Towards Linked Education Data: Metadata Extraction Projects for Education Network Australia (edna)

Sarah Hayman
Education.au, Australia
shayman@educationau.edu.au

Nick Lothian
Education.au, Australia
nlothian@educationau.edu.au

Abstract

This paper describes some different but related projects for edna (Education Network Australia), all of which undertake metadata extraction from a range of sources to facilitate the semantic annotation of collections of learning resources for Australian educators. A team of Information Officers at edna is responsible for building and maintaining a large collection of web-based learning resources with associated metadata. This process has been largely a manual approach at the individual item level. In an attempt to enhance the relevance and efficiency of edna's own collection, a two-pronged approach has been taken: on the one hand harvesting user selection, evaluation and metadata through social bookmarking tools and on the other employing some automated metadata creation tools to increase efficiency of description. It is envisaged that user engagement will augment the quality, relevance and currency of the resources in edna for the community of users and the rich metadata collected from users and automated tools will enhance the metadata and extend the discoverability of the collection.

Keywords: Education Network Australia; edna; me.edu.au; metadata; social bookmarking; metadata extraction; automated metadata; collection building; learning resources; resource discovery; machine learning; user-created metadata .

1. edna Semantic Collection (ESC) Project

1.1. Overview

ESC is a new tool designed to take advantage of the web 2.0 environment and harvest bookmarked items from key educators in Australia and worldwide. The tool at this stage is for internal use, to enhance the edna collection. It takes selected quality RSS feeds and displays their items together with descriptions, subject tags and other useful metadata (some created through automation). Information management staff at edna then evaluate and select the items, together with their metadata, and add those selected straight to the edna collection. This has resulted in efficiency gains as well as a unique perspective on what edna users value. Through this we can continue to maintain the quality, relevance and timeliness of edna's collection of educational resources. The tool is known as ESC (edna semantic collection). It is a building block for further developments in harnessing user contributions to edna and employing features of the semantic web.

1.2. Background

The edna semantic collection project aimed to enhance the quality, size and diversity of the edna digital resource collection through user engagement, automation and improved metadata tools.

edna was established in 1996 as an online service to support and promote the benefits of technology for education and training in Australia. It is a collaborative information service which is funded and developed in partnership with the Australian education and training community. edna is an aggregator service. It investigates, compiles, filters, evaluates, annotates and consolidates quality information and provides access to online resources, news, networks, events, projects and research for educators. The project supports a set of websites, collaborative workspaces, discussion lists, professional networking services and xml-based information

https://doi.org/10.23106/dcmi.952109636

services which are used by stakeholders on their own websites, portals, RSS readers and handheld devices. Content syndication via RSS, federated search and federated security is a key feature of the edna project. The edna metadata profile is based on the Dublin Core Metadata Initiative (DCMI, 1998) and the edna metadata standard v1.1 (education.au, 2001).

Developing and maintaining a collection involves policies, procedures, a collection mechanism, a storage and retrieval system and communication with the audience and stakeholders. The current Web 2.0 environment offers opportunities in all these areas of collection development and management.

A series of software applications and enhancements delivered within this project addressed three fundamental processes of digital collection development:

1. Discovery of resources
2. Evaluation of resources
3. Description of resources

From edna's inception, all three of these tasks for the edna collection have been undertaken by Information Officers at Education.au. Many aspects of all three tasks outlined are repetitive and lend themselves to automation. The semantic collection project is part of a broader platform of digital content management efficiencies, enhancements and software tools to position edna and associated services for the future. The knowledge and tools developed will be shared with stakeholders (state and territory education jurisdictions and institutions) who have expressed interest in these approaches for their own collection purposes.

1.3. Release 1: senda

Release 1 of this project developed a bookmarklet tool (called senda) for Information Officers at Education.au to use in a web browser. This enables them to capture appropriate resources for edna, together with some harvested metadata (for example, a highlighted description from the resource itself) and send this directly to the edna DSpace repository for final editing and approval.

