Semantic annotation and harvesting of federated scholarly data using ontologies

Dimitrios A. Koutsomitropoulos
High Performance Information Systems Laboratory (HPCLab), Computer Engineering and Informatics Department, School of Engineering, University of Patras, Patras-Rio, Greece

Received 9 December 2018
Revised 7 April 2019
Accepted 28 July 2019

Abstract

Purpose – Effective synthesis of learning material is a multidimensional problem, which often relies on handpicking approaches and human expertise. Sources of educational content exist in a variety of forms, each offering proprietary metadata information and search facilities. This paper aims to show that it is possible to harvest scholarly resources from various repositories of open educational resources (OERs) in a federated manner. In addition, their subject can be automatically annotated using ontology inference and standard thematic terminologies.

Design/methodology/approach – Based on a semantic interpretation of their metadata, authors can align external collections and maintain them in a shared knowledge pool known as the Learning Object Ontology Repository (LOOR). The author leverages the LOOR and show that it is possible to search through various educational repositories' metadata and amalgamate their semantics into a common learning object (LO) ontology. The author then proceeds with automatic subject classification of LOs using keyword expansion and referencing standard taxonomic vocabularies for thematic classification, expressed in SKOS.

Findings – The approach for automatic subject classification simply takes advantage of the implicit information in the searching and selection process and combines them with expert knowledge in the domain of reference (SKOS thesauri). This is shown to improve recall by a considerable factor, while precision remains unaffected.

Originality/value – To the best of the author's knowledge, the idea of subject classification of LOs through the reuse of search query terms combined with SKOS-based matching and expansion has not been investigated before in a federated scholarly setting.

Keywords Thesauri, Open educational repositories, Ontologies, Learning objects, Subject classification, Keywords expansion

Paper type Research paper

1. Introduction

Today's libraries and institutions are striving to repurpose their role from centralized information providers to curators and mediators of educational content. The inflation of open courses and the Massive Online Open Courses (MOOCs) trend are tantalizing business-model seekers looking for monetization (Kerres and Heinen, 2015).

Yet, there is still a lot of potential in educational content providers, including OER repositories and aggregators worldwide, that can form the "gold-standard" of reference for digital learning-object consumers. For this kind of applications, we see three major standing problems:

- Content searching and reuse requires careful and elaborate selection, more often than not backed by a manual process.
- As rich as the object's metadata may be, there is still a considerable semantic gap between annotations originating from different sources.

Digital Library Perspectives

2059-5816
DOI 10.1108/DLP-12-2018-0038

- The act of content selection, for example, as an extra-curricular suggestion for a university course, remains an intangible intellectual asset and is usually evaded, in the machine-readable sense.

In this paper, we primarily tackle the latter two problems. By leveraging the notion of a Learning Object Ontology Repository (LOOR) (Koutsomitropoulos and Solomou, 2018), we show that it is possible to search through various educational repositories' metadata and amalgamate their semantics into a common learning object (LO) ontology. We then proceed with automatic subject classification of LOs using keyword expansion and referencing standard taxonomic vocabularies for thematic classification, expressed in SKOS (Miles *et al.*, 2009). In essence, original user search keywords can be reused to provide subject annotations for selected learning objects by curators or instructors. However, merely supplying these keywords directly as subjects would be impractical, since they can be ad-hoc and uncontrolled. Instead, the keywords can be first matched against an authoritative, formal thematic thesauri, and the matches be injected as semantic annotations into the selected LOs. It is shown that this approach can lead to improved retrieval of annotated objects originating from various repositories.

To our knowledge, the idea of subject classification of LOs through the reuse of search query terms combined with SKOS-based matching and expansion has not been investigated before in a federated scholarly setting.

Our work stems from the collaboration with the Library and Information Services of the University of Thrace, where we have successfully deployed a LOOR to support online courses. Data sources include *MERLOT II*, a large archive of OERs (McMartin, 2006), *Europe PubMed Central*, a major repository of biomedical literature (Europe PMC Consortium, 2015), *ARIADNE finder*, a European infrastructure for accessing and sharing learning resources (Ternier *et al.*, 2009) and *openarchives.gr*, the entry point for Greek scholarly content. The current prototype is available at: https://github.com/swigroup/federated-semantic-search

In the following, we examine some related work involving scholarly dataset federation and approaches for semantic annotation and classification of learning material in repositories. Next, we introduce the establishment of a LOOR that aggregates search results from other primary educational sources. In the following section, we present the process for automatic subject annotation of aggregated resources by expanding original search terms and matching against standard thematic vocabularies. Then, we discuss evaluation results in terms of improved retrieval and outline usage examples. Finally, the last section summarizes our conclusions and future work.

2. Related work

There are many efforts for federated access to learning objects in the literature. For example, in De la Prieta *et al.* (2014), the necessity and significance of aligning learning content providers is recognized, and the authors describe a cloud-based architecture for the integration of federated search services. ASPECT (ASPECT Project, 2009) and, more recently, OpenAIRE (Atzori *et al.*, 2018) and Europeana (Petras *et al.*, 2017) are examples of European projects whose purpose is the aggregation of resources from multiple providers. ASPECT focuses on learning objects whereas OpenAIRE provides a unique access point for European research outcomes. Similarly, Europeana combines data from thousands of institutions across Europe while these data are transformed into Linked Data. One can query the integrated content by using SPARQL queries.

