Towards Requirements for the Dynamic Sourcing of Open Corpus Learning Content

Seamus Lawless, Declan Dagger, Vincent Wade
Knowledge and Data Engineering Group, Trinity College, Dublin, Ireland
{slawless@cs.tcd.ie, Declan.Dagger@cs.tcd.ie, Vincent.Wade@cs.tcd.ie}

Abstract. The dynamic sourcing of Open Corpus Learning Content is a complex and resource intensive process. However, modern technologies and methods can facilitate an efficient content identification and acquisition process. Several bodies of research have begun investigating the application of modern technologies to problems in this area, such as automatic semantic tagging of content, however none have managed to successfully implement a scenario whereby the requirements of elearning environments can be leveraged to accurately source and retrieve learning content. This paper describes the initial requirements for successfully sourcing and harvesting open corpus learning content, the architecture for an Open Corpus Content Sourcing service (OCCS) and a scenario for its use in Personalised eLearning.

1. Introduction

The term "Open Corpus Content" is applied to content which is freely available and accessible to the general public and educational institutions. This content may exist in numerous locations and in countless different formats. There are many existing open corpus content repositories which provide learning resources free of charge to the online community. The Worldwide Web has a vast array of content, ideal for incorporation into educational experiences. The content may be contained in websites, whitepapers, blogs, images, video, forums and many other manifestations, but share a common trait; it could be used in the creation of eLearning offerings.

However there are a number of issues which restrict and in some cases prevent such content from being discovered, harvested and reused. One of the main issues involved is the "discoverability" of such content. The vast quantities of content available via the Worldwide Web are seldom organised and vary in purpose, topic, format and exposure methods. How the content is stored can dictate how accurately machines can classify and interpret the content, this in turn relates to how accurately content requirements can be matched to suitable candidate content. Several methods of addressing these issues have been identified, which are discussed in this paper, such as content aggregation, automatic metadata generation, semantic and vocabulary mapping, web crawling, etc. However a scenario has yet to be created whereby all these methods are united and evolved to produce a content discovery and harvesting mechanism that can deliver open corpus content to eLearning environments that accurately matches learning content requirements.

This paper sets out to address the research question of dynamically sourcing open corpus learning content. The following sections describe the initial requirements for content sourcing, the process of dynamically sourcing appropriate open corpus content and the architecture of a proposed open corpus content service (OCCS). The paper then presents a usage scenario for this OCCS in the area of enhancing course developer support in Personalised eLearning Development Environments. Finally the paper concludes with a discussion on future work in this and related fields of research.

2. Requirements for Content Sourcing

This section describes the semantic and technical requirements that must be satisfied if learning content harvesting services are to be sufficiently flexible and effective in the sourcing and provision of suitable learning content to elearning environments. Semantic requirements refer to those that are knowledge related, for example the pedagogy, subject matter area, activity, personalisation and context. Another key semantic requirement is the inherent strategic flow that exists within the learning content. Technical requirements reflect the syntactic layers of the learning content, namely the representation layer (meta-model) and descriptive layer (metadata) of the learning content.

Pedagogical and activity information plays a critical role in the selection of appropriate learning content. It forms the strategical and educational foundations of a learning experience. However, learning content is seldom pedagogically neutral. In essence, the majority of learning content is created with instructional design principal(s) in mind and usually contains some inherent pedagogy. In this case, it is often difficult to identify the learning content and sometimes impossible to reuse that learning content outside of the pedagogical context(s)

in which it was created. Moreover, the learning content descriptions seldom identify the inherent pedagogical knowledge represented within (through its metadata information) and in some open corpus learning content, the meta-model information may not exist. The absence of any pedagogically-related semantic information makes it very difficult to effectively and efficiently identify appropriate learning content.

