Citation for published version:

Mouriño-García, M., Pérez-Rodríguez, R., Anido-Rifón, L., Fernández-Iglesias, M. J., & Darriba-Bilbao, V. M. (2018). Cross-repository aggregation of educational resources. *Computers & Education, 117*, 31-49. doi:10.1016/j.compedu.2017.09.014

Accepted Manuscript

Link to published version: https://doi.org/10.1016/j.compedu.2017.09.014

Cross-repository Aggregation of Educational Resources

Marcos Mouriño-García[a], Roberto Pérez-Rodríguez[a], Luis Anido-Rifón[a], Manuel J. Fernández-Iglesias[a,*], Víctor M. Darriba-Bilbao[b]

University of Vigo, Campus Universitario, 36310 Vigo, Spain.

[a]*Dept. of Telematics Engineering, University of Vigo, Spain*
[b]*Dept. of Informatics, University of Vigo, Spain*

Abstract

The proliferation of educational resource repositories promoted the development of aggregators to facilitate interoperability, that is, a unified access that would allow users to fetch a given resource independently of its origin. The CROERA system is a repository aggregator that provides access to educational resources independently of the classification taxonomy utilized in the hosting repository. For that, an automated classification algorithm is trained using the information extracted from the metadata of a collection of educational resources hosted in different repositories, which in turn depends on the classification taxonomy used in each case. Then, every resource will be automatically classified on demand independently of the original classification scheme. As a consequence, resources can be retrieved independently of the original taxonomy utilized using any taxonomy supported by the aggregator, and exploratory searches can be made without a previous taxonomy mapping. This approach overcomes one of the recurring problems in taxonomy mapping, namely the one-to-none matching situation. To evaluate the performance of this proposal two methods were used. Resource classification in categories existing in all repositories was automatically evaluated, and a maximum performance (F_1 score) of 84% was achieved. In

*Corresponding author

Email addresses: marcos@gist.uvigo.es (Marcos Mouriño-García), roberto.perez@gist.uvigo.es (Roberto Pérez-Rodríguez), lanido@det.uvigo.es (Luis Anido-Rifón), manolo@det.uvigo.es (Manuel J. Fernández-Iglesias), darriba@uvigo.es (Víctor M. Darriba-Bilbao)

the case of resources not belonging to one of the common categories, human inspection was used as a reference to compute classification performance. In this case, an F_1 score close to 70% was obtained. These results demonstrate the potential of this approach as a tool to facilitate resource classification, for example to provide a preliminary classification that would require just minor corrections from human classifiers.
Keywords: architectures for educational technology systems, cooperative/collaborative learning, distributed learning environments, media in education

---

1. Introduction

Open educational resources (OER) are educational materials in digital form that are freely available to educators, students and self-learners to be used and re-used in learning, teaching and research (UNESCO, 2002). One of the biggest challenges about OER is access (D'antoni, 2006), that is, how to make potential users aware of the existence of open educational resources to be utilized in their educational projects. Instruments to promote the use of open educational resources and to facilitate access to them were developed, including virtual learning environments, thematic portals, virtual communities, wikis, open magazines, social networks and repositories. Educational resource repositories are the most widespread platform because they offer benefits such as the preservation and reuse of content, permanent access, visibility, and ease of search and retrieval using metadata (Gibbons, 2009). There is a great number and variety of online repositories (Roy et al., 2010) such as OERCommons (Open Educational Resources, 2013), MERLOT (Cafolla, 2006), Open Stax CNX, Edna or Lornet among others (Ternier et al., 2009).

Metadata are the key elements for repositories to represent and organize educational resources. Due to the huge amount of learning objects available, manual generation of metadata is not feasible as a general solution. Note that this would be a resource-consuming process, and metadata created by humans are bound to include errors (Meire et al., 2007). As a consequence, automatic metadata generation techniques were proposed. Although some metadata is relatively straightforward to obtain (e.g., date of creation, original source of the resource) other metadata elements are more complicated to create. Thus, significant efforts have been made in order to automatically generate high quality metadata (Meire et al., 2007; Rodriguez et al., 2009; Broisin et al., 2005).

Each of the repositories above collects resources according to a particular metadata schema. Examples of these schemas (Anido-Rifón et al., 2014) are IEEE LOM, ISO / IEC MLR, or DublinCore Metadata Initiative, as well as several application profiles such as LRE Metadata Application Profile (Massart et al., 2011), Open Discovery Space Application Profile (Niemann et al., 2013) or UK LOM Core (Campbell, 2011) among others.

Due to the large number of existing repositories and providers, several repository alliances, networks or aggregators were deployed to promote the sharing and reuse of educational materials, such as ARIADNE (Ternier et al., 2009), MACE (Boeykens et al., 2009), MELT (Kurilovas and Dagiene, 2009), Edutella (Nejdl et al., 2002), GLOBE, ELENA (Dolog et al., 2004), LRE, Open Discovery Space (Nikolas et al., 2014) or PROLEARN (Wolpers and Grohmann, 2005). The main challenge in repository aggregation is interoperability, that is, a unified and integrated access to the collected resources independently of the underlying metadata scheme or application profile (Stefaner et al., 2007).

Indeed, the heterogeneity of the classification approaches in existing repositories (Dietze et al., 2012) led to the emergence of several methods to overcome it, including the most commonly used metadata mapping techniques. One of the recurring problems is related to the different degrees of equivalence encountered when mapping individual elements, namely one-to-one, many-to-one and one-to-none. As a consequence, in many cases there do not exist exact equivalents between elements, and meaning and scope superpositions occur in some cases (Chan and Zeng, 2006), while in others an existing equivalence will not be found by the mapping engine (Hillmann and Westbrooks, 2004).

Aggregators allow users to search and retrieve resources in different ways.

Simple search is used to fetch resources according to the keywords provided, while advanced search supports resource filtering according to specific values of metadata elements. Finally, browsing or exploratory search enables users to navigate a category tree to access specific elements (Neven and Duval, 2002; Roy et al., 2010).

Despite being a desirable feature, not all aggregators implement exploratory search or browsing, as it would require all resources being classified according to a common set of categories or taxonomy. This becomes an issue when trying to aggregate repositories implementing different taxonomies (e.g., OERCommons and MERLOT). Approaches exist that try to address this situation, such us the definition of a new common taxonomy (Kawase et al., 2013), or the application of ontology mapping, ontology matching or ontology alignment techniques (Doan et al., 2004), which in turn have several drawbacks. For example, they use to be performed manually, a laborious and error-prone process (Doan et al., 2004). To try to overcome this, several authors have proposed the introduction of machine learning (ML) techniques (Sebastiani, 2002) to automatically generate mappings between ontologies (Doan et al., 2004; Nezhadi et al., 2011; Shvaiko, 2013). However, in a similar way to metadata mapping, ontology mapping techniques perform mappings between individual elements, so the different degrees of equivalence mentioned above (i.e., one-to-one, many-to-one, and one-to-none) are not addressed.

The proposal discussed in this paper follows a different approach. While ontology mapping techniques act at the taxonomy level by computing equivalences between different nodes (cf. Fig. 1.a), the proposed solution processes educational resource metadata elements in each of the repositories (e.g., title, description, keywords) to classify the target resource again according to each of the taxonomies in existing repositories (Figure 1.b).

Thus, CROERA (Cross-Repository Open Educational Resources Aggregation) is an approach to the aggregation of repositories that has the features below:

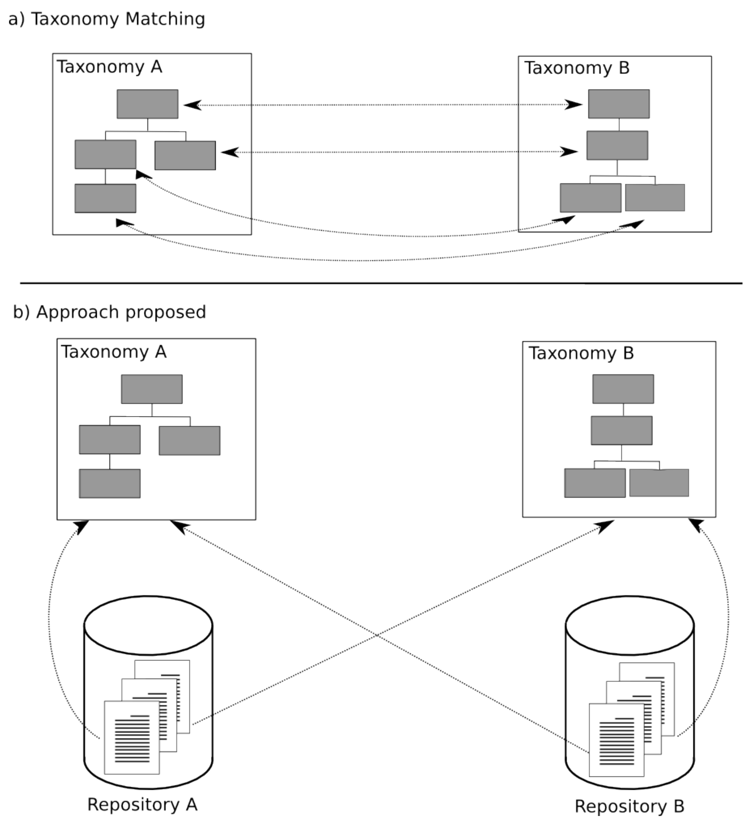

Figure 1: Taxonomy matching vs. the approach proposed.

- It provides access to educational resources independently of the taxonomy used by each of the integrated repositories. The aggregated resources are classified automatically using ML techniques according to each of the taxonomies of the repositories integrated. This allows users to access resources in several repositories no matter their taxonomies using a taxonomy with which the user is most familiar, or that is considered as the most appropriate or useful in that situation.

- It addresses the heterogeneity of categories, thus enabling browsing (i.e., exploratory searching) through any of taxonomies included in the aggregator.

- It does not require any mapping or matching technique for metadata or ontologies, thus avoiding the drawbacks of such techniques, in particular the one-to-none matching situation.

To evaluate the performance of the resource classifier, two approaches were followed. On the one hand, the quality of the classification of the resources classified under common categories is addressed automatically. On the other hand, in order to assess the quality of the classification of those resources not belonging to the set of common categories, human experts were used.

