IRRODL
THE INTERNATIONAL REVIEW OF RESEARCH IN OPEN AND DISTANCE LEARNING

OERScout Technology Framework: A Novel Approach to Open Educational Resources Search

Ishan Sudeera Abeywardena[1], Chee Seng Chan[2,] and Choy Yoong Tham[1]
[1]Wawasan Open University, Malaysia, [2]University of Malaya, Malaysia

Abstract

The open educational resources (OER) movement has gained momentum in the past few years. With this new drive towards making knowledge open and accessible, a large number of OER repositories have been established and made available online throughout the world. However, the inability of existing search engines such as Google, Yahoo!, and Bing to effectively search for useful OER which are of acceptable academic standard for teaching purposes is a major factor contributing to the slow uptake of the entire movement. As a major step towards solving this issue, this paper proposes *OERScout*, a technology framework based on text mining solutions. The objectives of our work are to (i) develop a technology framework which will parametrically measure the usefulness of an OER for a particular academic purpose based on the openness, accessibility, and relevance attributes; and (ii) provide academics with a mechanism to locate OER which are of an acceptable academic standard. From our user tests, we have identified that OERScout is a sound solution for effectively zeroing in on OER which can be readily used for teaching and learning purposes.

Keywords: OERScout; open educational resources; OER; OER search; desirability of OER; OER metadata

Introduction

Open educational resources (OER) have the potential to become a major source of freely reusable teaching and learning resources, especially in higher education (HE). The UNESCO Paris OER Declaration (2012) defines OER as

> teaching, learning and research materials in any medium, digital or otherwise, that reside in the public domain or have been released under an open license that permits no-cost access, use, adaptation and redistribution by others with no or limited restrictions. Open licensing is built within the existing framework of intellectual property rights as defined by relevant international conventions and respects the authorship of the work.

Claims have also been made by Caswell, Henson, Jenson, and Wiley (2008) that the move towards OER can significantly reduce the costs of learning. Thus, OER has the potential to broaden access and provide equity in education. This is especially important for countries in the Global South.

The recently concluded "OER Asia" study (Dhanarajan & Abeywardena, 2013) surveyed 420 junior to senior academics from public and private HE institutions in nine countries representing a majority of sub-regions in Asia. Based on this study, Abeywardena, Dhanarajan, and Chan (2012) state that 57.4% of the academics feel the lack of ability to locate specific and relevant resources using existing search engines to be a serious inhibitor of the use of OER. It is further pointed out that, in general, academics search and locate OER which are freely available on the Internet. However, many of these resources have not been subjected to academic quality assurance (QA) procedures imposed by degree accrediting organisations such as the Malaysian Qualifications Agency (MQA)[1]. In contrast, institutional and peer-reviewed OER repositories maintain an acceptable level of academic quality of material. These materials can be readily used and reused for teaching purposes. Furthermore, these repositories are equipped with native search mechanisms which facilitate the searching of relevant OER for a particular teaching need. Unfortunately, according to the study, only 43.2% of the academics use native search facilities of OER repositories. On the other hand, generic search engines such as Google, Yahoo!, and Bing are found to be used by 96.9% of the academics for OER search.

From this comparison, it is apparent that many academics depend on generic search mechanisms to locate the required OER for their teaching purposes. As a result, the inability of these generic mechanisms to locate useful OER for a particular teaching need, as will be discussed, has in fact become an inhibitor to the wider adoption of OER for teaching in Asia. In order to overcome this barrier, a centralised search mechanism

[1] http://www.mqa.gov.my

which can locate academically useful OER needs to be introduced. As a major step towards solving this issue, in this paper, we propose *OERScout*, a technology framework based on text mining solutions. The objectives of our work are to (i) develop a technology framework which will parametrically measure the usefulness of an OER for a particular academic purpose based on the openness, accessibility, and relevance attributes; and (ii) provide a search mechanism to effectively zero in on OER which are of an acceptable academic standard.

The rest of the paper is structured as follows: The Literature Review section gives an overview of the current solutions available to search for OER; the Methodology section details the proposed method; the Results and Discussion sections provide the expert user test results and discussion respectively; and the Conclusion concludes the work and discusses some future work. Overall, the paper provides a holistic view of the complete project.

Literature Review

Most current OER initiatives are based on established web technology platforms and have accumulated large volumes of quality resources. However, one limitation inhibiting the wider adoption of OER is the current inability to effectively search for academically useful OER from a diversity of sources (Yergler, 2010). This limitation of locating “fit-for-purpose” (Calverley & Shephard, 2003) resources is further heightened by the disconnectedness of the vast array of OER repositories currently available online. As a result, West and Victor (2011) argue that there is no single search engine which is able to locate resources from all the OER repositories. Furthermore, according to Dichev and Dicheva (2012), one of the major barriers to the use and reuse of OER is the difficulty in finding quality OER matching a specific context as it takes an amount of time comparable with creating one’s own materials. Unwin (2005) argues that the problem with open content is not the lack of available resources on the Internet but the inability to effectively locate suitable resources for academic use. The UNESCO Paris OER Declaration (2012) states the need for more research in this area to “encourage the development of user-friendly tools to locate and retrieve OER that are specific and relevant to particular needs”. Thus, the necessity for a system which could effectively search the numerous OER repositories with the aim of locating usable materials has taken centre stage.

The most common method of searching for OER is to use generic search engines such as Google, Yahoo!, or Bing. Even though this method is the most commonly used, it is not the most effective as discussed by Pirkkalainen and Pawlowski (2010, p. 2) who argue that “searching this way might be a long and painful process as most of the results are not usable for educational purposes”.

Alternative methods for OER search can be broadly categorised into federated search and semantic search. Federated search is achieved either by searching across different repositories at runtime or by periodically harvesting metadata for offline searching.

Recent examples of federated search include (i) BRENHET[2] proposed by De la Prieta, Gil, Rodríguez, and Martín (2011), which is a multi agent system (MAS) which facilitates federated search between learning object repositories (LOR); (ii) OpeScout (Ha, et al., 2011), which copies metadata from existing repositories to create an index of resources accessible through a faceted search approach; (iii) Global Learning Object Brokered Exchange (GLOBE), which acts as a central repository of IEEE LOM educational metadata harvested from various member institutional repositories (Yamada, 2013); and (iv) Pearson's Project Blue Sky (Kolowich, 2012), which is a custom search engine specifically concentrating on searching for OER with an academic focus. Semantic search is derived from semantic web technologies where people are considered as producers or consumers and machines as enablers. Some of the recent semantic search initiatives are (i) the OER-CC ontology which describes various accessibility levels (Piedra, et al., 2010, 2011); (ii) the "Assistant" prototype proposed by Casali, Deco, Romano, and Tomé (2013), which helps users with respect to loading metadata through automation; (iii) the "Folksemantic" project which is a hybrid search system combining OCW Finder and OER Recommender (Shelton, Duffin, Wang, & Ball, 2010); and (iv) "Agrotags", a project concentrating on tagging resources in the agriculture domain (Balaji, et al., 2010). However, despite showing initial promise, only a handful of these solutions have proceeded beyond the prototype stage. Out of these, the ones which have become global players are mainly commercial ventures or global federations backed by philanthropic funding. One reason underpinning the relatively low success rate of these initiatives can be attributed to the current lack of a search methodology which takes into consideration the level of openness, the level of access, and the relevance of a resource for one's needs (Abeywardena, Raviraja, & Tham, 2012). Though one might argue that popular search engines provide advanced facilities to define various filter criteria which would refine the searches, these search engines however are not tailored to effectively locate OER material which are the most useful for a particular academic purpose. As such, OER consumers will need to resort to frequenting OER repositories to search for the resources they are after. Pirkkalainen and Pawlowski (2010) argue that native search mechanisms of repositories are relatively better at locating resources with increased usefulness. However, the problem is which repositories to choose within the large and constantly expanding global pool. Furthermore, users would be spending an extended amount of time on these repositories conducting multiple searches (Figure 1), making it an inefficient method for locating resources.