Furthermore, various approaches exploit Semantic Web technologies and propose integration architectures for digital repositories, based on Linked Data and federated queries (Mosharraf and Taghiyareh, 2016; Piedra *et al.*, 2014; Segarra *et al.*, 2016). These architectures are leveraged to integrate a group of institutional repositories belonging to universities, thus building a federated query environment in this context.

Most approaches for improving LOs discoverability and subject characterization of LOs can be roughly grouped in two categories: Methods using some form of training or probabilistic techniques and methods leveraging metadata semantics and content descriptions using ontologies.

Regarding learning techniques, López *et al.* (2012) are based on multi-label classification methods to achieve the automated classification of LO collections in multiple type-based queries (classes). Query terms identified by teachers and pupils as necessary to support their learning discovery activities are reused to label search results from two different OER repositories. This set of LOs is then used as a training set to obtain predictions for an unlabeled dataset. In Lama *et al.* (2012), authors study the automated classification of LOs through their integration with a set of DBpedia categories. According to this approach, the LO subject – represented by the corresponding text-based field of the IEEE LOM standard (Hodgins and Duval, 2002) – is correlated with a set of categories which are semantically described in DBpedia. However, the accuracy of the above approaches in the characterization of LOs usually depends on an exhaustive and elaborate description of specific knowledge domains, while the latter can be naturally modeled with authority lists and thesauri.

Training methods using large lexical corpora for subject categorization of LOs have also been investigated (Gabrilovich and Markovitch, 2007; Meyer *et al.*, 2007) and show that Wikipedia can be used for improving the classification accuracy in various knowledge domains. Proposed methods use machine learning techniques to explicitly represent the meaning of any text as a weighted vector of Wikipedia-based concepts. However, a drawback of these approaches is that having available the appropriate corpus – such as that of Wikipedia – is often not an option within the context of an OER repository. On the other hand, the alternative to manually create a specialized training corpus requires a considerable effort.

Methods using semantic Web ontologies and Linked Data are the closest to our approach. The problem of free-text and unstructured metadata annotations of LOs can be addressed by exploiting existing thematic vocabularies. In Dietze *et al.* (2012), the authors try to further extend LOs discoverability and sharing by linking LOs descriptions with well-defined terms inside various established thesauri. As a result, data from a number of OER repositories can be integrated, exposed and finally enriched. Another approach towards interlinking of learning resources is presented in Rajabi *et al.* (2015). This work describes how to transform IEEE LOM metadata into XML representation and RDF triples and finally link them to other datasets, e.g. DBpedia.

SKOS has been recognized as a standard for promoting semantic interoperability between learning objects and thematic taxonomies early on. In Dehors and Faron-Zucker (2006), a semantic Web-based learning system is presented, where educational resources are described by subject ontologies and domain concepts are expressed using SKOS. Similarly, in Miranda *et al.* (2006), authors are based on SKOS for modelling their subject ontologies – which are used for organizing various aspects of an e-learning system – and propose a framework with improved recommendation processes.

To our knowledge, the idea of subject classification of LOs through the reuse of search query terms combined with SKOS-based matching and expansion has not been investigated

before in a federated scholarly setting. However, the effects of term expansion are generally acknowledged to enhance resource retrieval, especially when combined with expert knowledge. SKOS-based query expansion can achieve improved results in Web search (Haslhofer *et al.*, 2013) or recommendation (Wenige *et al.*, 2018). In Segura *et al.* (2011), the authors confirm expanded queries allow the user to retrieve relevant objects, which might not be obtained without expansion, but by using a non-SKOS ontology and addressing only a single repository.

3. Federated harversting into the LOOR

Learning resources residing within OER repositories are frequently well-documented and annotated with a rich set of meta-information. It is often the case however that these metadata follow different standards or even worse, proprietary metadata formats. Another common problem is that these metadata annotations are simple textual fields, given as inputs to flat XML elements and thus any structure or information that may exist in the semantics of descriptions is not maintained. This becomes even more critical in situations such as aided annotation and discovery of educational resources in repositories, where semantic matches can provide considerable added-value, for example, by considering semantic networks and thesauri. To address these concerns, the creation and maintenance of a semantics-aware learning repository has been proposed (Koutsomitropoulos and Solomou, 2018). Such a repository would tap into ontologies and thesauri and allow LO metadata instances be assigned machine-understandable semantic annotations.

Building on this premise, in this section, we discuss how LO can be ingested, and different schemata be aligned within the LOOR into a common LO ontology. First, however, we outline the architecture of our approach, which would allow us to annotate LOs with automatic subject suggestions.

3.1 Architecture and process flow

To design an architectural framework for federated harvesting and metadata annotation of OERs we have taken into account two main principles:

(1) *Use Web ontologies as a general semantic framework*: Ontologies are rich conceptual schemas, designed exactly to effectively capture knowledge and hence capable of optimizing the management, searching and discovery among repository resources which are represented in the form of LOs (Jensen, 2019).

(2) *Maintain a level of semantic interoperability between repositories*: To achieve this, we proceed by designing a unifying ontology for harvested LOs, that is based on broadly adopted standards in the e-learning domain, such as LOM and Dublin Core. Moreover, we link resources to formal thematic thesauri, a fact that reduces the chance for ambiguous interpretation of subject terms. Thesauri are expressed in the SKOS standard, thus facilitating integration and interoperability with other discovery mechanisms, digital repositories and the Web of Linked and Open Data (LOD) (Binding and Tudhope, 2016).

The overall architecture of the proposed approach, from federated searching of the remote repositories to automatic subject annotation to the population of the LO ontology is shown in Figure 1.