The rest of this paper is organized as follows. Section 2 introduces CROERA and the CROERA system[1], an aggregator of repositories based on CROERA, while Sect. 3 is devoted to the evaluation of the proposed system. Section 4 discusses the results obtained, and Sect. 5 identifies the limitations of this solution. Finally, Sect. 6 offers some concluding remarks together with an enumeration of present and future working lines in relation to this research.

2. Materials and Methods

Existing ontology mapping solutions include GLUE (Doan et al., 2004), a system that takes as input two taxonomies and uses multi-strategy learning and the Naïve Bayes algorithm to compute the distribution of joint probabilities between each pair of nodes in the taxonomies. Then, using these values and a similarity function (Jaccard's or Most-Specific-Parent) the similarity matrix is computed for the nodes in the two taxonomies, which in turn is used for the configuration of the mapping that best meets the restrictions of the domain, providing high levels of accuracy.

Straccia and Troncy (2005) introduced OMAP, a framework that aims to automatically align two ontologies looking for the best mappings between entities defined in the target ontologies, together with their weights. Final mappings are obtained by means of predictions from different classifiers (terminology-based,

[1] http://croera.gist.det.uvigo.es

ML-based, and based on the structure and semantics of ontologies). The main limitations of OMAP are its high computational complexity and that it only addresses ontologies described in the same knowledge representation language.

Nezhadi et al. (2011) trained a classification algorithm (K-Nearest Neighbor (KNN), Support Vector Machines (SVM), Decision Tree (DT) (Yang and Liu, 1999) or AdaBoost) from pairs of ontologies, where the relations between nodes in both ontologies was marked by experts as *Aligned* or *Not Aligned*. Once the training phase is completed, a fresh pair of ontologies is fed the algorithm to decide which nodes of one of the taxonomies are aligned to nodes in the other taxonomy.

Ngo and Bellahsene (2012) proposed the YAM++ tool. Firstly, a classification algorithm (Decision Tree, or Naïve Bayes or SVM) is trained with pairs of ontologies parsed and converted to graphs. Once trained, the tool computes the mappings in a new pair of non-aligned ontologies, also in the form of graphs. Finally, YAM++ displays them to the user through a graphical interface for them to assess the correctness of the mapping according to their knowledge about the domain.

All the aforementioned proposals experience the one-to-none matching problem. Thus, when this situation occurs, elements classified under unmatched categories or nodes will not be accessible from a taxonomy different from the original one. In the case of Straccia and Troncy (2005), it also refers as limitations the high computational complexity and the limitation to taxonomies described in the same knowledge representation language. Section 4 discusses how CROERA solves the limitations above.

The CROERA approach (cf. Fig. 2.b) is quite different from previous proposals no matter it is also ML-based, as it automatically classifies each of the resources collected according to all taxonomies in the repositories aggregated, thus facilitating exploratory search among all resources regardless of the taxonomy used. For this, an ML algorithm is trained with a set of educational resources from different repositories labeled according to the taxonomy of the repository of origin. The information used for training is extracted from the

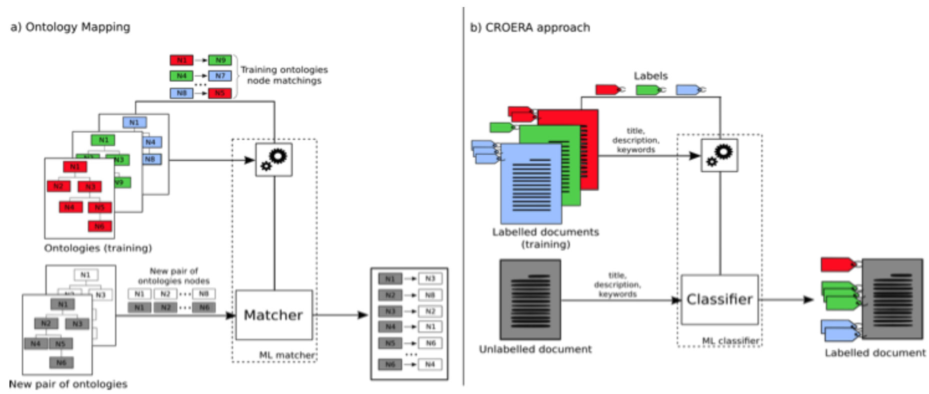

Figure 2: Ontology mapping vs. the approach proposed.

metadata of the resources, and more specifically from their title, description and keywords. After training is completed, a new educational resource will be classified according to each of the taxonomies of the repositories used to train the algorithm.

2.1. Background

The terminology utilized and some background on techniques for automatic document classification are introduced below, together with the classification algorithm selected, the representation of the documents used, and the metrics used for evaluation.

2.1.1. Multi-label Classification and the Support Vector Machines Algorithm

Classification is a typical machine learning technique. In particular, CROERA utilizes classification techniques based on supervised ML. The classifier is trained with several documents whose category is known and then the algorithm is applied to documents whose category is unknown (Sebastiani, 2002). Classification problems are divided into two groups: single-label and multi-label. In the first case, each document is associated to a unique label from a disjoint set of labels. In multi-label classification, each document can be associated to one or more labels from the set of labels (Tsoumakas and Katakis, 2007).

To implement the classifier, the Scikit-learn (Pedregosa et al., 2012) library was used, a suite of algorithms for supervised and unsupervised machine learning providing a rich environment implementing the most relevant algorithms from the state of the art. The one-vs-rest (Hsu and Lin, 2002) strategy was applied, which divides the multi-label classification problem into several single-label problems. This strategy offers a high computational efficiency and interpretability, and can be combined with different classification algorithms.

A large variety of classification algorithms exists, KNN, DT, Neural Networks, Bayesian or SVM (Yang, 1999) among them. For our proposal, the SVM algorithm was selected. SVM is a set of algorithms to solve clustering, regression and classification problems. It was selected because it is one of the most successful machine learning models, with a proven applicability to tasks related to automatic text classification (Rigutini et al., 2005). Besides, it offers a better performance than other relevant state of the art alternatives such as KNN or Naïve Bayes (Yang and Liu, 1999). Given a set of elements belonging to one of two possible categories, the algorithm will construct a model that can predict whether a new element belongs to one category or another (Hearst et al., 1998; Joachims, 1998).

2.1.2. Document representation

The operation of the automatic classifier in this research is based on the application of natural language processing (NLP) techniques to the documents to be classified. A software agent will recognize the category to which a document belongs by analyzing some NLP feature of its content, such as the frequency of occurrence of words or the language structure used (Settles, 2010). The Vector Space Model (VSM) (Salton et al., 1975) is the most often used representation, where each document in a collection is represented as a point in an N-dimensional space, N being the number of distinct features in the collection, using as weights the frequency of occurrence of such features. When words are used as features, the model is known as Bag-of-Words (BoW), and it is one of the traditionally used representations in document classification and information retrieval (Täckström, 2005). A bag or multiset is a set of elements where each element may occur more than once (Blizard, 1988). Thus, in the approach proposed in this paper, a document is represented as a Bag-of-Words collecting all the meaningful words appearing in the title and description of the document, and also the keywords. Unigrams (i.e., individual words) are used as features when they are extracted from the title and description of the educational resource, while n-grams (with $n \geq 1$) are used as features when they are extracted from the keywords of the educational resource (e.g., "Social Sciences" or "History about DNA").

2.1.3. Performance Assessment Metrics

Different metrics exist for evaluating multi-label classification (Tsoumakas and Katakis, 2007; Sokolova and Lapalme, 2009). In our case, precision (P) and recall (R), F_1 score, the area under the receiver operator characteristic curve (A_{ROC} (Hanley and McNeil, 1982)), the area under the precision-recall curve (A_{PR} (Boyd et al., 2013)) and the Cohen's kappa coefficient (κ (Cohen, 1960)) are used.

Let L_T denote the correct set of labels for a given instance, and L_P the set of labels predicted by the classifier. Precision P is defined as the fraction of the labels correctly predicted with respect to the total set of labels predicted:

$$P = \frac{|L_T \cap L_P|}{|L_P|} \tag{1}$$

Recall R is defined as the fraction of the labels correctly predicted with respect to the total set of correct labels:

$$R = \frac{|L_T \cap L_P|}{|L_T|} \tag{2}$$

The F_1 score is a metric commonly used in text classification tasks that combines precision and recall to provide an indication of the global performance (Sokolova and Lapalme, 2009). It is defined as:

$$F_1 = \frac{2 \times P \times R}{P + R} \tag{3}$$

A receiver operator characteristic (ROC) curve is a graphical plot that illustrates the performance of a classifier vs. its discrimination threshold. In other words, ROC curves show how the number of correctly classified positive examples varies with the number of incorrectly classified negative examples. The ROC curve is generated by plotting the true positive rate (TPR) or *sensitivity* vs. the false positive rate (FPR) (i.e. $1 - specificity$) for several threshold settings. The true and false positive rates are defined as follows (TP is the number of true positives and FN the number of false negatives):

$$TPR = \frac{TP}{TP + FN} \tag{4}$$

$$FPR = \frac{FP}{TN + FP} \tag{5}$$

The area under the ROC curve (A_{ROC}) is a single-number summary of the information in the ROC curve.

Precision-recall (PR) curves facilitates the visualization of performance by depicting precision-recall value pairs for several probability thresholds. They illustrate the tradeoff between precision and recall. Thus, the area under the precision-recall curve (A_{PR}) is a single number summarizing the information in the precision-recall curve. A large area under the curve would mean both high recall and high precision values.

The Cohen's kappa coefficient is a statistic that measures inter-rater agreement for categorical items. In opther words, it expresses the level of agreement between two annotators in a classification problem. This indicator is more robust than percent agreement, as it takes into account the possibility of the agreement occurring by chance. It is defined as:

$$\kappa = \frac{(p_o - p_e)}{(1 - p_e)} \tag{6}$$

where p_o is the empirical probability of agreement on the label assigned to any sample, and p_e is the expected agreement when both annotators assign labels randomly.

2.2. Theoretical Formulation

As discussed in Sect. 2.1.2, documents (i.e., educational resources) will be represented according to the BoW paradigm. In other words, each resource will be represented according to the frequency of occurrence of the meaningful words extracted from its title and description, and also its keywords.

