JIDER
Journal of Distance Education & Research

July - December – 2017
ISSN 2521-2397
journal.vu.edu.pk

Making MOOC Slides Semantically Accessible: A Semantic Content Generation Tool for PowerPoint Slides

Imran Ihsan
Mohammad Ali Jinnah University, Islamabad, Pakistan
iimranihsan@gmail.com

Abstract

The concept of personalized web based e-learning is getting popular with the passage of time. MOOCs (Massive Open Online Courses) are a trending nowadays offered by platforms like Coursera & Udacity. Currently, Microsoft PowerPoint® is the biggest contributor of personalized web based e-learning. Search engines use natural language processing algorithms on text data for the search and retrieval of learning content. For universities which are now looking forward to providing web based e-learning facilities, such natural language processing algorithms can be difficult to use for such a massive production of PowerPoint based lectures by faculty members. There is an alternate approach which includes semantic based e-learning content publishing tools that can be used instead of PowerPoint for the development of lectures. However, the huge legacy data available has to be converted into a format on which semantic search and retrieval is possible. Conversion or development of e-learning lectures with semantic layer on top can help learners and educators to find the required lecture or even a slide within a lecture with high precision. This study has proposed a tool that can take PowerPoint lecture slides and convert them into a semantically accessible format where search on a slide level within the PowerPoint file is possible. It can also help in developing new courses where lectures of multiple courses can be accessed and integrated to form adaptive courseware. Using the proposed technique, the lecture slides can thus become more interoperable, open and reusable, making MOOC based e-learning more semantically accessible.

Key words:

E-learning, Openness, Interoperability, MOOC (Massive Open Online Course), Semantics, Semantically Meaning Unit (SMU)

Introduction

An educational process describes all the activities taking place between a teacher and a student. The detailed view of an educational process is shown in Figure 1. In order to reduce the cost involved in an educational process, all of the components involved within that process need to be automated. A lot of research has been going on in automating these components (Grossman & Frieder, 1998; de Vries, 1999). Most of the research is in the domain of quizzes, assignments, examination and its grading whereas the research in the area of preparation and delivery of lectures is extremely limited. Teachers prepare their lectures in an isolated manner using various presentational software e.g. Microsoft® PowerPoint. These lectures are mostly intended for a particular class or a subject and are hardly used again in the same context and by the same teacher. As the lecture is specific in nature, it may not be useful for any other teacher or student or anyone in any other context even if shared by the owner. This situation adds more to the dilemma that redundant information is being piled up all over the world rather being reused.

Around the world, there are numerous researches going on to achieve efficient and effective learning and professional development goals. The most popular method adopted is the development of learning objects' repositories for Massive Open Online Courses – MOOCs (2014) such as available at Coursera (2015) & Udacity (2015). These MOOCs provide educators and learners an easy access to learning resources. These MOOCs are developed on the basis of reusability of learning resources but 'reusability' is not defined by these systems. Learning objects can be described as building blocks of a MOOC. But there is no standard definition available for a learning object and is mostly referred as a reusable unit of e-learning (Ihsan I., Uddin Ahmed M., Rehman M., Qadir M. A., Iftikhar N., 2006). In order to make a learning object reusable, especially openly reusable, we need to understand the structure of a course and how its contents are placed. Today's web is dominated by unstructured and semi-structure documents. Intelligent queries on a deeper information level are not possible. If a learning object, available on a MOOC site, can be converted into a structured and semantic nature, a true open re-usability and intelligent querying on a deeper information level can be achieved. This study has proposed a tool for semantic content generation for PowerPoint slides thus making MOOC slides

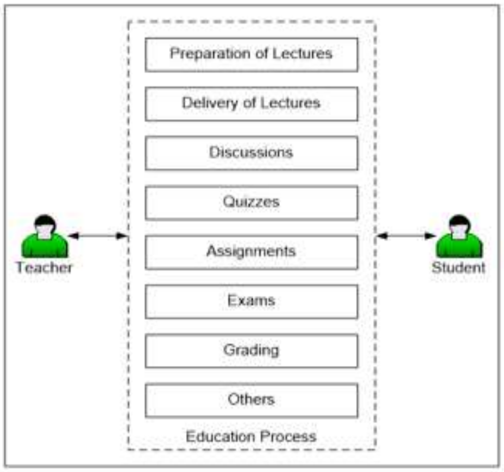

Figure 1. Educational Process

semantically accessible and re-usable.

