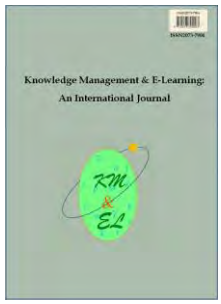
The open corpus challenge in eLearning

Mahantesh K. Pattanshetti
Sanjay Jasola
Graphic Era Hill University, Dehradun, India
Vivek Gupta
Akshay Rajput
Graphic Era (Deemed to be University), Dehradun, India



Knowledge Management & E-Learning: An International Journal (KM&EL)
ISSN 2073-7904

Recommended citation:
Pattanshetti, M. K., Jasola, S., Gupta, V., & Rajput, A. (2018). The open corpus challenge in eLearning. *Knowledge Management & E-Learning, 10*(1), 67–85.

Knowledge Management & E-Learning, 10(1), 67–85

The open corpus challenge in eLearning

Mahantesh K. Pattanshetti*

School of Computer Science and Engineering
Graphic Era Hill University, Dehradun, India
E-mail: mahant.india@gmail.com

Sanjay Jasola

Vice Chancellor
Graphic Era Hill University, Dehradun, India
E-mail: sjasola@yahoo.com

Vivek Gupta

School of Computer Science and Engineering
Graphic Era (Deemed to be University), Dehradun, India
E-mail: vivgupta95@gmail.com

Akshay Rajput

School of Computer Science and Engineering
Graphic Era (Deemed to be University), Dehradun, India
E-mail: akshay.rajput.1711@gmail.com

*Corresponding author

Abstract: Learning has transcended into a life-long endeavor in the information age. It is no longer restricted to confines of formal classrooms. Consequently, a student is not restricted to traditional learning resources like teachers, textbooks or printed content. Digital resources available on the Internet form a very significant component of self-learning. Copious volumes of learning resources without legal barriers to self-learning reside in digital repositories, educational institution portals and on numerous websites. Learners wishing to utilize the web for personalized learning are faced with a daunting array of content to wade through and select the suitable ones to fulfill his/her learning objectives. Therefore, it is not a question of availability; it is one of relevance and suitability. Typically, in addition to time constraints, learners lack the expertise to screen content for effective eLearning. Adaptive hypermedia systems (AHSs) offer a path to harnessing this large volume of learning resources for personalized learning. This *review paper* provides a concise and coherent discussion about the evolution of AHSs along with the challenges that need to be addressed for effectively harnessing openly available educational resources referred to as open corpus resources (OCRs).

Keywords: Open corpus; Open educational resources; Cost effective learning; eLearning; Open corpus adaptive hypermedia systems

Biographical notes: Mahantesh K. Pattanshetti is an engineer with over two

decades of international experience in consulting for fortune 500 companies and working in higher education institutions. Currently, he is working as a training officer in the School of Computer Science and Engineering at Graphic Era Hill University, Dehradun. He is passionate about research attempting to bridge the knowledge divide by providing cost-effective solutions for eLearning by utilizing existing free content on the web. His research interests are in applications of the semantic web, data science methods and AI techniques like machine learning and natural language processing (NLP) for eLearning.

Sanjay Jasola is a professor in the School of Computer Science and Engineering and founding Vice Chancellor at Graphic Era Hill University Dehradun, India since 2011. He received his Ph.D. in Computer Science from the School of Computer and System Sciences, Jawaharlal Nehru University (JNU), New Delhi in 2006, and his M.Tech. and B.Tech. from the University of Roorkee (presently IIT Roorkee) in 1996 and Kamla Nehru Institute of Technology in 1988, respectively. His research interests are in development of collaborative environments for eLearning and in approaches to the proliferation of OERs and MOOCs. In his prior job, he was a founding Dean of School of ICT in Gautam Buddha University, Greater NOIDA, India. During his international assignment, he served as Associate Professor in the School of Science and Technology at Wawasan Open University, Penang, Malaysia. He was awarded Gold Medal by Indira Gandhi National Open University (IGNOU), New Delhi for establishing EDUSAT (satellite) based education network in India in 2006. He is a member of the Editorial Board of several International Journals and the author of more than 80 publications including conference and journal papers.

Vivek Gupta is a research assistant pursuing Computer Science at Graphic Era University, Dehradun. His interests span big data analysis, technology-enabled learning, applications of machine learning, natural language processing (NLP), computer vision, and intelligent web-based systems.

Akshay Rajput is an Assistant Professor in the School of Computer Science and Engineering, Graphic Era University, Dehradun, India. He has completed his post-graduation from the prestigious IIT-Delhi and has previously worked with Works Application Singapore PTE LTD. His research interests include machine learning and IOT.

1. Introduction

The confluence of information technology coupled with the ubiquity of the web has led to a proliferation of learning resources on the Internet. A treasure trove of freely available educational learning resources exists on the web. Educational resources exist in the form of Open Educational Resources (OERs), research articles, tutorials, quizzes, videos, overviews, podcast, slides, lecture notes, tools, applications and other digital media formats (Mosharraf & Taghiyareh, 2016; Sosnovsky, Hsiao, & Brusilovsky, 2012). Finding high-quality learning resources is not such a major problem with sophisticated search engines like Google. The lacuna is that search results delivered by search engines are user agnostic. A user needs to sift through the search results to find suitable ones for their learning requirements. User attributes like varying background knowledge, goals, prerequisites, learning styles are ignored in favor of "one-size-fits-all" results. As a consequence, a user feels overburdened and adrift in the hyperspace due to lack of

individualized results. Thus, it can be inferred that the issue is not of accessibility to content, but one of *personalized access* to content (Atenas & Havemann, 2014; Brusilovsky, Kobsa, & Nejdl, 2007; Lawless, Hederman, & Wade, 2008).

A significant majority of the systems achieve personalization in what are commonly referred to as closed corpus systems (CCSs). In CCSs, the content is proprietary, known in advance and the relationship among the documents and documents to reference models for adaptation are defined at design time using manually created metadata (Lawless & Wade, 2006). In contrast, open corpus systems (OCSs) source learning resources from the web, which have minimal or no metadata, thereby implying a lack of knowledge about the content and their relationships. Due to lack of metadata, an *a priori* relationship among documents or documents to models cannot be inferred. This lack of knowledge about documents and the underlying relationships necessitates alternate paths to personalization and adaptivity (Brusilovsky, Kobsa, & Nejdl, 2007).

