Harvesting metadata from Open Educational Resources for semantic annotation of online educational content

Vijendra Kushwaha
vijendr@iitk.ac.in

T.V Prabhakar
tvp@iitk.ac.in

Abstract

The massive and ever increasing amount of information on the web has made it difficult to perform a faster and more relevant "search and discovery" for educational content. The current keyword-based search model fails to take into consideration the relevance of an educational resource from various perspectives. Tagging the online educational resources with metadata allows much faster and more accurate searching. The Semantic Web community and the Learning Resource Metadata Initiative (LRMI) have come up with a list of properties suitable for tagging educational resources. However, the challenge is the absence of a vocabulary for possible values of LRMI-recommended properties. In this work, we propose a method for building an educational vocabulary for online learning resources such as OERs, by harvesting and collating metadata from multiple open educational contents. The novelty of this method, over other works, lies in using metadata from educational resources already tagged by education community, which is indicative of the usefulness of the metadata. We further propose a semi-automatic framework for tagging online educational resources with values for the LRMI-recommended properties. This provides an innovative tool that can be used by educators and students alike, for creating and consuming learning resources enhanced with metadata, for mass adoption of OERs.

1. INTRODUCTION

1.1. BACKGROUND

The recent revolution in online education space, such as MOOCs and OERs, coupled with innovations in technological space, such as smartphones and mobile communication, have immensely eased the process of creating and sharing content (Collins & Halverson, 2018) (Altbach, Reisberg, & Rumbley, 2019). There has been a massive upsurge in the amount of information available on the web and the number of educational resources available on the web has risen dramatically in the last decade. This has created difficulties for users to do faster and more relevant "search and discovery" for educational content.

OERs, or Open Educational Resources, are digital learning resources freely and openly available on the Internet, for educators, students and self-learners to use and reuse for teaching, learning and research. They can help faculty members by reducing the amount of time and work they have to dedicate towards preparation of lectures. (Wenk, 2010).

The suggested methods for locating an appropriate resource is by combining multiple search strategies over many different platforms (Butcher, 2015). More than half of the faculty members using or aware of OERs, have cited the lack of a comprehensive catalog and difficulty of finding appropriate material as the most important barrier to using OERs (Allen & Seaman, 2014). The faculty and students, depending on Google-type searches for finding OERs, are increasingly demanding search granularity (Carey & Hanley, 2008).

Search is fundamental for finding a relevant educational resource. Users employ search engines such as Google, DuckDuckGo, Bing, Yandex etc. to query in the *keywords* that are related to the resource in context. Search engines maintain a mapping between keywords and the documents of their occurrence (Croft, Metzler, & Strohman, 2010). Using this mapping, search engines return a list of documents that match the keywords present in the user query; which is usually ranked by the order of relevance of the documents to the user query (Zobel & Moffat, 2006).

1.2. MOTIVATION

Search engines suffer from some drawbacks - they provide results that have high recall but low precision, and that they are highly sensitive to keywords or vocabulary (Antoniou & Van Harmelen, 2004). The search engines return hundreds, if not thousands, of results - all containing the keywords from the query (Croft et al., 2010). From these results, the users have to manually filter out the relevant documents that match their need in a particular context (Mohan & Brooks, 2003). This activity is time-intensive and laborious, and also does not scale.

The general purpose search engines, that many OERs locators use (Butcher, 2015), fail to take into consideration the nuances of the learner's requirement for a precise learning object, arising within particular context of the learner - such as the relevance of an educational resource from the perspective of its *type, objective, use, educational level, subject or domain, affiliation to an educational authority, certification and many other important factors*.

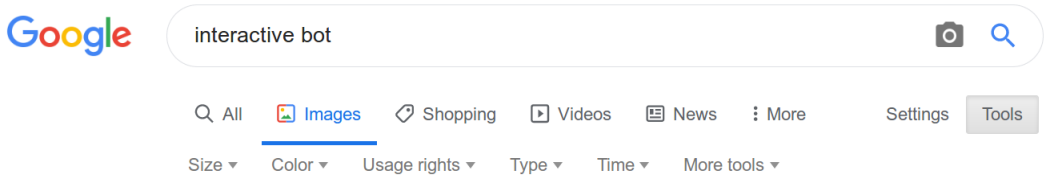

Figure 1: Google Image search provides tools for precisely specifying some of the features of an image

As an example, consider an individual, working in industry and wanting to self-educate oneself in the subjects of *Data Science* and *Machine Learning*, in order to upgrade their skills; with the intention of securing an employment in these domains. They can put in a query into an AI-enabled personal agent, such as *IBM Watson* or *Apple Siri*, inquiring for an upcoming MOOC or an open course work. Further, the individual might also want the coursework to help them match with requirements in multiple job profiles in these domains; and check whether their current work profile, academic background, skill set, and knowledge of technology are aligned with the job profiles.