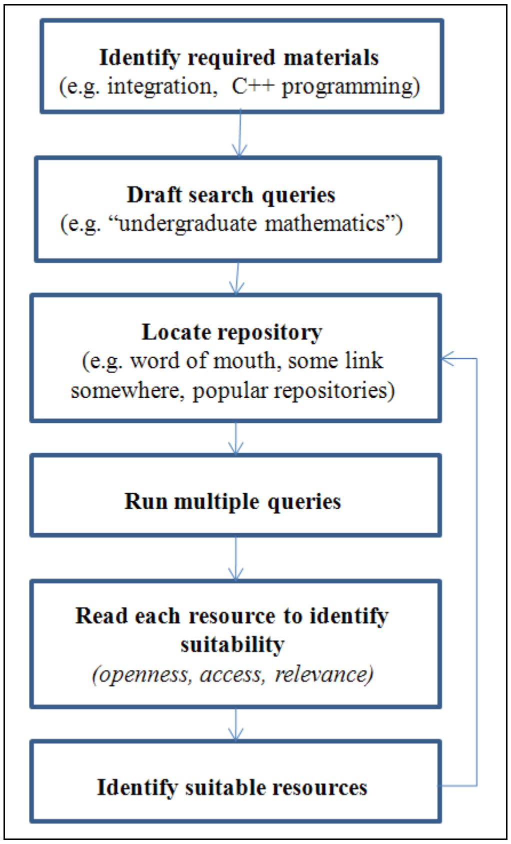

Identify required materials
(e.g. integration, C++ programming)
Draft search queries
(e.g. "undergraduate mathematics")
Locate repository
(e.g. word of mouth, some link somewhere, popular repositories)
Run multiple queries
Read each resource to identify suitability
(openness, access, relevance)
Identify suitable resources

Figure 1. The flow of activities in searching for suitable OER on heterogenous repositories based on personal experience (Abeywardena, 2013). These activities will need to be repeated on multiple repositories until the required resources are located.

Another factor inhibiting effective OER search is the heterogeneity of OER repositories. Within the context of parametric web based search, this disparity can be broadly attributed to (i) the lack of a single metadata standard; (ii) the lack of a centralised search mechanism; and (iii) the inability to indicate the usefulness of an OER returned as a search result.

Metadata provides a standard and efficient way to conveniently characterize educational resource properties (Anido, Fernández, Caeiro, Santos, Rodriguez, & Llamas, 2002). The majority of existing search methodologies, including mainstream search engines, such as Google, work on the concept of metadata for locating educational resources. However, it can be argued that the annotation of resources with metadata cannot be made 100% accurate or uniform if done by the creator(s) of the resource (Barton, Currier, & Hey, 2003; Tello, 2007; Devedzic, Jovanovic, & Gasevic, 2007; Brooks & McCalla, 2006; Cechinel, Sánchez-Alonso, & Sicilia, 2009). Therefore the use of human annotated metadata in performing objective searches becomes subjective and inaccurate. A possible way to overcome this inaccuracy and to ensure uniformity of metadata is to utilise a computer based methodology which would consider the content, domain, and locality of the resources, among others, for autonomously annotating metadata.

As a solution to these issues , this paper proposes the OERScout technology framework which accurately clusters text based OER by building a *keyword-document matrix* (KDM) using autonomously mined domain specific keywords. The advantage of our work is, using the KDM, the system generates ranked lists of relevant OER from heterogenous repositories to suit a given search query. The contribution of our work is two-fold: Firstly, we propose a technology framework for locating OER, which are useful for academic needs. In this regard, the advantage of OERScout over existing OER search methodologies is the incorporation of the *desirability* framework (Abeywardena, Raviraja, & Tham, 2012) in parametrically measuring the usefulness of an OER with respect to openness, accessibility, and relevance. Secondly, we introduce a novel methodology which allows academics to effectively zero in on OER which can be readily used for their teaching and learning purposes . We strongly believe that the OERScout system will broaden access and provide equity in education, particularly for countries in the Global South such as India, Pakistan, Afghanistan, Myanmar, and Sri Lanka to name a few.

Methodology

As discussed in the Literature Review, mainstream search engines, federated search, and semantic search are the key OER search methodologies adopted at present. However, all of these methodologies depend on human annotated metadata for approximating the usefulness of a resource for a particular need. Given the limitations of human annotated metadata with respect to accurately and uniformly describing resources, the accuracy of search becomes a function of the content creators' ability to accurately annotate resources. Therefore, the OERScout system uses text mining techniques to annotate resources using autonomously mined keywords.

The Algorithm

The OERScout text mining algorithm is designed to “read” text based OER documents and “learn” which academic domain(s) and sub-domain(s) they belong to. To achieve this, a *bag-of-words* approach is used due to its effectiveness with unstructured data (Feldman & Sanger, 2006). The algorithm extracts all the individual words from a particular document by removing noise such as formatting and punctuation to form the *corpus*. The corpus is then *tokenised* into the *list of terms* using the *stop words* found in the Onix Text Retrieval Toolkit[2] as shown in Figure 2.