Let $\mathbb{F}$ be the feature domain, that is, the set of words in the training sequence, except stop words, previously processed according to the Porter algorithm (Porter, 1980). A learning object $\vec{lo_m}$ is defined as a vector:

$$\vec{lo}_m = (w_{1m}, w_{2m}, ..., w_{|\mathbb{F}|m}),\ 1 \leq \text{m} \leq |LO| \tag{7}$$

where w_{im} represents the weight of feature f_i in $\vec{lo}_m$. $|S|$ represents the cardinality of set S.

Let LO represent the domain of learning objects. The weighting problem is defined as the problem of approximating a function

$$\check{\mathcal{W}}:\ LO{\times}\mathbb{F} \to \mathbb{R}^+ \tag{8}$$

provided that $\check{\mathcal{W}}(\vec{lo_m}, f_i) = w_{im}$, where w_{im} is the weight of feature f_i in object $\vec{lo_m}$ defined as the frequency of the occurrence of f_i in $\vec{lo_m}$.

To increase the performance of the classifier, instead of using just the features (i.e., words) extracted from the description and the title of the resource, the resource's keywords will also be included among the features when creating the BoW of the document. Besides, to avoid loss of semantics, spaces are replaced by underscores in entities composed of more than one word (e.g., "Social Sciences" → "Social_Sciences"). Finally, a parameter tuning exercise will be conducted to determine the relevance or optimal weight k to be assigned to each feature in the BoW of each document. Thus, a learning object $\vec{lo_m}$ will be defined as

$$\vec{lo_m} = (w_{1m}, w_{2m}, ..., w_{|\mathbb{F}|m}, k_{1m}, k_{2m}, ..., k_{nm}),$$
$$1 \leq \text{m} \leq |LO|,\ 0 \leq \text{n} \leq |N| \quad (9)$$

where k_{im} represents the weight of keyword k_i in $\vec{lo_m}$ and N the total number of keywords in the document.

A repository $Rep = \{\vec{lo_1}, \vec{lo_2}, ..., \vec{lo_{|Rep|}}\}$ is a set of $|Rep|$ educational objects, where the $\vec{lo_i}$ represent the objects in the repository.

A repository aggregator $AggRep = \{r_1, r_2, ..., r_{|A|}\}$ is a set of $|AggRep|$ repositories, where the r_i represent individual repositories.

Let c_{ji} be a category belonging to repository r_i. The taxonomy of r_i is the set of all categories c_{ji} in r_i, that is,

$$C_i = \{c_{1i}, c_{2i,}, ..., c_{|C_i|i}\},\ 1 \leq i \leq |A| \quad (10)$$

Note that domain of learning objects $LO = r_1 \cup r_2 \cup ... \cup r_{|AggRep|}$ is the set of all learning objects belonging to any of the repositories integrated by the aggregator.

The classification problem is defined as the problem of fitting a set of functions

$$\check{\mathcal{C}}_i :\ LO \times \mathrm{C_i} \rightarrow \{\mathrm{True}, \mathrm{False}\},\ 1 \leq i \leq |A| \quad (11)$$

such that $\check{\mathcal{C}}_i(\vec{lo_m}, c_{ji}) = True$ when educational object $\vec{lo_m}$ belongs to category c_j in repository r_i, and $\check{\mathcal{C}}_i(\vec{lo_m}, c_{ji}) = False$ otherwise.

The fitting of the $\check{\mathcal{C}}_i$ is performed by means of classification techniques based on supervised ML. Firstly, the training sequence is obtained from r_i. The training sequence for repository r_i is defined as

$$TS_i = \{\vec{lo_j} \mid \vec{lo_j} \in rand\_r_i,\ 1 \leq j \leq L\},\ 1 \leq i \leq |\mathrm{A}| \quad (12)$$

where L is the length of the sequence and $rand\_r_i$ the outcome of randomizing repository r_i.

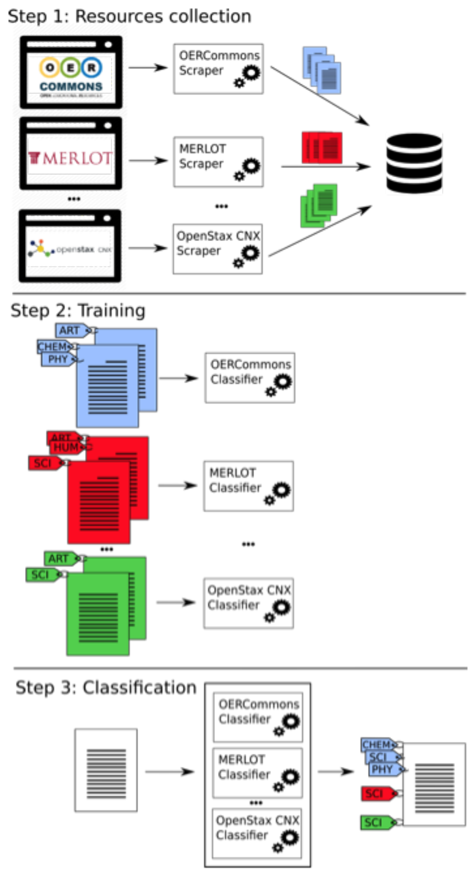

Figure 3: Classification Strategy

Once the classification algorithm has been trained, a fresh object $\vec{lo}$ will be automatically classified according to each of the taxonomies of the $|AggRep|$ repositories in the aggregator $AggRep$ as directed by the $\check{\mathcal{C}}_i$

2.3. Qualitative Description

The study presented in this article consists on the implementation and validation of an aggregator of educational resource repositories that operates according to the steps below (cf. Fig. 3):

1. Fetching of educational resources in aggregated repositories. For this, web scraping techniques are used (cf. Sect 2.4) on each of the repositories

integrated.

2. Training. For each of the integrated repositories, an SVM classifier is trained (cf. Sect. 2.1.1) with information from metadata from the resources being classified, namely their title, description and keywords, and the category to which they belong.
3. Classification. Finally, all educational resources obtained in step 1 are classified according to all the classifiers trained in step 2, so that the algorithm is able to predict to which category each resource belongs, according to the information provided by the metadata.

As a result of performing the above three steps, each educational resource will be eventually classified according to all taxonomies of the different repositories integrated. Then, users will be able to perform exploratory searches across all available resources in the aggregator using the taxonomy considered most appropriate, useful or convenient.

2.4. Architecture

The architecture of the proposed aggregator (cf. Fig. 4) is based on a collection of cooperating software components, namely a web application, an indexer, the database system and the web scrappers.

The web application was developed on Ruby on Rails (RoR), the combination of the Ruby programming language and the Rails framework, especially designed for web application development (Jazayeri, 2007). In recent years, an important developers' community grew around this framework. This, together with the large amount of programming libraries available (known as *gems*) make it an attractive option to serve as the foundation of complex web applications. RoR follows the model-view-controller paradigm (MVC) (Krasner et al., 1988).

The database system collects all the (relevant metadata and references of the) aggregated educational resources. MySQL was selected as the technological solution to implement it.

Performing queries directly on a large database is a time consuming task. In our case, inverse indexing techniques are applied to increase the performance

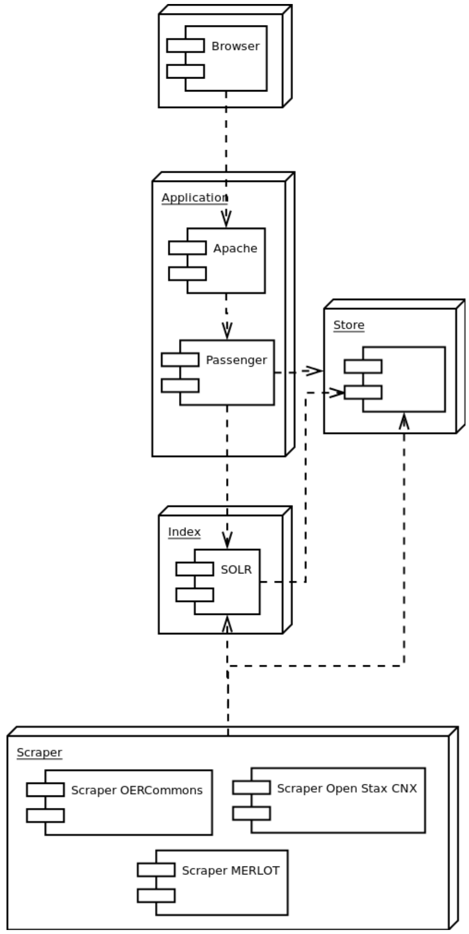

Figure 4: System Architecture

of database access (Zobel et al., 1998). To implement the indexer we selected Apache Solr (Smiley and Pugh, 2009), a mature technology that ensures stability and scalability together with fast response time.

The so-called Web scrapers are responsible for obtaining the educational resources from the target repositories to be stored in the database and indexed. These are software modules able to scan web sites and automatically extract relevant information from them. For this research, a different scraper was programmed for each target repository. Information extraction from the source repository is carried out according to an ontology that defines which fields are

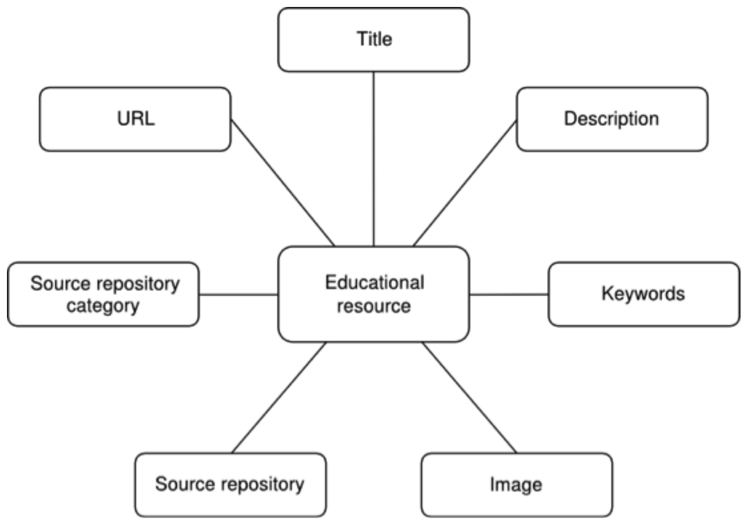

Figure 5: Ontology defining what fields are to be extracted.

to be extracted (cf. Fig. 5).