The individual might have a very particular requirement for courses - *offered by an accredited university, for online distance learning, within a certain number of hours required per week and total duration of the course, active discussion forums where teaching assistants frequently reply to student queries, whether the course has a project and exams, number of assignments and if it clashes with their schedule, whether video lectures have subtitles in Spanish language, if the lecture presentations are annotated, whether the individual satisfies the pre-requisites for the course, whether certificate of achievement are given out on course completion* etc.

While the current generation of keyword-based search engines may be able to search within the contents of the documents, they cannot capture the multiple contextual aspects that this query is requesting. Answering these requests requires an understanding of both, the content and context of the educational resources. McGreal, Kinuthia, Marshall, and McNamara, 2013 states that "unless

OER are consistently and adequately described, they cannot easily be located in online searches". Building this understanding will allow more granular queries to be performed by the users for locating the relevant resources.

Building systems that allow users to quickly converge on the most relevant resource using detailed description of its properties, will save user's time and reduce exhaustion from manually filtering results. It enables them to focus their cognitive faculties towards learning. Further, this opens up directions in personalized learning.

1.3. EXPLORING THE PROBLEM SPACE

Currently, the machines have no understanding of the content they are serving. The current design of the web and the associated practices, do not support the method for building context around the educational resources. However, it is possible to build an understanding of a learning object. One way to achieve this is by describing any educational resource as a precise set of features relevant to the context of education. This feature set of the learning object is derived from the content that it carries and the context that it is set in.

Berners-Lee, Hendler, Lassila, et al., 2001 identified this central problem and proposed a new architecture for the web, known as **semantic web**. They recognized the requirement for knowledge representation as solution to build semantic web using two things - structured collections of information and sets of inference rules for reasoning with the structured information. They further proposed a third component of the solution, **collections of information called ontologies**. Ontologies can incorporate the structured information and also provide methods of logical reasoning between relationships.

In the space of learning resources, Robson, 2001 identified the same problem and came up with a solution of showcasing the features of a learning resource by attaching pedagogic metadata for faster searching and discovery. The purpose of metadata is to facilitate search and discovery of relevant information by "allowing resources to be found by relevant criteria, identifying resources, bringing similar resources together, distinguishing dissimilar resources, and giving location information" (Press, 2004).

This could be achieved either by attaching metadata tags with the files, or embedding metadata tags within the webpage as HTML tags. Searching for a keyword in educational resources tagged with metadata is patently faster and more accurate over searching through entire document. Fischer, 2001 provide work that supports using metadata based methods for faster and more accurate retrieval of educational content. They further the work by using metadata standards such as Dublin Core and providing ways for knowledge management.

Mohan and Brooks, 2003 identifies the problem as central to the domain of learning and education, and plots methods to engineer for the future of web-based learning, that is grounded in metadata and their standards. This is further taken up by Roy, Sarkar, and Ghose, 2008 where they work on automatic extracting educational metadata from resources.

However, the recognition of metadata as a solution brought many metadata standards for education over the years (IEEE Learning Object Metadata (ieeexplore.ieee.org/document/1032843), Dublin Core (dublincore.org/), ISO Metadata for Learning Resources(www.iso.org/standard/60613.html)). Yet, there are no clear criteria to decide which schema is better for a particular case (Barker & Campbell, 2010). Kurilovas, 2009 identify this problem and provide methods for interoperability between the metadata standards, the number of which keep increasing and starting to become a

problem of which standard to follow.