The semantics of the subject area also provide vital information when sourcing learning content. They provide contextual information regarding the finite knowledge space of the learning content. One of the core issues that arises here is the mismatch in vocabularies and taxonomies that are used to describe the subject area of the learning content, if such vocabularies and taxonomies even exist. Cataloguing taxonomies such as the Library of Congress Subject Headings (LCSH) provide an up to date and comprehensive list of subject headings which could provide a solid distinct taxonomy for describing the subject area information.

A key issue which influences requirements for content sourcing across both the semantic and syntactic layers is granularity; knowledge granularity and technical granularity. Granularity of learning content typically refers to its size (conceptual or technical), its aggregation and its ability to be reused or repurposed. Conceptually coarse learning content is usually difficult to reuse outside of the context in which it was originally created. An example of a piece of conceptually coarse learning content would be something that illustrates relational database management systems (RDBMS). This would contain sub-concepts such as the relational model, RDBMS architectures and structured query language (SQL). Each of those concepts would consist of several sub concepts and so on. By reusing this piece of learning content in another learning experience, the learner would be exposed to not only the crux concept but all subsequent concepts as well. Characteristically, conceptually coarse learning content tends to be also technically coarse learning content. The inherent aggregation and flow of this coarse learning content also restricts its usability properties. There exists, therefore, a requirement for some level of granular mapping (mapping between knowledge granularity and learning content granularity) and granular interoperability (mechanisms to reason across the vocabularies and taxonomies of the knowledge and of the learning content).

The technical requirements for content sourcing emulate from the different syntactic layers, namely the representation layer (meta-model) and the descriptive layer (metadata). Requirements from the representation layer arise from the use of different standards schemas to tag the learning content and in a lot of cases regarding open corpus content, the absence of any metadata information. Requirements originating at the descriptive layer include the different types and different formats of the metadata, and the different vocabularies and different taxonomies used to describe the learning content. From this we can extrapolate a requirement for syntax mapping and syntax interoperability mechanisms.

3. Dynamic Sourcing of Open Corpus Content

Open corpus content can be defined as content that is freely available for use by any educational institution or system. Open corpus content can be sourced from various locations (See Figure 1). The content may be available via open source or commercial digital repositories or from the Worldwide Web.

Such content is available in public digital repositories such as, Connexions [Connexions] and JORUM [JORUM], commercial digital repositories such as, XanEdu [XanEdu] and LydiaLearn [LydiaLearn]. Countries such as the UK, USA, Canada, Ireland and Australia, are increasingly investing in national digital repositories to encourage the re-use of digital content. Examples of such initiatives include ARIADNE Foundation (European Union) [AriadneKP], Education Network Australia Online (Australia) [EDNA], eduSource (Canada) [eduSource], Multimedia Educational Resources for Learning and Online Teaching and Gateway to Educational Materials (USA) [MERLOT], NDLR (Ireland) [NDLR] and the National Institute of Multimedia Education (JAPAN) [NIME].

There are also both commercial and open source repository software solutions available that institutions are using to establish digital repositories of their own. Examples of open source repository software solutions are DSpace [DSpace] and Fedora [Fedora], while commercial offerings include IBM Workplace [Workplace] and Intrallect Intralibrary [Intralibrary].

Open corpus content may also be available from whitepapers or directly via the World Wide Web. In the case of content sourced on the Worldwide Web, no assumptions can be made regarding the structure of the content or the associated metadata information. The content may not be stored in any structured fashion and there may not be metadata descriptions of the content available.

Open corpus learning content can exist in numerous formats. It can be wrapped and tagged as a learning object or as an entire course; it can take the form of video, audio or graphical streaming; it can be in the form of documents, pdf files or PowerPoint presentations. Any text, regardless of its granularity, from single words or sentences up to entire books, novels or plays constitutes learning content. If the vast amount of open corpus

learning content that is available can be leveraged for eLearning environments, it becomes possible to facilitate Dynamic Contextual eLearning. The desired scenario is to facilitate the use of “any content, anywhere”.