2. Motivation, organization and scope

The technical motivation for this paper was the challenges faced by the authors in finding a single point of reference succinctly capturing the functional evolution of AHSs and the research challenges to be addressed in OCSs. The second and most significant research motivation is the belief that OCSs can immensely contribute to bridging the knowledge divide in developing nations by aiding in the development of cost-effective eLearning systems (Atenas & Havemann, 2014; Mosharraf & Taghiyareh, 2016).

There are an enormous number of high-quality research papers pertaining to AHSs. It is neither feasible nor advisable to have a discussion on every system developed. Therefore, the approach of this review paper has been to look at a few representative systems, in such a manner that the gist of functionality provided by AHSs is covered. Keeping this in mind, nearly eighty-six peer-reviewed journal articles in high-quality international publications and conferences were surveyed to scope out their relevance. Of these thirty-seven were selected for a detailed study.

In order to facilitate quick on-boarding of readers and to appreciate the potential and challenges of harnessing OCRs, the organization of this paper is as follows. Section 3 provides a conceptual framework for understanding AHSs. Section 4 covers the evolution of AHSs starting from systems using closed corpus content to mixed corpus content followed by, systems exclusively relying on OCRs for delivering functionality. Prior to concluding remarks, section 5 examines research challenges in the implementation of OCSs and forms a major contribution of this paper.

The scope of this paper is limited to an examination of the functionality of existing AHSs with the objective of furthering research in the utilization of OCRs to facilitate personalized eLearning. In addition, elements of the AHSs that assist in drawing inferences for personalization are also highlighted. Detailed discussion of AHSs architecture does not form a part of this paper, for a reference to AHSs architecture readers may refer (Knutov, De Bra, & Pechenizkiy, 2009). This paper does not to delve into security aspects, aesthetics of presentation, and legal barriers to access. Restrictions due to copyrights, digital rights management, intellectual property rights, licensing, and royalties though important are beyond the scope of this paper.

3. Overview of adaptive hypermedia systems

Conventional hypermedia offers pages with the same content, links, irrespective of user variations. This implies, it ignores individual differences in users and their varying backgrounds. In order to account for user diversity, AHSs come to the rescue. AHSs deliver *adaptation effect* by inferring from user models representing user knowledge, goals, preferences, learning styles, device characteristics and other individual attributes. AHSs have wide applications, wherever personalization of content is desirable as in the domain of information retrieval, eLearning, online help, and personalized views (Brusilovsky, 2001).

Origins of AHSs techniques can be traced to developments in personalized information retrieval and intelligent tutoring systems (ITS) (Mulwa, Lawless, Sharp, Arnedillo-Sanchez, & Wade, 2010). Some popular examples of applications having used these techniques are virtual museums, news personalization, and electronic commerce (Brusilovsky, Kobsa, & Nejdl, 2007). Significant application areas are however in education and medical domain. It is in these last two domains, where the scope for application and challenges are the maximum due to their unique nature of being highly sensitive to the quality of information (Brusilovsky, Kobsa, & Nejdl, 2007; Brusilovsky, 2012). The key to personalization (see Fig. 1) in AHSs are, index link typing to indicate the type of knowledge element (problem, solution, quiz etc.), a couple of techniques to provide adaptation effect taking into consideration mapping of domain and documents to concepts or ontology and a user model.

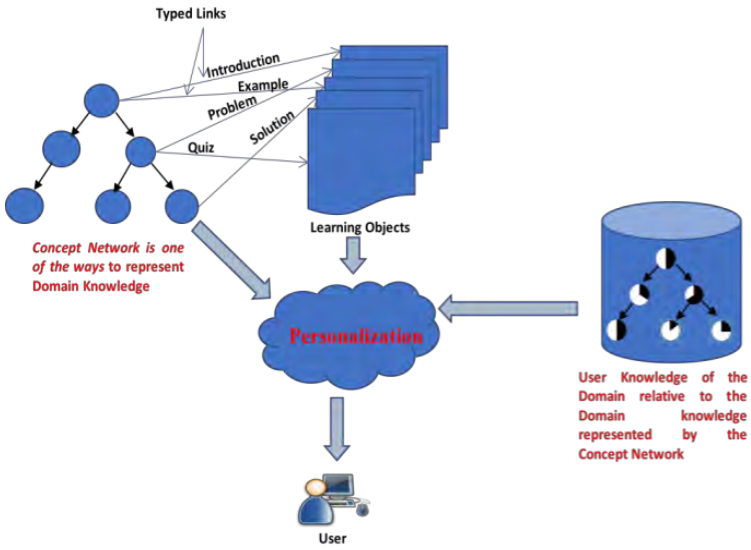


Fig. 1. Conceptual view of elements contributing to personalization

There are primarily two techniques that enable implementation of adaptation effect. They are namely, *adaptive presentation* and *adaptive navigation* support. An adaptive presentation may be achieved for example by natural language processing (NLP), or by techniques like adding/removing text, dimming text, sorting text fragments or using stretch text. Adaptive navigation support may be provided by, for example, using techniques like adaptive link sorting, link hiding/enabling/disabling, link annotation, or link generation (Brusilovsky, 1998).

Domain knowledge is represented in terms of a concept network or ontology. Documents are then mapped to a concept network or ontology representing domain knowledge. Networking of concepts helps in drawing inferences for determining prerequisites, providing problem-solving support, in learning path trail generation, curriculum sequencing, link annotation, and goal attainment (De Bra & Calvi, 1998).

In a similar fashion, for modeling user's knowledge of the domain, the user knowledge may be represented in terms of domain concepts or ontology. Knowledge of a concept is represented as value pair with one part representing the concept name and the second part its value. The value may simply be a Boolean value (T/F), or qualitative value (like novice, beginner, intermediate, advanced, or expert), or a quantitative value (say from 1 to 100) indicating the levels of knowledge (Brusilovsky, Kobsa, & Nejdl, 2007).

For example, consider a concept network or ontology representation to model and track the state of user knowledge of the domain. The model may use explicit and/or implicit mechanisms to maintain its state. An explicit mechanism may be through a survey administered to the user or through self-evaluation. Implicit mechanisms may include aspects like time spent by a user on a page, examples solved, pages read or through tests. Tracking user progress through tests or expert evaluation is generally preferred since other methods are subjective and thereby prone to errors. Knowledge of a concept may be inferred from a single source (page visit) or it may come from many sources (page visits, quizzes, problem solution, expert opinion, self-evaluation) (Brusilovsky, Kobsa, & Nejdl, 2007). User modeling is a significant discipline and only a basic conceptual idea is provided here for understanding the working of AHSs.

AHSs in addition to facilitating personalization based on modeling of user's knowledge of the domain can support personalization based on other reference models like device, goals, and learning style to deliver content. AHSs by offering customized learning content prevent information overload, lost in hyperspace syndrome, navigation guidance, and greater motivation among learners (Mulwa et al., 2010).

