interoperability between Learning Object Repositories (LOR) and Learning Management Systems (LMS) supported by two content packaging standards: SCORM and IMS Common Cartridge (CC). We set up a testing experiment and survey to find out 1) whether teachers would get any additional value from using these standards and 2) whether tools used to support these standards are reasonable for teachers to use in their everyday teaching.

According to The IEEE Learning Technology Standards Committee, a Learning Object is defined as "any entity, digital or non-digital, that may be used for learning, education or training" [1]. Learning Objects in this paper are defined as digital objects which can be used, re-used or referenced during a technology-supported learning process [2]. Educational Resources can be defined as Learning Objects that can be employed in technology-supported learning [3]. Learning Object Repositories are collections of Learning Objects that are accessible to users via a network without prior knowledge of the structure of the collections [4]. Distribution of Educational Resources is most commonly done via LORs. This study investigates the views of teachers re-using and sharing Learning Objects from LORs with the help of interoperability standards SCORM and IMS Common Cartridge. In their everyday

lesson preparation, teachers use Educational Resources in LMS such as Moodle [5]. LMS help teachers to manage digital contents, to aggregate blended learning and give homework to their students.

Recent research involving packaging standards SCORM and IMS Common Cartridge has been focusing on the technical specifications [6],[7],[8], without investigating the teachers' point of view. The assumed benefits of SCORM and CC have not been verified by the end user group, leading to the questions of this study: Are the teachers willing to use SCORM and CC packages in their busy schedule of designing Educational Resources for their classes. Is the use of content packaging yet another technical problem which the teachers will not be able to handle with a reasonable amount of training?

This study aims to examine what the real users, the teachers, think of using these Learning Object packaging standards and ask if they can see the interoperability between LOR and LMS facilitating the work they are doing in their everyday lesson plan creation process.

When trying to evaluate the impact of a learning technology standard to the users, it is important to realize that specifications cannot be evaluated by users as users do not work with them directly. Rather, interoperability specifications are implemented in software tools offering a set of functionalities to the end user. End users can then make use of the tool in practice and that use can be evaluated. The evaluation results need to be analyzed in detail to assess whether problems are caused by the interoperability specifications or by the functionality provided by the tool, or the user interface through which the functionality is made available to the end user. [9] However, it can be evaluated whether the users have understood and used the key concepts of a specification (such as packaging or metadata categories). To avoid this problem, we look at the key concepts and functions enabled by the standards. As SCORM and CC are both standards which are widely implemented in tools, we can map the standards' concepts and the resulting functionalities in the tools, which are then assessed in practical experiments. Therefore we tested the standards by using the tools keeping in mind that the interfaces of these players could affect the minds of the users.

2 Open Educational Resources Packaging Standards

In the following, we discuss the content packaging specifications SCORM (Sharable Content Object Reference Model) and IMS Common Cartridge (CC).

2.1. SCORM

The Sharable Content Object Reference Model (SCORM) was created to help the re-usability, interoperability, portability, access, maintenance and adaptation of Learning Objects. [10] SCORM is a collection of standards and specifications that enable learning platforms to find, import and deliver learning content in a standardized way. SCORM specifies how Learning Objects must be created in order to ensure interoperability across different platforms and tools. [7]

SCORM was originally intended for use in self-study, computer-based training scenarios rather than in interactive scenarios between teachers and students. SCORM's crucial functionality 'sequencing' supports the learner by allowing him/her only to navigate into parts of the package based on the previous learning assignments fulfilled. However, it can be said that SCORM was created to help teachers when transferring collections of Learning Objects, typically from an LOR to an LMS.

To date, most popular Learning Management Systems support SCORM objects [11] among these the system of Moodle. SCORM packages can be uploaded as single, unmodifiable entries to Moodle.