[2] lextek.com/manuals/onix/stopwords1.html

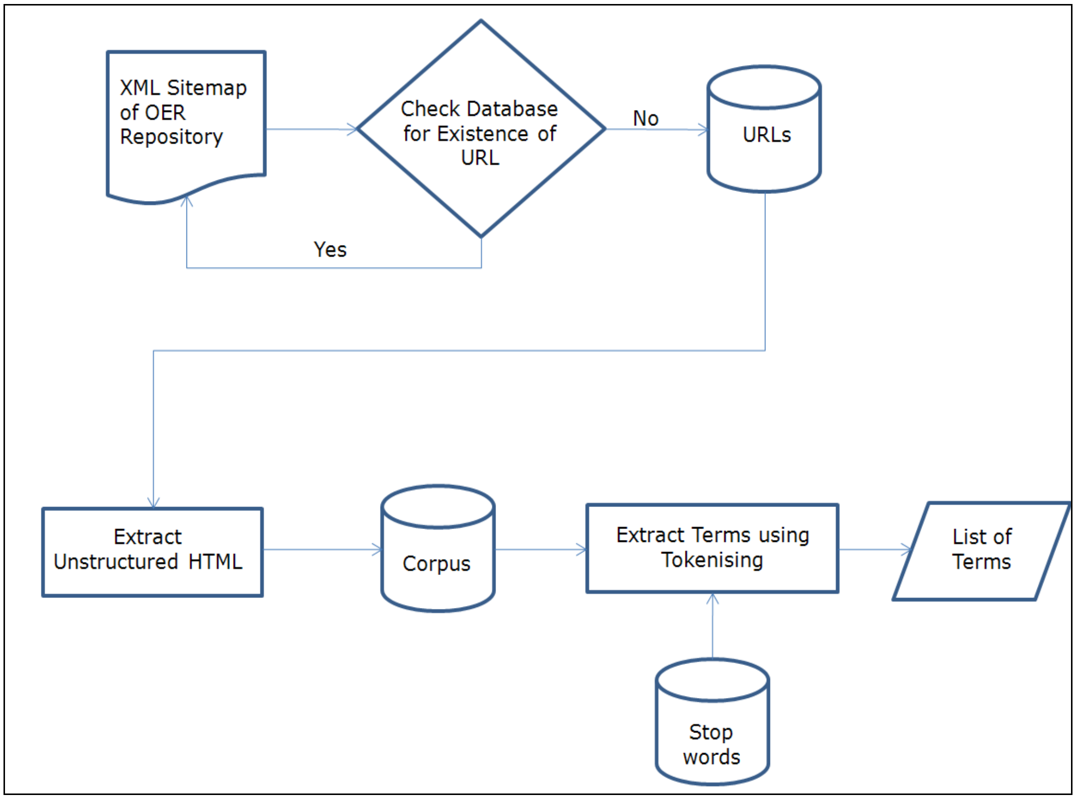

XML Sitemap of OER Repository
Check Database for Existence of URL
No
URLs
Yes
Extract Unstructured HTML
Corpus
Extract Terms using Tokenising
List of Terms
Stop words

Figure 2. The list of terms is created by tokenising the corpus using the stop words found in the Onix Text Retrieval Toolkit.

The extraction of the content describing terms from the list of terms for the formation of the term document matrix (TDM) is done using the term frequency–inverse document frequency (TF-IDF) weighting scheme. The weight of each term (TF-IDF) is calculated using Equation 1 (Feldman & Sanger, 2006):

$$(TF\text{-}IDF)_t = TF_t \; x \; IDF_t \quad (1)$$

TF_t denotes the frequency of a term *t* in a single document. IDF_t denotes the frequency of a term *t* in all the documents in the collection [IDF_t = Log (N/TF_t)] where *N* is the number of documents in the collection. The probability of a term *t* being able to accurately describe the content of a particular document as a keyword decreases with the number of times it occurs in other related and non-related documents. For example the term “introduction” would be found in many OER documents which discuss a variety of subject matter. As such the TF-IDF of the term “introduction” would be low compared to terms such as “operating systems” or “statistical methods” which are more likely to be keywords. Due to the large number of documents available in OER repositories and their document lengths, the TF value of certain words will be quite high. As a result, there will be a considerable amount of noise being picked up while identifying the keywords. However, the large number of documents will also increase the IDF value of words reducing the TF-IDF value which results in the reduction of noise picked up as keywords. As such, the TF-IDF weighting scheme allows the system to refine its set of identified keywords at each iteration. Therefore, the TF-IDF weighting scheme is found to be suitable for extracting keywords from the OER documents.

Keyword-Document Matrix (KDM)

The keyword-document matrix (KDM), a subset of the TDM, is created for the OERScout system by matching the autonomously identified keywords against the documents as shown in Figure 3.

	Keyword$_1$	Keyword$_2$	…………	Keyword$_n$
Document$_1$	√			√
Document$_2$	√		√	
…………..		√		
Document$_n$		√		√

Figure 3. The keyword-document matrix (KDM*)*, a subset of the TDM, is created for the OERScout system by matching the autonomously identified keywords against the documents.

The formation of the KDM (Figure 4) is done by (i) normalising the TF-IDF values for the terms in the TDM; and (ii) applying the Pareto principle (80:20) empirically for feature selection where the top 20% of the TF-IDF values are considered to be keywords describing 80% of the document.

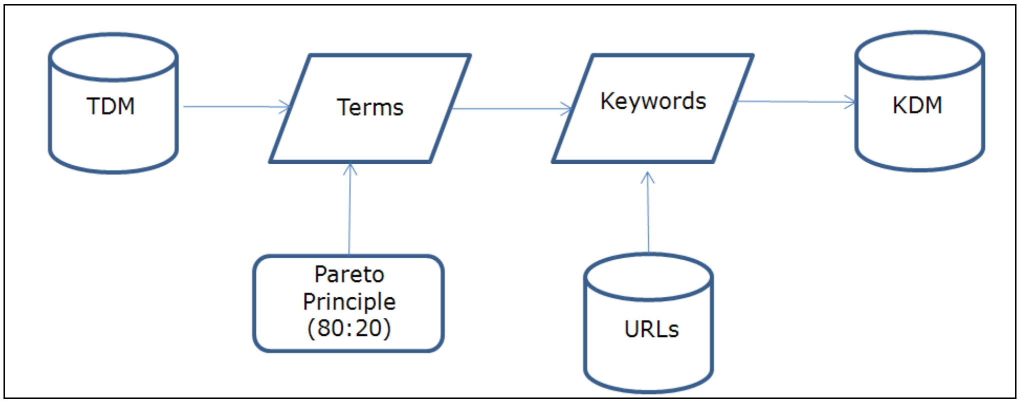


Figure 4. Formation of the KDM by normalizing the TF-IDF values of the terms in the TDM and applying the Pareto principle empirically for feature selection.

The OERScout user interface and algorithm are implemented using the Microsoft Visual Basic.NET 2010 (VB.NET, 2010) programming language. The corpus, list of terms, TDM, and KDM are implemented using the MySQL database platform. The OER resources are fed into the system using sitemaps based on extensible markup language (XML) which contain the uniform resource locators (URLs) of the resources. When implemented, new repositories will be identified for crawling based on referrals by end users. The sitemaps created by the crawlers will be input into the system to be processed. The server tools will continuously run at the server processing new documents and re-visiting processed documents to ensure accuracy. The KDM is accessed by end users through the OERScout Microsoft Windows based client interface. The deployment architecture of OERScout is shown in Figure 5.

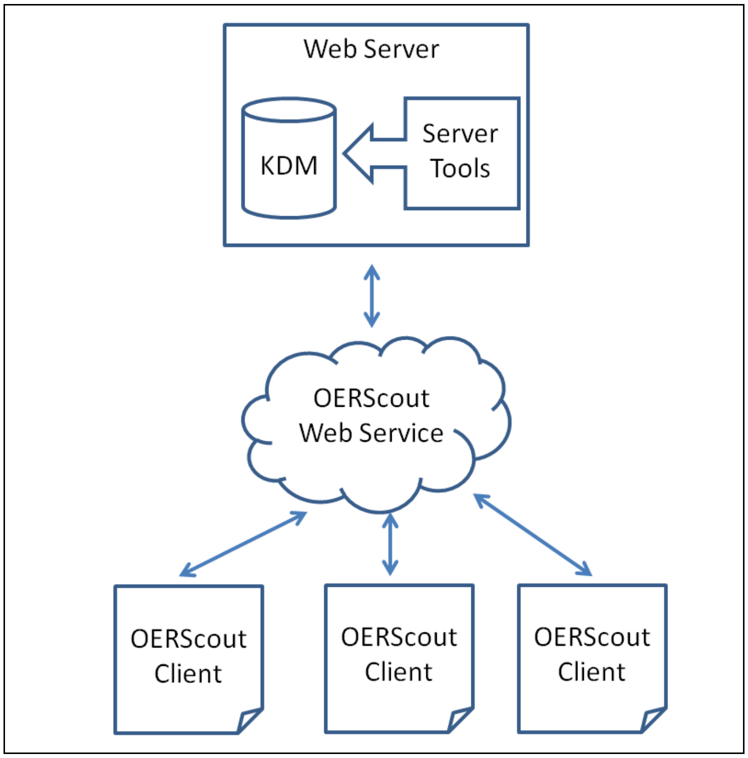


Figure 5. OERScout deployment architecture which has a web server hosting the KDM, a web service for accessing the KDM, and a Microsoft Windows based client interface.