4. Evolution of adaptive hypermedia systems

Prior to 1996, AHSs were constructed for use on standalone personal computers. In the latter half of 1990s advances in web technologies and increasing web access coupled with obvious advantages of the web led to the development and deployment of AHSs on the web. Due to increasing web penetration, starting from around 1996 researchers became well aware of each other's work leading to extensive collaboration among research teams. Since then, the educational domain has been the focus of research, simply due to the fact that most of the active researchers happen to be faculty having information technology background working in universities and experimenting using their courses as case studies. The time period from 1996 to 2004 resulted in the development of pioneer systems like ELM-ART, InterBook, KBS Hyperbook, and AHA! (Brusilovsky, 2012).

4.1. Pioneer systems using closed corpus resources

Initial research commenced with the display of physical textbooks as hypertext documents on the web, or as electronic copies of books, gradually evolving with rich media support paralleling developments in web technologies. One of the earliest systems (Schwarz, Brusilovsky, & Weber, 1996) ported from personal computer to the web the functionality of ITSs for learning LISP programming language. Interactive learning was

facilitated through examples, explanations, problems and solutions. This system provided problem-solving assistance, individualized learning paths, adaptive annotation, intelligent program analysis, and interactivity. A major objective of the research was porting standalone ITSs to the web and resolving the issues faced thereof. Porting onto the web ensured in providing access anytime, and anywhere as opposed to a ITSs that was tied to a particular computer.

ELM-ART (Brusilovsky, Schwarz, & Weber, 1996) is another example of ITSs inspired adaptive remote tutor providing most of the system functionality of (Schwarz, Brusilovsky, & Weber, 1996) and also attempts to reduce access tyranny imposed by geography and time. Adaptive techniques like link annotation using visual clues to indicate student readiness for a page content using a traffic light metaphor were implemented. As in a traffic light, colors are used to indicate user readiness to explore contents behind a link. For example, red color could indicate user does yet have the prerequisite knowledge. A concept graph was used to establish relationships between concepts and content was indexed with link typing indicating a learning resource as an explanation, a problem, a solution, an example or any other relevant knowledge unit. User modeling was done using a user modeling approach known as user overlay model. Here domain model captures the knowledge possessed by a domain expert and user knowledge is estimated as a subset of the domain model.

In KBS Hyperbook (Henze & Nejdl, 1999) the user identifies learning goals and the system generates a learning path. User guidance was provided for goal selection, project selection, and during project execution. A constructivist pedagogic approach was at the heart of the design by using a project-based approach to the attainment of learning goals. Domain concepts were modeled using knowledge dependency graph and a Bayesian network was used for maintaining the user model. User knowledge was estimated using self-evaluation by the user or by expert opinion and the user model was updated based on multiple evidence of learning.

4.2. Discussion on open corpus systems

OCRs incorporation and OCSs are not relatively new ideas. One of the earliest mentions of the open corpus problem was as early as in 2001 (Brusilovsky, 2001). Various attempts have been made since then to address the open corpus challenge. In the early systems, OCRs were incorporated into CCSs, and systems so designed are known as *mixed corpus* systems. System functionality was enhanced by linking closed corpus resources with open corpus content (Brusilovsky & Rizzo, 2002; Brusilovsky, Chavan, & Farzan, 2004; Dolog, Henze, Nejdl, & Sintek, 2004; Henze & Nejdl, 2001). Systems purely using OCR gained traction with the efforts of (Kravčík & Wan, 2013; Lawless & Wade, 2006; Lawless et al., 2008; Levacher, Hynes, Lawless, O'Connor, & Wade, 2009; Lin & Brusilovsky, 2011; Muntean & Muntean, 2009; Sosnovsky et al., 2012; Steichen, Lawless, O'Connor, & Wade, 2009) and is still an ongoing area of active research with huge potential and challenges.

A discussion of pioneer mixed corpus systems is followed by systems solely relying on OCRs.

4.2.1. Mixed corpus systems

One of the earliest systems (Henze & Nejdl, 2001) to have provided access to OCRs was java tutorial KBS Hyperbook. Semantic links from closed corpus resources (CCRs) that

were manually authored to Sun java tutorial were established. OCRs of the Sun java tutorial were provided as a means of enhanced functionality or alternate means to concept understanding. Indexing of concepts was done manually for both the corpus contents. Each user was given the impression that he/she was having access to a personalized electronic textbook. A constructivist approach to learning as in the earlier Hyperbook (Henze & Nejdl, 1999) was continued. In order to simulate real-world learning, examples, past projects by users and knowledge support to solve projects was provided. Adaptive annotation, learning trail generation, goal identification, project selection guidance and project development guidance were provided on the basis of a user model. Knowledge of a user was graded as a novice, beginner, advanced or expert. The grading was based on judgments of experts, self-evaluation, and tests of learning. In order to avoid causing confusion to the learner, separate individual trails for CCRs and OCRs were provided. An innovative aspect of the system was the usage of OCRs. But, the use of laborious manual indexing was an impediment. Manual methods of indexing are also impractical should there be additions, deletions, and modifications to the OCRs. Also, as it utilizes a single web portal the approach used by the researchers for manual linking of CCRs to OCRs renders it unsuitable for dynamically retrieved resources scattered over the web.

Connectivity between CCRs and OCRs was established using the metaphor of navigation through maps using landmarks (Brusilovsky & Rizzo, 2002). High-quality literature pertaining to 'C' programming language was manually identified from the web. Initially, the authors tried to provide a link from the closed corpus content to the root of a related web tutorial; but, they noticed that users were reluctant to utilize OCRs. Users did not relish the prospect of having to navigate through, to locate relevant contents. Semantic matching was then used for providing links between the two corpuses at the relevant section. This approach turned out to be far more useful and encouraged users to explore and learn from OCRs. Documents were clustered into a mixed corpus by performing page level keyword analysis using self-organizing maps (SOM) based on artificial neural networks. Keywords functioned as a legend for identifying cell contents. Contents of the two-dimensional SOMs are semantically related based on their relative positions in the cell. Documents in the same cell shared the maximum semantic similarity and relative distance from the cell accounted for their corresponding reduction in semantic similarity. Information density in the cells was shown using varying shades of blue similar to displaying depths of water in ocean maps. Hence the logically apt system name Knowledge Sea. The major advantage was that new content could be automatically incorporated into the system speedily, but a drawback was locating OCRs manually.