To make scraping as efficient as possible, several considerations were taken into account:

- The process is time-consuming, so scraping processes were run as background tasks.
- The number of requests per unit of time should be kept below a certain threshold on each server to avoid being incorrectly considered a denial-of-service attack.
- As information sources are dynamic, scrapers would periodically check for new information. As a side effect, special care had to be taken to filter out duplicate items. For this, hashing techniques are used (cf. Sect. 4).

2.5. Repositories Integrated

Although CROERA was developed to support the integration of any existing resource repository, for the initial version discussed in this paper three actual repositories were selected, namely OERCommons, MERLOT and Open Stax

CNX. These repositories are characterized by the high quality of the resources collected and their metadata, as communities of experts supervise content quality in all cases. Besides, each repository has its own taxonomy, which is a key aspect to evaluate one of the most relevant features of this study (i.e., taxonomy-independent content fetching).

- OERCommons (Open Educational Resources, 2013) is a repository created by the Institute for the Study of Knowledge Management in Education (ISKME) to support the deployment of a network of learning and teaching materials. Users can navigate and fetch content on a database with more than 98,000 resources organized around 21 categories: Arts, Education, Business, Humanities, Mathematics and Statistics, Physics, Geoscience, Computing and Information, Ecology, Engineering, Science and Technology, Forestry and Agriculture, Space Science, Mathematics, Life Science, Politics, Law, Technology, Social Sciences, Chemistry and History.

- MERLOT (Cafolla, 2006) is a community of users centered on the provision of open educational resources. It collects more than 47,000 OER organized around 9 broad categories: Academic Support Services, Arts, Business, Education, Humanities, Mathematics and Statistics, Science and Technology, Social Sciences and Workforce Development.

- Open Stax CNX (OpenStax, 1999) is a digital ecosystem created by the Rice University for the distribution and sharing of educational content to improve the learning experience of its users. It hosts tens of thousands of educational resources organized around 6 major categories: Arts, Business, Humanities, Mathematics and Statistics, Science and Technology and Social Sciences.

2.6. Usage Example

The next paragraphs illustrate the operation of the solution proposed with an exploratory search performed using the CROERA system. Figure 6 shows how a user accustomed to Open Stax CNX tries to locate educational resources on "Science and Technology". Let us suppose that this user is a regular user of the repository, so they will utilize the particular taxonomy or classification scheme of this repository.

First, the user selects the taxonomy for Open Stax CNX as the reference taxonomy (step 1), and then selects the category on which educational resources are to be fetched, "Science and Technology" in our case (step 2).

After this, the system offers the user a variety of educational resources about the category selected from all the repositories aggregated, OERCommons, MERLOT and Open Stax CNX in this example (step 3). Note that the user may obtain educational resources on a particular category from any repository and using any classification scheme available. CROERA supports this through the re-classification of every educational resource according to each of the taxonomies, so that a resource is always accessible regardless of the taxonomy used (step 4).

By accessing the detail of a particular resource, the user can view metadata elements such as title, description and keywords. In addition, information about the repository of origin is also displayed, OERCommons in this example (step 5), together with the original classification in the source repository (step 6) and the classification performed by the CROERA system (step 7) .

As pointed out above, a given resource is classified according to the three existing taxonomies in OERCommons, MERLOT and Open Stax CNX. Thus, although the resource in this example was originally classified in category "Geoscience", the system was able to correctly classify it in the most appropriate category in this case (i.e., "Science and Technology" within Open Stax CNX and MERLOT). Thus, category heterogeneity is overcome and access to all resources available in the aggregator is provided, solving the one-to-none mapping situation, that is, those cases where there is no correspondence among the elements in the different taxonomies.

Note that mainstream ontology-matching techniques may not be able to match category "Geoscience" in OERCommons to any of the categories in MER-

Figure 6: Usage example. Target repository: Open Stax CNX.

LOT or Open Stax CNX without the assistance of a human expert. In such cases, the resource object would be accessed using the taxonomy in OERCommons only, remaining hidden for the rest of the repositories (i.e., taxonomies).

3. Results

The evaluation of the repository aggregator is discussed along the next paragraphs. The experiments carried out to compute the optimal weight of the features extracted from the keywords associated to educational resources are described (cf. Sect. 2.2). Finally, the actual evaluation of the aggregator is discussed. Note that evaluation is focused on evaluating the performance of the classifier of the CROERA system. In this context, performance means how well the classifier classifies the educational resources contained in the platform. The performance is measured in terms of precision, recall, F_1 score, A_{ROC}, A_{PR}, and Cohen's κ (cf. Sect. 2.1.3).

3.1. Parameter tuning

To maximize the classifier's performance, a parameter tuning exercise was carried out to compute the optimal weights to be assigned to the features obtained from content keywords. To define the relevance or optimal weight to be assigned within each document's BoW, the process below was implemented for each repository:

1. 5,000 elements were randomly selected from each repository.
2. 500 additional elements were randomly selected as the testing sequence.
3. The range of values for the relevance is defined as the set including 0 and the first 8 values in the Fibonacci sequence (i.e., $\{0, 1, 2, 3, 5, 8, 10, 13, 21\}$).
4. The classification algorithm is trained and tested using the sets of elements in points 1 and 2 above. Features obtained from keywords are weighted according to the relevance values defined in point 3.

Figure 7 shows the F_1 values from the experiments carried out. This figure illustrates how high relevance values affect the classifier's performance due to the higher relative weight of features in the BoW obtained from keywords with respect to features extracted from document titles and descriptions. According to this figure, optimal weights for keywords depend on the target repository, which are 3 for OERCommons; 2 in the case of MERLOT, and 2 again for Open Stax CNX. As depicted in Figs. 8, 9 and 10, the performance offered by the classifier after enriching the BoW with (weighted) keywords is higher than the performance without them, especially in the case of OERCommons. Table 1 shows that p-values obtained by 10-fold cross validation and a two-tailed t-test (Sacchet et al., 2015) are below the significance level of 0.05, thus indicating that the enrichment of the BoW with keywords significantly improves classification performance.

3.2. Performance Assessment

Due to the different taxonomies of the repositories aggregated, two complementary strategies were considered to carry out the assessment of the performance of the classifier. First, the classification performance for resources in

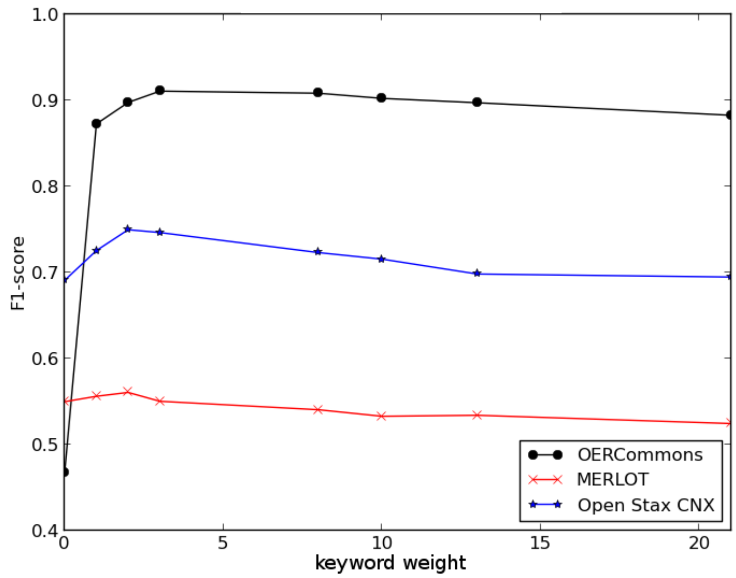

Figure 7: Performance of the classifier vs. keyword weight (F_1 values)

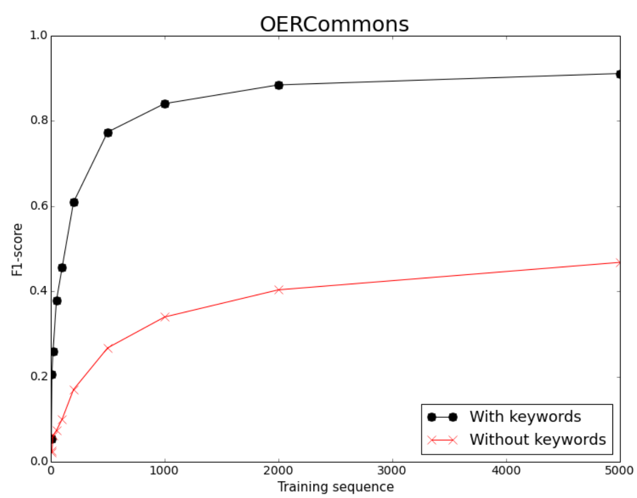

Figure 8: OERCommons: performance (F_1 values) of the classifier vs. training sequence length, with keywords (dots) and without keywords (crosses).

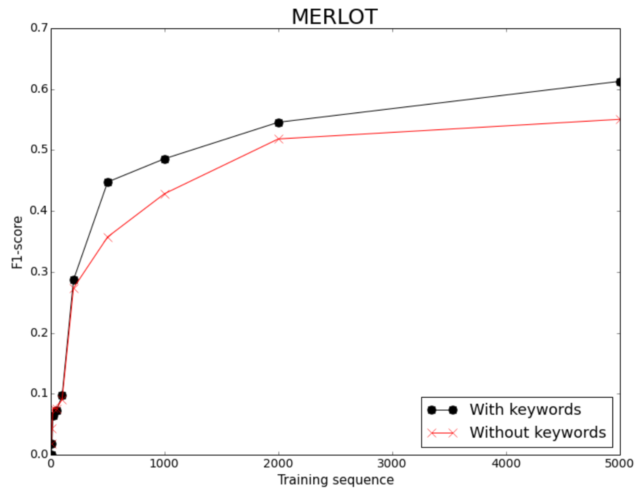

Figure 9: MERLOT: performance (F_1 values) of the classifier vs. training sequence length.

