Semantic Searches for Extracting Similarities in a Content Management System

Amirah Ismail
Centre for Information Technology
Universiti Kebangsaan Malaysia
Selangor, Malaysia
amira@ftsm.ukm.my

Mike Joy
Department of Computer Science
The University of Warwick
Coventry, United Kingdom
M.S.Joy@warwick.ac.uk

Abstract—**Recent content management systems have restricted means for organizing and inferring documents although much of an organization's knowledge can be created in text repositories. In the Semantic Web search emergence, inferring and understanding can be deal by ontology-based semantic mark-up and metadata management. Whilst in the educational domain, learning objects are a fundamental resource. Literally, Content Management Systems and repositories have restricted the means for organising and understanding the captured semantic relationships between the learning objects and other stored documents. To cater this situation, we propose the application of metametadata as a useful semantic based approach to address similarities in a domain to gather definite requirements. This paper focuses on the existing approaches for describing semantic relationships in Content Management Systems and how metametadata capture the pedagogic information which can be applied to enhance the semantic information stored within such a Content Management Systems or repository. It is understood that there is still lacking approaches to address similarities in a domain that meets certain requirements but the progress for the ongoing research in the area is active and shows potential advancement**.

Keywords-component; formatting; style; styling; insert (key words)

I. INTRODUCTION

These days, the amount of learning objects available and stored in several repositories is increasing significantly. A few coexisting trends made this happen. Current software made creating, sharing, and editing of multimedia resources a common and easy task. Several repositories either have been developed for commercial purposes or educational needs have been increased and are on demand from users. Online communities that allow sharing of their learning objects or learning materials from various domains are growing rapidly and users can easily access, organise and annotate their own or others content. These scenarios have added to quantity and multiplicity of content, metadata, and users, that involve such a tough task to manage, retrieve and share the learning objects content.

In this settings, Content Management Systems (CMSs) have been used for many years. CMSs are software system for creating, publishing, editing and managing content. They are extensively used by the news and broadcasting media organizations, e-commerce websites, as well as in film industry, libraries and academic institutions to manage their stored learning objects content. Learning objects content stored in a hierarchical manner in a content repository within a CMS that support for structured and unstructured data. Since CMSs are developed to organize learning objects content and to ensure their accessibility during instinctive queries, metadata task means are considered such a non-trivial characteristic of CMSs.

In this context, the metadata have been considered as a major role and show significant impact on the capability of CMSs to manage, retrieve, and describe content. Metadata serves some meanings generally vital to facilitate a learning object to be located and functional within systems as in [1].

The problem when searching for a learning object from repositories involves locating the resource and the content design that may restrict practical usability as in [1]. Metadata created for educational elements which implicate general meaning across learning contexts and disciplines are open to explanation. For instance, in higher education, learning requires certain objectives to be achieved and the learner to be assessed.

However, the main concern of instructors, designers, learners and academics is the nature of interactivity within a digital learning situation. Metadata as descriptors require explicit meaning which is apparent to, and is interpreted and used in the same way by potential end users (learner, author, teacher, manager) and society. A learning object metadata file may include certain types of information or pedagogical attributes about the learning objects such as the creator's name, organizational connection, learning objectives, prerequisites and keywords.

978-1-61284-353-7/11/$26.00 ©2011 IEEE

However, creating a CMS that can work with different types of metadata may be a tedious task. Metadata formats exist and generally their structure is defined in a metadata scheme (e.g., using XML Schema), and the definite meaning of the metadata fields is given in plain text but in our research context, metadata is created based on ILEELOM metadata standard. This has cause towards several interoperability issues since the absence for formal representation of the underlying semantics.

In this paper, we introduce a semantic approach to build a CMS that is metadata-driven. We propose a Metametadata model, representing system, in the context of a CMS. The model is used to support the created ontology (expressed by an OWL schema) that is linked to a set of controlled vocabularies by metadata ontologies.