Property	Description
educationalAlignment	An alignment to an established educational framework
educationalUse	The purpose of the work in the context of education. Ex. "Assignment", "Group Work"
useRightsUrl	The URL where the owner specifies permissions for using the resource.
timeRequired	Approximate or typical time it takes to work with or through this learning resource for the typical intended target audience.
typicalAgeRange	The typical range of ages the contents intended end user.
interactivityType	The predominant mode of learning supported by the learning resource. Acceptable values are active, expositive, or mixed.
learningResourceType	The predominant type or kind characterizing the learning resource.
useRightsUrl	The URL where the owner specifies permissions for using the resource.
isBasedOnUrl	A resource that was used in the creation of this resource. This term can be repeated for multiple sources.

Table 1: Education specific properties that LRMI added to Schema.org

Finally, in 2014, the Semantic Web community and the Learning Resource Metadata Initiative (LRMI) came up with a list of properties suitable for tagging educational resources. Learning Resource Metadata Initiative (LRMI) was a group that created a metadata extension that build on the works of Schema.org (schema.org) (Barker & Campbell, 2014). The purpose of LRMI is to support end-user search and discovery of educational resources. LRMI has developed a common metadata framework to provide rich, education-specific metadata properties for describing educational resources on the web. However, the absence of an *agreed-upon* common vocabulary for possible values of LRMI recommended properties is a barrier to mass adoption.

Summarily, at the core, the challenge that remains is the inability to converge to a standard set of features using which we could describe the educational resources. The many scholarly works, as described above, have tried to come up with standard vocabulary which allows inter-operability between different contexts, but remain far from adoption by the community.

1.4. PROPOSED SOLUTION

A solution lies in coming up with a metadata vocabulary which is already is use by the education community. Since the problem with metadata standards is lack of flexibility and poor adoption across communities with different needs, the solution should also provide a way for adding user-specific vocabulary words that they could use within their communities. This idea is akin to social annotations - folksonomy, which is an act of tagging done by the consumer of the information (Xu, Bao, Fei, Su, & Yu, 2008).

In this work, we propose a method for building an educational vocabulary for OERs, by harvesting and collating metadata from multiple open educational content. By using the already existing metadata values, we ensure that the vocabulary is in use by the community and is not artificially created. We also bypass the need for a domain expert to come up with a list of keywords used for describing educational resources.

We further organize this collated metadata vocabulary into taxonomical structures using the ontological structure provided by dictionaries such as WordNet or online encyclopedias such as

Wikipedia. Again, by using the already existing infrastructure for finding relationships between concepts, we bypass the need for a domain ontologist or taxonomist.

In this work, we will find the values for the '**educationalUse**' field in the LRMI specification. The 'educationalUse' field has been briefly described as 'The purpose of the work in context of the education' (see table 1). This is an important field for those stakeholders of the education domain, searching for resources by their use or the function the resource will serve.

2. METHODOLOGY

2.1. BUILDING VOCABULARY

2.1.1. Metadata Harvesting

To be searchable and discoverable, OER repositories must comply with the standards. The repositories must exchange the descriptions of the educational resources in an interoperable manner. The means of exchange of the metadata records is provided by Open Archives Initiative Protocol for Metadata Harvesting OAI-PMH.

The Open Archives Initiative (OAI) has given a method for inter-operable metadata exchange, known as Protocol for Metadata Harvesting (PMH). It is a well-defined protocol that allows *data providers*, such as educational institutions, to open up their educational resources as a *repository*, by exposing the necessary metadata on the resources in various formats. The exposed metadata can be harvested by *service providers* using a number of methods provided by the protocol, with the goal of developing services that enable search and discovery of educational resources (Lagoze & Van de Sompel, 2003) (Warner, 2001). The individual resources are disseminated as '*records*' within these repositories.

The following metadata fields are associated with each record in each of the repositories: *title, creator, subject, description, publisher, date (of creation of the work), type, format, identifier, source, language, relation, content rights*. We focused on the **educational resource type**, to collect all the instances of possible values for this field. In terms of numbers, we went through 2200 repositories, parsing a total of 450,677 records. From these, we obtained 3356 unique values for the '**resource type**' field. However, this number was reduced to 1632 by manually removing identifiable spurious, incorrectly added or near-duplicate values.