The tool checks the repository for existing items (by url already in the collection). It also allows for referral of an item to other Information Officers. It enhances speed, efficiency and collaboration and was a foundation for the second release ESC. It requires an administrative login to DSpace.

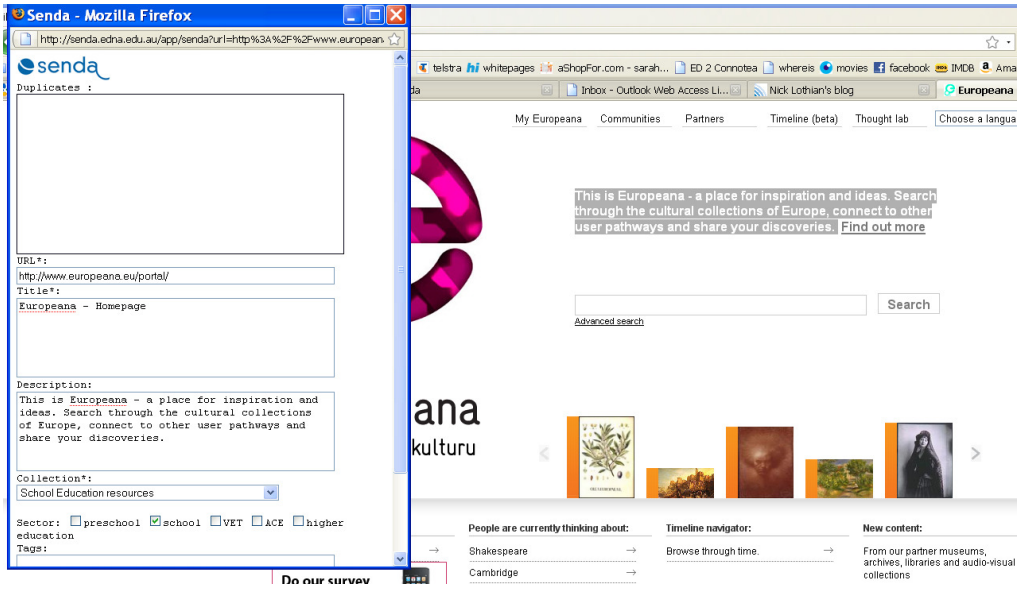

FIG. 1. Senda tool with Europeana front page.

https://doi.org/10.23106/dcmi.952109636

Figure 1 shows the senda tool activated for the Europeana website. One click (once logged in to DSpace) has brought up the screen showing captured metadata and no duplicates (at this point Europeana had not been added to the edna collection). Other fields allow for selection of the appropriate collection, entry of user tags and referral if required to a colleague.

Current anecdotal feedback suggests the senda tool is proving useful as:

- a quick way to check if a web resource has already been entered into edna
- a quick way of referring a web resource to a colleague within the DSpace metadata management system
- a quick way of saving a web resource to DSpace to work on immediately or later.

1.4. Release 2: ESC

Release 2 was the development of a tool (called ESC) to retrieve resources from a series of carefully selected sources within trusted applications by trusted contributors, via RSS feeds. The tool also extracts metadata from the feeds and the resources and provides some automated metadata (using OpenCalais[1]). Final selection, evaluation and description are undertaken by edna Information Officers to maintain the consistency and quality of the edna collection. Feeds can be managed, selected and customised by the Information Officers.