2.2. IMS Common Cartridge

IMS Common Cartridge (CC) was supposed to enhance SCORM, offering more flexibility and support for assessments, web 2.0 standards, content authorization, collaborative forums and outcomes reporting [7]. The aim of IMS Common Cartridge was not to compete with SCORM but to support to blended learning rather than self learning. According to IMS [12] the IMS Common Cartridge specification's objectives for facilitating teaching include increasing flexibility, sharing and re-use. CC enables teachers to assemble lesson plans taking only parts of an CC package or integrating the whole package into their LMSs. This study aimed to find out how teachers saw the difference of interoperability when using IMS Common Cartridge packages integrated into a Moodle LMS.

IMS Common Cartridge is not as widely supported by different LMSs as SCORM perhaps because there are not yet as many tools to support it. However, LMS Platforms have gained growing interest towards IMS Common Cartridge support. It is still gathering momentum, which should increase by the announcement [13] from the popular open source platform Moodle to start supporting IMS Common Cartridge packages in the spring of 2010 on. Users can upload CC packages as a whole into the Moodle system much in the same way as uploading SCORM packages.

3 Test Setting

Adopting Standards and Specifications for Educational Content (ASPECT) is a Best Practice Network for educational content that aims at improving the adoption of learning technology standards and specifications [14]. Standards and interoperability experts produce recommendations that are implemented by tools and content providers before being tested by teachers during school pilots. Tests were carried out in order to demonstrate in which way the implementation of Standards and Specifications leads to greater interoperability and cross-border re-use of content [15]. The Learning Resource Exchange (LRE) is a pan-European federation of Learning Object Repositories [16]. The service is offered to stakeholders providing digital content, such as ministries of education, commercial publishers, broadcasters, cultural institutions, and other non-profit organizations offering online content to schools [17]. The LRE was used as the testing LOR in the ASPECT project and it provided the possibility to get the same package available in all the different formats (web page,

SCORM package, CC package, SCORM in Icodeon player and CC in Icodeon platform) in its metadata (see section 3.2 for further information).

Our main research aim is the validation of artifacts: the standards SCORM and CC as well as corresponding tools. As these standards are widely implemented in tools and rely upon them in practical experiments, the use of these tools can be considered a valid evaluation of the standards [9]. This paper mainly looks at the design evaluation [18] in an experimental setting [19] and essentially the usability aspects [20]. Both SCORM and IMS Common Cartridge have different versions which support different features. This study focuses on the key feature of compatibility and portability of resources as it can be seen as a critical feature for teachers when re-using learning objects from repositories.

3.1. Validation Group

A teacher workshop was organized for a group of 44 mathematics and science teachers equally split, from Portugal, Lithuania, Romania and Belgium. The four countries were chosen randomly among the participating countries to the ASPECT project. Teachers were selected to have the following characteristics: teachers of mathematics, science and/or technology, some experience with ICT and Learning Resources and currently teaching students aged 6 to 21. Before selecting teachers on the basis of these profiles, the strategy for finding volunteers varied from specifically contacting teachers who, in one way or another, had collaborated with the selection team before (e.g., Portugal) to publishing an open call for volunteers (e.g., Lithuania).

Initial observations showed that the ICT skills of this group of teachers ranged from little knowledge (a few even unsure of what Google was) to advanced (users of LaTeX). Up to 80% of the teachers had advanced ICT skills. The group of advanced teachers was strongly represented. The validation results must be seen in this light: Teachers participating in European projects tend to be more motivated towards ICT and improving their teaching methods. These teachers have clearly higher levels of motivation and are very eager to share their knowledge and expertise. As a consequence, it must be taken into account that if these teachers find the project tasks and concepts too difficult or uninteresting, one can be reasonably certain that this will apply even more so to teachers with average levels of ICT competence.

3.2. Tests Description

The workshop combined straight-forward assignments and direct feedback gathering from the teachers in the form of interviews and two surveys. The test session was organized in May 2010, concentrated on the integration of Resources into Learning Management Systems and content packaging. In preparation for the tests, teachers had already learned how to create basic courses on Moodle platform. They also had learned how to browse the Learning Object Repository, LRE.