To show the extensibility of our system we produce pedagogical ontology based on the IEEELOM Metadata specification applied within the proposed model. Section 2 provides related work within the context of using semantic technologies to support different XML sources in CMSs. In Section 3, we describe the created CMS Metametadata model.

Following, in Section 3 we discuss the interoperability issues whenever we try to provide the definitions for metametadata that integrate the metadata schemes with our Metametadata model. Whilst Section 4 presents the developed Metametadata model service and discusses the used technologies and finally, conclusions are drawn in Section 5.

II. RELATED WORKS

Semantic technology is considered as the uprising area of current research and several available different approaches have been developed to create semantic learning object content management systems. However, to reach its full potential, this technology should provide more and diverse semantic relations between the commonly used metadata terms.

Semantic Search attempts to expand and improve traditional search and retrieval by using XML and RDF to support queries the learning objects. Complex queries can be formulated in order to find groups of objects with certain types that are linked by certain relationships.

In the Semantic Web queries are often applied to graph models in which the nodes represent entities (documents, web pages, learning objects) and the arcs represent relationships. For example, a query such as “Find all documents published in Nature from 2005 to 2007 about tea, where it has been cited by John Smith”, relies on nodes such as “publications”, “topics”, “time restrictions” and authors”, and “published in” and “cited by” (etc.), are required to model the links between those nodes.

Ontologies have been used to convert a natural language query to a formal query and then into a database query as in [2], [3] and [4]. Although this method appears to have been successful, it is not clear whether users actually want to use a full natural language query, and it is important to develop approaches that can respond to a keyword query on a database.

In this regard, research has been conducted on the translation of keywords to XML-based Queries on XML data as in [5], and this is related to our approach as the structure of XML elements infers the relations between given keywords. On the other hand, classify semantic search capabilities into six distinct categories with regards to research aims and objectives, approaches, and functionalities:· document-oriented;· entity- and knowledge-oriented;· multimedia information search;· mining-based· semantic analytics; and relation-centred as in [6].

Whilst this categorisation is just one view of present approaches to semantic search, it nevertheless provides a useful aid to describing current tools and algorithms and to placing our search paradigm in context. The first category – *document-oriented* – is based around document retrieval, where a document encompasses text documents, web pages, or other text-based entities, and retrieval may be via keyword search or query formulation methods as in [7], [8] and [9].

These retrieval technologies rely on entities being named and searches being performed based on the entity types and attributes, and the relations between them. Each of these may be annotated with semantic information within an automated framework, and a semantic web document is therefore different from a standard document since it has been pre-processed in order to facilitate automatic searching.

Many systems applied approaches which centers on restricting the type or form of query allowed. A similar approach is taken by IRIS as in [10], which uses ontology to extract semantically relevant documents from a restricted corpus of Computer Science documents. Search systems in the second category – *entity- and knowledge-oriented* – do more than simply return documents, they provide facilities for links between entities in the search domain to be exploited, for example for exploratory searches.

[11] suggest mapping keywords to matching WordNet synsets, but although they state that this approach is able to ascertain relations between keywords, certain questions have arisen as to how this is achieved, as WordNet itself does not contain any non-taxonomic relations. The difference in approaches relate to the ontology navigation phase, although word sense disambiguation of the terms (in the input query) and words in the document is known to be useful to enhance both precision and recall of an IR system as in [12].

One of the approaches which are similar to ours is SemSearch as in [13]. The idea behind SemSearch is to provide a simple “Google like” interface, allowing users to make simple keyword searches, and these search terms are refined into more complex formal RDF queries – keywords are thus interpreted as instances, properties or concepts.

978-1-61284-353-7/11/$26.00 ©2011 IEEE

Our approach is similar to SemSearch in that we convert each complex keyword query into a logical query. However our approach differs in how the query is computed. Each queries is represented by keywords or controlled vocabularies that need to be mapped from XML to ontologies form and sort the queries according to a few types of search query either by metadata search, ontology search or metametadata search to gather the semantic relationship for the selected learning objects from the corpora of knowledge.