First, a federated query is initiated towards the various repositories. Next, the metadata of items returned as responses to the query are harvested and aligned to a unified LO Ontology Schema, using a common set of elements and mapping rules. The subject of these

items is then automatically populated taking as basis the initial query keywords in accordance with thematic thesauri for specific knowledge domains. At this point, a curator or instructor may review the LO and decide to incorporate it into the LO ontology, thus letting it available for others to reuse.

3.2 Metadata alignment and harvesting

While different metadata schemata may pose barriers for direct integration, it is possible to identify a least common set of elements and use well known educational metadata standards such as LOM, as mediators. This set can form the basis for a common schema to be used for immediate ingestion of learning objects into the LOOR. Taking university courses as a use case, it would be useful to interoperate with course management systems, such as Open eClass (GUnet, 2019), which is heavily used by Greek universities. In the context of Open eClass, a *course link* represents external material that can be suggested by an instructor for their course and can be specified by setting the following four metadata values: a *URL*, a *URL Title*, a *URL Description* and a *Category*. Category is optional and can be used to provide an arbitrary header to group links, e.g. in a thematic manner.

Therefore, a solution for metadata alignment is to extract all these useful and essential details from the incoming collections and then map them to one unified ontology schema, which at least contains the above four fields (Figure 2). The rest of the metadata annotations are not lost. Rather, they are easy to retrieve directly from their sources, using their unique URL or harvest them through an OAI service provider. Further, a curator or instructor can manually review automatically assigned values, edit the rest of the fields of the unified schema and opt for the addition of the item into the LOOR.

However, incoming collections may contain metadata in proprietary structure. For example, in openarchives.gr, the main result node of the response is identified with the <entry> node and roughly follows RSS. For the Europe PubMed Central, each <result>

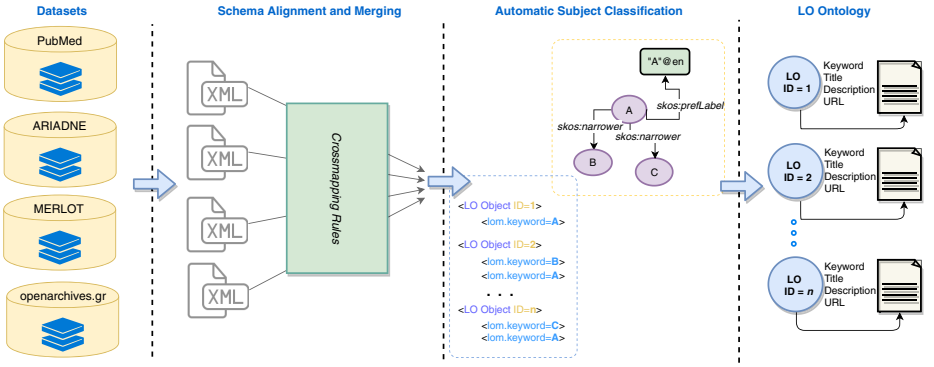

Figure 1. Architecture for federated harvesting and subject classification of learning objects

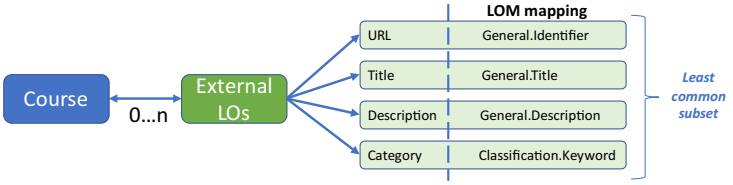

Figure 2. Metadata alignment of external repositories and mapping to LOM

node corresponds to a search result and so on. To align external OER repositories to our unified schema, we have designed a process that allows to express a set of cross-mapping rules between each repository's output and the LO Ontology Schema (see sub-section 3.3 below). This is implemented by a configuration file expressed in JSON or programmatically using enumerations. An example of such a configuration file including the query string parameter for two of the repositories in the federation, as well as their mapping rules is shown in Figure 3. Therefore, the accommodation of additional external repositories becomes parametric and straightforward: one has simply to express another set of mapping rules and add it to the configuration file.

3.3 Learning object ontology schema

The full metadata application profile for our implementation and the associated ontology schema have been detailed in (Koutsomitropoulos and Solomou, 2018). For the paper to be self-contained, we briefly describe part of this LO ontology and its main characteristics.

An essential step in the migration from a flat schema to a LOOR is a "semantification" process i.e. the transformation of the textual information captured by a metadata instance into a semantically enriched and thus machine-understandable format. The resulting *LO Ontology Schema* contains entities representing elements of the IEEE LOM schema and combines terminology with the Dublin Core metadata terms specification. This correlation helps control the values of fields for LOM properties and can increase interoperability with applications that are based on DC.

The *lom:LearningObject* class is a top class used to capture the notion of an LO, or an educational resource in general. The various characteristics of an educational resource are represented as either classes or properties in this ontological schema. The datatype properties *lom:description*, *lom:identifier*, *lom:language*, *lom:rights*, *lom:size*, and *lom:title* are used to declare a short description, a unique identifier, the LO's content language, the copyright policies, and finally LO's physical size and title, respectively. We chose to express these elements of the LOM schema as datatype- and not as object- properties given that they simply assign values to some of the resources' basic characteristics and convey no correlations among them.