Personal reader (Dolog et al., 2004) integrated the local CCRs consisting of learning objects like java topics, quizzes, summaries and articles to the OCRs containing resources like java FAQ's, tutorials, simulations, exercises, blogs, applets, and illustrations. The resources of open and closed corpus were manually annotated with RDF metadata adhering to Dublin core and IEEE Learning Object Metadata (LOM) standards. In order to satisfy query requirements, metadata along with the user model was used as the basis for mapping to ontology. Object link typing was done to identify whether a page was a tutorial, example or any other relevant knowledge unit.

A community driven approach akin to Wikipedia was creatively used in Knowledge Sea II (Brusilovsky, Chavan, & Farzan, 2004). In this system, social and collaborative approaches to learning resources annotation were successfully demonstrated. Communities of users provided explicit and implicit feedback about the content. The density of user traffic is visualized using a footprint metaphor. More frequently visited pages are indicated by a higher density of footprints. This approach allows users to navigate through content by following the footmarks of other users with similar learning

objectives. Each cell is identified by keywords and lecture slides acting as landmarks for history-based navigation to satisfy user learning requirements.

In the preceding paragraphs, novel approaches for incorporating OCRs by researchers into their learning systems were examined. The subsequent sub-section discusses systems utilizing only OCRs for eLearning.

4.2.2. Systems exclusively using open corpus resources

Delivery of standalone OCSs with minimal manual intervention may be regarded as one of the major research challenges facing researchers in the domain of AHSs. In fact, addressing the open corpus challenge has been the focus of major research groups worldwide over the past decade and a half. Researchers at top notch global universities have been aggressively working on finding ways and means to incorporate OCRs into their systems. The following paragraphs summarize the work in the past decade.

Researchers at the University of Dublin have been active over the past decade in incorporating OCRs for eLearning. Lawless and Wade (2006) have been investigating the gamut of issues involved in sourcing high-quality content, harvesting the content and finally providing useful semantic slices for end use (Levacher, Lawless, & Wade, 2012; Bayomi, 2015). The researchers at Dublin have used a service-oriented approach by having individual components for sourcing, harvesting, personalization and presentation. WebCrawler's are utilized for scouring the web for the creation of a metadata cache. Since a vast majority of learning resources lack any form of metadata, the author's utilized domain ontology for creation and mapping of metadata. Experts also assist in the process of metadata annotation of documents. The contents are then offered via a personalized search against the metadata. Metadata cache forms the link used for retrieving the content from the World Wide Web.

Muntean and Muntean (2009) propose the use of OCRs subject to them being cost effective, by optimizing learning cost with regard to network constraints and other resources. The focus is on cost effective access, managing device characteristics, formatting, and content presentation. It is assumed that the concepts are mapped to the learning objects in the digital educational repositories, and documents are annotated with metadata for use in their eLearning system.

Sosnovsky, Hsiao, and Brusilovsky (2012) use semantic web tools and techniques for content and user modeling. Authors publishing on the web, structure their pages in order to deliver knowledge to a user. Structuring in the form of chapters, sections, headings, tables, links forms an important source of topics (vital knowledge elements). These topics are then mapped onto a central ontology, which in turn facilitates reasoning and semantics. The system provides supplementary reading material along with recommendations to the students. Recommendations are ordered by relevance, semantic similarity, and the student model. The authors claim that during evaluation the system demonstrated learning outcomes comparable to CCSs.

Kravčík and Wan (2013) use domain ontology for document mapping. By considering domain ontology as a network of concepts represented in a formal and explicit manner, the author's exploit formalization afforded by ontology for standardized representation, mapping, querying, and reasoning. Federated search is performed for locating educational objects in digital learning repositories and the documents are annotated with ontology using automated tools. Concepts form the base layer and are linked to documents semantically in the presentation layer. On receiving a user query

input as keywords or concepts, the content is adaptively presented based on the user model.

Table 1 succinctly captures the developments in utilizing OCRs, techniques used and limitations.

Table 1
Evolution of open corpus resources-based systems

System/Authors	Overview	Techniques	Limitations
KBS Hyperbook; Henze & Nejdl, 2001	Closed corpus contents were manually linked to the Sun java tutorial to provide enhanced conceptual understanding.	Indexing and content fusing between closed and open corpus was manually done.	Open corpus content was pre-selected and also manual effort for indexing and linking content.
Knowledge Sea; Brusilovsky & Rizzo, 2002	Learning objects from both corpuses were clustered into cells on basis of semantic similarity.	Self-organizing maps (SOM) using artificial neural networks.	Manually locating open corpus resources.
Personal Reader; Dolog et al., 2004	Local contents were manually linked to open corpus contents for enhanced learning.	Corpus contents were manually annotated with RDF.	Significant manual component for annotation and linking.
Knowledge Sea II Brusilovsky et al., 2004	Social and collaborative approach to resource annotation.	Users provided explicit and implicit feedback about the content.	Cold start problem when new content is added.
Levacher et al., 2012	Complete tool chain from content locating to harvesting and delivering information slices.	WebCrawler's to locate content, ontology for document mapping, and expert assistance for annotation of content.	Complex process requires an enormous amount of time to deliver the content.
Sosnovsky et al., 2012	Utilizes document structuring to extract and map knowledge to ontology.	Semantic web technologies.	Pre-selected web content and problem domain. Difficult to scale to other domains.
Kravčík & Wan, 2013	Documents have been mapped to a central domain ontology.	Federated search for locating learning objects and automatic tools for annotation.	User interacts via query interface using a restricted vocabulary of keywords or concepts.

5. Critical examination of the challenges posed by open corpus resources

In the preceding sections, an overview of the functionality provided by AHSs using open and closed corpus resources was reviewed. It can be noticed that a wide range of support for effective eLearning is delivered by these systems. The sheer heterogeneity of web content, non-existent or sparse metadata, finding high-quality resources, maintaining pedagogical consistency, challenges of natural language processing and the resulting constraints offer a grand research opportunity (Brusilovsky, Kobsa, & Nejdl, 2007; Lawless & Wade, 2006). Readers may notice few papers cited of recent history (last five years or so). Two factors account for this. The first is the nascent stage of research in the utilization of OCRs and second is that only a handful of papers were found to be relevant to this paper.

Prior to delving into the challenges posed in constructing AHSs using OCRs, it would be instructive to examine the generic structure of an AHS (see Fig. 2) and the issues (refer Table 2) to be addressed prior to the deployment of an AHS (De Bra, Houben, & Wu, 1999).