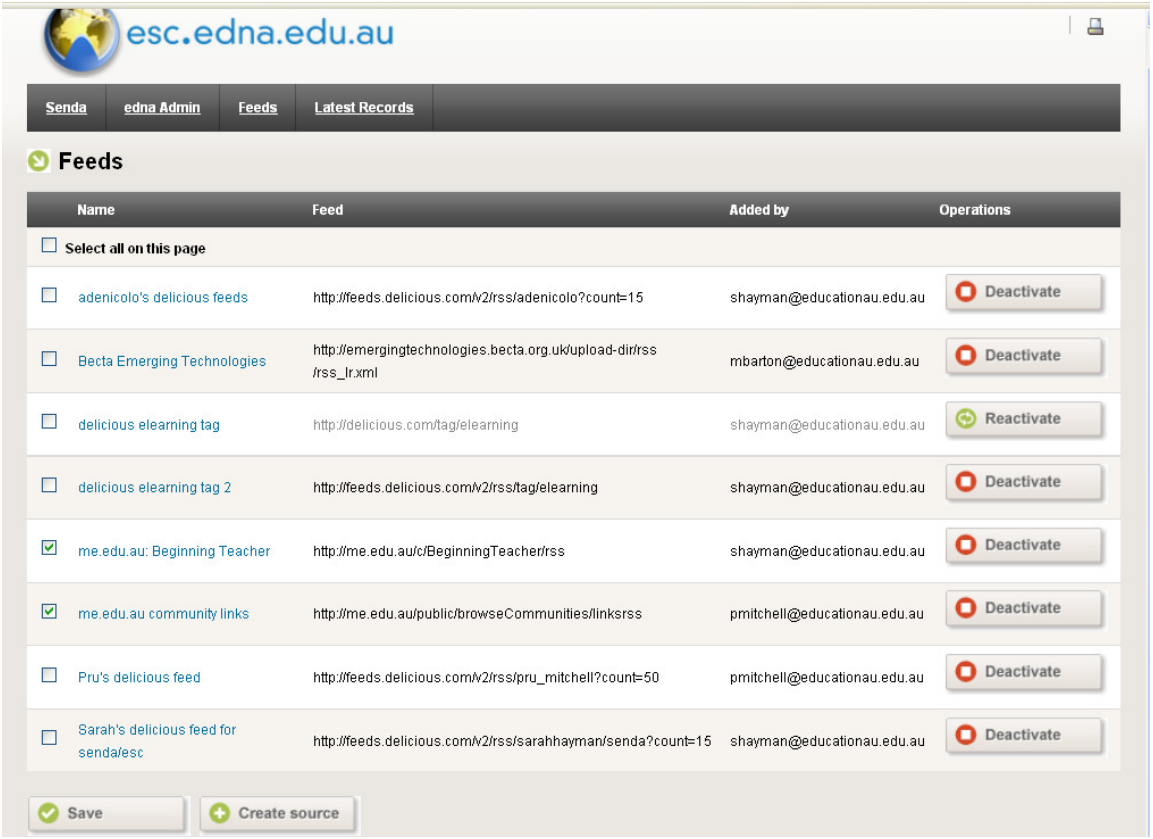

FIG. 2. ESC Feeds Subscriptions screen.

Figure 2 shows the Feed Subscriptions screen of the ESC tool as viewed by a logged in Information officer. This is the list of chosen feeds at May 2009. They can be activated or deactivated. Bookmarks from all activated feeds and all associated metadata are collected and stored in the ESC repository. Users (currently only internal) can choose from which feeds they

[1] http://www.opencalais.com/

https://doi.org/10.23106/dcmi.952109636

want to see items (but all items are still collected unless the feed is deactivated). Selecting one or more feeds and then clicking on **Save** takes the user to the next screen, shown in Figure 3 below:

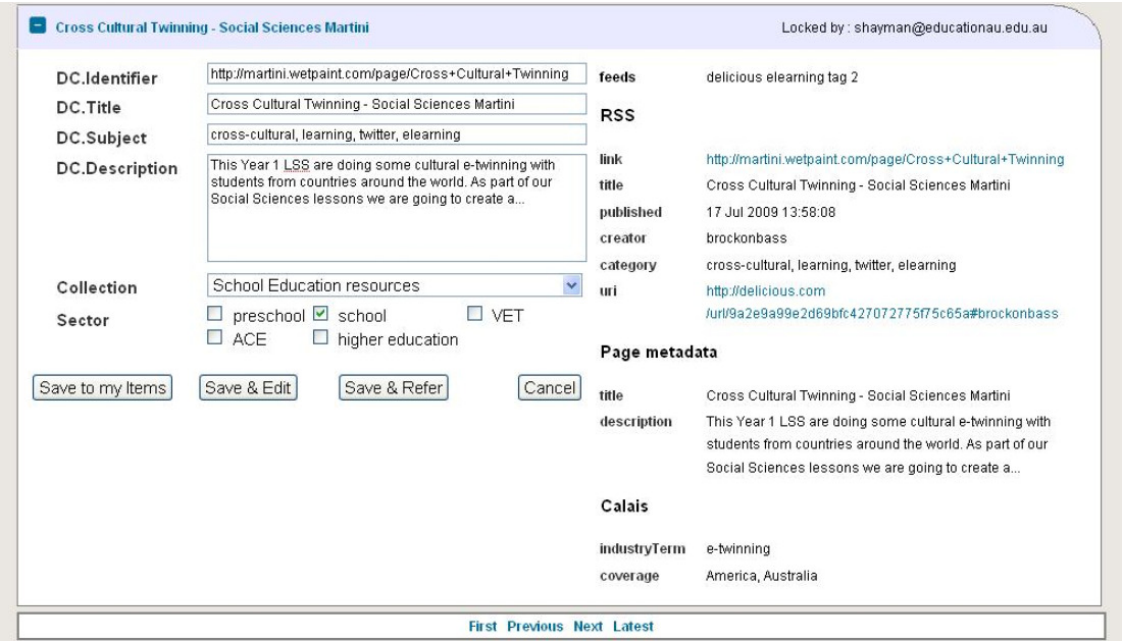

FIG. 3. Section of ESC Data Collection screen

This section of the ESC data collection screen shows an item with its associated metadata on the right hand side and the ESC edit screen on the left that will send the metadata to DSpace in the same way that the senda tool does. In this example, a website describing a cross-cultural twinning exercise for school students has been picked up from the delicious *elearning* tag feed[2]. ESC has already checked if it is in the edna collection (duplicates are not shown). ESC has collected metadata from three sources: delicious (including the bookmarker, the date bookmarked, the user's tags); the page metadata (including description and title; DC metadata fields would be collected if available); OpenCalais automated metadata (including industry terms and coverage). Note the Calais term e-twinning does not occur in the other sources. This metadata can be mapped to DSpace fields as desired by the Information Officer. It is a rich set of metadata that can be evaluated for use in edna. Meanwhile it is all being collected and stored in the ESC repository.

1.5. Outcomes

The following outcomes have been realised from early use of ESC:

- increased efficiency for Information Officers meaning that the building of the edna collection can be managed more quickly, adding more and better resources in a timely manner
- freeing of Information Officers to focus on the quality and higher end work that their expertise allows (automating some of the repetitive tasks)
- adding a new element of user-suggestion via the feeds for items to go into edna after final evaluation by internal IO experts (still a major and important point of difference for edna)
- provision of a wider range of resources that are being captured and evaluated eg online video, pictures, expert opinions and commentary on current educational issues.

[2] http://feeds.delicious.com/v2/rss/tag/elearning

https://doi.org/10.23106/dcmi.952109636

- a flexible system allowing for a range of feeds to be added; Information Officers can even build an individual feed by using an aggregator method such as Yahoo Pipes[3]. Business rules will be developed for managing the collection of feeds in ESC.

1.6. Future

In the near future, resources from the social bookmarking release of edna's professional bookmarking service, me.edu.au[4] can be collected via ESC and incorporated into the edna collection with automatically generated metadata and user-created metadata such as descriptions/comments and tags.

In the longer term we see the ESC tool as laying a foundation for a possible second repository to sit alongside edna. It would contain user-bookmarked items where the users might be all those educators who are bookmarking web resources in a range of places (such as delicious and diigo[5] as well as me.edu.au). The ESC tool collects all the items in a datastore which itself could be published as a more "liberal" edna – without the full evaluation that edna provides but with edna-like metadata and searchable – this repository could be seen as sitting somewhere in between Google and edna.