III. METAMETADATA DESCRIPTIONS

Metametadata are data about metadata which represent semantic relationships between items of metadata and between the metadata and one or more semantic domains. The relationships may be structural (physical and logical organization of metadata), behavioral (static or dynamic - change, view, modify semantics) or environmental (creator, revision history). Metametadata will use higher-level definitional associative keywords, or vocabularies from documents describing content, to capture those relationships.

For the purposes of this paper, we represent metametadata using XML, the data we are working with are learning objects, and the semantic domains capture pedagogic information, which will be explored in later in this paper. We consider that the relationships between items of metadata and between the metadata can logically be divided into pure metadata (information about the contents of the learning object), and metametadata (information about the metadata "wrapper"). Simple examples are shown in Figures 1 through 3.

```
ID: 12345
Title: Introduction to Java Programming
Author: John Smith
Learning Objective: Introduce Java
Programming concepts...
URL: www.example.com/javaparams/
Description: This course deals with
functions
concepts, principles and practices within
Java programming
ResourceType: InteractiveResource; Text
Pre-requisites: None
Keywords: Java, Programming, Concepts.
```

Figure 1. Metadata for learning objects

```
Reference: 12345
Modified by: James Jones
Modification date: 15:00-01-May-2008
```

Figure 2. Metametadata (modification details for the metadata.

```
Author Name: John Smith
Organization: Wiley Design Company
```

Figure 3. Metametadata author affiliation

A. Metametadata types

The information captured here may be categorized into two types. Firstly, information about the learning object itself (Figure 1) – this information ("metadata") is essentially static, that is, it will not change unless the learning object is itself altered. Secondly, information about the metadata (Figures 2 and 3) – this information ("metametadata") is dynamic, and may change or be added to; it refers not to the content of the learning object, but to its context. Figure 2 shows a simple example for the environmental relationship for metametadata tagged by the creator, James Jones, and the date when the metadata were last modified the creator.

Figure 3 may also be considered as environmental metametadata since it displays the organization with which the author is affiliated and which may change from time to time. It is accurate to describe it as metametadata since it refers to a field described in the metadata. For example, in Figure 3, the metadata for learning object number 12345 has been modified by James Jones on 1 May 2008. In Figure 3, the author (John Smith) for learning object 12345 is stated as being affiliated to the organization known as the Wiley Design Company – this is metametadata, since the affiliation is information about the author (metadata) for the learning object. Both of these metametadata may change – metadata for learning object 12345 may be edited, and John Smith my change his job.

The dynamic aspect for the metadata context can be done by using automatic extraction of terms from the documents and annotations. Processes involve incorporating extracted terms and linking them to a list of identified terms or to a control vocabulary to support a more semantically oriented search potential. Metadata tags that originate from a list of terms stored in a database handle different metadata transformations, and a list of such terms forms information which we can consider to be metametadata.

We should note that the distinction between metadata and metametadata may not always be simple. For example, a keyword may be used to tag a learning object, and if that keyword is unchanging it is clearly metadata (such as "Java" in Figure 2). However, if a set of keywords might change (perhaps as a result of the use of the learning object) then they may reasonably be considered as metametadata. This is because the changes to the metadata are information about the metadata and about the context of the learning object, which may be categorized as environmental changes to the description of the original metadata but not information about the learning object itself.

Another use of metametadata is to show the changes of information contained in the metadata. For example, a learning object whose metadata had been edited at various times might have that information recorded using environmental metametadata. An example is given in Figure 4, which may consider as the changes in the environmental information for the original metadata.

```
Reference: 12345
Contributor
role: Creator
entity: Wiley, J.
Contributor
role: Validator
entity: Meta project
date: 2007-08-08
Contributor
role: Publisher
```

978-1-61284-353-7/11/$26.00 ©2011 IEEE

```
entity: Western University
date: 2007-04-06
Educational
Intended end-user role: Learner; Author;
Teacher
Learning context: Higher Education
Typical age range: 17-25
Metadatascheme: IMS-IEEE LOM
Language: en
```