Table 2
Major issues to be addressed in the construction of an AHS

AHS Component	Issues to be addressed in the deployment of AHSs
Content	• Custom developed content or reuse content from web or repositories. If reusing web/repository content, then locating quality high content. • Support for different file and data formats. • Manually annotate content or use tools (metadata is required for adaptation). • Approaches to link typing & content indexing for speedy retrieval.
Domain Model	• Modeling knowledge. For example, to use a network of concepts or to use ontology. • Mapping of domain concepts to different learning objects.
User Model	• Static or a dynamic user model (UM). • Construction and maintaining the state of an UM. For example, to use explicit feedback to construct UM or by implicitly observing and tracking user behavior or a combination of both. • Representation of an UM. To use a concept network or ontology. • Estimating user knowledge through tests, implicit mechanisms or Bayesian methods.
Adaptation Model	• In addition to adaptation by drawing inferences from domain knowledge and user's knowledge of the domain, adaptation may also be done based on device characteristics, learning styles, interests, learning goals, group dynamics and pedagogical considerations among others. • Approaches to combining inputs from user and domain models plus any other models to provide adaptation. • Defining and implementing action rules for responding to various user actions triggering events. • Approaches to update the different models used to enable adaptation.
Presentation Layer	• Specifications for the display of content to the users based on their roles. For example, the teacher may be provided a different interface to the content as opposed to a student.
Runtime Layer	• Management of hypertext display, user interaction and error handling.

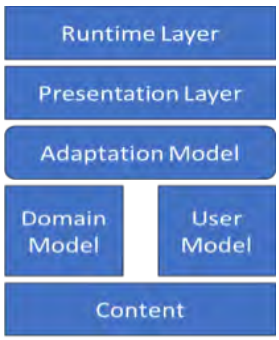


Fig. 2. Abstraction of the key components in an AHS

Existing CCSs and MCSs have been used for both formal and informal learning due to the fact that control can be exercised during all aspects of system construction and deployment. On the other hand, OCSs are being used for supplementing and complementing formal learning due to challenges posed in the utilization of OCRs. A gold standard objective would be to utilize OCRs to replicate the functionality provided by successful AHSs systems using CCRs. This aforementioned consideration forms the basis of the following discussion.

5.1. Content sourcing, resource diversity, access, and resource quality

Learning resources are exponentially growing and are distributed all over the web among digital repositories, on websites and blogs. Thereby, the first task would be *locating* learning objects. Focused crawlers (Levacher et al., 2009) can find highly valuable resources but they need to be configured to target the most valuable resources and also take an enormous amount of time to complete a crawl. Also, the setup of a crawl requires technical expertise and may act as an impediment to users without a technical background.

Documents on the web occur in a hue of *structural formats* (HTML, PDF, DOC, RTF/TXT, EPUB, and PPTs to name a few popular ones). For the application of any further processing techniques like natural language processing, or machine learning, documents have to be first scraped in order to obtain information in textual format. In HTML pages, for example, it can be quite challenging to obtain textual information with all the additional links for navigation, advertisements, and other items not relevant to page functionality (Levacher et al., 2009). In order to prevent automated tools from hogging web resources, a good number of popular sites have an exclusion policy in robots.txt and in addition, new barriers like *semantic captchas* are being deployed (Vikram, Fan, & Gu, 2011). This acts as a potential obstacle for automatic data extraction from web documents.

Once the learning resources (could be text-based or multimedia objects) are available for processing, issues of *quality, grading* resources in terms of relevance and screening content to match domain concepts and user models need to be addressed. A multi-pronged and interdisciplinary approach to learning resource analysis is imperative. Issues of quality, for example, may be discerned from peer review, user feedback, author reputation, page popularity, and page ranking (Atenas & Havemann, 2014). Other aspects

like concept matching, grading, and adaptation to user models shall need an application of techniques from conventional AHSs, semantic web, NLP, and machine learning.

Learning resources come with different presentation styles, depth of coverage with a target audience in mind (Brusilovsky, Kobsa, & Nejdl, 2007). This can lead to challenges in *content repurposing* for new contexts, user levels, and applications (Levacher et al., 2009). Manually it is easy to identify a learning resource as a problem statement, an example, an overview, an introduction or any other type of knowledge element. Extracting the type of knowledge element in OCRs using automated techniques poses challenges due to the flexibility offered by natural language representation.

5.2. Metadata and indexing

Lack of metadata and its standardization poses the single biggest obstacle to machine processing and thereby large-scale deployment of applications using OCRs. Existing approaches require a huge manual effort for the creation of content, indexing, annotation or metadata generation. Manual approaches are clearly not feasible for web scale projects due to lack of time, resources, expertise, and maintenance required. As a result, techniques used in existing systems are not necessarily applicable to the OCSs. Rather than engaging faculty in monotonous, time-consuming and laborious manual tasks, it would be more productive if faculty efforts are directed towards pedagogical and academic endeavors. (Brusilovsky, 2008; Lawless & Wade, 2006).

If all the learning resources had well-annotated metadata in a standardized format, the problem would have been akin to conventional AHSs using CCRs. Lack or insufficient metadata, varying metadata standards; problems of natural language in the description of learning resources pose barriers to automatic analysis of content. Some automated tools like Semtag, Metasaur (Kravčík & Wan, 2013) exist for automatic metadata annotation but are of little use to deal with diverse knowledge domains. As a consequence, there is no interoperability between systems and difficulty is experienced in deriving semantic knowledge by machine processing of content (Lawless & Wade, 2006).

A number of approaches exist to enhance the documents with the knowledge to facilitate adaptation and interlinking. A majority of advanced systems use indexing in reference to an external ontology and the same is used for adaptation (Brusilovsky, Kobsa, & Nejdl, 2007). Manual, community-based and automatic approaches may use keyword or concept indexing. Keyword based approaches possess low precision and adapt to content, but shall perform poorly for adaptation to goals and knowledge. Hence it may not be a suitable option in the educational context. Application of semantic web technologies is not always feasible. For example, there are domains without a defined ontology or with variations in the descriptions of domains using ontology. As a consequence, mapping of learning resources and reasoning using ontology is challenging. Ontology mapping itself is an ongoing research area and there is still a lot left to be achieved.

The focus of this paper is mostly on textual resource analysis, but the web offers an array of multimedia resources in the form of massive open online courses (MOOC's) (Knutov, De Bra, & Pechenizkiy, 2009). A vast majority of these have little to no metadata to be used for the purpose of analysis. Social media tools like tagging (Steichen et al., 2009) need to be investigated for meaningful utilization of these resources.