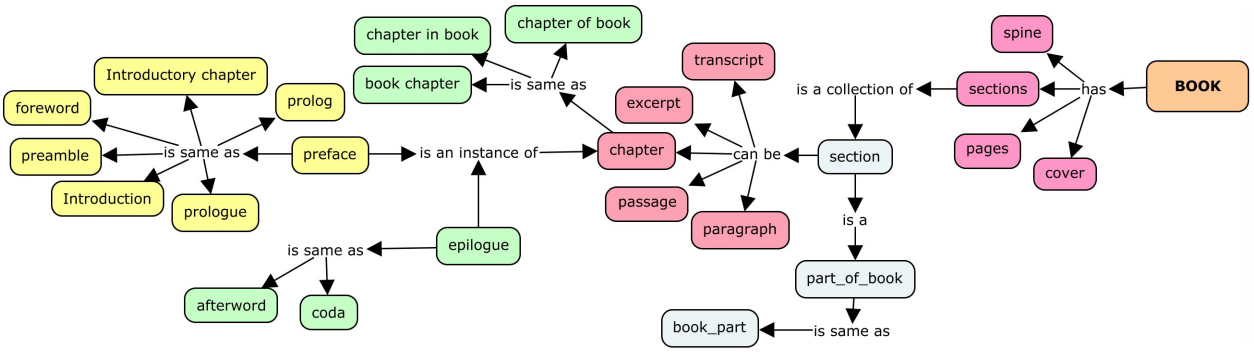

Figure 2: A concept map for parts of a book

2.1.2. Building List of Keywords

Not all 1600 terms so obtained, by above described process, were unique representations of concepts. ***For example***, *chapter in a book* was found to be represented as *'bookchapter'*, *'book-chapter'*, *'chapter in book'*, *'chapter of book'*; in Italian language this was represented as *'chapitre de livre'*.

Some syntactic variations on *chapters in a book* were present as *'bookchapters'*, *'book-chapters'*, *'book_chapters'*, *'book chapters'*, *'chapters_in_book'* and so on. Also, 'chapters' as a plural form of *'chapter'* was also taken into account. Another variation, semantic in nature, was representation of chapter as *'part of book'*. Both *'spine'* and *'cover'* are physical parts of book yet they are not same as *'chapter'*. Figure 2 tries to capture this idea; it shows the relationships between the components of a *'book'* object.

As we can see that there is a lot of ambiguity concerning the educational resource 'book chapter'. This is where the role of ontology or taxonomy comes into play. By charting out the relationships between concepts, we can arrive at a hierarchical network of concepts that will make them consumable by semantic agents to provide faster and more accurate search.

2.1.3. Finding Concept Hierarchies

We established the relationships between the concepts, extending some of the concepts obtained from the work through Metadata Harvesting, by using open publicly available information from **WordNet**, **Wikipedia** and other free online thesauri. Many of the relationships that have been established in this work are of broader term (captured by *'can be'*) and narrower term (captured by *'is an instance of'*); or that of synonyms (captured by *'is same as'* or *'is similar to'*).

For example, a *'reference book'* is known to users (in English language) as either of the words - *'reference'*, *'reference work'* or *'book of facts'*. Therefore, a search for a 'reference book' should include, in its results, all the documents that contain all the synonymous keywords (see figure 3)