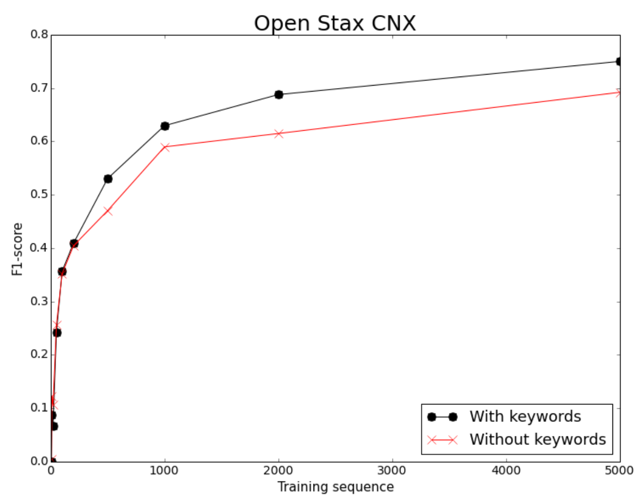

Figure 10: Open Stax CNX: performance (F_1 values) of the classifier vs. training sequence length.

common categories to all three repositories was evaluated. Then, in order to evaluate the classification performance for the resources not belonging to com-

Table 1: Precision (P), recall (R), F_1, area under the ROC curve (A_{ROC}), area under the precision-recall curve (A_{PR}) and Cohen's kappa coefficient (κ) (averaged over all classes) for the three repositories considered with and without keywords.

	OERCommons	MERLOT	OpenStaxCNX
F_1 w/o keywords	0.468	0.550	0.691
F_1 with keywords	0.911	0.613	0.750
F_1 improvement	94.00%	11.45%	8.54%
A_{ROC} w/o keywords	0.687	0.728	0.792
A_{ROC} with keywords	0.943	0.758	0.827
A_{ROC} improvement	37.26%	4.12%	4.42%
A_{PR} w/o keywords	0.570	0.640	0.810
A_{PR} with keywords	0.92	0.70	0.85
A_{PR} improvement	61.40%	9.38%	4.94%
κ w/o keywords	0.411	0.490	0.636
κ with keywords	0.900	0.563	0.706
κ improvement	118.9%	14.90%	11.01%
$p - value$	< 0.0001	0.0230	0,0335

mon categories, an alternate approach was followed based on human experts. The combination of both strategies offers a broader picture of classification performance.

3.2.1. Performance for Common Categories

Evaluation is carried out according to the steps below:

1. Selection of the common categories in the three aggregated repositories, namely Arts, Business, Humanities, Mathematics and Statistics, Science and Technology and Social Sciences.
2. Selection of those resources classified in just the common categories.
3. Training of the classification algorithm using 5,000 randomly selected documents (i.e., their metadata) from all three repositories.

4. Classification of 500 documents randomly selected from each of the repositories (i.e., OERCommons, MERLOT and Open Stax CNX).

Table 2 collects precision, recall, F_1, A_{ROC}, A_{PR} and κ (averaged over all classes) for the classification performed according to the steps above. It can be observed that average values for the six metrics considered are fairly high (e.g., around 80% precision, A_{ROC} and A_{PR}, and around 75%-77% recall and F_1 values). According to the widely used rating scale proposed in (Landis and Koch, 1977), Cohen's kappa values obtained show an agreement strength between *moderate* (0.41 - 0.60) and *substantial* (0.61 - 0.80). Besides, the best results are obtained when resources originally in OERCommons are classified, and the worst results correspond to resources in MERLOT.

Although the performance values offered by the classifier are high, a number of factors prevent this value from being optimal:

- Category scope. Although categories common to the three repositories were selected, this selection was made according to just the name of the category. Note that two categories having the same name may have different scope, that is, a category may include a number of resources in a repository that would not be an exact match of the resources in the same category in a different repository.

- Differences in the criteria applied by human experts. The classification of the resources of each repository was performed by different groups of human experts, and as a consequence the classification of a resource in one category or another depended on the judgment of the actual experts deciding upon the classification.

- Sub-optimal training. The computing resources available prevented the use of all the documents in the repositories to train the classification algorithm. This causes classifier training to be sub-optimal, which in turn has a negative impact on the performance of the classification.

Table 2: Precision P, recall R, F_1 , area under the ROC curve (A_{ROC}), area under the precision-recall curve (A_{PR}) and Cohen's κ (averaged over all classes) for the common categories in the three repositories considered.

Test elements	P	R	F_1	A_{ROC}	A_{PR}	κ
OERCommons	0.863	0.812	0.840	0.878	0.860	0.751
MERLOT	0.745	0.666	0.705	0.748	0.730	0.525
Open Stax CNX	0.781	0.746	0.764	0.788	0.790	0.583
Average	0.796	0.743	0.770	0.804	0.790	0.620

3.2.2. Performance for Disjoint Categories

Expert-based evaluation was carried out by 68 participants (cf. Table. 3) including university and secondary education teachers from several fields of study. Participants interacted with the platform and were asked to classify, in the most appropriate category for each of the three taxonomies available, a collection of educational resources. Participants should consider only the metadata elements utilized in automatic classification, that is title, description and keywords. Each participant classified a randomly assigned sample of the resources available in the platform. A total of 647 documents (475 unique documents) were classified. The classification made by human experts served as a reference for assessing the classification made by the CROERA system on the same resources, and thus for obtaining an indication of the performance of the classification system proposed.

Obviously, classification abilities were not uniform among human experts. Factors such as the level of expertise in different areas or their subjectivity had an influence on their decisions. To obtain an indication of a human expert's classification skills, the classification performed was compared to the classification in the repository of origin. This way, numeric values measuring the classification abilities of human experts can be computed. In our case, the values correspond to precision, recall, F_1, A_{ROC}, A_{PR} and κ, computed for the experts and their classification exercise. Table 4 collects the average values obtained. As expected, human experts were not infallible and perfect values were not obtained.

Table 3: Human experts participating in the evaluation of the CROERA system.

	CS	Ma	L	Me	Ed	Ch	El	CA	**Tot**
HE teachers	19	6	3	1	6	1	1	1	**38**
SE teachers	3	0	1	0	0	0	1	0	**5**
Total teach.	22	6	4	1	6	1	2	1	**43**
Other	20	1	0	0	3	0	1	0	**25**

Legend: CS=Computer Science, Ma=Maths, L=Languages, Me=Medicine, Ed=Education, Ch=Chemistry, El=Electronics, CA=Culture & Arts; HE=Higher Education, SE=Secondary Education.

The lowest values are those obtained for the OERCommons repository, followed by the MERLOT ones. Finally, the highest values are those obtained for the Open Stax CNX repository. This behavior is due to two factors, which are enumerated below.

- Each educational resource to be classified by experts may belong to more than one category, thus adding complexity to the classification problem. Indeed, complexity increases with the number of categories which can be applied to a given element. For instance, for the elements from the OERCommons repository, experts had to select the more appropriate category (or categories) from among a set of 21 different categories. In the case of MERLOT, experts had to select one or more categories from a pool of 9 categories. Finally, for the elements in Open Stax CNX repository, experts only have to decide among 6 categories. As a consequence, experts performed better when classifying elements belonging to Open Stax CNX, followed by MERLOT and OERCommons, as Table 4 indicates.
- OERCommons holds the highest average number of categories per educational resources, followed by MERLOT and Open Stax CNX. This also adds complexity to the classification problem. It is more difficult for ex-

Table 4: Average classification abilities of human experts for the three repositories considered.

Test elements	P	R	F_1	A_{ROC}	A_{PR}	κ
OERCommons	0.426	0.514	0.465	0.648	0.420	0.289
MERLOT	0.494	0.650	0.561	0.644	0.470	0.293
Open Stax CNX	0.748	0.886	0.811	0.829	0.77	0.554

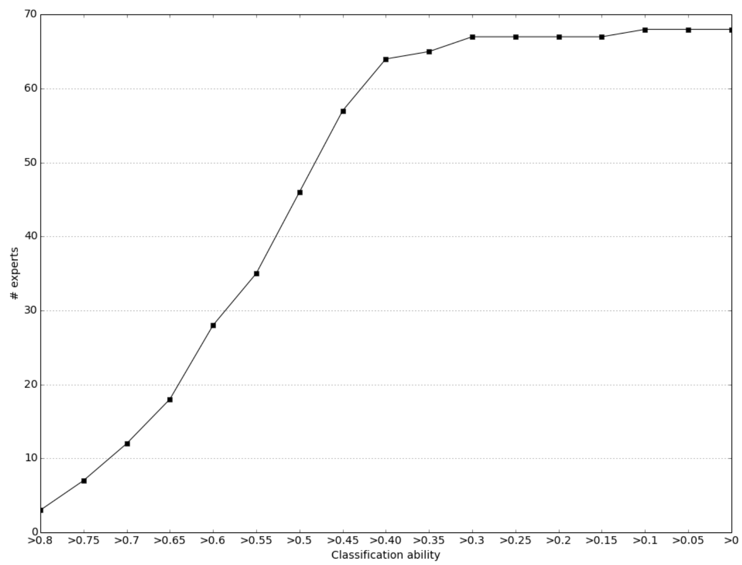

Figure 11: Distribution of human experts according to their classification skills (F_1 value).

perts to classify well elements belonging to repositories which hold a high number of categories per element.

Then, human experts were grouped according to their classification skills. Figure 11 summarizes the distribution of human experts according to this criterium. Note that there are relevant differences in the classification abilities of different experts.

Finally, the performance of CROERA is computed taking as a reference the elements manually classified by each of the experts' groups. Again, precision, recall, F_1, A_{ROC}, A_{PR} and κ (averaged over all classes) are computed.

As it can be observed in Tables 5 & 6, performance seems to increase with

Table 5: Performance of CROERA, taking as a reference experts' classification (I).