To facilitate the use of open corpus content in eLearning environments, mechanisms for sourcing, harvesting and delivering this content need to be identified and made accessible through flexible services. An example scenario in which these services would be beneficially used involves Personalised eLearning Development Environments (PEDE) [Dagger, 2006]. This describes a situation where the generation of semantic queries, based on the content requirements embedded in Personalised eLearning designs, can provide an Open Corpus Content Sourcing service (OCCS) with blueprints for content required to satisfy the educational outcomes of Personalised eLearning design. This scenario is further explored in section 4 of this paper.

3.1 Sourcing Open Corpus Content

In order to cope with the numerous sources, formats and attributes of open corpus content, an OCCS will need to perform a series of operations to ensure accuracy and consistency in content identification and harvesting.

A web crawler will traverse digital repositories and the Worldwide Web creating a metadata cache of sourced learning content. As discussed, inconsistencies in both the vocabulary used and the metadata standards implemented in sourced content need to be addressed.

The variety of metadata standards applied to content on the Worldwide Web and in repositories currently limits the interoperability of content and systems. The lack of a standard vocabulary in describing content also makes semantic matching of searches to relevant content a difficult task. These interoperability issues will not only effect the identification of suitable learning content but also the re-use of any learning content delivered to elearning environments. Content will be encountered with various metadata standards applied, SCORM, Dublin-Core, IEEE LOM or one of a variety of other metadata standards, some of which may be proprietary.

A mapping from the current metadata standard applied to the content to a canonical metadata model will be required to resolve this problem. A fixed vocabulary and taxonomy will be used to ensure consistency in the metadata descriptions of content. This mapping is essential to ensuring semantic differences do not cause searches of the metadata cache to overlook suitable content. The ability to ensure interoperability between metadata standards would be a major step in the development of the semantic web.

In the case of some open corpus content there may be no associated metadata descriptions. Automatically generating descriptions of learning content that has no, or insufficient, metadata information, will be a major challenge in leveraging open corpus content. Thoroughness and consistency in this metadata generation is essential to ensure accurate identification and retrieval of content.

The number of tools emerging in the research community that facilitate the automatic generation of semantic content descriptions is an indication of its importance in the progression and development of the Worldwide Web. Future systems will rely on machines being able to correctly comprehend and process the content contained in web pages. Annotea [Annotea], developed by the W3C project, uses an RDF based annotation schema to describe annotations as metadata. Metasaur [Metasaur], developed by the University of Sydney, supports the creation of metadata for learning objects, using a standard vocabulary ontology that is generated automatically. This ontology can then have terms added to improve accuracy. This ensures consistency in the terms used to tag content. Semtag [Dill et al., 2003] is a system that has been developed using Seeker, which is a platform for large-scale text analytics. Semtag was created by the IBM Almaden Research Centre. The system crawls through web pages and performs an automated semantic tagging of each page using the TAP ontology. Semtag uses a Taxonomy Based Disambiguation (TBD) algorithm to ensure the correct classification of content in its tagging. IBM’s LanguageWare [LWLE] is a text analytics tool than can be employed in the classification of content that has no associated descriptions. The tool analyses content for vocabulary found in an ontology to gain an understanding of the contents function, topic, structure etc.

Thorough and consistent generation of metadata descriptions for learning content will not only ensure the identification and harvesting of suitable learning content, it will also provide a major step towards and entirely machine-comprehendible “semantic web”.

3.2 Identification and Harvesting of Learning Content

Once the metadata cache has been created by the OCCS the identification of suitable learning content and retrieval of that content needs to take place. When searching for suitable candidate content to deliver to the eLearning environment, the search functionality must satisfy the semantic content requirements that have been provided by the environment. This creates a situation where content requirements generated by the eLearning system, can be used to ensure the accurate retrieval of applicable learning content.