2. me.edu.au

me.edu.au is a professional networking site for Australian educators. The site is under development and this paper discusses its functionality as at May 2009.

One of the main features of the site is the ability for users to share web based resources with their colleagues either directly or by entering an RSS/Atom feed of their own publishing activity as a source of links. While there are many other sites that let users share resources like that (delicious and diigo are two obvious ones) this has a few unique features.

When resources are entered into the system it extracts as much metadata information as possible.

edna developers have built a generalised metadata extraction and mapping system which can automatically extract and store metadata found in RSS/Atom feeds, HTML pages and MP3/podcasts. It stores metadata in a relational (Postgres) database, but in a schema based around the RDF data model.

Metadata is extracted into name-predicate-value triples.

Various types of metadata are exposed to users in standardised ways. For example, four different types of geographic metadata are commonly encountered:

1. webpage ICBM metadata
2. webpage geo.position metadata
3. georss metadata in RSS/Atom feeds; and
4. gml metadata in RSS/Atom feeds

It is planned to support geographic data in JPEG files the near future.

Semantically, this information is identical and needs to be presented to the user in the same way, that is, as a map. To support this, a mapping process has been built which takes values with a given name and predicate and outputs them into our standard display formats.

[3] http://pipes.yahoo.com/pipes
[4] http://me.edu.au
[5] http://www.diigo.com

https://doi.org/10.23106/dcmi.952109636

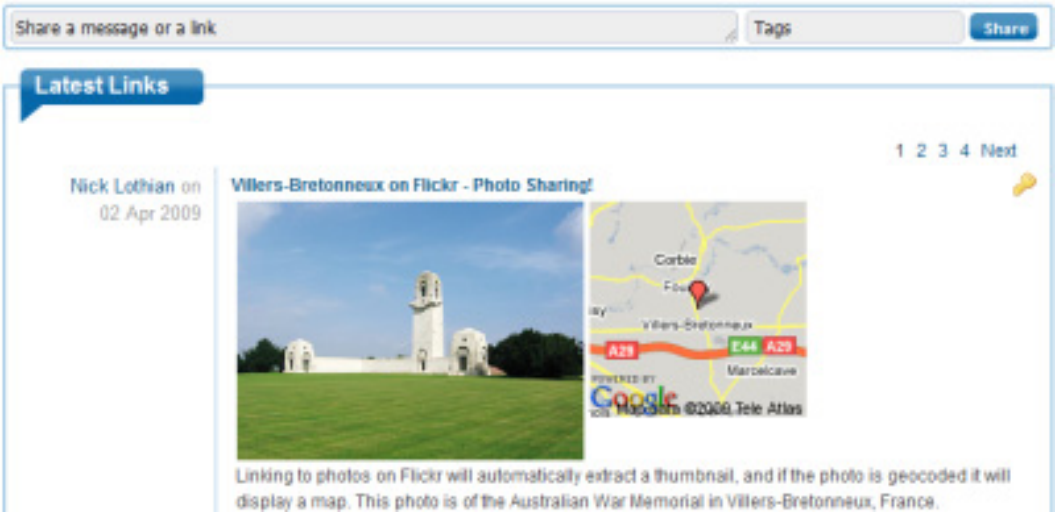

FIG. 4: View of the me.edu.au Share Interface

While this metadata isn't all exposed yet, Fig 4 above shows the beginnings of what can be done with it. In this case, a link to a Flickr photo page has been shared which was marked up using the geotag standard (me.edu.au supports both the ICBM and geo.position methods). Using that metadata enabled the location of the photo to be extracted and a map to be displayed. In the future there is potential to do a position-based search which shows other resources in the same area.

This process is useful for all forms of data in our triple store. Our semantic mapping process can handle situations where no name/predicate value exists and fall back to other options. For example it is always necessary to display a title for a resource. MP3 Podcasts often include ID3V1 metadata from which a title can be extracted, but in the case where that metadata does not exist it is possible to fall back to other options (in this case file title).