Figure 4. *Environmental Metametadata Record*

A pedagogical context for behavioral metametadata may be considered as a semantic structure or network whereby pedagogical entities are assembled. A pedagogical document contains a pedagogical context together with links such as prerequisites. In other words, this type of metametadata assists in coordinating the use and storage of learning objects by connecting and describing the metadata and the metadata sources. For example, in Figure 5, behavioral metametadata may identify connectedness relations between certain learning objects and the contexts of those learning objects.

```
Reference: 12345
IsAPrerequisiteFor: 67890
UsedBy
University: Warwick
Module: CS456
UsedBy
University: Birmingham
Module: CS/200813
```

Figure 5. Behavioral Metametadata

Behavioral metametadata can be considered as knowledge about the metadata itself, and can be used to express similarities between items of metadata. Metametadata formats are supported by IMS as CORBA and XML bindings, and in RDF.

This work addresses the problem raised as in [15], whereby metametadata can be created using a process known as *reification* which contains a vocabulary to allow RDF statements to refer to other RDF statements Structural metametadata can be described as data that describe the *process* of metadata. For example, the description of a modelling language such as UML can be considered as metametadata.

Structural metametadata can be used to specify the types of metadata for aparticular information source. Behavioral metametadata semantics support information extraction, contextual metadata presentation, editing, and relation. Behavioral metametadata use XML bindings for typed object instances, gather metadata subclass meanings from metametadata instances which will then bind metadata XML to metadata subclass instances respectively. Behavioral metametadata are attached to metadata to produce the semantic definitions, for instance, in information visualization composition.

B. Other Definitions of Metametadata

The definition of metametadata used for our research context differs from the common concept for metametadata. In order to place our metametadata definition in context, we compare it with two other notions of metametadata which are commonly used.

The definition of metametadata we are using here is different to that used in the IEEE LOM schema because metametadata in our context focus on the method to support ontologies by providing such semantic identifiers identified by their pedagogic attributes to assist in capturing pedagogic information on the learning object's educational attributes within the educational learning domains.

These metametadata identifiers are created and developed by producing a taxonomy derived from the educational metadata in the IEEE LOM schema which could be able interpret concepts captured in a data model (metadata formats or semantic tags) and relate them to an abstract model that characterizes the various entities involved in the research process.

C. Metametadata Concept

Our work on the Metametadata taxonomy is focused on the identification of the required metadata elements consisting of *Class, Property* and *Representation*.

Metametadata Element Concept (MeMeC) =
ObjectClass + Property
Metametadata element (MeMe) =
Metametadata Element Concept + [Representation]

Figure 6. Metametadata Concept

Figure 6 presents the Metametadata Element Concept to show the relationship between *metametadata element, representation, object classes, property* and *value domain*. A class is a set of clearly defined ideas, abstractions, or "things" in the real world which have common behaviour and properties. A property is an attribute common to all members of a class. A representation of data describes a value domain, data type, and a character set, etc. This representation of terms will be realized in the form of ontologies.

IV. METAMETADATA TAXONOMY

A. Descriptions of a Novel Taxonomy

The metametadata concept is based on pedagogical selection by having typebased logical representations that will be used as vocabularies for the common kinds of learning object features. However, the educational category does not describe the significant connections or relationships between each of the following metadata e Interactivity level, Intended end user role, Context, Difficulty, Typical learning time, Description and Language of the typical intended user (IEEE, 2002).

The proposed metametadata relationship defines the semantic relationship between pedagogical metadata elements.

978-1-61284-353-7/11/$26.00 ©2011 IEEE

Educational metadata from one category in the IEEE LOM specification cover the pedagogical aspects or elements for the learning objects. Other elements listed – the interactivity type or level, semantic density and difficulty – have not been elaborated further here.

There is a need to improve the semantic relationships between metadata under the educational metadata category in LOM in order to improve learning object reusability. Therefore, it is necessary to find a semantic definition by describing each metametadata type that would link pedagogical aspects of chosen learning objects.