Literature Review

Microsoft® PowerPoint is the most popular e-learning content development tool. Websites providing MOOCs use Microsoft® PowerPoint as one of the biggest source of content generation and information used by almost every student, teacher & businessman in the whole world. The vision of the Semantic Web is to extend principles of the web from documents to data (Hitzler & Harmelen, 2010). Data need to be accessed using the general web architecture, for example, by using URLs related to one another just as documents (or portions of documents). It requires development of a common framework that allows data to be shared and reused across various applications, enterprises and communities. This framework must process content both automatically by tools and manually revealing possible new relationships among pieces of data.

Semantic Web technologies can be used in a variety of applications, for example, data integration in a seamless way, resource discovery and classification for better domain specific search and retrieval, cataloguing the content and its relationship to the outside world, developing intelligent software agents to facilitate knowledge sharing and exchange, content rating and intelligent querying (Naeve, 2005). To achieve these goals, data should be defined in a semantic manner with its relationships that are more intelligent than a mere hyperlink. Semantic Web allows intelligent relationships between any two resources (Yu, Dong Liu, Dietze, & Domingue, 2011). The definition of those relations allows a better and automatic interchange of data. XML and RDF are the fundamental building blocks of the Semantic Web. On the basis of these building blocks, intelligent queries can be performed for high precision and low recall results by exploiting these relationships using SPARQL (Hartig, 2012).

It is worth consulting the list of Semantic Web case studies and using those cases in order to have a good overview of existing applications. The list is often updated when new application examples come in. The Semantic Web is an extension of the current web and not its replacement. Major application areas (like Health Care and Life Sciences) adopt Semantic Web technologies locally and then spread over the web in general. Personalized web based learning is a trending mode nowadays & MOOCs are also getting popular. Most of the MOOCs use Microsoft® PowerPoint as their source of content and this is the main reason to select Microsoft® PowerPoint as the source of content for the tool development. Another challenge in this tool development was to choose among XML, RDF & OWL as the representation of Microsoft® PowerPoint information. After a careful analysis, XML was selected to be our first step for representing Microsoft® PowerPoint extracted data in a structured format. Microsoft® PowerPoint is unable to run on a variety of platforms such as UNIX & Linux whereas XML can run on almost any platforms and operating systems. Thus the first step towards the implementation part of the tool is to extract data from Microsoft® PowerPoint and store it in XML format. In this paper, the described tool covers this first step providing semantic content generation for Microsoft® PowerPoint slides. In order to do so, a study of Microsoft® PowerPoint slides in detail is required to understand its structure, content and presentation.

There is no software system available that can generate the semantic content from Microsoft® PowerPoint documents. There exist websites that use semantic web technologies such as DBPedia[1] that represents semantic version of WikiPedia[2]. Various ontologies are also available such as FOAF[3] (Friend of a Friend) to describe persons and their social network in a semantic way. Semantic web technologies are growing day by day; however, none of the works in semantic domain has been seen previously that can generate a semantic content from existing Microsoft® PowerPoint documents. There has been an effort to generate semantic based e-learning data in a similar fashion in which Microsoft® PowerPoint slides are made. This tool is described as S-Point – Semantic Point (Ihsan, Rehman, Ahmed, & Qadir, 2008) but still there is need for semantic content generation tool for legacy data.

Methodology

Learning objects are defined as building blocks of any MOOC based system. An LMS (Learning Management System) is an environment where educators can create, store, manage, reuse and deliver learning content using a central data repository. The LMS generally requires content that is based on a model of a learning object. A learning object comprises two elements: the content itself and the metadata used to describe and categorize it. A learning object in practice may be a piece of text or sound, an image, a video clip, a flash animation, a Java applet, a web page or an executable program each with some associated metadata (Jehad, Stefaan & Erik., 2003). The term 'learning object' can refer to any unit of learning that can be used alone or combined into a greater variety of particular programs of instruction for each learner.