Metadata is also consolidated from multiple sources. For example, Slideshare and Flickr RSS feeds both support the Media RSS format, which allows extraction and use of thumbnail images. Here, for example (Figure 5) Julian Ridden has posted his excellent moodle themeing slidedeck on Slideshare, which shows up in me.edu.au via RSS:

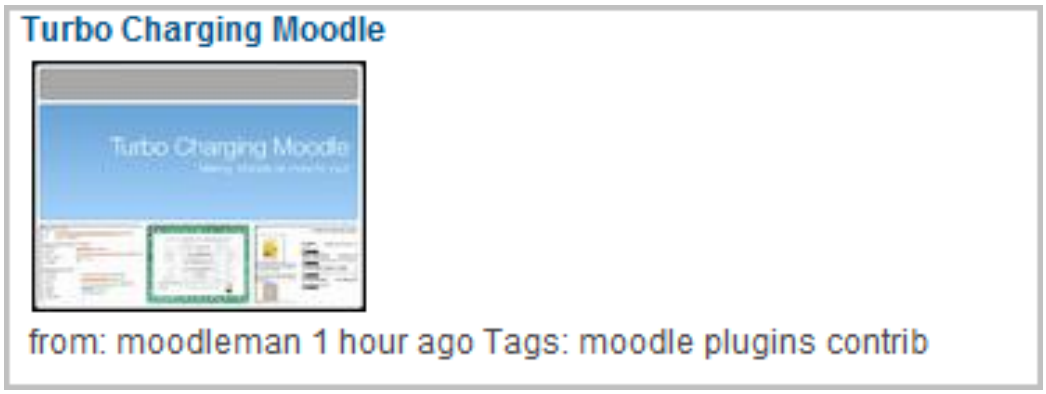

FIG. 5: Thumbnail from Slideshare in me.edu.au

When someone bookmarks the same item later in diigo or delicious the RSS will not contain the thumbnail. Fortunately the system can look up the thumbnail for the resource and enhance the

https://doi.org/10.23106/dcmi.952109636

display of that item with it. Here (Figure 6), Kerry Johnson has bookmarked the same slideshow on diigo, and me.edu.au has displayed her caption with the original thumbnail.

FIG. 6 Thumbnail enhancement of diigo link

Presently the collected metadata is exposed in very limited ways (primarily in HTML and RSS feeds). However, further system functionality currently under development will expose a much more comprehensive view of the metadata we are collecting for each item. As well as providing a page for each collected resource displaying human readable metadata, it is intended to provide users with ways to edit and organise the items they have shared with me.edu.au.

3. Collaborative artificial intelligence research project

In 2008-2009, a joint project has been undertaken by researchers from Flinders University Artificial Intelligence Laboratory in collaboration with information managers from Education.au, the agency that manages and produces Education Network Australia (edna).

The project is aimed at partial automation of the selection, categorisation and annotation of web pages for inclusion in edna. It investigates the possibility of analysing the text of new resources in order to automatically classify them into edna topics, and suggesting controlled vocabulary subject terms for them. Among other things, the project investigates the production of a classification data capture tool to capture the reasons for classification decisions made by Information Officers and a topic classifier to classify new documents into edna categories.

In preliminary work on mapping subject terms from a controlled vocabulary (the ScOT ontology (http://scot.curriculum.edu.au/), extensively used in edna) to edna categories, results have indicated that a knowledge engineering approach may be more successful than a standard Machine Learning approach. A mapping based on semantic similarity in WordNet (see Yang and Powers, 2005) was able to attain 60% accuracy, compared to using common classical machine learning methods such as SVM and Naïve Bayesian classifier.

This work is preliminary but we believe merits mention here because of its relationship to the other projects.

4. Conclusion

Experience gained through involvement in this suite of projects highlighted some new directions that we believe must be considered in managing online resources collections in 21st century collections and particularly to position such collections for imminent developments in resource discovery and collection optimisation. We anticipate a movement towards a collection built to a far greater extent by users and managed and facilitated to a far greater extent by smart tools exploiting interoperable metadata.