In the tests, teachers were initially asked to create a normal lesson plan using the Moodle learning platform in a "traditional" way by combining different Resources. Then they were asked to repeat the same task using a Resource on the same topic that

had been 'packaged' by ASPECT content developers using both the SCORM and IMS Common Cartridge packages. The test task was to create a simple lesson plan made up of some text, an image, a quiz and a forum, on the topic of thermodynamics for Moodle, using four different approaches:

1) Using non-packaged content
2) Using the entire SCORM package (created from the non-packaged content)
3) Using an entire IMS Common Cartridge package (created from the non-packaged content, with a forum added)
4) Picking up parts from the IMS Common Cartridge package

The tests were designed to serve both as a basic training on the use of different types of packaged content and their features (necessary as the teachers had no previous experience with this kind of content) and at the same time obtain their reactions as rewards to usefulness in their everyday teaching, interest and facility to use. Each teacher had an empty Moodle course and editing rights. Each teacher designed the same lesson plan four times as described before. Teachers searched for the resource in question in the LRE, where it was provided in the 3 formats, both to view and download. For the non-packaged lesson plan, teachers had to use Moodle options to create the quiz and the forum themselves. Both the SCORM and Common Cartridge packages had the quiz included in the package, and the latter also contained a forum.

Additionally, teachers were presented with a dozen additional Resources packaged as SCORM and Common Cartridge to browse through and see their benefits, independently of the topic of the Resources. In all cases, to view the packaged content, Icodeon's Common Cartridge Platform and SCORM player were used as the tools to show the packaged content.

4 Results Analysis

As the sample size of the survey (n=44) was small, the results of the statistical analysis can only give us some indication on the teachers' attitudes. The survey results were backed up by a qualitative analysis of interviews of the teachers.

Generally, the teachers reacted to SCORM Resources in much the same way that they treated unpackaged content; for example, they did not see much difference between having a SCORM Resource and a PowerPoint presentation. While they saw that a SCORM package could include more than one resource, they did not use it any differently than PowerPoint; both types of content were integrated into an LMS as a single, unmodifiable entity. In comparison, the teachers were very enthusiastic about CC content packaging. After importing a CC package into Moodle, the teachers could remove parts that they did not need, edit the content and change the order of different resources. Many teachers requested instructions on how to adapt Moodle to use CC packages and even some teachers expressed an interest in using CC to package their own content in order to share it with other teachers.

Most teachers had little interest in simply viewing and playing SCORM or CC packages. On the other hand, apart from the option to upload the complete packages into an LMS and have the different parts of the package converted into Moodle format, the teachers liked the possibility of being able to embed only parts of a

Common Cartridge package in the LMS, or even blogs or websites, which is not possible to do with items from within SCORM packages. While this second option did not take advantage of the LMS's features, teachers liked to have the possibility to only integrate the parts of the cartridge into their courses that they liked or thought were relevant to their lesson. In the survey,

- 25% said that taking an entire course in SCORM format and using it in Moodle (or their school's own System) would be extremely helpful
- 39% said that taking an entire course in CC format and using it in Moodle (or their school's own System) would be extremely helpful
- 43% said that taking a piece of the Learning Resource from one of the CC packages and using it with their other teaching materials would be extremely helpful

Most of the teachers who did not see the approaches as extremely helpful, saw that the packages could be useful for them in limited cases like when giving homework or teaching an online course. The survey results indicate that teachers see the interoperability between LORs and LMSs created by the specifications SCORM and CC as useful for their everyday teaching – especially when the packaging allows them to alter the content after it was uploaded or selecting only bits of the package before uploading to LMS. IMS Common Cartridge specifies this interoperability.