The *lom:learningResourceType* object property aims at specifying the different educational types that can be assigned to LOs and it is associated with a predefined list of terms (Exercise, Experiment, Figure, Lecture, etc.). In a similar way, concepts met in our LO metadata profile, like the groups of end-users to which a LO applies, the intended

```
{
    "repositoryA": {
        "repository_url": "http://www.ebi.ac.uk/europepmc/webservices/rest/search/query=",
        "repository_extra_parameter": "&resultType=core",
        "metadata": {
            "result": "result",
            "url": "url",
            "url_title": "title",
            "url_description": "abstractText"
        }
    },
    "repositoryB": {
        "repository_url": "http://openarchives.gr/opensearch/",
        "repository_extra_parameter": "/limit:25",
        "metadata": {
            "result": "entry",
            "url": "content",
            "url_title": "title",
            "url_description": "dc:identifier"
        }
    }
}
```


Figure 3.
Cross-mapping rules and configuration parameters for the alignment of federated repositories

instructional context, LO's level of difficulty, average learning time, level of completeness (draft, revised or final) and type of interaction (active, expositive, etc.) are captured using the appropriate object properties *lom:intendedEndUserRole*, *lom:context*, *lom:difficulty*, *lom:typicalLearningTime*, *lom:status*, *lom:interactivityType* respectively (Figure 4).

Potential relationships among LOs can be captured via the object property *lom:relation*, which is used exactly to correlate between instances of the *lom:LearningObject* class. In addition, we use the *dcterms:Agent* class to include any person or organization responsible for the creation (or other modifications) of an educational resource. The object property *lom:contributor* comes to implement this type of relation.

Finally, it is important to note that the *lom:keyword* property, used in our LO profile to express the thematic subject of the LO's content, is represented as an object- rather than a datatype- property. This is to directly correlate the subject keywords of a LO to SKOS ontology concepts, thus increasing the value of our LO ontology when used in the context of knowledge discovery applications.

3.4 SKOS thesauri

As a thesaurus of reference, we have used the *Thesaurus of Greek Terms* (TGT) a bilingual (Greek, English) controlled vocabulary published by the National Documentation Center in

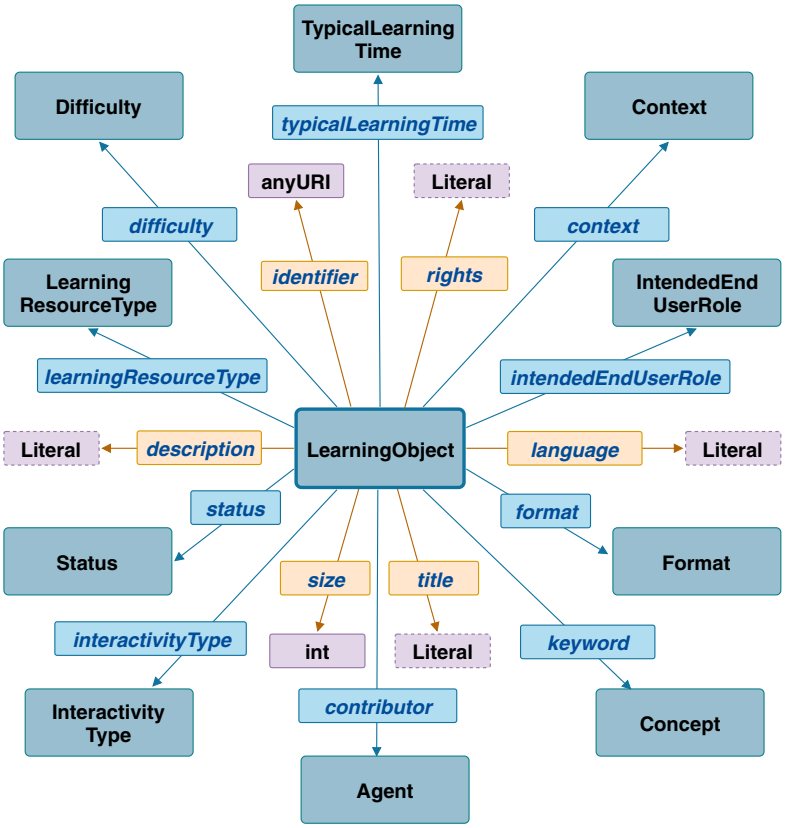

Figure 4.
Visualization of the LO ontology schema classes and properties

Greece (National Documentation Center, 2019). As would be appropriate for a scholarly setting, we chose to bind to specific knowledge domains. To this end, we extracted two thematic micro-thesauri out of TGT for the fields of mathematics and medicine: The *Maths Thesaurus* comprises 76 terms, making reference to another 17 related terms, whereas the *Medicine Thesaurus* contains 54 terms and makes reference to 71 additional terms.

SKOS is a Semantic Web standard that facilitates publication and use of thesauri as Linked Data (Miles *et al.*, 2009). SKOS is expressed in machine readable format, namely, RDF(S) and OWL. Neither these micro-thesauri nor TGT were expressed in SKOS format. Through an appropriate mapping process, we achieved the SKOS transformation of these two thesauri, from their initial XML format into RDF/OWL. In addition, we experimented with larger thesauri, for example *MeSH* (Medical Subject Headings), already converted into SKOS (Van Assem *et al.*, 2006) and comprising of 23,883 concepts.

4. Subject classification of learning objects

The purpose of query expansion when harvesting OERs has been investigated before and its coupling with standard thematic vocabularies and ontology-based reasoning has been shown to improve the searching process in an educational setting (Koutsomitropoulos *et al.*, 2017). This meta-information caused by the unfolding of the thesauri term hierarchy may however be lost once an item has been selected and ingested into the LOOR. We can maintain and reuse this information to thematically annotate such an item when it is added into the LOOR but taking a slightly different stance this time.