In reviewing the early literature on semantic collections in education, several projects based on LOM metadata (IEEE, 2002) were found (Nilsson, M. et al. , 2003, Porto, F. et al, 2007, and

https://doi.org/10.23106/dcmi.952109636

Prolearn D4.1, 2004), however the edna projects described use Dublin Core based metadata in line with the edna standard.

The edna collection is not yet a semantic web (or linked data) service but we aim through developments like those described above to position the collection, its mechanisms and its content (and indeed its creators and its user community), for the linked data environment that is undoubtedly on its way. The future of semantic collections is uncertain (even contentious) but investigation of the issues is important for those whose services are based on traditional metadata practice, in order to maximise their existing investment as well as continue to provide the best possible service to users. Further information about any of the projects described in this paper is available from the authors and their colleagues and we would welcome comments, questions and suggestions

References

DCMI. (1998). Dublin Core Metadata Element Set, version 1.0: Reference description. Retrieved, May 15, 2009, from http://www.dublincore.org/documents/1998/09/dces/.

education.au. (2001). EdNA metadata standard v1.1. Retrieved, May 15, 2009, from http://www.edna.edu.au/edna/go/resources/metadata.

education.au. (2006). edna metadata application profile. Retrieved, May 15, 2009, from http://www.edna.edu.au/edna/go/resources/metadata/edna_metadata_profile/.

education.au. (2009). edna Labs. Retrieved, May 15, 2009, from http://labs.edna.edu.au/.

Hunter, Jane, Khan, Imran, and Gerber, Anna. (2008). Harvana: harvesting community tags to enrich collection metadata. Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries (2008), 147-156. Retrieved, May 16, 2009, from DOI= http://doi.acm.org/10.1145/1378889.1378916

IEEE. (2002). Draft standard for learning object metadata. Retrieved, May 16, 2009 from http://ltsc.ieee.org/wg12/files/LOM_1484_12_1_v1_Final_Draft.pdf

Lothian, Nick. (2009). Nick@Education.au: Nick's Education.au weblog. Retrieved, May 16, 2009 from http://blogs.educationau.edu.au/nlothian/author/nlothian/ .

Mitchell, Pru. (2007). Education.au and metadata for events. Proceedings of the International Conference on Dublin Core and Metadata Applications, 2007, 106-115.

Nilsson, Mikael, Palm, Matthias, and Brase, Jan. (2003). The LOM RDF binding - principles and implementation. Proceedings of the Third Annual Ariadne Conference. Retrieved, May 16, 2009 from http://kmr.nada.kth.se/papers/SemanticWeb/LOMRDFBinding-ARIADNE.pdf

Porto, Fabio, Moura, Ana Maria de Carvalho, da Silva, Fábio José Coutinho, and Fernandez, Adriana Pereira. (2007). The ROSA project: leveraging e-learning to a semantic layer. International Journal of Knowledge and Learning, 3 (1). Retrieved, May 16, 2009, from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.64.2976&rep=rep1&type=pdf Porto, F.,

ProLearn D4.1. (2004). Infrastructure for (semi-)automatic generation of Learning Object Metadata. Retrieved May 16, 2009 from http://knowgate.nada.kth.se:8080/portfolio/files/Prolearn-KUL/KUL/Participation/WorkPackages/WP_04_-_KUL__leader_/Deliverable_4.1/D4.1.pdf

Stanelis, Nancye. (2009). Web 3.0!: What do you mean Web 3.0? Adelaide: Education.au. Retrieved, May 15, 2009 from http://www.educationau.edu.au/jahia/webdav/site/myjahiasite/shared/papers/Innovate08.pdf.

Yang, Dongqiang and Powers, David M. W. (2005). Measuring semantic similarity in the taxonomy of WordNet. Proceedings of the Twenty-Eighth Australasian Computer Science Conference, 315 - 322.

https://doi.org/10.23106/dcmi.952109636