Part of the objective of this research was not only to find out whether or not the teachers could see content packaging useful, but to find out if they could actually manage working with these standards with the ICT skills that they possess. Our initial hypothesis was that the tools supporting these standards have not yet developed enough to be easy enough for the teachers to use. However, according to the survey, normal web pages were unsurprisingly the easiest to use. 75% of the teachers found using entire SCORM packages really easy or reasonable, whereas they admitted having some problems when using CC packages, whether it was the package as a whole, or taking parts of it. However, interestingly none of the teachers evaluated that any of these methods as impossible to use in their every day teaching (see Fig. 1).
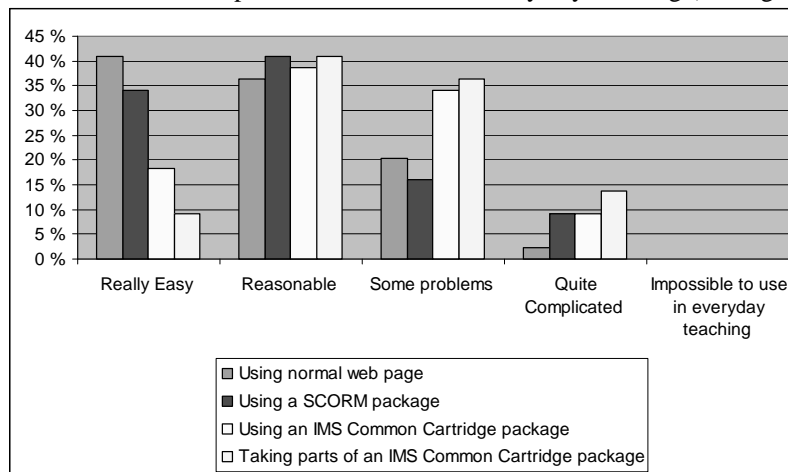

Fig.1. "How easy/difficult was it to create a lesson plan...".

This study indicates that even though SCORM packages might not be the preferred solution for teachers, the longer period of development seems to have helped it to be easier to use than IMS Common Cartridge packages, which was not supported by Moodle before the spring of 2010. There seems to be a need to develop the tools for these standards to be more usable for teachers in the future. However this result would also indicate that the teachers did not like the easiest option best, which would suggest that they were able to look beyond the interfaces of the players into the ideas of interoperability and the standards.

Part of testing SCORM and IMS Common Cartridge with users, we looked at the opinions regarding the technical interfaces which show the contents of the packages. Teachers were asked to think of three typical use cases from their everyday teaching life:

A)   Showing Educational Resources to students in their class rooms
B)   Teaching an online course
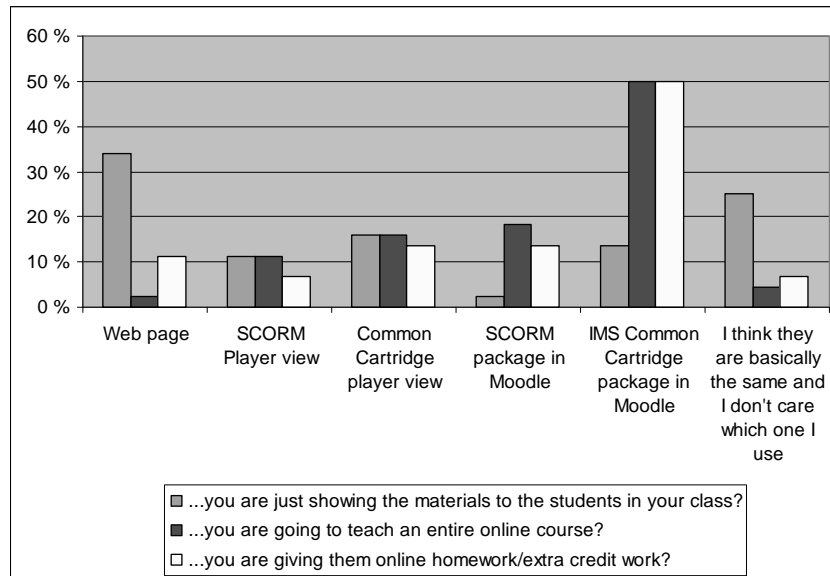C)   Giving online homework/extra credit work to the students

Fig. 2. Interface preferences of teachers in three use scenarios.