AHSs derive their power based on the quality of content indexing with respect to a model. Indexing strength is a function of link expressiveness, granularity, and cardinality. A flat link may just express a relation, whereas a more complex link can

define the type of content, whether it is an introduction or an example or a summary or a prerequisite or any other knowledge element attribute required for its effective usage. Finer granularity of indexing permits adaptation to be more specific. In OCSs, it would be ideal if there is a one to one match between a learning resource and a concept. This may not always be the case, as it is quite possible that a page can describe multiple concepts or each concept may refer to multiple pages forming concept hubs. Atomic or composite concepts pose new challenges in identifying the most suitable pages and their sequencing (Brusilovsky, 2012; Knutov, De Bra, & Pechenizkiy, 2009).

5.3. Pedagogical issues

Domain knowledge presented must be coherent, maintain aesthetic flow and not cause *pedagogical surprises* to users (Lawless & Wade, 2006). For this, the system needs to ensure that the user has the prerequisites and subsequent information sequenced in a manner to facilitate learning (Knutov, De Bra, & Pechenizkiy, 2009). As content may originate from multiple sources, this shall lead to challenges in maintaining pedagogical consistency and aesthetic flow.

CCSs like KBS Hyperbook (Henze, & Nejdl, 2001) have utilized constructivist pedagogy in designing the learning systems. Most of the systems have been designed by people with a background in technology and evaluated with a technical bent of mind. Systems need to have the end user at the core of the process of evaluation. Different students learn in different ways. Students may have varied learning preferences for information ranging from concrete vs. abstract to hands on vs. reflective, to visual vs. verbal, adaptation to these requirements is going to be challenging. OCSs to be successful shall need to incorporate successful learning theories and pedagogy (Mulwa et al., 2010). Open corpus has the pool of resources to cater to this diversity of learners/students but providing personalized learning is no trivial task.

All domains may not easily lend themselves to the application of techniques like machine learning for the adaptation of OCRs. For example, literary texts have high entropy and can be quite difficult to determine concepts for learning (Levacher et al., 2009). In addition, there are known problems posed by the richness of natural language such as *ambiguity* and of inferring *contextual semantics*. Improvements in utilizing OCRs shall mirror progress with concurrent developments in natural language processing tools and techniques.

A well-written textbook commences with a high *novelty* initially since the reader lacks any background knowledge about the domain in the book. Over the following chapters, the novelty decreases as the reader becomes familiar with the domain being discussed (Lin & Brusilovsky, 2011). A well-designed learning system utilizing OCRs also needs to ensure that novelty gradually tapers. Failing to maintain gradual novelty decline may result in the user being presented with content which he/she may find it difficult to comprehend. A gradual reduction in novelty along with proper conceptual flow to facilitate learning can be a major hurdle in implementation.

Structuring an eLearning system to maintain gradual decline in novelty can benefit from Lev Vygotsky's insight (see Fig. 3) about the socio-cultural (*collaborative*) context to learning known as the zone of proximal development (ZPD). ZPD's implications for the development of an eLearning system are that, to achieve maximum success in learning, tasks need to be neither too easy resulting in boredom nor too difficult leading to frustration but just challenging enough to enable the student to complete the tasks with the assistance of teachers or peers (Wass & Golding, 2014).

Scaffolding in eLearning environments using OCRs in terms of guidance, problem-solving support, feedback, peer interaction, social collaboration offers a possible way to achieve the ZPD.

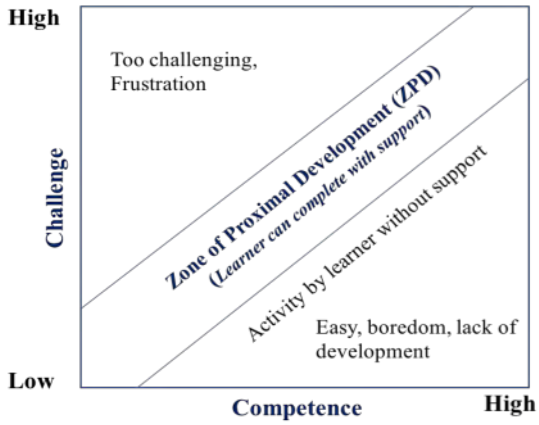


Fig. 3. Lev Vygotsky's insight to facilitate learning in the zone of proximal development

To maintain a proper *conceptual flow* in learning, inputs from Mihaly Csikszentmihalyi's theory, regarding the concept of flow are worth considering. In the *state of flow* (zone) a person experiences profound enjoyment, high concentration and a complete state of pleasure while immersed in an activity at hand. Activities that are neither too simple nor too difficult aid the students to remain in the zone. As an implication, activities for students need to be designed in a manner that they have a realistic chance of completing, by providing instantaneous feedback and with clearly defined objectives (Csikszentmihalyi, 2014). In CCSs or MCSs the construction of the eLearning application is majorly a manual effort and thereby implementing activities to ensure students remain in the zone is quite feasible. But in systems using OCRs ensuring that learning environment is structured to ensure that the learner remains in the flow is going to be challenging and requires study.

AHSs maintain user model in an application using tests, observations, expert opinions, and self-appraisals (Knutov, De Bra, & Pechenizkiy, 2009). In case the system decides to use tests for determining user knowledge, the system designed should be capable of conducting a test which adequately determines user knowledge of the domain from the corpus resources itself. Failing to maintain the test scope with reference to learning objects used by the user shall render the system to provide faulty evaluations and consequent user dissatisfaction.

5.4. Access barriers and domain neutral solutions

It is vital that developing nations merely do not turn out to be consumers of OCRs but also contribute significantly to the advancement of knowledge. As of now most of the resources are contributed by developed nations and also in internationally dominant languages like English. Lack of resources in local languages coupled with relevance to *cultural contexts* can form a significant barrier to adoption (Atenas & Havemann, 2014).

There are many adaptive educational hypermedia systems in use, but almost all of them are built from scratch. This can be explained by the lack of *interoperability* and functional *re-usability* (Brusilovsky, Kobsa, & Nejdl, 2007). It would be necessary that

generic and widely accepted service-oriented models are agreed upon to facilitate interoperability and reuse.

The systems discussed (Kravčík & Wan, 2013; Steichen et al., 2009; Sosnovsky et al., 2012) have a process which is customized to specific problem types or domains. Enlarging the scope to handle diverse knowledge domains is needed. Also, the systems have been so designed that they require teachers who have necessary technical expertise to use them. In order to gain wider acceptance, it is of paramount importance to have simpler mechanisms for utilizing OCRs from basic schooling to all other levels, without significantly compromising on the quality of the output.