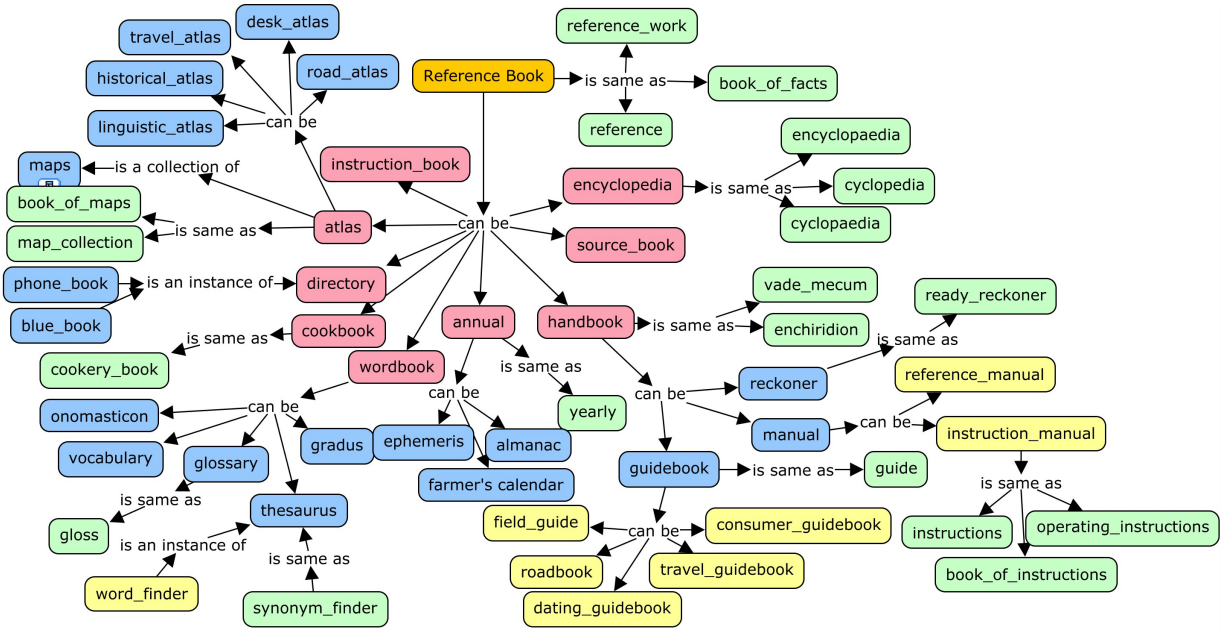

Figure 3: A concept map for Reference Work, showing its relations with other book concepts

2.2. FRAMEWORK FOR AUTOTAGGER

In this section, we propose a semi-automatic framework for tagging online educational resources with the possible values for LRMI-recommended semantic properties. For this we used heuristics to extract possible metadata values from the contents of online open educational resources. By identifying the semantics associated with these resources we are able to make them available for more accurate and faster search.

2.2.1. Identification of Resource Use

In an educational setting, a resource can have some of the following uses: *lecture, homework, reading, assignment, project, research, game, problem solving, solved problems, laboratory, presentation* etc. These uses will have an objective to achieve, along with a structure that is necessary for the resource to be called so and a format in which they are presented to a user.

For example, a *lecture* is used by a person with the objective to impart knowledge or convey ideas to an audience pertaining to a topic, usually in a classroom setting. Lectures can be imparted in form of a traditional person-to-person, a video recording or live video feed, or as lecture notes or presentation slides. As to the structure, lectures will have an lecturer, a subject or a topic, some audience, a time duration in which the lecture will convene, the date and location where lecture happens and so many other such properties.

2.2.2. Semantics of resources

As each of the educational resource has a structure that it embodies, this information about the structure of the resource could be used to identify the type of resource in consideration. When an educational resource is created, certain semantics have to be included into the making of that resource, for a resource to be called so. For example, a homework assignment would usually carry the submission due date.

We identify the resources based on the purpose they serve in an educational setting, like in a classroom setting, and some of the structural components within the document. The figure 4 shows how many values *resource object use* can take up.

2.2.3. Extracting metadata from documents

In this section, we detail a method for identifying and extracting educational metadata from Portable Document File (PDF). The portability across different platforms and immutability to modifications, make PDF a popular means for exchanging information. Like most other documents present on web, they are meant for human consumption. We used the observation that keywords of importance, such as headings or titles etc, are usually formatted to appear larger, emphasized, bold, underlined, placed distant from other text etc.

Using this observation, we built some heuristic approaches to extract information of our necessity from online PDFs, particularly used in educational settings of school or universities. These are **assignments, homework, exams, solutions to assignment or exam, lecture notes or study notes, quiz, lab work and tests**.