Overall, Fig. 2 shows that when it comes to just showing Resources in the classroom, that is easiest to do from a normal webpage, or even that it does not matter so much what the interface is. However, when you are creating an online course or giving homework to students, normal web page is no longer enough. Half of the teachers thought that using IMS Common Cartridge packages in Moodle would be the ideal way to teach online or give assignments in the form of homework or extra credit to the students. In the interviews with teachers it became obvious that Learning Management Systems like Moodle are widely used in schools and therefore standards

that create interoperability between content and these LMSs, have additional value for teachers.

If we want students to study independently, you can give them a SCORM package that they unzip and then use. But if you want, for example, to integrate a package in Moodle, because in Portugal we use Moodle a lot, probably Common Cartridge is good, because we can prepare everything and import it into Moodle." (ICT Teacher from Portugal) "

After the workshop, teachers were asked, what would they prefer to use after learning about the functionalities of SCORM and IMS Common Cartridge packages and interoperability between LORs and LMSs. Most of the teachers ended up preferring IMS Common Cartridge packages (87%), leaving only 7% to prefer normal web pages and 9% to prefer SCORM packages. This result supports the finding that teachers were genuinely excited about the prospects of IMS Common Cartridge after one day of training and lesson plan making. The teachers also seemed to support the solution that was the most adaptable to their own needs rather than the solution that was easiest to use.

5 Conclusions

Open educational materials and other web-based resources lead to new opportunities for sharing and re-using content. [21] European teachers are seldom aware of the content packaging standards SCORM and IMS Common Cartridge but understand the added value of re-using Educational Resources in their everyday teaching.

In this paper, we have presented the teachers' view on the interoperability between LORs and LMSs while using SCORM and CC standards. Teachers showed special interest towards CC packages and in particular their use in Moodle system. In detail, the teachers enjoyed the possibility of editing a package, taking some elements and mixing them with their own teaching Resources very much in the same way as they do in with the non-digital Resources in their classrooms. CC supports this kind of interoperability, which is the key finding of this research as it aims to solve one of the biggest problems in the field [22]. In this study, the teachers said that they were more willing to use a solution that would be suitable for them than the solution which seemed the easiest to use. Also no teacher believes that packaging standards are impossible to use in everyday teaching, even though they can see some problems and complications in the process. After a standard is finalized, it takes a long time before tools are developed that actually deliver the functionality to end users in a way that is useful and usable [23]. Hence, for the success of the standard among users, it is crucial to develop CC tools and improve their usability.

We also need to recognize that many teachers still struggle to obtain the basic IT skills which are more essential for their day-to-day work even if these teachers were optimistic about content packaging – teachers with lower ICT skills might have a different point of view. In our opinion, training in content packaging standards may be something that remains of interest to a fairly small number of European teachers. However, it is not imperative for the teachers to know that these standards exist. What is vital for them is that the process of lesson preparation using Educational Resources

will be smooth and quick. Further development of these standards and the tools around them is the way of assuring re-use of Educational Resources.