We propose a taxonomy as shown in Figure 7 for pedagogic metametadata which uses the IEEE LOM metadata specification elements, together with key pedagogic characteristics, and metametadata elements for relational and classification purposes.

The distinction between data and metadata is well understood, and metadata models may be described by classes, relationships and properties, known collectively as *types*. Our proposed taxonomy consists of a collection of types of metametadata, analogous to types of metadata, which we refer to as *connectors*.

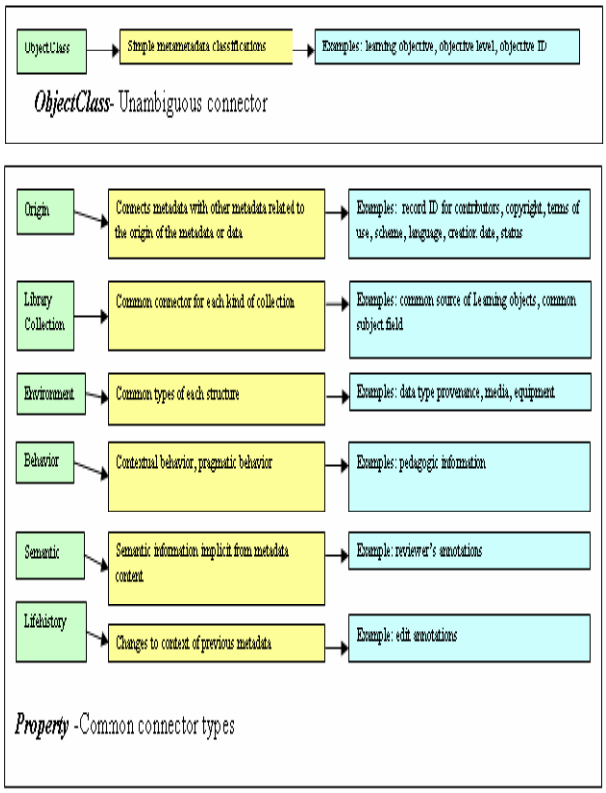

Figure 7. Metametadata Taxonomy

B. *Metametadata Capturing*

Metadata schemas such as the IEEE LOM provide a semantic description of an learning object that is characterized by using key values. This description may be questioned to verify how the learning object has changed and which requires appropriate data types to be included in the metadata schema together with the understanding produced by the authoring tools, adaptive systems and repository interfaces.

A catalogue record using such elements to describe resources or learning objects is a "metadata instance", an example of "instance metadata". These metadata instances could be viewed as particular s of metadata elements connected or linked with a set of "values" for those elements. For instance, "Author: William Shakespeare", "Title: Mark Antony and Cleopatra", and "Subject: Roman Empire".

In IEEE-LOM, the metadata author is identified in the metametadata. Other contributors could also be included in this part. Attributing element authors may be identified by placing a 'source' attribute within the tag pointing to the original metadata (where the first declaration was made).

However, capturing changes needs a state change vocabulary. The intrinsic data typing of metadata keys suggests that different vocabularies are needed to express the relationships between various keys for each tagged metadata item for each learning object. Figure 8 shows the semantic changes for history record for metadata using a metametadata taxonomy or tag, *LifeHistory* metametadata. Thus, this suggests increasing the semantic density level to be set higher from very low to low level for the learning object content.

```
<LifeHistory
xmlns="http://www.thedataweb.org/mif/
MetadataXMLtransactions.html/meta/xml/1
5.08.2006">
<identifier="http://www.thedataweb.org/
mif/MetadataXML
transactions.html>
<change>
<semantic value="very low"
transformation="general:Addition" />
<perspective="semantic-density" />
<record>Initial semantic density level
set</record>
</change>
<change>
<semantic value="low"
transformation="enum:GreaterThan">
<perspective="semantic-density"/>
<record>Increased the level of semantic
density.</record>
</change>
</LifeHistory>
```

Figure 8. Metametadata Structure for Historical Changes in Pedagogical Attributes

Metadata fields, such as URLs matching an author's page on the IEEE portal, characterize valuable semantic relationships. Such URLs are commonly displayed as text to the user. In its place, the semantic architecture permits a metadata field to be specified within the metametadata as related to another metadata field, to be described to the user in the navigational visualization.