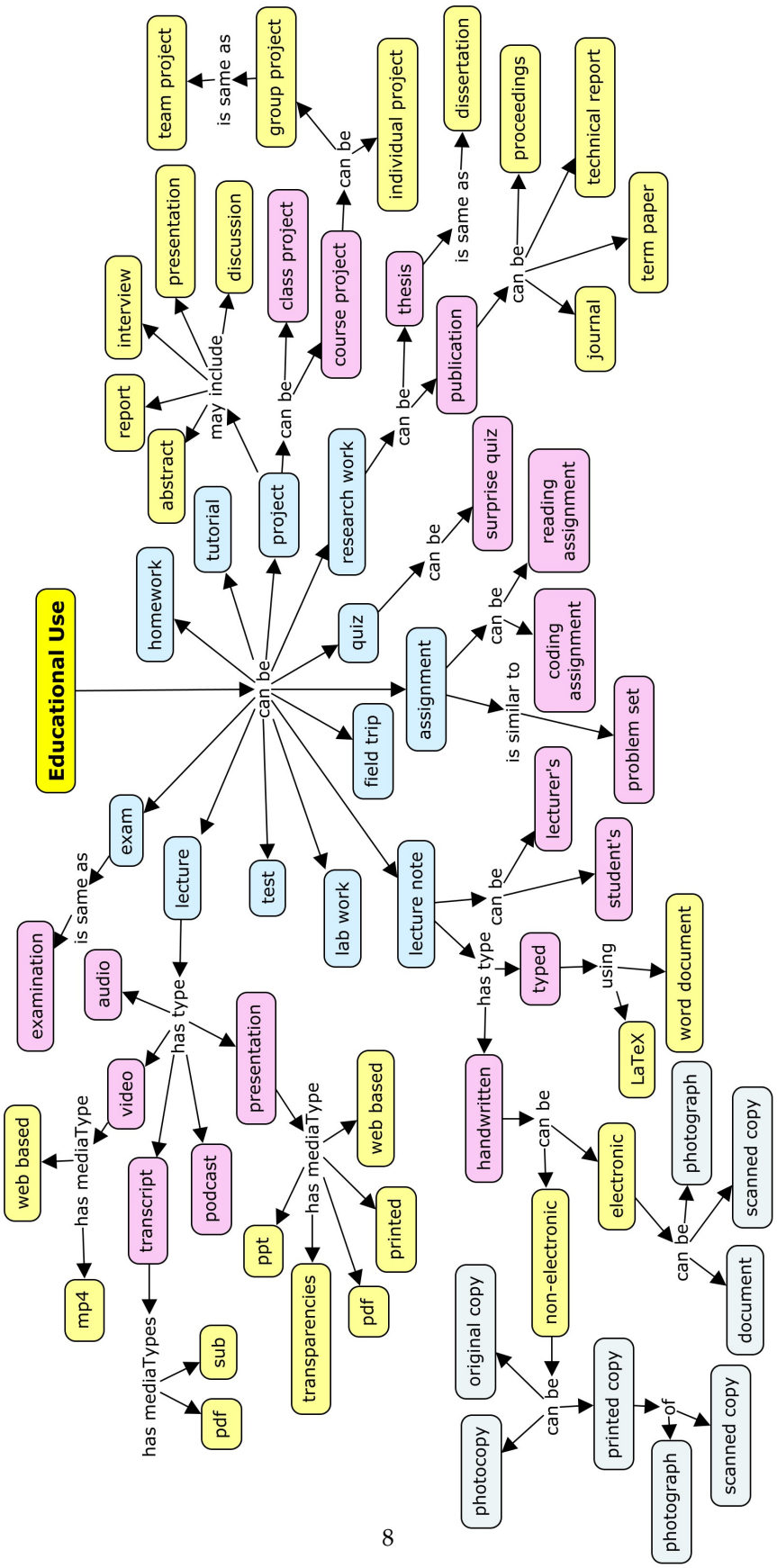

Figure 4: A concept map for LRMI property educationalUse.

3. CONCLUSION AND FUTURE WORK

In this work, we reinforce the advantages of using metadata to enable faster and more accurate "search and discovery" of educational resources by tagging them with appropriate values. Metadata provides methods for introducing more granularity when describing learning resources. Given the immense number of open educational resources available on the web and the increasing difficulty in locating relevant learning resource, we identified this as a problem central to education; the solution to which is keenly sought out by the educators, students and self-learners.

Instead of coming up with new standards for metadata exchange, we banked on the recent efforts by semantic web community and Learning Resource Metadata Initiative who have provided education specific metadata properties. However, this work is only one side of the coin; the other side being the tags that are values for these metadata properties.

This absence of an *agreed-upon* common vocabulary for possible values of LRMI recommended properties poses as a barrier to mass adoption. For populating the metadata values, we used metadata from already tagged documents, made available through metadata harvesting using the protocol given by OAI-PMH. We organize the vocabulary list so obtained into a hierarchy of relationships, so that it can enable Semantic web to showcase the understanding of the educational resources. To achieve that we again used the already existing ontological structures provided by WordNet and other thesauri.