Calculation of the Desirability

The desirability of OER (Abeywardena, Raviraja, & Tham, 2012) is a parametric measure of the usefulness of an OER for a particular academic need based on (i) level of openness, the permission to use and reuse the resource; (ii) level of access, the technical keys required to unlock the resource; and (iii) relevance, the level of match between the resource and the needs of the user. The desirability is calculated using Equation 2 and is denoted as the *D-index*, which is a value between 0 and 1. The higher the D-index, the more desirable an OER is for a particular academic need. The value 256 is used to normalise the access, openness, and relevance parameters. It is the product of the values

16, 4, and 4, respectively, which correspond to the highest value assigned to each parameter.

D-index = (level of access x level of openness x relevance) / 256 (2)

The desirability of each document in the KDM is calculated using the openness, accessibility, and relevance of the document. As suggested by Abeywardena, Raviraja, and Tham (2012), the openness of the document is calculated using the Creative Commons (CC) license of the document (Table 1). A maximum value of 4 is assigned to the most open CC license with respect to permission to reuse, redistribute, revise, and remix (Hilton, Wiley, Stein, & Johnson, 2010). A value of 2 is assigned to the least open license as the CC license starts at the redistribute level.

The accessibility is calculated by extracting the file type of each document as shown in Table 2. This version of OERScout is built only to index documents of type PDF (.pdf), webpage (static and dynamic web pages which include .htm, .html, .jsp, . asp, .aspx, .php etc.), TEXT (.txt), and MS Word (.doc, .docx) as these file types were found to be the most commonly used for text based OER (Wiley, 2006). The value for each file type was calculated based on the ALMS analysis proposed by Hilton, Wiley, Stein, and Johnson (2010) which builds on the parameters (i) **A**ccess to editing tools; (ii) **L**evel of expertise required to revise or remix; (iii) ability to **M**eaningfully edit; and (iv) **S**ource-file access.

The relevance of a document to a particular search query is calculated using the TF-IDF values of the keywords which are stored as additional parameters of the KDM. As shown in Table 3, building on the work by Vaughan (2004), the top 10 search results based on the TF-IDF value are assigned a maximum value of 4, the top 11-20 search results are assigned a value of 3, the top 21-30 results are assigned a value of 2, and search results below 30 are assigned a minimum value of 1.

The D-index of each document is then calculated using Equation 2 and the desirable resources for a particular need are presented to the user in descending order.

Table 1

Openness Based on the CC License (Abeywardena, Raviraja, & Tham, 2012).

Permission	Creative Commons (CC) licence	Value
Reuse	None	1
Redistribute	Attribution-NonCommercial-NoDerivs (CC BY-NC-ND) Attribution-NoDerivs (CC BY-ND)	2
Revise	Attribution-NonCommercial-ShareAlike (CC BY-NC-SA) Attribution-ShareAlike (CC BY-SA)	3
Remix	Attribution-NonCommercial (CC BY-NC) Attribution (CC BY)	4

Table 2

Accessibility Based on the File Type (Abeywardena, Raviraja, & Tham, 2012)

File type	Access (ALMS)				
	A	L	M	S	Value
PDF	Low	High	No	No	1
MS Word	Low	Low	Yes	Yes	8
Webpage	High	Low	Yes	Yes	16
TEXT	High	Low	Yes	Yes	16

Table 3

The Level of Relevance Based on Search Rank

Search rank	Value
Below the top 30 ranks of the search results	1
Within the top 21-30 ranks of the search results	2
Within the top 11-20 ranks of the search results	3
Within the top 10 ranks of the search results	4

One of the key observations made during the calculation of the desirability is that certain OER repositories do not specify or use the CC licensing scheme as the standard for defining the intellectual property rights. However, these repositories explicitly or implicitly mention that the resources are freely and openly available for use and reuse. Due to the inability of the current OERScout system to determine the level of openness of these resources, a value of zero was assigned to any resource which did not implement the CC licensing scheme. As such the desirability of these resources was reduced to zero due to the ambiguity in the license definition. This feature spares the user from legal complications attached to the use and reuse of resources which do not clearly indicate the permissions granted.

Results

The application of the system in a real world scenario was done using the *Directory of Open Educational Resources* (DOER)[3] of the Commonwealth of Learning (COL). DOER is a fledgling *portal OER repository* (McGreal, 2010) which provides an easily navigable central catalogue of OER distributed globally. At present, the OER available through DOER are manually classified into 20 main categories and 1,158 sub-categories. However, despite covering most of the major subject categories, this particular ontology would need to expand by a large degree due to the variety of OER available in an array of subject areas. This expansion, in turn, becomes a tedious and laborious task which needs to be accomplished manually on an ongoing basis. As a possible solution to this issue, a mechanism was needed for autonomously identifying the subject area(s) covered in a particular OER, in the form of keywords, in order for it to be accurately catalogued.

Given this requirement, DOER was used as the training dataset for OERScout. In addition to the resources categorised in DOER, 1,536 resources from the Rice University's Connexions[4] repository were also included in the training dataset due to (i) the large number of OER materials available; and (ii) the relatively high popularity and usage rates. An XML sitemap containing a total of 1,999 URLs belonging to the domains of arts, business, humanities, mathematics, and statistics, science and technology, and social sciences was created as the initial input. The system was run with the initial input and was allowed to autonomously create the KDM. This training process was critical to the functioning of the algorithm as it had to learn a large number of academic domains and sub-domains before being able to accurately cluster resources according to the domain.

On average, each document required 15-90 minutes to be downloaded, read, and learnt by the system depending on the size and file type. The system took approximately five days to process all the documents in the training dataset. Although the training process required a considerable amount of time due to the lack of optimisation and enterprise scale infrastructure, this process takes place as a background operation at the server. Therefore, once the KDM is created, the end user does not experience any delays during the search process.

After completion of the run, the system had processed documents of various size, file types, and licenses from 11 repositories representing many regions of the world (Table 4). It was noted that there was a certain amount of noise in the keywords identified due to the limited number of resources indexed in a given domain. However, with more documents being indexed, the expansion of the list of terms will result in larger IDF values which will decrease the TF-IDF value for noise words. This will result in the algorithm rejecting these noise words as keywords, that is, the reduction of noise.