5.5. Scaffolding informal learning

The paper approaches learning through an AHSs model, but value offered by informal learning merits mention. Consider a scenario, where a student is using an AHSs application to supplement his/her learning. It is quite possible the student fails to grasp a concept or requires further clarification on a particular problem. Most likely in the current networked world, the student shall search, browse Wikipedia and/or run through some video(s) on YouTube (Hao, Barnes, Branch, & Wright, 2017). An inference from the above discussion is that no learning system is complete and infallible (Labrović, Bijelić, & Milosavljević, 2014). Informal learning through personal learning environments (PLEs) inevitably adds value to an eLearning application. Among significant motivations for informal learning are learner control over learning, learner-centered process, availability of abundant resources, freedom of choice, peer support, and engagement (Song & Bonk, 2016). Table 3 provides an overview of the tools used in PLEs and the challenges for integration of PLE support.

Table 3
Popular tools and challenges to personal learning environments

Popular tools used in PLEs	Challenges to PLEs / Scope for Research
Search: Google, Yahoo, Google Scholar **Visual Media**: YouTube, Khan Academy, Slide share **Networking**: Facebook, Twitter **RSS**: Wikis, Blogs, Delicious **Communication**: Skype, Email, WhatsApp	• Structuring PLEs in terms of guidance towards quality resources, personalization (search, content, appropriateness to user goals), reducing information overload, developing custom dashboards to manage learning resources, annotation support with respect to resource relevance and usefulness. • The possibility of providing context-sensitive interface to PLE tools on the web with regard to a user model.

5.6. Learning theories and learning analytics

For learning to be effective, it is imperative that design of an eLearning system is rooted in well regarded educational theories. A good amount of eLearning design is based on the classical theories of learning namely Behaviorism, Cognitivism, and Constructivism.

Most of the eLearning systems have not given the required attention to learning theories. A few which have attempted have mainly used Constructivism (Henze & Nejdl, 1999, 2001) to enable learners to discover meaning through activities. The rapid proliferation of technological tools, the necessity of lifelong learning, knowledge from multiple sources, non-linear knowledge acquisition, reduced knowledge lifecycle, and networking has influenced the development of *Connectivism* theory of learning (Siemens, 2005).

For the learning objectives to be attained and for learning to be effective a feedback mechanism is essential to provide inputs to various stakeholders (designers, teachers, users) of the system. Various stakeholders interact with the system generating invaluable data for analysis. Learning analytics provides invaluable feedback which can be utilized to improve learning outcomes and meeting learning objectives. Tools of data mining and/or machine learning are being used for log analysis and interactive behavior with the system (Agudo-Peregrina, Iglesias-Pradas, Conde-González, & Hernández-García, 2014).

In an eLearning system (CCSs & MCSs) where the design and operation are under the complete control of the developers applying learning theory and tools of learning analytics is feasible with some effort. Implementation of OCSs adhering to educational theories of learning and providing feedback using learning analytics shall pose challenges not currently addressed in any research paper.

AHSs are not a silver bullet to resolve all the problems pertaining to eLearning. Areas AHSs require a significant amount of effort and work is in the development of tools for collaborative learning, problem-solving support and harnessing social media tools for enhanced learning outcomes (Brusilovsky, 2012). Since we can rule out a significant manual effort in incorporating OCRs, it is but inevitable that a computational model using AI tools and techniques shall play a major role in the solution of OCRs research problems (Brusilovsky, Kobsa, & Nejdl, 2007).

6. Conclusion

There is a copious volume of learning resources on the web. Thereby, the issue is not one of availability, but one of personalized access. The paper commenced with a discussion on the functionality of traditional closed corpus AHSs, followed by mixed corpus AHSs, and finally systems exclusively utilizing OCRs. This was followed by a critical analysis of the vast challenges and research opportunities offered by OCRs for personalized learning.

Implementation of OCSs can bridge the knowledge divide especially for the deprived sections of the society in a cost-effective manner. In the coming time research on open corpus systems are bound to increase and speed up due to parallel developments in artificial intelligence techniques like machine learning, natural language processing, and the semantic web.

The vast number of challenges in the development of open corpus system implies that it is not feasible to work on all the issues at one go. Research challenges need to be addressed in a sequential manner. A vast majority of students in developing nations lack access to educational resources due to economic and social barriers. Simplified access to learning resources would be hugely beneficial to them. In the first stage of research, the authors shall attempt to develop a generic framework for personalized learning using OCRs to work in broad domains rather than any targeted prototypes. Focus shall be on providing access to systems for even teachers without a significant background in

technology to harness the power of OCRs. Research trends indicated by state-of-the-art papers points to research moving in the direction of using semantic web technologies, machine learning tools, and NLP techniques.

There is little doubt OCRs have the potential to enhance learning beyond the classroom. The question is how much OCRs can be harnessed using the current state of knowledge?

Acknowledgements

The authors would like to express their profound gratitude to the chancellor of Graphic Era Hill University, Dr. Kamal Ghanshala for providing excellent ambiance for research. In addition, authors are also highly indebted to Dr. R.C. Joshi former Professor and Head Department of Electronics and Computer Engineering at the prestigious Indian Institute of Technology (IIT), Roorkee and now Chancellor of Graphic Era University for his invaluable suggestions and guidance. Finally, the authors are indebted to the anonymous reviewers for their effort in meticulously reviewing our manuscript and for their valuable comments, insight's and suggestions.

References

Agudo-Peregrina, Á. F., Iglesias-Pradas, S., Conde-González, M. Á., & Hernández-García, Á. (2014). Can we predict success from log data in VLEs? Classification of interactions for learning analytics and their relation with performance in VLE-supported F2F and online learning. *Computers in Human Behavior, 31*(4), 542–550.

Atenas, J., & Havemann, L. (2014). Questions of quality in repositories of open educational resources: A literature review. *Research in Learning Technology, 22*: 14.

Bayomi, M. (2015, August). A framework to provide customized reuse of open corpus content for adaptive systems. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media* (pp. 315–318). ACM.

Brusilovsky, P. (1998, August). Adaptive educational systems on the world-wide-web: A review of available technologies. In *Proceedings of Workshop on WWW-Based Tutoring – the 4th International Conference on Intelligent Tutoring Systems (ITS'98)*. San Antonio, TX.

Brusilovsky, P. (2001). Adaptive hypermedia. *User Modeling and User-Adapted Interaction, 11*(1/2), 87–110.

Brusilovsky, P. (2008, July). Adaptive navigation support for open corpus hypermedia systems. In *Proceedings of International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems* (pp. 6–8). Springer Berlin Heidelberg.