As previously discussed, there are both semantic and syntactic requirements that must be fulfilled by any candidate content provided. These requirements will fall under the following categories

- Semantic
    - Pedagogical structuring
    - Subject Matter descriptions
        - Vocabulary
        - Context
    - Activity
    - Personalisation
    - Strategic Flow
- Syntactic
    - Meta-Model (Representation Layer)
        - Standards
    - Metadata (Descriptive Layer)
        - Taxonomies
        - Format
- Both Semantic and Syntactic
    - Granularity

Semantically the content must be compatible with the concept granularity in all areas, pedagogical, subject matter, activity and personalisation granularity. Syntactically the content must meet the technical requirements of the environment, it must be structured correctly and a metadata mapping must have taken place to ensure accuracy and consistency in the search performed.

Once content has been sourced, its granularity can dictate how useful it may be to the eLearning environment. The OCCS will catalogue all sourced learning content regardless of its conceptual and technical granularity. eLearning environments that perform personalisation will require content that is both conceptually and technically fine grained. Small pieces of content such as paragraphs of text or single images can be more easily inserted, reused and repurposed during personalisation sequencing. The content requirements provided by the elearning environment need to accurately identify the concept, subject matter and detail desired to ensure that the content provided is granularly appropriate.

A key issue that arises in the harvesting of content refers to intellectual property rights and digital rights management. The ownership of digital content and permissions around its harvesting and use is a major issue in all areas of society. Some public repositories, such as the Maricopa Learning Exchange [20] are trying to address this by applying licensing to all its stored content. However content residing openly on the World Wide Web currently has no such standards applied.

3.3 Content Delivery

Once the required content has been retrieved it needs to be delivered to the eLearning environment in a format that can be readily consumed. The candidate content can be passed back in its raw state or it can be restructured into a learning object for ease of incorporation into a learning offering and for ease of reuse. This means the provision of appropriately grained and suitably tagged content to a learning object generation environment such as the iClass [iClass] project's LOGenerator [Brady et al., 2005]. The LO generator is responsible for generating appropriate learning objects based on the pedagogy being used, the subject area concepts and the sourced learning content. It also ensures the correct strategic flow of the learning content within the LO.

3.5 Proposed Architecture

It is proposed to provide the open corpus content sourcing function as a stand-alone service that can replace the current method of content sourcing in the architecture of elearning environments. This will minimize the impact on the current architecture of such systems.

The OCCS generates a content cache of all sourced learning content. When a web crawl is performed and content identified an analysis is performed to ascertain the condition of the content. The associated metadata descriptions are investigated to ensure that they are sufficient. A mapping to the canonical metadata model is then performed. The new tags are created using a standard vocabulary and taxonomy. Once this mapping is complete, an entry can be made in the OCCS metadata cache.

The Semantic and Syntactic content requirements that have been generated by the system or extracted from the elearning designs are passed to OCCS. These requirements are then used to generate a semantic search query.

This search query is used to identify suitable learning content from the metadata cache stored within the content service.

Once the search query has been performed and candidate content has been identified and harvested from its physical location, the content is passed back to the elearning environment or to an LO Generation service. This service structures and sequences the content into and LO. This LO is then passed back to the elearning environment. The developer of the elearning offering can then assess the content and decide on its suitability. If they are not satisfied, the semantic search requirements that were passed to the OCCS can be manually refined and the search query regenerated to improve the suitability of the content that is returned.