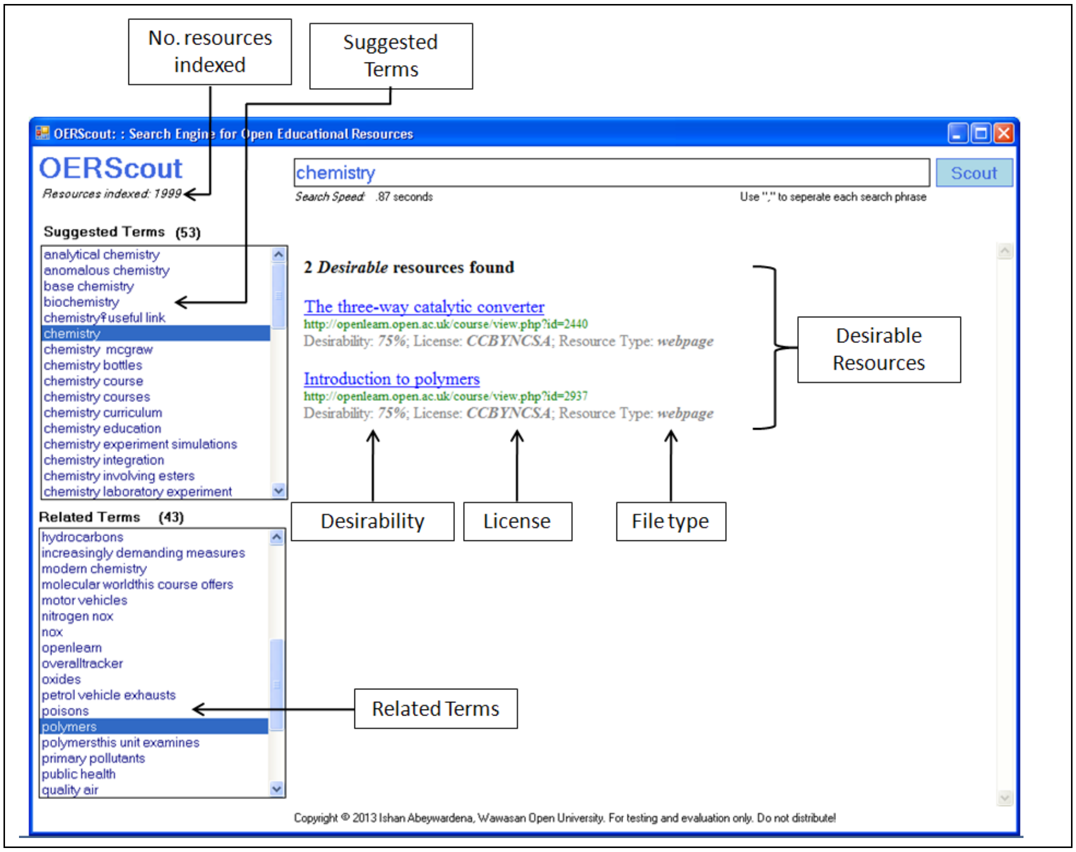
[3] http://doer.col.org/
[4] http://www.cnx.org

Table 4

Resources Indexed in the KDM Based on the Initial Input

	Repository	Host institution	Region	License	File type	No. resources indexed
1	Connexions	Rice University	USA	CC BY	Webpage	1536
2	OCW Athebasca	Athebasca University	Canada	CC BY	Webpage	07
3	OCW Capilano	Capilano University		CC BY-NC-SA	Webpage	19
4	OCW USQ	University of Southern Queensland	Australia	CC BY-NC-SA	Webpage	10
5	UCT Open Content	University of Cape Town	South Africa	CC BY-NC-SA	Webpage	63
6	OpenLearn	The Open University	UK	CC BY-NC-SA	Webpage	242
7	WikiEducator	COL & Ottago Polytechnic	New Zealand	CC BY-SA	Webpage	38
8	Unow	University of Nottingham	UK	CC BY-NC-SA	Webpage	27
9	TESSA	Multiple African Universities	Africa	CC BY-SA	PDF	15
10	OER AVU	African Virtual University	Africa	CC BY-SA	DOC DOCX PDF	40
11	WOU OER	Wawasan Open University	Malaysia	Various	PDF	02
	Total					1999

In order to test the functionality of the system from a real-world user's perspective, 27 academics who have at least 3-5 years of experience in OER advocacy, creation, use, and reuse were invited to test the system. Out of the 27 experts invited, 19, including six professors, five associate professors, three PhD holders, and four mid career academics, agreed to test the system and provide feedback. This group of users represented Australia, Brazil, Cambodia, Canada, China, Hong Kong SAR, Indonesia, Malaysia, Pakistan, and Vietnam. They comprised of varied backgrounds such as engineering, computer science, electronics, instructional design, distance education, agriculture, biology, law, and library science. The KDM was made available to this group through the OERScout client interface shown in Figure 6.



chemistry

Resources indexed: 1999

Search Speed: .87 seconds

Use "," to seperate each search phrase

Suggested Terms (53)

- analytical chemistry
- anomalous chemistry
- base chemistry
- biochemistry
- chemistry?useful link
- chemistry
- chemistry mcgraw
- chemistry bottles
- chemistry course
- chemistry courses
- chemistry curriculum
- chemistry education
- chemistry experiment simulations
- chemistry integration
- chemistry involving esters
- chemistry laboratory experiment

Related Terms (43)

- hydrocarbons
- increasingly demanding measures
- modern chemistry
- molecular worldthis course offers
- motor vehicles
- nitrogen nox
- nox
- openlearn
- overalltracker
- oxides
- petrol vehicle exhausts
- poisons
- polymers
- polymersthis unit examines
- primary pollutants
- public health
- quality air

2 *Desirable* resources found

The three-way catalytic converter
http://openlearn.open.ac.uk/course/view.php?id=2440
Desirability: *75%*; License: *CCBYNCSA*; Resource Type: ***webpage***

Introduction to polymers
http://openlearn.open.ac.uk/course/view.php?id=2937
Desirability: *75%*; License: *CCBYNCSA*; Resource Type: ***webpage***

Copyright © 2013 Ishan Abeywardena, Wawasan Open University. For testing and evaluation only. Do not distribute!

Figure 6. OERScout client interface used for testing the system. The figure shows a search result for resources on “chemistry: polymers”.

A comprehensive user manual was provided to the users which outlined how OERScout searched for the most desirable resources. The testing was conducted for a duration of seven days. The users tested the system by searching for OER for their day-to-day academic needs. At the end of the test period, the users provided qualitative feedback through a web based feedback form on various aspects of the OERScout framework. The general feedback which holistically critiques the OERScout technology framework is consolidated in Table 5.