Class.	OERCommons			MERLOT			CNX			Avg.
Skills	P	R	F_1	P	R	F_1	P	R	F_1	F_1
>0.80	0.563	0.308	0.436	0.579	0.567	0.573	0.710	0.687	0.698	0.569
>0.75	0.453	0.305	0.379	0.596	0.636	0.615	0.700	0.686	0.693	0.563
>0.70	0.363	0.280	0.322	0.581	0.619	0.599	0.667	0.668	0.667	0.530
>0.65	0.370	0.224	0.297	0.594	0.597	0.595	0.664	0.656	0.654	0.518
>0.60	0.388	0.216	0.302	0.596	0.536	0.564	0.682	0.635	0.658	0.509
>0.55	0.381	0.228	0.305	0.582	0.538	0.559	0.667	0.627	0.646	0.504
>0.50	0.351	0.226	0.289	0.575	0.533	0.554	0.658	0.626	0.642	0.495
>0.45	0.337	0.233	0.285	0.558	0.526	0.542	0.617	0.581	0.598	0.475
>0.40	0.331	0.234	0.283	0.545	0.512	0.528	0.614	0.571	0.592	0.468
>0.35	0.332	0.237	0.285	0.545	0.515	0.530	0.614	0.570	0.591	0.469
>0.30	0.326	0.235	0.281	0.538	0.511	0.524	0.608	0.564	0.585	0.464
>0.25	0.326	0.235	0.281	0.538	0.511	0.524	0.608	0.564	0.585	0.464
>0.20	0.326	0.235	0.281	0.538	0.511	0.524	0.608	0.564	0.585	0.464
>0.15	0.326	0.235	0.281	0.538	0.511	0.524	0.608	0.564	0.585	0.464
>0.10	0.322	0.232	0.277	0.535	0.510	0.522	0.604	0.562	0.582	0.461
>0.05	0.322	0.232	0.277	0.535	0.510	0.522	0.604	0.562	0.582	0.461
> 0.00	0.322	0.232	0.277	0.535	0.510	0.522	0.604	0.562	0.582	0.461

experts' classification skills, except in the case of A_{ROC}, A_{PR} and κ for the OERCommons repository, which do not increase or slightly increase. In order to validate this hypothesis, regression analysis techniques (Rawlings et al., 1998; Kahane, 2007), namely a 2-degree polynomial regression, was applied to F_1 values in Table 5. The results obtained are depicted in Figs. 12, 13 and 14 for the three repositories considered. Figure 15 depicts the average performance. Tables 5 & 6 and regression curves in the four figures show that the performance of the classifier of the proposed system increases with the threshold that determines the classifying ability of human experts. In other words, performance

Table 6: Performance of CROERA, taking as a reference experts' classification (II).

Class.	OERCommons			MERLOT			CNX		
Skills	κ	A_{ROC}	A_{PR}	κ	A_{ROC}	A_{PR}	κ	A_{ROC}	A_{PR}
>0.80	0.117	$*$	0.290	0,435	$*$	0.610	0,543	$*$	0.720
>0.75	0.211	$*$	0.320	0.432	0.710	0.670	0.472	0.736	0.750
>0.70	0.228	0.642	0.290	0.385	0.682	0.630	0.485	0.738	0.740
>0.65	0.197	0.631	0.260	0.350	0.673	0.590	0.433	0.710	0.730
>0.60	0.197	0.644	0.240	0.303	0.658	0.540	0.425	0.708	0.690
>0.55	0.215	0.640	0.260	0.330	0.666	0.540	0.439	0.718	0.690
>0.50	0.239	0.643	0.280	0.311	0.660	0.530	0.445	0.722	0.680
>0.45	0.242	0.641	0.290	0.322	0.660	0.540	0.403	0.700	0.650
>0.40	0.240	0.639	0.290	0.306	0.654	0.520	0.393	0.693	0.630
>0.35	0.245	0.642	0.290	0.305	0.653	0.520	0.391	0.692	0.630
>0.30	0.244	0.641	0.290	0.306	0.654	0.530	0.390	0.691	0.630
>0.25	0.244	0.641	0.290	0.306	0.654	0.530	0.390	0.691	0.630
>0.20	0.244	0.641	0.290	0.306	0.654	0.530	0.390	0.691	0.630
>0.15	0.244	0.641	0.290	0.306	0.654	0.530	0.390	0.691	0.630
>0.10	0.242	0.640	0.290	0.307	0.654	0.530	0.390	0.691	0.630
>0.05	0.242	0.640	0.290	0.307	0.654	0.530	0.390	0.691	0.630
> 0.00	0.242	0.640	0.290	0.307	0.654	0.530	0.390	0.691	0.630

$*$ Only one class present in ground truth. A_{ROC} is not defined in that case.

increases when the elements taken as a reference are those classified by experts with better classification abilities. Besides, the values of the four coefficients of determination R^2 are very high (i.e., 0.954, 0.851, 0.810, and 0.981) confirming the high quality of the regression models. Furthermore, in 4 cases the p-value is less than 0.05, eliminating the null hypothesis and providing evidence on the statistical relationship between data and models. All evidence confirm that, as the quality of the base truth increases, the performance of the classification made by the CROERA system increases.

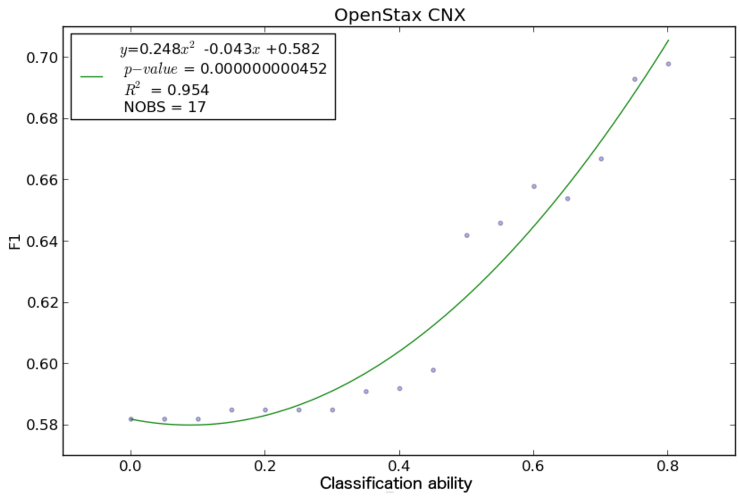

Figure 12: Regression Analysis. CROERA vs. human experts. CNX case.

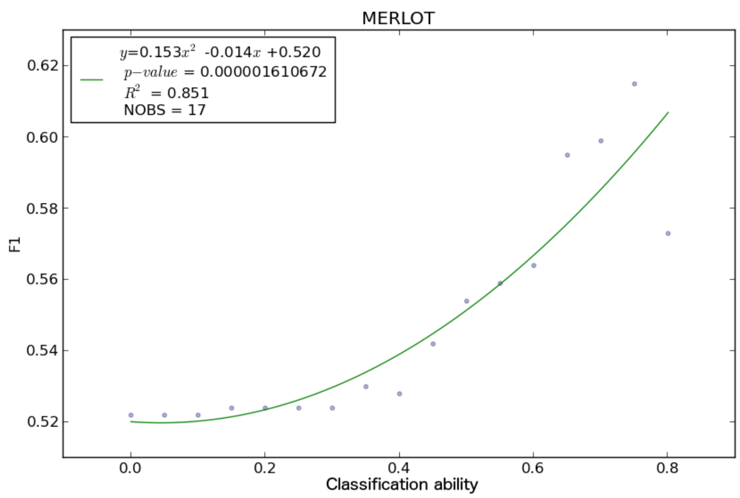

Figure 13: Regression Analysis. CROERA vs. human experts. MERLOT case.

4. Discussion

The most relevant proposals using ML techniques for conducting matchings between ontologies do not address repositories, or do not refer to taxonomies or ontologies on educational resources. Therefore, no relevant contributions were found in the literature in relation to the research discussed in this paper. Fur-

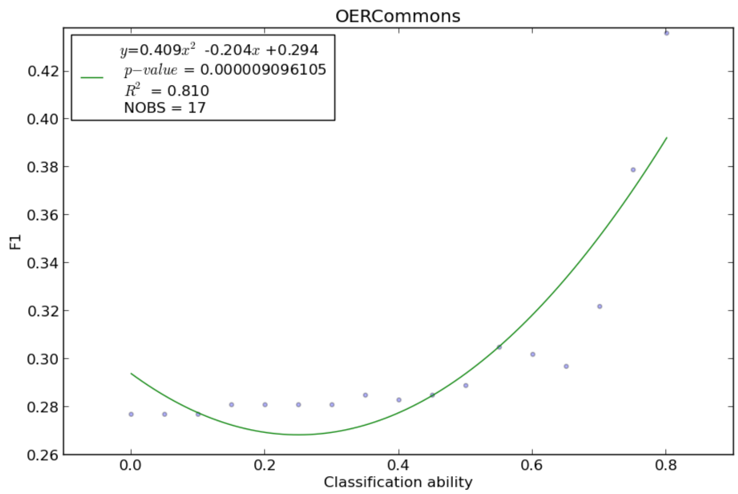

Figure 14: Regression Analysis. CROERA vs. human experts. OERCommons case.

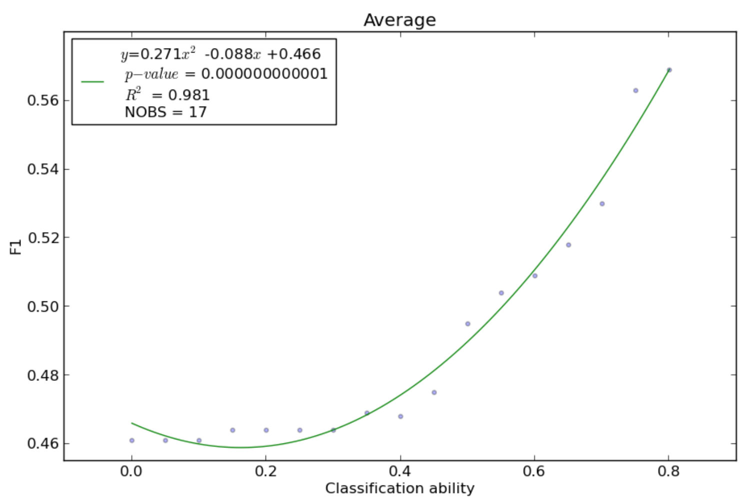

Figure 15: Regression Analysis. CROERA vs. human experts. Average results.

thermore, the development of the most promising initiatives was halted at the time of writing this paper, and the related software tools are not available (Doan et al., 2004; Straccia and Troncy, 2005; Otero-Cerdeira et al., 2015). The only system found to serve as a reference is YAM++ (Ngo and Bellahsene, 2012). This system was under development between 2009 and 2013 (Otero-Cerdeira et al., 2015) and there is functional open-source software available. Therefore, the results obtained by our system will be compared with those of YAM++. Note that automatic tools for creating mappings between taxonomies, and particularly YAM++, are intended to be used with complex taxonomies taking advantage of all the features offered, such as node, element, structural and semantic-level analysis (Doan et al., 2003; Straccia and Troncy, 2005; Ngo and Bellahsene, 2012).