Acknowledgements

References

1. IEEE Learning Technology Standards Commity.: IEEE Standard for Learning Object Metadata. Available at http://ltsc.ieee.org/wg12/ site visited 4.6.2010.
2. WILEY, D.: Connecting Learning Objects to Instructional Design Theory: a Definition, a Metaphor, and a Taxonomy. Utah State University: Digital Learning Environments Research Group, The Edumetrics Institute (2000)
3. McGreal, R.: Learning Objects: A Practical Definition. Int'l J. Instructional Technology and Distance Learning, vol. 1, no. 9, pp. 9 (2004)
4. Mohan, P.: Learning Object Repositories. Proceedings of the Informing Science and IT Education Joint Conference, Flagstaff, Arizona, USA (2005)
5. Dougiamas, M., Taylor, P.: Moodle: Using Learning Communities to Create an Open Source Course Management System. In D. Lassner & C. McNaught (Eds.), Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications, pp. 171-178. Chesapeake, VA: AACE. (2003)
6. Qu, W., Nejdl, W.: Towards Interoperability and Reusability of Learning Resources: a SCORM-conformant Courseware for Computer Science Education. IEEE International Conference on Advanced learning (2002)
7. Gonzalez-Barbone, V., Anido-Rifon, L.: From SCORM to Common Cartridge: A step forward. Computers & Education Volume 54, Issue 1, January 2010, Pages 88-102 (2009)
8. Rey-López, M., Díaz-Redondoa, R., Fernández-Vilasa, A., Pazos-Ariasa, J., García-Duquea, J., Gil-Sollaa, A., Ramos-Cabrera, M: An extension to the ADL SCORM standard to support adaptivity: The t-learning case-study. Computer Standards & Interfaces, Volume 31, Issue 2, February 2009, pp 309-318 (2008)
9. Duval, E.: Learning Technology Standardization: Making Sense of it All. ComSIS Vol 1, No. 1, (2004)
10. ADL Technical Team.: ADL Guidelines for Creating Reusable Content with SCORM 2004 (2008) Available at: http://www.adlnet.org. site visited 4.6.2010.

11. Gonzalez-Barbone, V., Anido-Rifon, L.: Creating the first SCORM object. Computers & Education, pp. 1634–1647 (2008)

12. IMS Global Learning Consortium: IMS Common Cartridge Specification. Available at: http://www.imsglobal.org/commoncartridge.html Site visited 7.6.2010.

13. Moodle.: Development: IMS common cartridge. Available at: http://docs.moodle.org/en/Development:IMS_common_cartridge. Site visited 8.6.2010.

14. Massart, D., Chaudron, S., Ayre, J., Klerkx, J., Heckmann, P., Ravet, S., Ingesman, L., Gras-Velázquez, À., Pawlowski J.: D-1.2.1 ASPECT First progress report. ASPECT deliverable. Confidential (2009)

15. Gras-Velázquez, À. Clements, K., Yalaho, A., Ayre J., Ingesman, L., Van Assche, F., Blamire R., Vuorikari, R., De Four, H.: D-6.2 Protocol of Experimentation with National Adaptations. ASPECT deliverable. Confidential (2009)

16. Massart, D.: The EUN Learning Resource Exchange (LRE). The 15th International Conference on Computers in Education (ICCE2007) Supplementary Proceedings, vol. 1, pp. 170-174 (2007)

17. Ternier, S., Massart. D., Campi, A., Guinea, S., Ceri, S., Duval, E.: Interoperability for Searching Learning Object Repositories - The ProLearn Query Language, D-LIB Magazine, vol. 14 (2008) Available at http://dlib.ejournal.ascc.net/dlib/january08/ceri/01ceri.html Site visited 4th June 2010.

18. Hevner, A.R., March, S.T., Park, J., Ram, S.: Design Science in Information Systems Research. MIS Quarterly, vol. 28, 75-105 (2004)

19. Zelkowitz M. V., Wallace, D.: Experimental models for validating computer technology, IEEE Computer, vol. 31, pp. 23-31. (1998)

20. Nielsen, J.: Usability Engineering. Academic Press, Boston (1993)

21. Vuorikari, R.: Insight Special Report: Why Europe Needs Free and Open Source Software and Content in Schools, 2004.

22. Johnstone S.: Open Educational Resources Serve the World: Sharing Educational Resources over the Internet. Educause Quarterly, Vol. 28, No. 3, 2005.

23. Duval, E., & Verbert, K. On the Role of Technical Standards for Learning Technologies. IEEE Transactions on Learning Technologies, vol. 1, no. 4 (2008)