Table 5

Consolidated Feedback Gathered from the OERScout Test Users

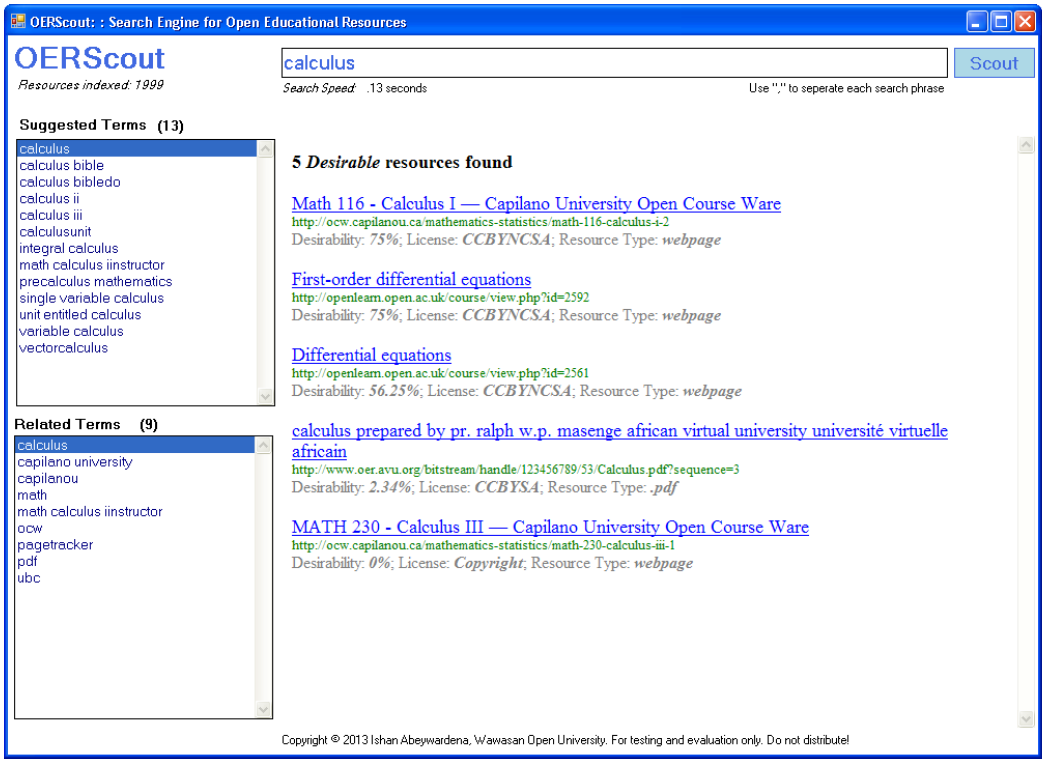
	Criteria	Advantages of OERScout	Weaknesses of the prototype
1.	User interface	The user interface is quite simple, friendly, intuitive, un-cluttered and easy to operate. It avoids the hassle of shifting between search modes.	Add advanced search features such as year, language, author and type of resources are not available.
2.	"Faceted search" approach which allows users to dynamically generate search results based on suggested and related terms	The ability to drill down using "faceted search" is very useful. It helps to locate resources faster.	As the number of resources grows the list of suggested and related terms will be quite long. Some noise terms are generated along with the keywords.
3.	Ease of use	It is a powerful tool which allows users to easily locate relevant resources.	The number of resources indexed is quite small.
4.	Relevance of the suggested terms generated according to the search query	The suggested terms are quite relevant and covers the scope of the search adequately.	Some unfamiliar noise words were generated as suggested terms.
5.	Use of related terms to effectively zero in on the resources being searched for	The feature is very useful and performs well. The functionality is similar to a thesaurus used by librarians for cataloging.	Many different terms point to the same resource due to the small dataset. Some terms are not related to the domain. Too many terms are generated.
6.	Usefulness of the resources returned with respect to Openness (the ability to use, reuse, revise and remix)	The use of the CC license to locate the most open resources is a useful feature. The value of this feature will increase along with the increase of quantity and quality of OER available.	The licensing scheme needs to be indicated in a more user-friendly manner.
7.	Usefulness of the resources returned with respect to Access (the ease of reuse and remix of resource type)	The resources returned met the criteria of access with respect to use and reuse. Based on the resource type, users can immediately identify how they can use the resource.	This might not be important as the licensing type defines the reuse and remix capabilities.
8.	Usefulness of the resources returned with respect to Relevance (the match between the results and your query)	Currently quite accurate and very useful.	The small size of the dataset limits the relevance.
9.	Effectiveness with respect to identifying the academic domain(s) of a resource	The autonomous identification of academic domains increases the focus of the search and the quality of the resources returned.	The technology shows promise but the number of domains identified are limited due to the size of the dataset.
10.	Use of the desirability for filtering the most useful resources for ones needs	The desirability framework is an interesting idea which will help in identifying resources appropriate for specific needs.	The concept of desirability needs to be explained to the user through the interface.
11.	Effectiveness with respect to locating desirable resources in comparison to mainstream search engines or native search	A comparison between the OERScout and conventional search engines cannot be made as they serve different purposes. OERScout is	Search engines such as Google have large databases of indexed resources. In this sense they cannot be compared to OERScout.

	engines of OER repositories[a]	much more focused and addresses some key issues in OER search.	
12.	Innovativeness of the technology framework	The technology framework is quite innovative and can bridge the gap between different metadata standards. The simplicity of the user interface complements the scale of innovation.	The scope of the framework needs to be refined. The system needs to be made available as an online service.
13.	How the wider OER community will be benefited	The technology will benefit the wider OER community as a tool for thought provoking discussion on adopting and adapting resources. It will be very beneficial for the novice user with respect to ease of use and affordability.	At the moment it is only a prototype. More resources need to be indexed before it can benefit the community.

Discussion

Empirical Evidence

Figure 6 shows a search conducted for the term “chemistry” on OERScout based on the KDM. In contrast to the static list of search results produced by generic search engines, OERScout employs a “faceted search” (Tunkelang, 2009) approach by providing a dynamic list of *suggested terms* which are related to “chemistry”. The user is then able to click on any of the suggested terms to access the most desirable OER from all the repositories indexed by OERScout. Furthermore, based on the selection by the user, the system will provide a list of *related terms* which will enable the user to drill down further to zero in on the most suitable OER for his/her teaching needs. In this particular example (Figure 6), the user has selected “polymers” as the related term to locate two desirable resources from the OpenLearn repository of The Open University which is known to host OER of high academic standard. Furthermore, Figure 7 shows the search results returned by OERScout for the search query on “calculus”. The desirable resources returned are from the open course ware OCW Capilano of Canada, OpenLearn of UK, and OER AVU of Africa. As such, it can be seen that OERScout is a more focused and dynamic system for effectively searching for desirable OER. This becomes one of the major benefits to ODL practitioners as the system spares the user from conducting repeated keyword searches in OER repositories to identify suitable material for use. It also allows users to quickly zero in on OER suitable for their needs without reading through all the search results returned by a generic search mechanism such as Google. Table 6 summarises some of the key features of OERScout in contrast to the generic search engines Google, Yahoo!, and Bing.