Thus, the three taxonomies introduced in this work, (i.e., OERCommons', MERLOT's and Open Stax CNX's) were fed to YAM++ to compute mappings between each pair of them. Results are summarized in Table 7. Note that YAM++ could only match a few elements in each pair of taxonomies. More specifically, seven nodes were mapped between OERCommons and MERLOT, six nodes were matched between OERCommons and Open Stax CNX, and the same number of nodes between MERLOT and Open Stax CNX. Besides, there are many existing relations that YAM++ was unable to compute. Specifically, 16 relations between OERCommons and MERLOT, 15 between OERCommons and Open Stax CNX, and 3 between MERLOT and Open Stax CNX.

Note that the use of automated tools does not provide dramatic advantages in the case of simple, single-level taxonomies as the ones tackled in this work. This is because the information available on the taxonomy allows element-level mappings only, that is, mappings based on the information in each node provided by categories' labels. Thus, when the tool does not find a match between nodes in two categories, a one-to-none mapping situation occurs, so that some elements will not be matched and others would not be matched optimally. As a consequence, in the context discussed in this paper, items classified under categories or nodes that could not be matched will not be accessible from a repository different from the original one.

For example, the "Physics" category in OERCommons has no correspondence with any category in MERLOT or Open Stax CNX, and therefore items classified under category "Physics" cannot be accessed using the Open Stax CNX or MERLOT taxonomies. The same applies to all the elements in cate-

Table 7: Mappings computed by YAM++.

OERC, MERLOT, CNX	Arts, Business, Humanities, Math&Stat, Sci&Tech, Social Sci
OERC, MERLOT	Education
OERC	Physics, Geoscience, Comp&Inf, Ecology, Engineering, Fore&Agr, Space Sci, Math, Life Sci, Politics, Law, Tech, Chemistry, History
MERLOT only	Ac.Sup.Serv, Web Deb

gories or nodes that could not be matched. Indeed, using YAM ++ to compute mappings between elements in the taxonomies of the repositories referenced in this work, only 79,855 from the 149,856 accumulated resources at the time of writing this paper would be accessible, that is, more than 46% of the resources would not be accessible from all taxonomies.

CROERA solves the situation discussed above and enables access to 100% of the resources available regardless of the selected taxonomy, as all of them would be classified according to the three taxonomies. The results obtained in Sect. 3 show that the CROERA system provides access to all resources regardless of the taxonomy used, with an average performance (F_1 score) of 77% (maximum of 84%) for resources belonging to the common categories in the three repositories. In the case of resources not classified in common categories, average F_1 is reduced to 56.9%, with a maximum close to 70%.

Insofar execution time is concerned, the training of the classifier using 5,000 documents as discussed in Sect. 3.1 took 33 minutes in an office-grade personal computer (i.e., Intel® Core$^{\mathrm{TM}}$ 7-4770 CPU @ 3.40GHz $\times$ 8 with 16GB RAM).

Note that training will be performed only once. The classification of 500 new resources according to the three taxonomies involved took between 1'17" and 1'38", that is, an average of 168-196 ms. per resource. CROERA will store the new classifications, so once a resource is re-classified, it would not be required to be classified again unless a new repository using a new taxonomy or classification scheme is aggregated.

It has to be taken into account that the classifier selected in this study is a multi-label classifier, which means that an element may belong to one or more categories. In other words, an element may be labeled with one or more labels. Therefore, to achieve a performance of 100% all elements would have to be classified in all of the categories to which it belongs. A performance value below 100% does not necessarily mean that the resource was misclassified, but that it was not classified in all applicable categories. For example, a performance of 60% would not mean that 40% of the resources were misclassified, but 60% of the resources were classified optimally for each of the applicable categories, while 40% of the remaining resources were classified sub-optimally.

The fact that each element may belong to more than one category adds further complexity to the classification problem, which increases with the number of categories applicable to a given element. In the case of CROERA, 38% of the educational resources are classified in at least two categories; 18% in three or more categories; 8.5% is classified in five or more categories, and there are more than 1,000 resources that are labeled with 10 or more labels. This causes that the perceived performance is better than the performance expected from the values obtained in the evaluation process, as the CROERA system provides access to all the elements available, either optimally (average F_1 of 56.9%) or sub-optimally (43.1% average F_1).

Although it can be argued that the usability and relevance of exploratory search and browsing may not be as high as expected (Ochoa, 2005) this is still an open field. For example, in some scenarios exploratory search-based visit strategies are perceived as both relevant and usable (Phang et al., 2010). On the other side, previous work (Pérez-Rodríguez et al., 2016) indicated that the exploratory search of educational resources was perceived as a very convenient way of interacting with a system that integrates a great number of educational resources. In particular, 85% of teachers who interacted with an educational resource repository considered exploratory search as “excellent” or “good”; 10% of them considered it as “average”, and only the 5% considered exploratory search as not useful.

With respect to the scalability of the CROERA system, the educational resources managed are classified only once, at the time they are stored in the system. When a new repository is aggregated, its resources are also classified only once according to the existing taxonomies. Although this preliminary version of the CROERA system does not update the resources it has previously scrapped, this issue has been addressed in the previous work referenced (Pérez-Rodríguez et al., 2016). The approach used is the following: the first time the system gets an educational resource, it creates a hash of the resource’s metadata. As each educational resource is identified by a unique identifier (in this case the URL), the next times the system gets an already fetched resource, it computes again the hash of the corresponding metadata. If it is different from the hash stored in the database, the system will update the resource and will classify it again according all taxonomies available in the system; if the hash value has not changed, the resource is assumed to be unchanged and it will not be necessary to classify it again. As a consequence, in our opinion, there are no relevant concerns insofar its scalability is concerned. Note that one of the key objectives pursued with CROERA is to facilitate access to all resources aggregated by providing alternate ways to access them, independently of the taxonomy originally used to classify them. On the other side, resource aggregation has its main impact on the enrichment of the metadata of existing resources, and not necessarily on the final number of resources obtained. For example, most relevant resources in a given field would have been discovered and classified by all relevant repositories.

5. Limitations

No matter that CROERA could contribute in a relevant way to facilitate the classification, location and fetching of educational resources independently of the repository of origin, the approach introduced has some limitations.

The CROERA system considers only the top level categories of the taxonomies of the repositories aggregated. Automatic classification in CROERA is based on information provided by the metadata attached to educational resources (i.e., title, description, keywords). However, in most cases educational resources include some sort of textual content that may provide information relevant to classification beyond the one extracted from metadata. As CROERA does not consider actual textual content, this may affect performance. This situation would become more relevant in the case of low-quality metadata (e.g., when the description or keywords of a given resource do not correctly characterize its content). Besides, the actual version of the system only considers repositories and educational resources written in English.

On the other side, state-of-the-art computing equipment prevented us from using all the items available for training the classification algorithm at a reasonable cost. This causes classifier training to be sub-optimal, which in turn negatively affects the classification performance.

Finally, the lack of taxonomy-matching ML-based systems different from YAM++ that could be used to compare our approach with other equivalent solutions did not contribute to provide the clearest picture possible about the actual benefits of CROERA in terms of performance or classification power.

6. Conclusions and Future Work

CROERA is an aggregator of educational resource repositories that enables access to resources independently of the taxonomy utilized to classify the content in each repository. The aggregated resources are classified automatically using ML techniques according to each of the taxonomies of the repositories integrated. This enables users accustomed to the descriptions and classification strategy of a particular repository to discover resources in other repositories without needing to switch to another taxonomy or classification scheme.

This platform was designed to integrate any existing repository. However, for the discussion in this paper three of the most relevant OER repositories were utilized (i.e., OERCommons, MERLOT and Open Stax CNX) due to the high quality of their resources and metadata, and the fact of being backed by communities of experts and active users.

CROERA provides a solution to the heterogeneity of taxonomies and enables exploratory searches (i.e., browsing) through any of the taxonomies integrated without requiring any metadata or ontology mapping, thus eliminating one of its main drawbacks, that is, the one-to-none mapping situation. This problem of nonexistence of relations or matchings between nodes in different taxonomies has as a consequence that items classified under categories or nodes that could not be mapped would not be accessible from a different repository.

CROERA is most useful tool to facilitate the work of human experts, as CROERA may be used to perform an initial pre-classification of educational resources, which in turn will contribute to save time and resources.

Present and future work include addressing the limitations discussed in Sect. 5, extending the user interface to support simple searches, the integration of additional repositories, the classification of resources according to the target educational level, and the definition of a taxonomy based on Wikipedia categories that would facilitate the use of Wikipedia categories to navigate any of the repositories integrated.

Acknowledgments

This research was partially funded by the Xunta de Galicia through projects R2014/034 (RedPlir) and R2014/029 (TELGalicia).

References

Anido-Rifón, L. E., Fernández-Iglesias, M. J., Caeiro-Rodríguez, M., Santos-Gago, J. M., Llamas-Nistal, M., Sabucedo, L. Á., Pérez, R. M., 2014. Standardization in computer-based education. Computer Standards & Interfaces 36 (3), 604–625.

Blizard, W. D., 1988. Multiset theory. Notre Dame Journal of Formal Logic 30 (1), 36–66.
URL `http://projecteuclid.org/euclid.ndjfl/1093634995`

Boeykens, S., Santana Quintero, M., Neuckermans, H., 2009. Metadata for Architectural Contents in Europe. In: Book of Abstracts.

Boyd, K., Eng, K. H., Page, C. D., 2013. Area under the precision-recall curve: Point estimates and confidence intervals. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, pp. 451––466.

Broisin, J., Vidal, P., Meire, M., Duval, E., 2005. Bridging the gap between learning management systems and learning object repositories: exploiting learning context information. In: Telecommunications, 2005. Advanced Industrial Conference on Telecommunications/Service Assurance with Partial and Intermittent Resources.

Cafolla, R., 2006. Project MERLOT: Bringing Peer Review to Web-Based Educational Resources. Journal of Technology and Teacher Education 14, 313–323.
URL `http://go.editlib.org/a/4623`

Campbell, L. M., 2011. UK Learning object metadata core: working draft, version 0.3 _12042.