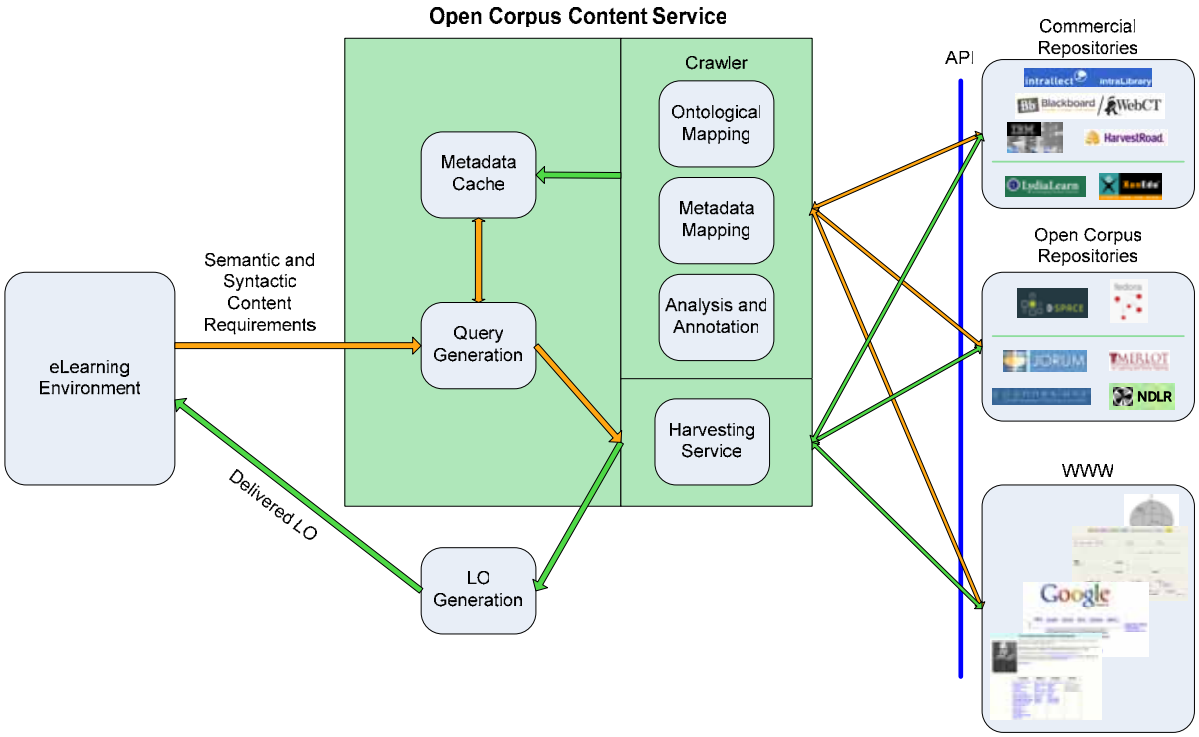

Figure 1, Open Corpus Content Service Architecture

4. Scenario for Requirements Gathering

Personalised eLearning Development Environments provide course developers with pedagogical scaffolding for the effective development of Personalised eLearning experiences. They provide various levels of support during the adaptive course composition process through pedagogic modelling, activity modelling, subject area modelling and personalisation axes modelling. However, the sourcing of appropriate learning resources to satisfy the learning outcomes of Personalised eLearning Design has a number of prerequisite requirements, namely; [i)] the content must already exist, [ii)] the content must be marked up appropriately and [iii)] the content must be accessible. However, inherent to the Personalised eLearning Design are the blueprints for the content which are required to satisfy its learning outcomes. These blueprints describe a range of invaluable information which could be reused to source appropriate learning content. For example, within a Personalised eLearning Design a section applying the pedagogic principle of “introduction” to the subject area of “XML” while adapting based on the users “VARK” learning style value may describe the following blueprints; “all learning content (text, images, audio, video, etc.) covering the subject area of XML based on the WebQuest pedagogic principal of Introduction”. As illustrated in figure 2, these blueprints can then be used to form the semantic queries which an Open Corpus Content harvesting service can use for sourcing appropriate learning resource candidates. The harvesting service then returns collections of candidate learning resources which satisfy the requirements defined in the blueprints.