Resources indexed: 1999

calculus

Search Speed: .13 seconds

Use "," to seperate each search phrase

Suggested Terms (13)

calculus
calculus bible
calculus bibledo
calculus ii
calculus iii
calculusunit
integral calculus
math calculus iinstructor
precalculus mathematics
single variable calculus
unit entitled calculus
variable calculus
vectorcalculus

Related Terms (9)

calculus
capilano university
capilanou
math
math calculus iinstructor
ocw
pagetracker
pdf
ubc

5 *Desirable* resources found

Math 116 - Calculus I — Capilano University Open Course Ware
http://ocw.capilanou.ca/mathematics-statistics/math-116-calculus-i-2
Desirability: ***75%***; License: ***CCBYNCSA***; Resource Type: ***webpage***

First-order differential equations
http://openlearn.open.ac.uk/course/view.php?id=2592
Desirability: ***75%***; License: ***CCBYNCSA***; Resource Type: ***webpage***

Differential equations
http://openlearn.open.ac.uk/course/view.php?id=2561
Desirability: ***56.25%***; License: ***CCBYNCSA***; Resource Type: ***webpage***

calculus prepared by pr. ralph w.p. masenge african virtual university université virtuelle africain
http://www.oer.avu.org/bitstream/handle/123456789/53/Calculus.pdf?sequence=3
Desirability: ***2.34%***; License: ***CCBYSA***; Resource Type: ***.pdf***

MATH 230 - Calculus III — Capilano University Open Course Ware
http://ocw.capilanou.ca/mathematics-statistics/math-230-calculus-iii-1
Desirability: ***0%***; License: ***Copyright***; Resource Type: ***webpage***

Copyright © 2013 Ishan Abeywardena, Wawasan Open University. For testing and evaluation only. Do not distribute!

Figure 7. Search results generated by OERScout for the term “calculus”. The desirable resources returned are from Capilano University, The Open University, and African Virtual University.

Table 6

Key Features of OERScout in Contrast to Google, Yahoo!, and Bing

	Key Feature	*OERScout*	Google	Yahoo!	Bing
1.	Provides a centralised mechanism to search for OER	Yes	Yes	No	No
2.	Searches for only the most *desirable* resources for academic purposes	Yes	No	No	No
3.	Effectively locates and presents resources from the distributed repositories	Yes	No	No	No
4.	Provides a dynamic mechanism instead of a static list of search results which can be used to zero in on the required resources	Yes	No	No	No
5.	Uses autonomously identified keywords for locating the most relevant resources	Yes	No	No	No
6.	Uniformly annotates resources with the relevant keywords to facilitate accurate searching	Yes	No	No	No
7.	Removes human error in the annotation of keywords	Yes	No	No	No

User Feedback

Based on the expert user feedback summarised in Table 5, the key strengths of the system include the ease of use, the specific focus on OER, the ability to quickly zero in on the required resource, and the use of desirability in the identification of the resource. The ability to autonomously identify academic domains and locate resources from heterogenous repositories regardless of the metadata standard are also found to be strengths of the system. The users felt that OERScout will especially benefit academics who are novices to OER.

One of the major weaknesses of the current prototype version was the limited number of resources indexed. This contributes to noise in the identified keywords and results in long lists of suggested and related terms. However, as the number of resources indexed grows, the noise words will be reduced giving way to more focused suggested and related terms. The users also felt that more advanced filters need to be added onto the search interface to allow filtering of properties such as file types and licences. However, the fundamental concept behind the desirability framework is to parametrically identify the most useful resources without the user's intervention. This observation suggests that a change in mindset with respect to search engines needs to take place before users can get accustomed to OERScout. The users also felt that the licensing scheme needs to be explained in non technical terms such as “can reuse, redistribute, revise and remix even commercially” instead of “CCBY”. They further suggested that the calculation of the desirability be explained to the user.

The technology framework used was also found to be a limitation of the system. The current Microsoft Windows based client interface limits the users to Microsoft PC consumers. However, the real world implementation of the system will be done on a web based platform which will provide wider access regardless of device or operating system. Another limitation is that this version of OERScout is not designed to cluster non-text based materials such as audio, video, and animations which is a drawback considering the growing number of multimedia based OER. However, it is noted from the initial results that the system will accurately index multimedia based material using the textual descriptions provided. One more design limitation is its inability to cluster resources written in languages other than English. Despite this current limitation, the OERScout algorithm has a level of abstraction which allows it to be customised to suit other languages in the future.

Considering the opportunities, the system was found to be thought provoking with respect to finding, adopting, and adapting OER. It also appeals to the novice OER users in terms of training, affordability, teaching, and learning. This in turn will promote further research and development in the field of OER. Analysing the threats, one of the major threats to OERScout is the scale of the resource databases available to mainstream search engines such as Google. In this respect, the users felt that OERScout will be unable to compete with these search engines. However, the users also felt that OERScout addresses a few focused issues related to OER and need not be compared to mainstream search engines which are more general in nature. It is also worth noting that the system will need to continuously update its resource database to ensure accuracy. Among the threats identified, the change in mindset with respect to this new search approach remains the greatest challenge to overcome.

Based on the above discussion, we strongly feel that the OERScout technology framework addresses the key deficiencies with respect to OER search. In sum, the provision of a centralised system which allows academics to effectively zero in on desirable resources hidden away in heterogenous repositories makes OERScout a viable alternative to existing OER search methodologies.

Conclusion

With more and more OER repositories mushrooming across the globe and with the expansion of existing repositories due to increased contributions, the task of searching for useful OER has become a daunting one. As discussed in the literature, a compounding factor to this current predicament is the inability of present day OER search methodologies to effectively locate resources which are desirable in terms of openness, access, and relevance. As a potential solution to this issue we propose the OERScout technology framework.

OERScout uses text mining techniques to cluster OER using autonomously mined domain specific keywords. It is developed with a view of providing OER creators and users a centralised system which will enable effective searching of desirable OER for academic use. The benefits of OERScout to content creators include (i) elimination of the need for manually annotating resources with metadata used in search; (ii) elimination of the need for publicising the availability of a repository and the need for native search mechanisms; and (iii) reach of material to a wider audience. The system benefits OER users by (i) providing a central location for finding resources of acceptable academic standard; (ii) locating only the most desirable resources for a particular teaching and learning need; and (iii) allowing the user to effectively zero in on the resources they are after. Based on the initial expert user test results, OERScout shows promise as a viable solution to the global OER search dilemma.The ultimate benefit of OERScout is that both content creators and users will only need to concentrate on the actual content and not the process of searching for desirable OER.

It is our intention to make OERScout available as a public service via www.oerscout.org which would allow academics to search desirable OER for their specific teaching and learning needs. We also intend to transfer the system onto a free and open source software (FOSS) platform in the spirit of openness and accessibility. Considering the limitation of the current system with respect to searching resources written in languages other than English, we are currently designing a further extension to OERScout which will facilitate searching of resources written in other languages. Furthermore, we are exploring the possibility of autonomously extracting some important IEEE LOM metadata from OER to provide better recommendations.

Acknowledgements

This research project is funded as part of a doctoral research study through Grant # 102791 generously made by the International Development Research Centre (IDRC) of Canada through an umbrella study on Openness and Quality in Asian Distance Education.

The authors acknowledge the support provided by

- Emeritus Professor G. Dhanarajan of the Institute for Research and Innovation (IRI), Wawasan Open University with respect to facilitating the project,
- participants of the user test.

Ishan Sudeera Abeywardena acknowledges the support provided by

- Professor A. Kanwar and Dr. V. Balaji of the Commonwealth of Learning (COL), Vancouver, Canada through an Executive Secondment (4th – 25th May 2012),
- Faculty of Computer Science and Information Technology, University of Malaya where he is currently pursuing his doctoral research in Computer Science,
- Wawasan Open University where he is currently employed.

References

Abeywardena, I. S. (2013). Development of OER-based undergraduate technology course material: “TCC242/05 web database application” delivered using ODL at Wawasan Open University. In G. Dhanarajan & D. Porter (Eds.), *Open educational resources: An Asian perspective* (pp. 173-184). Vancouver: Commonwealth of Learning and OER Asia.