In its most basic form, a learning object is made up of two components: metadata and content (Downes & Stephen, 2001). Importantly, each learning object consists of two components, especially to represent the schema of a learning object, but it requires a third component as well that is its presentation. When a lecturer develops a lecture, some presentational tool is used to achieve the desired results. While designing the lecture, bold typeface, underline or changed colour,

[1] http://wiki.dbpedia.org
[2] https://en.wikipedia.org
[3] http://www.foaf-project.org

change in font or size is used to emphasize a particular concept. The lecture designed this way becomes a learning object and its template used to emphasize concepts can increase the semantics involved in that particular learning object. Making design templates a vital part of a learning object make them Semantically Meaning Units - SMUs (Ihsan I., Rehman M., Uddin Ahmed M., Qadir M. A., 2008) and are helpful in making them openly reusable and semantically accessible.

Metadata

Metadata standards associated with learning objects are focused on minimal set of attributes that are required to manage, locate, and evaluate them effectively and efficiently. These standards are extensible in nature and can hold obligatory or optional values. Learning object metadata should also include pedagogical attributes (Salomon, 1984; Kozma, 1991). These attributes can be related to teaching or interaction style, grade level, etc. It is also possible that a single learning object can be associated with more than one set of metadata attributes. Generally, these standards are concerned with security, privacy, and evaluation and are not focused on the fact, how these features are actually implemented (IEEE, 2005).

There exits various metadata standards defined for learning objects. Out of many different ones, a few standards were studied and analyzed. These standards are “Dublin Core[4]”, “E-Government Metadata Standards e-GMS (GMS, 2002)”, “Gateway to Educational Material GEM[5]”, “IEEE Learning Object Metadata (IEEE, 2005)”, “IMS Consortium[6]”, “Machine Readable Cataloguing MARC[7]”, “Metadata Encoding & Transmission Standards METS[8]”, “Metadata Object Description Schema MODS[9]”, “Online Information Exchange ONIX[10]” and “SCORM (ADL-I, 2001; ADL-II, 2001; ADL-III, 2001)”.

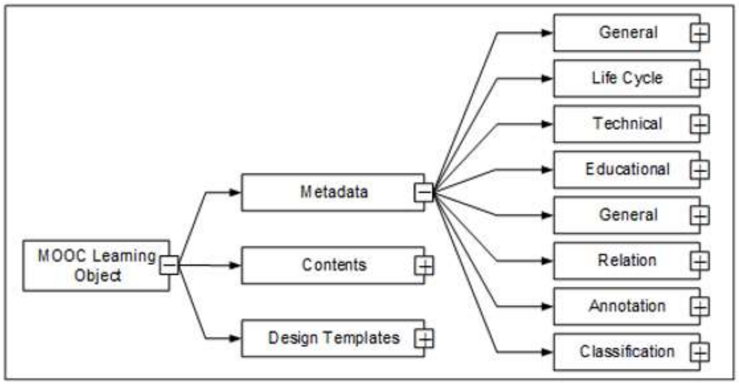

Figure 2. Metadata Schema (IEEE Compliant)

Content

Content in a learning object such as MOOC slides is a collection of materials to teach a single concept in a particular domain. Content material used in online courses/tutorials can be

[4] http://www.dublincore.org
[5] http://www.geminfo.org
[6] http://www.imsglobal.com
[7] http://www.loc.gov/marc/marc.html
[8] http://www.loc.gov/standards/mets
[9] http://www.loc.gov/standards/mods
[10] http://www.editeur.org/onix.html

an explanation, instructions, definition, image, animation, programs, quiz, etc. Tools used to author an SMU must follow the schema that can be defined using mark-up languages such as "SGML (Standard Generalized Mark-up Language)" or "XML (eXtensible Mark-up Language)" or a language that can be converted to XML using external XML converters. Once authored, content of SMU can be stored in a content repository in SGML/XML files that provide rich capabilities to search, retrieve, revise, and control on versioning etc. The SGML/XML file is also easy to access, interoperate and exchange between various repositories.