Brusilovsky, P. (2012). Adaptive hypermedia for education and training. In P. Durlach & A. Lesgold (Eds.), *Adaptive Technologies for Training and Education* (pp. 46–68). Cambridge: Cambridge University.

Brusilovsky, P., Chavan, G., & Farzan, R. (2004, August). Social adaptive navigation support for open corpus electronic textbooks. In *Proceedings of the International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems* (pp. 24–33). Springer Berlin Heidelberg.

Brusilovsky, P., Kobsa, A., & Nejdl, W. (Eds.). (2007). *The adaptive web: Methods and strategies of web personalization*. Springer Science & Business Media.

Brusilovsky, P., & Rizzo, R. (2002). Using maps and landmarks for navigation between closed and open corpus hyperspace in Web-based education. *New Review of*

Hypermedia and Multimedia, 8(1), 59–82.

Brusilovsky, P., Schwarz, E., & Weber, G. (1996, June). ELM-ART: An intelligent tutoring system on World Wide Web. In *Proceedings of the International Conference on Intelligent Tutoring Systems* (pp. 261–269). Springer Berlin Heidelberg.

Csikszentmihalyi, M., (2014). Toward a psychology of optimal experience. In M. Csikszentmihalyi (Ed.), *Flow and the Foundations of Positive Psychology* (pp. 209–226). Springer Netherlands.

De Bra, P., & Calvi, L. (1998). AHA! An open adaptive hypermedia architecture. *New Review of Hypermedia and Multimedia, 4*(1), 115–139.

De Bra, P., Houben, G. J., & Wu, H. (1999, February). AHAM: A Dexter-based reference model for adaptive hypermedia. In *Proceedings of the tenth ACM Conference on Hypertext and Hypermedia* (pp. 147–156). ACM.

Dolog, P., Henze, N., Nejdl, W., & Sintek, M. (2004, August). The personal reader: Personalizing and enriching learning resources using semantic web technologies. In *Proceedings of the International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems* (pp. 85–94). Springer Berlin Heidelberg.

Hao, Q., Barnes, B., Branch, R. M., & Wright, E. (2017). Predicting computer science students' online help-seeking tendencies. *Knowledge Management & E-Learning, 9*(1), 19–32.

Henze, N., & Nejdl, W. (1999, May). Adaptivity in the KBS hyperbook system. In *Proceedings of the 2nd Workshop on Adaptive Systems and User Modeling on the WWW*.

Henze, N., & Nejdl, W. (2001). Adaptation in open corpus hypermedia. *International Journal of Artificial Intelligence in Education, 12*(4), 325–350.

Knutov, E., De Bra, P., & Pechenizkiy, M. (2009). AH 12 years later: A comprehensive survey of adaptive hypermedia methods and techniques. *New Review of Hypermedia and Multimedia, 15*(1), 5–38.

Kravčík, M., & Wan, J. (2013, October). Towards open corpus adaptive e-learning systems on the web. In *Proceedings of the International Conference on Web-Based Learning* (pp. 111–120). Springer Berlin Heidelberg.

Labrović, J. A., Bijelić, A., & Milosavljević, G. (2014). Mapping students' informal learning using personal learning environment. *Management, 19*(71), 73–80.

Lawless, S., Hederman, L., & Wade, V. (2008, July). OCCS: Enabling the dynamic discovery, harvesting and delivery of educational content from open corpus sources. In *Proceedings of the Eighth IEEE International Conference on Advanced Learning Technologies* (pp. 676–678). IEEE.

Lawless, S., & Wade, V. (2006, June). Dynamic content discovery, harvesting and delivery, from open corpus sources, for adaptive systems. In *Proceedings of the International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems* (pp. 445–451). Springer Berlin Heidelberg.

Levacher, K., Hynes, É., Lawless, S., O'Connor, A., & Wade, V. (2009). A framework for content preparation to support open-corpus adaptive hypermedia. In *Proceedings of the International Workshop on Dynamic and Adaptive Hypertext: Generic Frameworks, Approaches and Techniques* (pp. 1–11).

Levacher, K., Lawless, S., & Wade, V. (2012, September). Slicepedia: Automating the production of educational resources from open corpus content. In *Proceedings of the European Conference on Technology Enhanced Learning* (pp. 407–412). Springer Berlin Heidelberg.

Lin, Y. L., & Brusilovsky, P. (2011, July). Towards open corpus adaptive hypermedia: A study of novelty detection approaches. In *Proceedings of the International Conference on User Modeling, Adaptation, and Personalization* (pp. 353–358). Springer Berlin Heidelberg.

Mosharraf, M., & Taghiyareh, F. (2016). The role of open educational resources in the eLearning movement. *Knowledge Management & E-Learning (KM&EL), 8*(1), 10–21.

Mulwa, C., Lawless, S., Sharp, M., Arnedillo-Sanchez, I., & Wade, V. (2010, October). Adaptive educational hypermedia systems in technology enhanced learning: A literature review. In *Proceedings of the 2010 ACM conference on Information Technology Education* (pp. 73–84). ACM.

Muntean, C. H., & Muntean, G. M. (2009). Open corpus architecture for personalised ubiquitous e-learning. *Personal and Ubiquitous Computing,13*(3), 197–205.

Schwarz, E., Brusilovsky, P., & Weber, G. (1996). World-wide intelligent textbooks. In *Proceedings of the ED-TELECOM'96 - World Conference on Educational Telecommunications* (pp. 302–207). AACE.

Siemens, G. (2005). Connectivism: Learning as network-creation. *ASTD Learning News, 10*(1), 1–28.

Song, D., & Bonk, C. J. (2016). Motivational factors in self-directed informal learning from online learning resources. *Cogent Education, 3*(1): 1205838.

Sosnovsky, S., Hsiao, I. H., & Brusilovsky, P. (2012, September). Adaptation "in the Wild": Ontology-based personalization of open-corpus learning material. In *Proceedings of the European Conference on Technology Enhanced Learning* (pp. 425–431). Springer Berlin Heidelberg.

Steichen, B., Lawless, S., O'Connor, A., & Wade, V. (2009, June). Dynamic hypertext generation for reusing open corpus content. In *Proceedings of the 20th ACM Conference on Hypertext and Hypermedia* (pp. 119–128). ACM.

Vikram, S., Fan, Y., & Gu, G. (2011, December). SEMAGE: A new image-based two-factor CAPTCHA. In *Proceedings of the 27th Annual Computer Security Applications Conference* (pp. 237–246). ACM.

Wass, R., & Golding, C. (2014). Sharpening a tool for teaching: The zone of proximal development. *Teaching in Higher Education, 19*(6), 671–684.