When performing query expansion, items matching narrower topics are indeed exact matches - therefore query expansion on narrower/related terms is effective. But one cannot assign narrower terms as a subject to items matching only higher-level concepts. Normally, an item should be assigned its matching concept and at most, the latter's broader path to the parent concept, but no narrower concepts.

So, searching for a broad concept A should also fetch all items matching narrower concepts (similar to the way query expansion currently works). Searching for a narrower concept B however, should not return the whole set of items (for example the ones matching A) but only the items that match B's narrower concept tree.

As a result, not only this information is retained but also the subject of an LO can be assigned a *concept* of a thematic thesaurus, rather than a mere text keyword, thus improving interoperability and retrieval. For once, resources with content characterized by related, narrower or broader in meaning concepts (and captured through the corresponding SKOS properties) can also be retrieved. In addition, concept descriptions themselves can be accessed as Linked Data. Each thesaurus term in the LOOR is linked to its source structured data in the parent institution, where one can navigate the term hierarchy and explore further relations (Georgiadis *et al.*, 2016).

More formally, the problem of assigning subject annotations to learning objects can be expressed as follows, using model-theoretic semantics:

Let $i_1, i_2, \ldots, i_n \in I$ individuals corresponding to each learning object searched for and finally selected by the instructor for inclusion into the LOOR and I the individuals domain. Also, let P an object property, with $P \in I \times I$ and $R \equiv P^-$ the inverse of P. P represents the *transitive closure* of a broader/narrower relationship between terms of the thesaurus (e.g. skos:broader). Then, if k is the original keyword (it is trivial to generalize for more keywords, by iteration) the problem of finding subject annotations can be reduced to finding all SKOS concepts y such that:

$T \sqsubseteq P<x, y> \cup R<y, x>$ where $x, y \in$ skos:Concept and $<x, k> \in$ skos:prefLabel $\cup$ skos:altLabel

Now, subject annotations can be assigned to learning objects with the assertions $K(i_1,x)$, $K(i_1,y)$, $K(i_2,x)$, $K(i_2,y)$, …, $K(i_n,x)$, $K(i_n,y)$ for all x and y, where K represents the lom: keyword property.

The selection of subject annotations for search results follows the simple algorithm shown below.

```
∀ keyword k:
        //exact matches put first
        C_k ← find_matching_concepts (k)
∀ C_k, ∀ c ∈ C_k:
        D_k ← C_k ⋃ find_narrower_concepts (c)
∀ label l: <c, l> ∈ skos:prefLabel ⋃ skos:altLabel, c ∈ D_k
        // c is the matching concept for l
        C_l ← C_l ⋃ {c} ⋃ C_k[1]
```

find_matching_concepts looks within the thesaurus to find concepts that have a matching lexical representation, or label, with the search keyword(s). A concept is considered a match when one or more of its labels matches the provided keyword.

find_narrower_concepts takes advantage of broader/narrower relationships in the thesaurus ontology and fetches concepts that refine or specialize the initially discovered matching concepts up to a certain depth. These refinements are added to the initial set C_k for the particular keyword k. However, because of term expansion, k would be expanded to include also the labels of its matching *and* narrower concepts (set D_k). Therefore, a new concept set is also created for each label that a concept belonging in D_k may have. This new set C_l contains the concept c owing the label and the uppermost parent concept that led to c. This is reasonable, since a broader concept is also a valid thematic subject for the resource in a classification hierarchy. In case l has been already visited (i.e. more than one concept having the same label) the set is simply augmented.

The labels set would then be used for querying. This includes also terms translations, supported in multilingual thesauri such as TGT. The concepts of C_l will be used as subject annotations to every learning object satisfying the query for term l.

For a matching concept c, the thesaurus is being traversed in *BFS* order in the direction of the skos:narrower property, either asserted or inferred, to find the refining concepts (i.e. compute the transitive closure). Regarding performance, it is easily seen that the complexity of our algorithm, given the BFS traversing is, in the case of thesauri trees, polynomial to the number of axioms in the ontology.

```
<skos:Concept rdf:about="http://thesaurus.duth.gr/medicine/medicine">
  <skos:prefLabel xml:lang="en">medicine</skos:prefLabel>
  <skos:prefLabel xml:lang="el">ιατρική</skos:prefLabel>
  <skos:narrower rdf:resource="http://thesaurus.duth.gr/medicine/surgery"/>
  <skos:narrower rdf:resource="http://thesaurus.duth.gr/medicine/ophthalmology"/>
</skos:Concept>
<skos:Concept rdf:about="http://thesaurus.duth.gr/medicine/surgery">
  <skos:prefLabel xml:lang="en">surgery</skos:prefLabel>
  <skos:prefLabel xml:lang="el">χειρουργική</skos:prefLabel>
  <skos:broader rdf:resource="http://thesaurus.duth.gr/medicine/medicine"/>
</skos:Concept>
<skos:Concept rdf:about="http://thesaurus.duth.gr/medicine/ophthalmology">
  <skos:prefLabel xml:lang="en">ophthalmology</skos:prefLabel>
  <skos:prefLabel xml:lang="el">οφθαλμολογία</skos:prefLabel>
  <skos:broader rdf:resource="http://thesaurus.duth.gr/medicine/medicine"/>
</skos:Concept>
```

Figure 5.
Excerpt of the medicine thesaurus in SKOS, modeling the concept medicine and two narrower terms

As an example, consider the following snippet from the medicine thesaurus (Figure 5) and the input keyword "medicine". The algorithm would first fetch the concept *medicine* having found an exact match with its prefLabel. Next, the narrower concepts of *medicine* that is, *surgery* and *ophthalmology* would be added to the set. Their labels will be used for query expansion in addition to the labels of the original matching concept. Thus, search results for the original keyword will be annotated with the SKOS concept *medicine*, as will be search results for the other labels of this concept. Results matching the various labels of *surgery* will be automatically classified with the concept *surgery* as well as the concept *medicine*, since this is the first (and the only) matching concept for the initial keyword. Likewise, results matching labels of the concept *ophthalmology* will be assigned both *ophthalmology* and *medicine*.