Mark-up languages used to define the content schema of a learning object provide tags that can define learning outcomes along with associated resources. Identifying content and what it is meant for, under lays foundation to intelligently query the content to be fully or partially reused. For example, a lecturer can post an intelligent query on a learning object repository for an efficient search and retrieval, afterwards making it personalized in the way he/she wants that content to be displayed. In order to achieve this goal, the following content schema based on Microsoft® PowerPoint was adopted where content is distributed among slides or sections and is shown in the

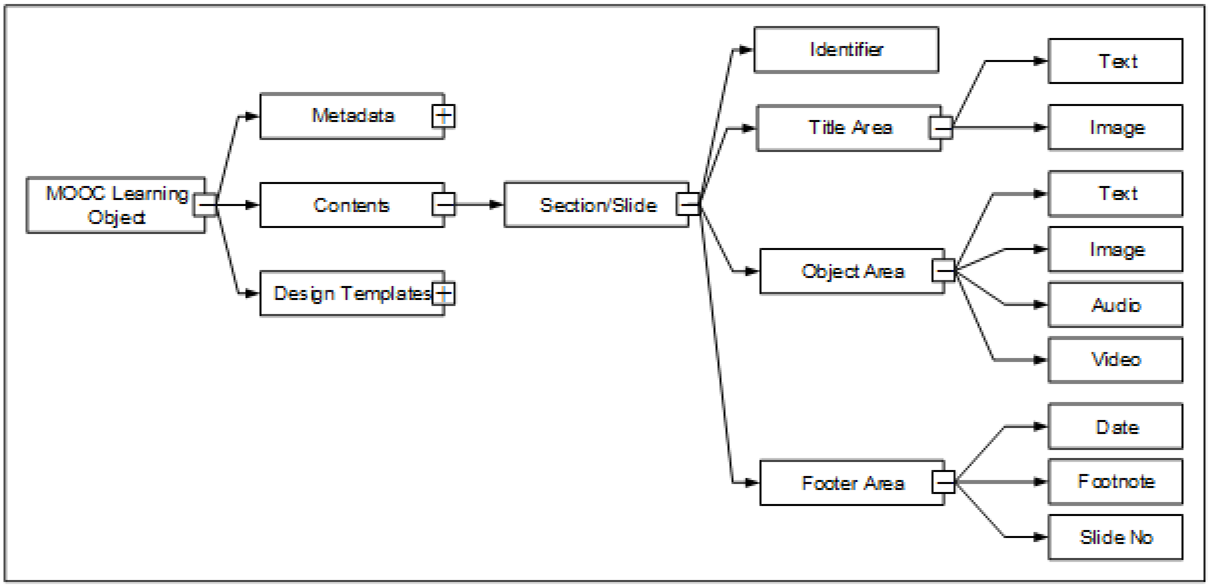

Figure 3. SMU Schema (Slide/Section Breakup)

Design Template

Styles used in a particular learning object that include type or size of bullets and fonts, backgrounds and colour schemes are knows as design templates. In a single learning object, two distinct types of design templates, having two different levels of semantic importance, can be used. These two types of design templates can be termed as "Master Template" and "Secondary

Template". Apart from these two templates, color scheme can also be associated with design templates and defined in detail as:

- **Color Scheme:** A color scheme is a set of colors used while designing a learning object such as colors used as background, foreground, content, heading and hyperlinks etc.
- **Master Template:** Master template is the information about the layout of the content such as font faces, backgrounds and color schemes. Any formatting change in the master slide will updated automatically all slides that are based on the master template. Master template is a fast and effective way to enhance the look and feel of the learning object.
- **Secondary Template:** Being similar to master template, secondary templates have an important difference. The difference is that it is not applied across slides. It has local scope and only deals with any format change that is applied within a particular slide such as to emphasize a particular word or a sentence. Such emphasis can be created by simply modifying the font, size, or colour or by converting the word to bold, italic or underlining. Such information can play a vital role while preserving the semantics integrated with a learning object.