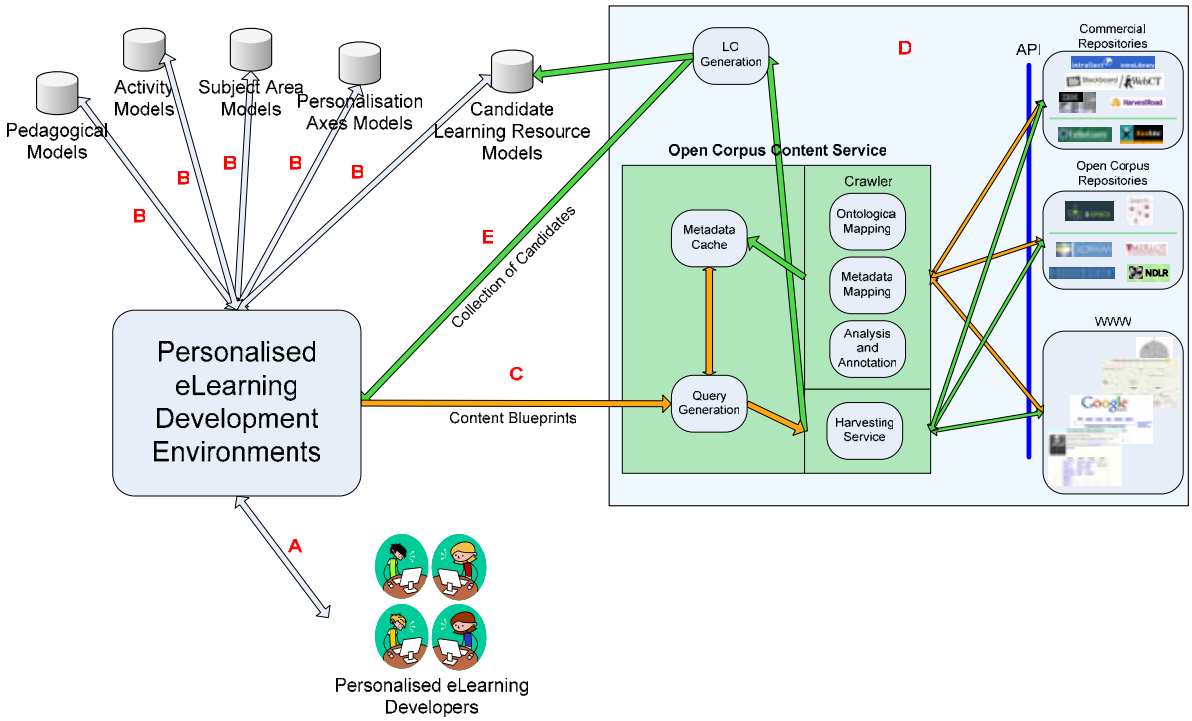

Figure 2, Satisfying the Content Requirements of Personalised eLearning Development Environments

Figure 2, depicts the following scenario. The Personalised eLearning Developers (PED) interact with the support-oriented Personalised eLearning Development Environments (PEDE) to design their Personalised eLearning experiences (A). The PEDE uses a range of information models (pedagogy, activity, subject area, personalisation axes and candidate learning resource) and the Adaptive Course Construction Methodology to support the PED in creating their Personalised eLearning experiences (B). Noticing that appropriate learning content is not currently available, through the candidate learning resources, the PED indicates to the PEDE that it should seek alternative candidates. The PEDE then sends off the semantic queries to the Open Corpus Content harvesting service to find appropriate content based on the supplied blueprints (C). The harvesting service then initiates the sourcing process and begins assembling appropriate candidate learning resource (D). When the harvesting service has finished it returns the collections of appropriate candidates to the PEDE which in turn updates its candidate learning resource models (E). Now the course developer can complete the creation of their instantiated Personalised eLearning design and begin the testing phase through the publication support of the PEDE.

This proposed functionality would significantly increase the support offered to the course developer, through Personalised eLearning Development Environments, and further reduce the complexities and development time involved with designing Personalised eLearning experiences.

5. Conclusions

A number of key issues in the dynamic sourcing of open corpus content were addressed in this paper. The requirements needed to accurately describe a desired piece of content were identified. These requirements are both semantic and technical in nature and all need to be satisfied if the sourcing of content is to be efficient and precise.