Lastly, we built heuristic approaches to identify educational resources by their use in classroom settings, a property captured by LRMI metadata field *'educationalUse'*. We provided a semi-automatic framework for tagging online educational content with the corresponding metadata that was extracted from the content and its context.

For future work, we propose using Deep Learning algorithms for pattern recognition, such as to substitute the heuristics we used to find the mappings between the structure of the resources and their semantics in the "educational resources" space. This work could also be replicated for finding values for other LRMI properties such as *learningResourceType* or *interactivityType*. Another extension to the present work could be the identification of action-verbs from Bloom's taxonomy for resources like assignments, homework and exams. This will allow searching for problems directed towards assessing a cognitive domain of a user, such as analysis or understanding.

REFERENCES

Allen, I. E. & Seaman, J. (2014). Opening the curriculum: Open educational resources in us higher education, 2014. *Babson Survey Research Group*.

Altbach, P. G., Reisberg, L., & Rumbley, L. E. (2019). *Trends in global higher education: Tracking an academic revolution*. Brill Sense.

Antoniou, G. & Van Harmelen, F. (2004). *A semantic web primer*. MIT press.

Barker, P. & Campbell, L. M. (2010). Metadata for learning materials: An overview of existing standards and current developments. *Technology, Instruction, Cognition and Learning*, *7*(3-4), 225–243.

Barker, P. & Campbell, L. M. (2014). Learning resource metadata initiative: Using schema. org to describe open educational resources.

Berners-Lee, T., Hendler, J., Lassila, O., et al. (2001). The semantic web. *Scientific american*, *284*(5), 28–37.

Butcher, N. (2015). *A basic guide to open educational resources (oer)*. Commonwealth of Learning (COL);

Carey, T. & Hanley, G. L. (2008). Extending the impact of open educational resources through alignment with pedagogical content knowledge and institutional strategy: Lessons learned from the merlot community experience. *Opening up education*.

Collins, A. & Halverson, R. (2018). *Rethinking education in the age of technology: The digital revolution and schooling in america*. Teachers College Press.

Croft, W. B., Metzler, D., & Strohman, T. (2010). *Search engines: Information retrieval in practice*. Addison-Wesley Reading.

Fischer, S. (2001). Course and exercise sequencing using metadata in adaptive hypermedia learning systems. *Journal on Educational Resources in Computing (JERIC)*, *1*(1es), 5.

Kurilovas, E. (2009). Interoperability, standards and metadata for e-learning. In *Intelligent distributed computing iii* (pp. 121–130). Springer.

Lagoze, C. & Van de Sompel, H. (2003). The making of the open archives initiative protocol for metadata harvesting. *Library hi tech*, *21*(2), 118–128.

McGreal, R., Kinuthia, W., Marshall, S., & McNamara, T. (2013). *Open educational resources: Innovation, research and practice*. Commonwealth of Learning (COL);

Mohan, P. & Brooks, C. (2003). Engineering a future for web-based learning objects. In *International conference on web engineering* (pp. 120–123). Springer.

Press, N. (2004). Understanding metadata. *National Information Standards*, *20*.

Robson, R. (2001). Pedagogic metadata. *Interactive Learning Environments*, *9*(3), 207–218.

Roy, D., Sarkar, S., & Ghose, S. (2008). Automatic extraction of pedagogic metadata from learning content. *International Journal of Artificial Intelligence in Education*, *18*(2), 97–118.

Warner, S. (2001). Exposing and harvesting metadata using the oai metadata harvesting protocol: A tutorial. *arXiv preprint cs/0106057*.

Wenk, B. (2010). Open educational resources (oer) inspire teaching and learning. In *Ieee educon 2010 conference* (pp. 435–442). IEEE.

Xu, S., Bao, S., Fei, B., Su, Z., & Yu, Y. (2008). Exploring folksonomy for personalized search. In *Proceedings of the 31st annual international acm sigir conference on research and development in information retrieval* (pp. 155–162). ACM.

Zobel, J. & Moffat, A. (2006). Inverted files for text search engines. *ACM computing surveys (CSUR)*, *38*(2), 6.