Chan, L. M., Zeng, M. L., 2006. Metadata interoperability and standardization - A study of methodology part I: Achieving interoperability at the schema level. D-Lib Magazine 12 (6), 23–41.

Cohen, J., 1960. A coefficient of agreement for nominal scales. Educational and psychological measurement 20 (1), 37—46.

D'antoni, S., 2006. The virtual university: Models and messages, lessons from case studies. Vol. 819. Unesco Paris.

Dietze, S., Yu, H. Q., Giordano, D., Kaldoudi, E., Dovrolis, N., Taibi, D., 2012. Linked Education: interlinking educational Resources and the Web of Data. In: Proceedings of the 27th annual ACM symposium on applied computing. ACM, pp. 366–371.

Doan, A., Domingos, P., Halevy, A., 2003. Learning to match the schemas of data sources: A multistrategy approach. Machine Learning 50 (3), 279–301.

Doan, A., Madhavan, J., Domingos, P., Halevy, A., 2004. Ontology matching: A machine learning approach. In: Handbook on ontologies. Springer, pp. 385–403.

Dolog, P., Henze, N., Nejdl, W., Sintek, M., 2004. Personalization in distributed e-learning environments. Alternate track papers posters of the 13th international conference on World Wide Web WWW Alt 04, 170–179.
URL http://portal.acm.org/citation.cfm?doid=1010432.1010460

Gibbons, S., 2009. Benefits of an institutional repository. Library Technology Reports 40 (4), 11–16.

Hanley, J. A., McNeil, B. J., 1982. The meaning and use of the area under a receiver operating characteristic (roc) curve. Radiology 143 (1), 29—36.

Hearst, M., Dumais, S., Osman, E., Platt, J., Scholkopf, B., 1998. Support vector machines.

Hillmann, D. I., Westbrooks, E. L., 2004. Metadata in practice. American Library Association.

Hsu, C.-W., Lin, C.-J., 2002. A comparison of methods for multiclass support vector machines. Neural Networks, IEEE Transactions on 13 (2), 415–425.

Jazayeri, M., 2007. Some Trends in Web Application Development. In: Future of Software Engineering (FOSE '07). pp. 199–213.

Joachims, T., 1998. Text categorization with support vector machines: learning with many relevant features. In: Proceedings of ECML-98, 10th European Conference of Machine Learning. Vol. 1398. pp. 137–142.
URL `http://www.springerlink.com/index/10.1007/BFb0026683`

Kahane, L. H., 2007. Regression basics. Sage publications.

Kawase, R., Fisichella, M., Niemann, K., Pitsilis, V., Vidalis, A., Holtkamp, P., Nunes, B., 2013. Openscout: harvesting business and management learning objects from the web of data. In: Proceedings of the 22nd international conference on World Wide Web companion. International World Wide Web Conferences Steering Committee, pp. 445–450.

Krasner, G. E., Pope, S. T., Others, 1988. A description of the model-view-controller user interface paradigm in the smalltalk-80 system. Journal of object oriented programming 1 (3), 26–49.

Kurilovas, E., Dagiene, V., 2009. Learning Objects and Virtual Learning Environments Technical Evaluation Criteria. Electronic Journal of e-Learning 7 (Issue 2), 127–136.
URL `http://www.ejel.org/Volume-7/v7-i2/v7-i2-art-6.htm`

Landis, J. R., Koch, G. G., 1977. The measurement of observer agreement for categorical data. biometrics, 159–174.

Massart, D., Shulman, E., Van Assche, F., 2011. LRE Metadata Application Profile version 4.5. European Schoolnet.

Meire, M., Ochoa, X., Duval, E., June 2007. Samgi: Automatic metadata generation v2. 0. In: World Conference on Educational Multimedia, Hypermedia and Telecommunications. pp. 1195–1204.

Nejdl, W., Wolf, B., Qu, C., Decker, S., Sintek, M., Naeve, A., Nilsson, M., Palmér, M., Risch, T., 2002. EDUTELLA: a P2P networking infrastructure based on RDF. In: Proceedings of the 11th international conference on World Wide Web. pp. 604–615.
URL http://portal.acm.org/citation.cfm?id=511446.511525

Neven, F., Duval, E., 2002. Reusable Learning Objects : a Survey of LOM-Based Repositories. Evaluation 68 (4), 291–294.
URL https://lirias.kuleuven.be/handle/123456789/134106

Nezhadi, A. H., Shadgar, B., Osareh, A., 2011. Ontology alignment using machine learning techniques. International Journal of Computer Science & Information Technology 3 (2), 139.

Ngo, D., Bellahsene, Z., 2012. YAM++: A multi-strategy based approach for ontology matching task. Vol. 7603 LNAI. Springer Verlag.

Niemann, K., Moumoutzis, N., Rajabi, E., Manolis, N., Stoitsis, G., Simon, B., Zimmer, K., Megalou, E., Kyriazos, G., Petalas, G., 2013. Open Learning Content Architecture & Interoperability. Tech. rep., Open Discovery Space.

Nikolas, A., Sotiriou, S., Zervas, P., Sampson, D. G., 2014. The open discovery space portal: A socially-powered and open federated infrastructure. In: Digital Systems for Open Access to Formal and Informal Learning. Springer, pp. 11–23.

Ochoa, X., 2005. Learning object repositories are useful, but are they usable? In: IADIS AC. pp. 138–144.

Open Educational Resources, 2013. OER Commons.
URL http://www.oercommons.org/

OpenStax, 1999. OpenStax CNX. https://cnx.org/about, Last accessed: July, 15th 2016.

Otero-Cerdeira, L., Rodríguez-Martínez, F. J., Gómez-Rodríguez, A., 2015. Ontology matching: A literature review. Expert Systems with Applications 42 (2), 949–971.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2012. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 12, 2825–2830.

Pérez-Rodríguez, R., Anido-Rifón, L., Gómez-Carballa, M., Mouriño-García, M., 2016. Architecture of a concept-based information retrieval system for educational resources. Science of Computing Programming 129, 72–91.

Phang, C., Kankanhalli, A., Ramakrishnan, K., Raman, K. S., 2010. Customers' preference of online store visit strategies: an investigation of demographic variables. European Journal of Information Systems 19, 344–358.

Porter, M. F., 1980. An algorithm for suffix stripping. Program: electronic library and information systems 14 (3), 130–137.

Rawlings, J. O., Pantula, S. G., Dickey, D. A., 1998. Applied regression analysis: a research tool. Springer Science & Business Media.

Rigutini, L., Maggini, M., Liu, B., 2005. An EM based training algorithm for Cross-language Text Categorization. In: Proceedings - 2005 IEEE/WIC/ACM InternationalConference on Web Intelligence, WI 2005. Vol. 2005. pp. 529–535.

Rodriguez, M. A., Bollen, J., Sompel, H. V. D., Mar. 2009. Automatic metadata generation using associative networks. ACM Trans. Inf. Syst. 27 (2), 7:1–7:20. URL http://doi.acm.org/10.1145/1462198.1462199

Roy, D., Sarkar, S., Ghose, S., 2010. A comparative study of learning object metadata, learning material repositories, metadata annotation & an automatic metadata annotation tool. Advances in Semantic Computing 2, 103–126.

Sacchet, M. D., Prasad, G., Foland-Ross, L. C., Thompson, P. M., Gotlib, I. H., 2015. Support vector machine classification of major depressive disorder using diffusion-weighted neuroimaging and graph theory. Frontiers in psychiatry 6.

Salton, G., Wong, A., Yang, C. S., 1975. A vector space model for automatic indexing. Communications of the ACM 18 (11), 613–620.

Sebastiani, F., Mar. 2002. Machine learning in automated text categorization. ACM Computing Surveys 34 (1), 1–47.
URL `http://portal.acm.org/citation.cfm?doid=505282.505283`

Settles, B., 2010. Active Learning Literature Survey. Machine Learning 15 (2), 201–221.

Shvaiko, P., 2013. Ontology Matching : State of the Art and Future Challenges. IEEE Transactions on Knowledge and Data Engineering 25 (1), 158–176.

Smiley, D., Pugh, E., 2009. Apache Solr 3 Enterprise Search Server. PACKT Books.

Sokolova, M., Lapalme, G., 2009. A systematic analysis of performance measures for classification tasks.

Stefaner, M., Dalla Vecchia, E., Condotta, M., Wolpers, M., Specht, M., Apelt, S., Duval, E., 2007. MACE-enriching architectural learning objects for experience multiplication. In: European Conference on Technology Enhanced Learning, EC-TEL 2007: Creating New Learning Experiences on a Global Scale. pp. 322–336.
URL `http://www.springerlink.com/index/2036201625m1rq21.pdf`

Straccia, U., Troncy, R., 2005. omap: Results of the ontology alignment contest. In: Workshop on Integrating Ontologies. pp. 92–96.

Täckström, O., 2005. An Evaluation of Bag-of-Concepts Representations in Automatic Text Classification. Master's thesis TRITA-NA-E05150, KTH Numerical Analysis and Computer Science.

Ternier, S., Verbert, K., Parra, G., Vandeputte, B., Klerkx, J., Duval, E., Ordonez, V., Ochoa, X., Jul. 2009. The Ariadne Infrastructure for Managing and Storing Metadata. IEEE Internet Computing 13 (4), 18–25.

Tsoumakas, G., Katakis, I., 2007. Multi-label classification: An overview. International Journal of Data Warehousing and Mining 3, 1–13.

UNESCO, 2002. Forum on the impact of open courseware for higher education developing countries. Final report. Tech. rep., UNESCO, Paris.

Wolpers, M., Grohmann, G., 2005. PROLEARN: technology-enhanced learning and knowledge distribution for the corporate world. International Journal of Knowledge and Learning 1 (1/2), 44.

Yang, Y., 1999. An Evaluation of Statistical Approaches to Text Categorization. Information Retrieval 1 (1), 69–90.

Yang, Y., Liu, X., 1999. A re-examination of text categorization methods. In: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval SIGIR 99. pp. 42–49.

Zobel, J., Moffat, A., Ramamohanarao, K., 1998. Inverted files versus signature files for text indexing. ACM Transactions on Database Systems (TODS) 23 (4), 453–490.