5. Evaluation and usage

Automated subject annotation naturally helps with search and discovery of learning content. Especially in fielded search or faceted browsing, correct subject classification can enhance recall of LOs when the subject of an LO is used as the primary dimension for search and navigation (English *et al.*, 2002).

In summary, this approach has the following benefits:

- *enables triple queries*: the triplification of resources' metadata and their alignment into a unifying RDF/OWL schema makes it possible to perform SPARQL queries by virtue of the corresponding lom:keyword triples that are being injected.
- *enables semantic search*: The fillers of the lom:keyword property are individual objects themselves (SKOS concepts) rather than mere keywords. Therefore, queries can be combined with reasoning results, like relationships between concepts, transitivity etc. to achieve semantic searches. Additionally, concepts can have their own properties as well as a variety of lexical representations (alternate labels, translations), which are considered during term expansion.
- *improves search recall*: expanded query terms refine initial search keywords and thus fetch additional LOs which can greatly expand the initial result set with resources of guaranteed relevance.
- *facilitates semantic interoperability and authority control*: SKOS concepts selected for subject annotation originate from standardized thematic taxonomies that are established, documented and curated by professional documentation organizations.

In the following we first discuss some qualitative aspects and examples of this approach regarding LO retrieval. Then we present the results of experiments that compare recall performance against searching remote repositories directly.

5.1 Qualitative account and usage examples

Compared with term expansion for generic keyword search, the subject annotation effect is complementary: when performing federated search over remote repositories and using their own search mechanisms, term expansion fetches many more LOs that are guaranteed to be of related or narrower meaning to the original keyword. Conversely, when searching LOs by subject within the LOOR or another system that interoperates with it (a course management system, a SPARQL endpoint), augmenting their subject with SKOS concepts can fetch LOs that would be impossible to get otherwise, i.e. if it was not for their automatic subject classification presented previously.

As a side note, keeping explicitly the top-level SKOS concept in the subject (lom: keyword) of an LO can further improve retrieval. This virtually amounts to the amalgamation of broader/narrower inferences of matching SKOS concepts into the LO ontology. Although a reasoning-based search mechanism could make the appropriate inferences, not all systems are reasoning-enabled and inference operations are often deemed to be costly (Koutsomitropoulos *et al.*, 2010).

The original search keyword, the one that seeded the expansion process, is also retained into the LOs metadata by keeping it as an additional lom:keyword. This is helpful for queries that may rely solely on text search or that are SKOS oblivious, and is a useful fallback measure.

Finally, search precision remains unaffected. LOs imported into the LOOR match search keywords (original or expanded irrespectively) based on the remote repository search capabilities. If search results are correct for the remote premises, they would still be correct for our own LOOR and vice-versa. A detailed discussion of this issue and corresponding proof can be sought in (Koutsomitropoulos *et al.*, 2017).

To see an example, consider the following SPARQL query performed on a SPARQL endpoint connected to the LOOR, which asks for LOs having the skos:Concept *medicine* as their subject:

```
SELECT? lo WHERE {
  ?lo a lom:LearningObject .
  ?lo lom:keyword [skos:preflabel "medicine"@en; a skos:Concept] }
```

Given that subject annotations have been already automatically injected into the LOs during the term expansion and subject classification phase, this would fetch not only direct matches but also narrower and related ones (e.g. alternate labels, translations etc.). What is more, all LOs are *guaranteed* to have a lom:keyword property and, even more so one that is connected to a *skos:Concept*, rather than just text.

In contrast, consider a user performing a free text search for the keyword "medicine" in the remote repositories (federated search with no query expansion). Then, just the results containing this specific keyword within their metadata, and only them, would be fetched, ordered using the repository's relevance metric. Needless to say, if we were to perform a SPARQL query, similar to the one above, directly to the remote repositories that is, skipping the LOOR entirely, this would only work for LOs that have their subject already filled by a SKOS concept; even then, only results that contain this concept (medicine) explicitly would be returned.

5.2 Search recall

For the purposes of evaluation, we compare recall performance between two cases:

(1) Searching remote repositories directly, without any term expansion or subject annotation. This is the standard approach followed in current practice (Mosharraf and Taghiyareh, 2016; Piedra *et al.*, 2014) when not considering implicit semantic annotations of resources. Within the LOOR context, searching LOs without subject annotation means that resources get annotated only with the original search keyword that retrieved them. Therefore, for evaluating retrieval, this can form a baseline for comparison.

(2) Searching the collection of LOs harvested into the LOOR after term expansion and subject classification has occurred.