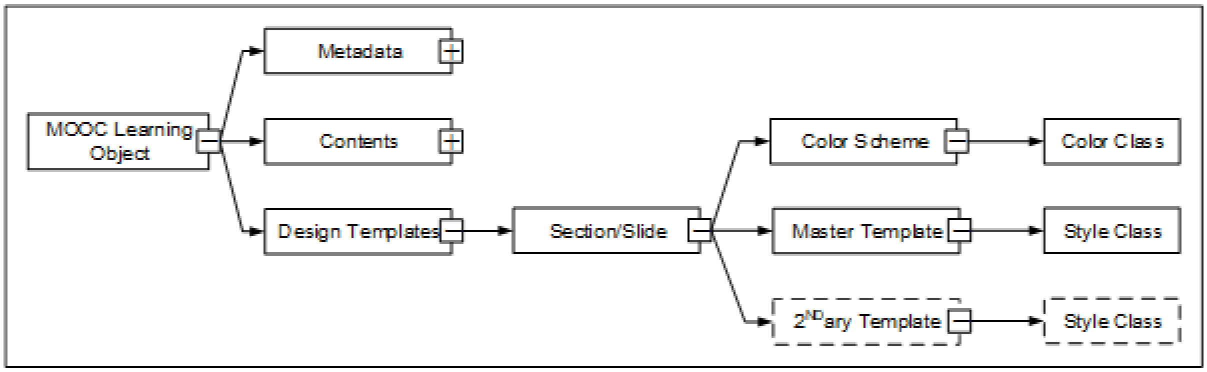

Figure 4. Design Template Schema

After a critical review, the definitions of a learning object have to be restructured to make it openly reusable. An SMU comprises three sections: metadata, content and design template. Contents and metadata of an SMU are stored in XML format whereas the design template is stored in XSL format making it interoperable and personalizable.

Discussion and Implementation

For the development of the proposed tool, Microsoft.Office.Interop.PowerPoint library open PowerPoint Slides have been used. Stop words are removed using dictionaries (Hui Xiong. Pandey, G. Steinbach, M. Kumar, V., 2006). A counter is maintained for every word that is extracted from Microsoft® PowerPoint documents for calculating TF/IDF scores and weights. Using the DTD

(Document Type Definition) for generating semantic XML document, implementation for XML Transformation phase is performed. During this extraction process, design template used in Microsoft® PowerPoint documents are also extracted and stored in an XSLT document attached with the extracted XML documents. Afterwards all of these extracted XML documents can be indexed for efficient search and retrieval. There exist various indexing techniques for XML documents such as 2-Index and 4-Index schemes. However for our tool, we have adopted Semantic Indexing technique (Ihsan I., Qadir M. A., 2015) that uses a modified version of 4-Index scheme. Next phase is visualizing the results of our XPATH queries for a better understanding of generated semantic content. DevExpress, Infragistics and Telerik are used for visualizing information for the results of queries.

Figure 5 describes the system flow diagram for two processes of the developed tool. One is the generation of semantic XML store along with its metadata and design templates for Microsoft® PowerPoint based MOOC slides and other is semantic searching and visualization of results for intelligent querying. In the first process, contents are extracted on slides. After removing stop words, TF/IDF of occurring terms are calculated. Afterwards using the specified DTD or XML schema, semantic XML is generated for the fed MOOC document. This newly formed XML file has connection with its corresponding design template XSLT and metadata XML file. A sample XML file generated is shown in Figure 6. Afterwards, semantics indexes are generated on extracted XML files for efficient search and retrieval process.