V. CONCLUSIONS

We have presented a semantic-based approach using a novel approach to capture the relationships between tagged metadata for learning object stored from the repository. The novel aspects of the research have been motivated by these essential needs as to extend the educational metadata elements to identify the semantic relationships between metadata tags for each learning object or pedagogical resource.

Moreover , we need to have certain mechanisms to extract and capture semantic definitions from each metadata by enhancing and extending vocabularies specifically designed for pedagogical resource purposes to tailor the user's needs. As a result, a new scheme for presenting the metadata is offered by applying a metadata development tool suitable for searching learning objects in the content management system.

REFERENCES

[1] R. Robson. Pedagogic Metadata. *Interactive Learning Environments, 9*(3), 207 - 218. Retrieved October 7, 2002 from http://www.eduworks.com/robby/papers/metadata.pdf

[2] V. Lopez, M. Pasin and E. Motta. "Aqualog: An ontology-portable question answering system for the semantic web". ESWC 2005, pp. 546–562.2005.

[3] A.-M. Popescu, O. Etzioni and H.A. Kautz. Towards a theory of natural language interfaces to databases, *Intelligent User Interfaces*, ACM (2003), pp. 149–157, 2003.

[4] P, Cimiano and J,Völker. Text2Onto : A Framework for Ontology Learning and Data-Driven Change Discovery. Proceedings of the 10th International Conference on Applications of Natural Language to Information Systems (NLDB), 2005.

[5] J. Li, D. Gasevic, J.C. Nesbit, and G. Richards. Ontology Mappings Enable Interoperation of Knowledge Domain Taxonomies. 2nd LORNET international annual conference. Vancouver, 2005.

[6] W. Wei, P.M. Barnaghi, and A. Bargiela.. "Search with Meanings: An Overview of Semantic Search Systems", International journal of Communications of SIWN, Vol. 3, pp. 76-82, 2008.

[7] Heflin, J. and Hendler, J. "Dynamic Ontologies on the Web". Seventeenth National Conference on Artificial Intelligence (AAAI-2000), California, AAAI/MIT Press, 2000.

[8] J. Mayfield. and T. Finin. "Information retrieval on the semantic web: Integrating inference and retrieval. In proceedings of the Workshop on Semantic Web at SIGIR 2003.

[9] A. Kiryakov, B. Popov, I. Terziev, D.Manov, and D. Ognyanoff, D. "Semantic annotation, indexing, and retrieval", Journal of Web Semantics, Vol. 2, No. 1, 2004, pp. 49–79.2004.

[10] W. Wei, P.M. Barnaghi, and A. Bargiela. "The Anatomy and Design of A Semantic Search Engine", Tech. rep. UNMC-CS-200712-1,, School of Computer Science, University of Nottingham Malaysia Campus, 2007.

[11] J. Royo, E. Mena, J. Bernard, A. llarramendi. "Searching the web:From keywords to semantic queries". In: Proceedings of the Third International Conference on Information. (ICITA'05), IEEE Computer Society . 244–249, 2005.

[12] S K Dwivedi and P. Verma. "Web Information Retrieval: Exploring New Dimensions with Word Sense Disambiguation(WSD) for Naïve users", CSI Communications, October 2007.

[13] Y. Lei, V.S. Uren and E Motta, E. "Semsearch: A search engine for the semantic web". EKAW 2006, pp. 238–245.2006.

[14] RDF (2004). "RDF/A Syntax: A collection of attributes for layering RDF on XML languages". 2004-10-11

[15] RDF (2004). "RDF/A Syntax: A collection of attributes for layering RDF on XML languages". 2004-10-11