Both cases can be represented by the following SPARQL query, given a search keyword *k*:

```
SELECT? lo WHERE {
  ?lo a lom:LearningObject .
  ?lo lom:keyword? c .
  {{?c skos:prefLabel | skos:altLabel "k" .}
  UNION {FILTER(?c="k")}}}
```

The search keywords used for the experiments have been chosen to be evenly distributed among high-level thesauri terms (lots of children in the hierarchy), mid-level terms (fewer children in the hierarchy), bottom-level terms (no children) and keywords having no matches whatsoever in the thesauri used. All experiments were conducted on standard commodity hardware. A four-core laptop CPU was used, and Java was assigned 4GB for heap memory.

Figure 6 summarizes the results of these experiments for the three thesauri used: medicine thesaurus, math thesaurus and MeSH. We have allowed a keyword to expand up to four, eight and twelve terms respectively. We notice a boost in recall almost by a factor of *8* when allowing the search keyword to expand up to 12 terms, i.e. we are able to retrieve up to eight times more LOs. But even with fewer terms, recall is considerably increased, between 3 and 4 times, depending on the thesaurus.

Another important effect this approach has on retrieval is multilingualism. The use of the two bilingual (Greek-English) thesauri about medicine and mathematics can help to correctly classify matching LOs, because a term in these thesauri includes labels in both languages (translations). Therefore, where once would be no concept match for a keyword in a different language, now resulting LOs get annotated with the correct SKOS concept, regardless of their language or the language of the initial keyword, provided it matches one of the concept's labels.

6. Conclusions and future work

Subject classification of LOs can play an important role when addressing the multitude of available OER repositories. On the one hand, harvesting remote sources can offer various possibilities to curators and instructors to seamlessly integrate additional educational material into their workflows for example, by interoperating with a course management system. On the other hand, automatic subject annotation can alleviate the manual handpicking of items and enhance reuse of LOs within the LOOR.

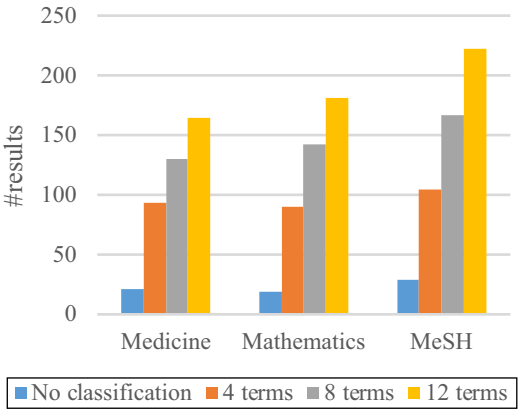

Figure 6.
Number of results and recall increase when applying automatic subject classification using term expansion

Indeed, the LOOR can then act as a local-to-the-institution, cached pool of LOs, each referencing its original source and be ready for retrieval (*lom:identifier* property). In addition, resources included in the repository are already semantically indexed by other experts. This facilitates the job of another instructor to easily and quickly discover additional material. It also has the added benefit of collaborative intelligence, by exposing material already trusted by colleagues.

The algorithm for automatic subject classification simply takes advantage of the implicit information in the searching and selection process and combines them with expert knowledge in the domain of reference (SKOS thesauri). This has been shown to greatly improve recall, but may have a cost, when search results are not already within the LOOR and need to be fetched from the remote repositories.

Although these metrics are promising, we intend to further evaluate our system in more "real-word" scenarios, by surveying instructors' opinion and learners' satisfaction when the system is employed into everyday practice. As an additional step, the algorithm can learn from previous instructor queries, analyze content and reuse potentially existing taxonomic information within a LO's metadata, to make more accurate subject suggestions and propose correlations to other material available in the web of knowledge.

References

ASPECT Project (2009), "ASPECT approach to federated search and harvesting of learning object repositories", Deliverable D2.1, ECP 2007 EDU 417008, available at: http://aspect-project.org/sites/default/files/docs/ASPECT_D2p1x.pdf

Atzori, C., Manghi, P. and Bardi, A. (2018), "De-duplicating the OpenAIRE scholarly communication big graph", *eScience*, Vol. 2018, pp. 372-373.

Binding, C. and Tudhope, D. (2016), "Improving interoperability using vocabulary linked data", *International Journal on Digital Libraries*, Vol. 17 No. 1, pp. 5-21.

De la Prieta, F., Gil, A.B., Martín, A.J.S. and Zato, C. (2014), "Learning object repositories with federated searcher over the cloud", *Methodologies and Intelligent Systems for Technology Enhanced Learning*, Springer International Publishing, pp. 93-100.

Dehors, S. and Faron-Zucker, C. (2006), "QBLS: a semantic web based learning system", *Proceedings of EdMedia: World Conference on Educational Media and Technology 2006*, *Association for the Advancement of Computing in Education (AACE)*, pp. 2795-2802.

Dietze, S., Yu, H.Q., Giordano, D., Kaldoudi, E., Dovrolis, N. and Taibi, D. (2012), "Linked education: interlinking educational resources and the web of data", *The 27th ACM Symposium On Applied Computing (SAC-2012), Special Track on Semantic Web and Applications*.

English, J., Hearst, M., Sinha, R., Swearingen, K. and Lee, K.P. (2002), "Flexible search and navigation using faceted metadata", Technical report, University of Berkeley, School of Information Management and Systems, 2003. Submitted for publication.

Europe PMC Consortium (2015), "Europe PMC: a Full-Text literature database for the life sciences and platform for innovation", *Nucleic Acids Research*, Vol. 43, pp. D1042-D1048, Database issue PMC. Web. 11 Aug. 2017.