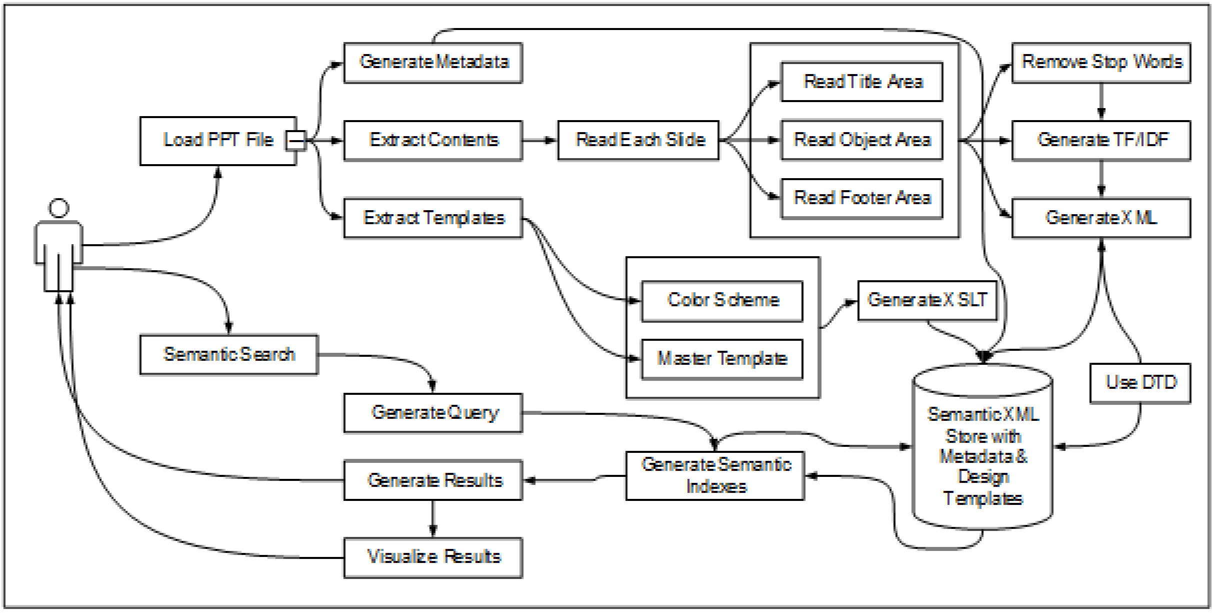

Figure 5. System Flow Diagram for Semantic Content Generation and Semantic Querying

```
<SMU id=”U-113311”>
        <Slide id="1" >
                <TitleArea>
                        <text> Computer Organization and Architecture 6th Edition </text>
                </TitleArea>
                <ObjectArea>
                        <text> Introduction </text>
                        <text> Chapter 1 </text>
                </ObjectArea>
                <FooterArea>
                        <text> 12-AUG-2005 04:25:35 </text>
                </FooterArea>
        </Slide>
        <Slide id="2" >
                <TitleArea>
                        <text> Architecture &amp; Organization 1 </text>
                </TitleArea>
                <ObjectArea>
                        <text> Organization is how features are implemented</text>
                        <text> Architecture is those attributes visible to the programmer </text>
                </ObjectArea>
                <FooterArea>
                        <text> 12-AUG-2005 04:27:05 </text>
                </FooterArea>
        </Slide>
        <Slide id="3" >
                <TitleArea>
                        <text> Architecture &amp; Organization 2 </text>
                </TitleArea>
                <ObjectArea>
                        <text> Organization differs between different versions </text>
                        <text> This gives code compatibility - At least backwards </text>
                </ObjectArea>
                <FooterArea>
                        <text> 12-AUG-2005 04:28:21 </text>
                </FooterArea>
        </Slide>
        <Slide id="4" >
                <TitleArea>
                        <text> Structure &amp; Function </text>
                </TitleArea>
                <ObjectArea>
                        <img src="\\Hiperdcom1\ureka\Objects\Images\mm_Image1.gif" />
                        <text> Structure is the way in which components relate to each other </text>
                </ObjectArea>
                <FooterArea>
                        <text> 12-AUG-2005 04:29:38 </text>
                </FooterArea>
        </Slide>
</SMU>
```

Figure 6. Sample XML File Generated by Tool

The second part is provision of semantic search on the generated content. Formulation of query, generation and visualization of results are the key components. The interface provides a deeper level search on slides and even on the title, object and footer area. Some of the sample queries are listed below:

- Search by keyword in a particular slide.
- Search by keyword that appears in the title of a slide.
- Search by keyword that appears in an object area of a slide.
- Search by keyword that appears in a footer of a slide.

Conclusion

MOOC (Massive Open Online Course) providing websites such as Coursera & Udacity use Microsoft PowerPoint® as the biggest contributor for content generation for personalized web based e-learning. A huge number of PowerPoint based lectures has been produced by faculty members of different universities across the globe. Semantic based e-learning content publishing tool that can be used instead of PowerPoint for the development of lectures is a requirement of the day. However, huge legacy data available needs to be converted into a format on which semantic search and retrieval is possible. This study has proposed a tool that converts PowerPoint lecture slides into a semantically accessible format where search on a slide level within the PowerPoint file is possible. Conversion of e-learning lectures with semantic layer on top can help learners and educators to find the required lecture or even a slide within a lecture with high precision. By virtue of this tool, the lecture slides can thus become more interoperable, open and reusable making MOOC based e-learning more semantically accessible.