Open corpus content and the issues surrounding its identification, harvesting and delivery were discussed in detail. The numerous locations in which the content can reside and the many formats which it may adopt were identified; issues which can affect both the sourcing and categorization of content. Once content is identified there are numerous issues which need to be addressed; metadata and vocabulary mappings need to be applied; metadata generation may need to occur if insufficient descriptions exist; accurate searching across sourced content needs to be ensured; and the effective delivery of content to the relevant eLearning environment needs to be implemented. All of these issues are addressed throughout the paper.

A proposed architecture for the OCCS is then presented. The individual elements of the architecture are discussed and the functionality of the service is explained. The paper concludes with an investigation into a

usage scenario for an OCCS. The scenario examines the generation of content requirements by a PEDE and the subsequent incorporation of an OCCS into the architecture of the PEDE to satisfy content requirements.

6. References

[Annotea] Annotea – A W3C LEAD project. Available at http://www.annotea.org/

[AriadneKP] Ariadne foundation for the European Knowledge Pool. Available at http://www.ariadne-eu.org/

[Brady et al., 2005] Brady, A., Conlan, O., Wade, V. Towards the Dynamic Personalized Selection and Creation of Learning Objects. In: Proceedings of World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education, Vancouver, CA, 2005 E-Learn 2005, 1903-1909

[Connexions] Connexions - Sharing Knowledge and Building Communities. Available at http://cnx.rice.edu/

[Dagger, 2006] Dagger, D., (2006), "*Personalised eLearning Development Environments*", a thesis submitted to the University of Dublin, Trinity College, for the Degree of Doctor in Philosophy (November 2005),

[DSpace] DSpace Open Source Digital Repository Solution. Available at http://www.dspace.org/

[Dill et al., 2003] Dill, S., Eiron, N., Gibson, D., Gruhl, D., Guha, R., Jhingran, A., Kanungo, T., Rajagopalan, S., Tomkins, A., Tomlin, J., Zien, J. (2003) SemTag and Seeker: Bootstrapping the semantic web via automated semantic annotation. In: 12th International Conference on World Wide Web, Budapest, Hungary, 2003, 178-186.

[EDNA] Education Network Australia Online. Available at http://www.edna.edu.au

[eduSource] eduSource - Canadian network of learning object repositories. Available at http://www.edusource.ca

[Fedora] Fedora – Flexible Extensible Digital Object Repository Architecture. Available at http://www.fedora.info/

[iClass] Intelligent Distributed Cognitive-based Open Learning System for Schools (iClass), European Commission FP6 IST Project. http://www.iclass.info

[IntraLibrary] Intrallect *intra*library – Learning Object Repository System – Available at http://www.intrallect.com/

[Jorum] Jorum - UK HE Institutions Digital Repository. Available at http://www.jorum.ac.uk/

[LydiaLearn] LydiaLearn provides global access to eLearning resources. Available at http://www.lydialearn.com/

[LWLE] IBM LanguageWare Linguistic Engine – Available at http://www-306.ibm.com/software/globalization/topics/languageware/index.jsp

[Maricopa] Maricopa Learning Exchange. Available at http://www.mcli.dist.maricopa.edu/mlx/

[MERLOT] Multimedia Educational Resource for Learning and Online Teaching. Available at http://www.merlot.org

[MetaSaur] University of Sydney's Metasaur Project. Available at www.it.usyd.edu.au/~alum/demos/metasaur_hci/

[NDLR] NDLR – Irish National Digital Learning Repository. Available at http://www.learningcontent.edu.ie/

[NIME] National Institute of Multimedia Education. Japan. Available at http://www.nime.ac.jp/index-e.html

[Workplace] IBM Workplace. Available at http://www.ibm.com/software/workplace/collaborativelearning

[XanEdu] XanEdu - Digital Repository Application Service Provider. Available at http://www.xanedu.com/