Abeywardena, I. S., Dhanarajan, G., & Chan, C. (2012). Searching and locating OER: Barriers to the wider adoption of OER for teaching in Asia. *Proceedings of the Regional Symposium on Open Educational Resources: An Asian Perspective on Policies and Practice.* Penang, Malaysia.

Abeywardena, I. S., Raviraja, R., & Tham, C. Y. (2012). Conceptual framework for parametrically measuring the desirability of open educational resources using D-index. *International Review of Research in Open and Distance Learning ,* *13*(2), 104-121.

Anido, L. E., Fernández, M. J., Caeiro, M., Santos, J. M., Rodriguez, J. S., & Llamas, M. (2002). Educational metadata and brokerage for learning resources. *Computers & Education ,* *38*(4), 351-374.

Balaji, V., Bhatia, M. B., Kumar, R., Neelam, L. K., Panja, S., Prabhakar, T. V., et al. (2010). Agrotags–a tagging scheme for agricultural digital objects. *In Metadata and Semantic Research* (pp. 36-45). Berlin Heidelberg: Springer.

Barton, J., Currier, S., & Hey, J. (2003). Building quality assurance into metadata creation: An analysis based on the learning objects and e-prints communities of practice. *In 2003 Dublin Core Conference: Supporting Communities of Discourse and Practice* - *Metadata Research and Applications*. Seattle, Washington.

Brooks, C., & McCalla, G. (2006). Towards flexible learning object metadata. *International Journal of Continuing Engineering Education and Life Long Learning ,* *16*(1), 50-63.

Calverley, G., & Shephard, K. (2003). Assisting the uptake of on-line resources: Why good learning resources are not enough. *Computers & Education ,* *41*(3), 205-224.

Casali, A., Deco, C., Romano, A., & Tomé, G. (2013). An assistant for loading learning object metadata: An ontology based approach. *Interdisciplinary Journal of E-Learning and Learning Objects (IJELLO)*, *9*, 11.

Caswell, T., Henson, S., Jenson, M., & Wiley, D. (2008). Open educational resources: Enabling universal education. *International Review of Research in Open and Distance Learning* , *9*(1), 1-11.

Cechinel, C., Sánchez-Alonso, S., & Sicilia, M. Á. (2009). Empirical analysis of errors on human-generated learning objects metadata. In *Metadata and Semantic Research* (pp. 60-70). Berlin Heidelberg: Springer.

De la Prieta, F., Gil, A., Rodríguez, S., & Martín, B. (2011). BRENHET2, A MAS to facilitate the reutilization of LOs through federated search. *In Trends in Practical Applications of Agents and Multiagent Systems* (pp. 177-184). Berlin Heidelberg: Springer.

Devedzic, V., Jovanovic, J., & Gasevic, D. (2007). The pragmatics of current e-learning standards. *Internet Computing, 11*(3), 19-27.

Dhanarajan, G., & Abeywardena, I. (2013). Higher education and open educational resources in Asia: An overview. In G. Dhanarajan & D. Porter (Eds.), *Open educational resources: An Asian perspective* (pp. 3-10). Vancouver: Commonwealth of Learning and OER Asia.

Dichev, C., & Dicheva, D. (2012). Open educational resources in computer science teaching. *Proceedings of the 43rd ACM technical symposium on Computer Science Education* (pp. 619-624). ACM.

Feldman, R., & Sanger, J. (2006). *The text mining handbook: Advanced approaches in analyzing unstructured data*. Cambridge University Press.

Ha, K. H., Niemann, K., Schwertel, U., Holtkamp, P., Pirkkalainen, H., Boerner, D., et al. (2011). A novel approach towards skill-based search and services of open educational resources. *Proceedings of Metadata and Semantic Research* (pp. 312-323). Berlin Heidelberg: Springer.

Hilton, J., Wiley, D., Stein, J., & Johnson, A. (2010). The four R‘s of openness and ALMS Analysis: Frameworks for open educational resources. *Open Learning: The Journal of Open and Distance Learning,25* (1), 37-44.

Kolowich, S. (2012, November 5). *Pearson's open book*. Retrieved from http://www.insidehighered.com/news/2012/11/05/pearson-unveils-oer-search-engine

Lextek. (n.d.). *Onix Text Retrieval Toolkit API reference*. Retrieved from lextek.com: lextek.com/manuals/onix/stopwords1.html

McGreal, R. (2010). Open educational resource repositories: An analysis. *Proceedings of the 3rd Annual Forum on e-Learning Excellence*. Dubai, UAE.

Piedra, N., Chicaiza, J., López, J., Martínez, O., & Caro, E. T. (2010). An approach for description of open educational resources based on semantic technologies. *In Education Engineering (EDUCON)* (pp. 1111-1119). IEEE.

Piedra, N., Chicaiza, J., López, J., Tovar, E., & Martinez, O. (2011). Finding OERs with social-semantic search. *Proceedings of the 2011 IEEE Global Engineering Education Conference (EDUCON)* (pp. 1195-1200). Amman, Jordan: IEEE.

Pirkkalainen, H., & Pawlowski, J. (2010). Open educational resources and social software in global e-learning settings. In P. Yliluoma (Ed.), *Sosiaalinen verkko-oppiminen* (pp. 23-40). Naantali: IMDL.

Shelton, B. E., Duffin, J., Wang, Y., & Ball, J. (2010). Linking opencoursewares and open education resources: Creating an effective search and recommendation system. *Procedia Computer Science, 1*(2), 2865-2870.

Tello, J. (2007). Estudio exploratorio de defectos en registros de meta-datos IEEE LOM de objetos de aprendizaje. *In Post-Proceedings of SPDECE 2007 - IV Simposio Pluridisciplinar sobre Diseno, Evaluacion y Desarrollo de Contenidos Educativos Reutilizables* (pp. 19-21). Bilbao, Spain.

Tunkelang, D. (2009). Faceted search. In G. Marchionini (Ed.), *Synthesis lectures on information concepts, retrieval, and services* (Vol. 5, pp. 1-80). Morgan & Claypool.

UNESCO. (2012, June 22). *2012 PARIS OER Declaration*. Retrieved from unesco.org: http://www.unesco.org/new/fileadmin/MULTIMEDIA/HQ/CI/WPFD2009/English_Declaration.html

Unwin, T. (2005). Towards a framework for the use of ICT in teacher training in Africa. *Open Learning: The Journal of Open, Distance and e-Learning, 20*(2), 113-129.

Vaughan, L. (2004). New measurements for search engine evaluation proposed and tested. *Information Processing and Management, 40*, 677-691.

West, P., & Victor, L. (2011). *Background and action paper on OER*. Report prepared for The William and Flora Hewlett Foundation.

Wiley, D. (2006). *On the sustainability of open educational resource initiatives in higher education*. Retrieved from oecd.org:http://www1.oecd.org/edu/ceri/38645447.pdf

Yamada, T. (2013). Open educational resources in Japan. In G. Dhanarajan & D. Porter (Eds.), *Open educational resources: An Asian perspective* (pp. 85-105). Vancouver: Commonwealth of Learning and OER Asia.

Yergler, N. R. (2010). Search and ciscovery: OER's open loop. *Proceedings of Open Ed 2010*. Barcelona.