References

ADL-I. (2001). Advanced Distributed Learning (ADL) Initiative. Sharable Courseware Object Reference Model (SCORM). The SCORM Overview, http://www.adlnet.org/Scrom/

ADL-II. (2001). Advanced Distributed Learning (ADL) Initiative. (2001) Sharable Courseware Object Reference Model (SCORM). The SCORM Content Aggregation Model, http://www.adlnet.org/Scrom/

ADL-III. (2001). Advanced Distributed Learning (ADL) Initiative. (2001). Sharable Courseware Object Reference Model (SCORM). The SCORM Run-Time Environment, http://www.adlnet.org/Scrom/

Coursera. (2015). Free Online Course from Top Universities: https://www.coursera.org/.

De Vries, A. P. (1999). *Content and multimedia database management systems*. University of Twente, Centre for Telematics and Information Technology (CTIT).

Downes, S. (2001). Learning objects: resources for distance education worldwide. *The International Review of Research in Open and Distributed Learning*, *2*(1).

GMS. (2002). e-Government Metadata Standard (e-GMS). Retrieved from http://www.govtalk.gov.uk/documents/e-Government_Metadata_Standard_v1.pdf

Grossman, D. A., & Frieder, O. (2012). *Information retrieval: Algorithms and heuristics* (Vol. 15). Springer Science & Business Media.

Hartig, O. (2012). SPARQL for a Web of Linked Data: Semantics and computability. *The Semantic Web: Research and Applications*, 8-23.

Hitzler, P., & van Harmelen, F. (2010). A reasonable semantic web. Semantic Web – Interoperability, Usability, Applicability 0 (0) 1. IOS Press.

IEEE. (2005). IEEE Learning Technology Standards Committee ©2005, http://ltsc.ieee.org/wg12, *IEEE Standards for Learning Object Metadata* (1484.12.1)

Ihsan, I., Rehman, M., Ahmed, M. U., & Qadir, M. A. (2008). S-Point a semantic based e-learning content development tool. *Journal of Applied Sciences*, *8*(1), 127-133.

Ihsan, I., Ahmed, M., Qadir, M. A., & Iftikhar, N. (2006, March). Semantically Meaningful Unit-SMU; An Openly Reusable Learning Object for UREKA Learning-Object Taxonomy & Repository Architecture-ULTRA. In *Computer Systems and Applications, 2006. IEEE International Conference on.* (pp. 1011-1018). IEEE.

Ihsan, I., & Qadir, M. A (2015). Querying Semantically Related Items using modified 4-Index Scheme for XML Documents. In *Frontiers of Information Technology (FIT), 2015 13th International Conference on* (pp. 47-52). IEEE.

Kozma, R. B. (1991). Learning with media. *Review of educational research*, *61*(2), 179-211.

MOOC. (2014). MOOC Trends and Implementation at Community Colleges, Hanover Research, September 2014.

Naeve, A. (2005). The human Semantic Web shifting from knowledge push to knowledge pull. *International Journal on Semantic Web and Information Systems (IJSWIS)*, *1*(3), 1-30.

Salomon, G. (1984). Television is" easy" and print is" tough": The differential investment of mental effort in learning as a function of perceptions and attributions. *Journal of educational psychology*, *76*(4), 647.

Udacity. (2015). Online Courses and Nondegree Programs: https://www.udacity.com/.

Xiong, H., Pandey, G., Steinbach, M., & Kumar, V. (2006). Enhancing data analysis with noise removal. *IEEE Transactions on Knowledge and Data Engineering*, *18*(3), 304-319.

Yu, H. Q., Liu, D., Dietze, S., & Domingue, J. (2011, October). Developing RDF-based Web services for supporting runtime matchmaking and invocation. In *Next Generation Web Services Practices (NWeSP), 2011 7th International Conference on* (pp. 392-397). IEEE.