Gabrilovich, E. and Markovitch, S. (2007), "Computing semantic relatedness using wikipedia based explicit semantic analysis", *Proceedings of The 20th International Joint Conference on Artificial Intelligence (IJCAI)*.

Georgiadis, H., Papanoti, A., Paschou, M., Roubani, A., Pelekanou, D., Chardouveli, D. and Sachini, E. (2016), "Semantics. gr: a self-improving service to repositories and aggregators for massively enriching their content", *10th International Conference on Metadata and Semantics Research, Digital Humanities and Digital Curation (DHC) workshop*, *Göttingen*.

GUnet (2019), "Open eClass e-learning platform", available at: www.openeclass.org/en/

Haslhofer, B., Martins, F. and Magalhães, J. (2013), "Using SKOS vocabularies for improving web search", *Proceedings of the 22nd international conference on World Wide Web companion*, International World Wide Web Conferences Steering Committee, pp. 1253-1258.

Hodgins, W. and Duval, E. (Eds). (2002), "Draft standard for learning object metadata", Institute of Electrical and Electronics Engineers, 2002, available at: http://ltsc.ieee.org/wg12/files/LOM_1484_12_1_v1_Final_Draft.pdf

Jensen, J. (2019), "A systematic literature review of the use of semantic web technologies in formal education", *British Journal of Educational Technology*, Vol. 50 No. 2, pp. 505-517.

Kerres, M. and Heinen, R. (2015), "Open informational ecosystems: the missing link for sharing resources for education", *The International Review of Research in Open and Distributed Learning*, Vol. 16 No. 1.

Koutsomitropoulos, D.A. and Solomou, G.D. (2018), "A learning object ontology repository to support annotation and discovery of educational resources using semantic thesauri", *IFLA Journal*, Vol. 44 No. 1, pp. 4-22.

Koutsomitropoulos, D., Solomou, G. and Kalou, K. (2017), "Federated semantic search using terminological thesauri for learning object discovery", *Journal of Enterprise Information Management*, Vol. 30 No. 5, pp. 795-808.

Koutsomitropoulos, D., Solomou, G., Pomonis, T., Aggelopoulos, P. and Papatheodorou, T. (2010), "Developing distributed reasoning-based applications for the semantic web", *2010 IEEE 24th International Conference on Advanced Information Networking and Applications Workshops (WAINA)*, *IEEE*, pp. 593-598.

Lama, M., Vidal, J.C., Otero-García, E., Bugarín, A. and Barro, S. (2012), "Semantic linking of learning object repositories to DBpedia", *Educational Technology and Society*, Vol. 15 No. 4, pp. 47-61.

López, V.F., de La Prieta, F., Ogihara, M. and Wong, D.D. (2012), "A model for multi-label classification and ranking of learning objects", *Expert Systems with Applications*, Vol. 39 No. 10, pp. 8878-8884.

McMartin, F. (2006), "MERLOT: a model for user involvement in digital library design and implementation", *Journal of Digital Information*, Vol. 5 No. 3.

Meyer, M., Rensing, C. and Steinmetz, R. (2007), "Categorizing learning objects based on wikipedia as substitute corpus", in Massart, D., Colin, J.-N. and Assche, F.V. (Eds), *CEUR Workshop Proceedings*, Vol. 311

Miles, A. and Bechhofer, S. (Eds) (2009), "SKOS simple knowledge organization system reference", W3C Recommendation, available at: www.w3.org/TR/skos-reference

Miranda, S., Orciuoli, F. and Sampson, D. (2006), "A SKOS-based framework for subject ontologies to improve learning experiences", *Computers in Human Behavior*, Vol. 61, pp. 609-621, doi: 10.1016/j.chb.2016.03.066.

Mosharraf, M. and Taghiyareh, F. (2016), "Federated search engine for open educational linked data", *Bulletin of IEEE Technical Committee on Learning Technology*, Vol. 18 No. 6

National Documentation Center (2019), "Thesaurus of greek terms", available at: http://general-terms.thesaurus.ekt.gr/vocab/index.php

Petras, V., Hill, T., Stiller, J. and Gade, M. (2017), "Europeana, a search engine for digitised cultural heritage material", *Datenbank-Spektrum*, Vol. 17 No. 1, pp. 41-46.

Piedra, N., Chicaiza, J., Lopez, J. and Tovar, E. (2014), "An architecture based on linked data technologies for the integration and reuse of OER in MOOCs context", *Open Praxis*, Vol. 6 No. 2, pp. 171-187.

Rajabi, E., Alonso, S.S. and Sicilia, M. (2015), "Interlinking educational resources to web of data through IEEE LOM", *Computer Science and Information Systems*, Vol. 12 No. 1, pp. 233-255.

Segarra, J., Ortiz, J., Espinoza, M. and Saquicela, V. (2016), "Integration of digital repositories through federated queries using semantic technologies", *Computing Conference (CLEI), 2016 XLII Latin American*, pp. 1-9.

Segura, N.A., García-Barriocanal, E. and Prieto, M. (2011), "An empirical analysis of ontology-based query expansion for learning resource searches using MERLOT and the gene ontology", *Knowledge-Based Systems*, Vol. 24 No. 1, pp. 119-133.

Ternier, S., Verbert, K., Parra, G., Vandeputte, B., Klerkx, J., Duval, E., Ordonez, V. and Ochoa, X. (2009), "The ariadne infrastructure for managing and storing metadata", *IEEE Internet Computing*, Vol. 13 No. 4.

Van Assem, M., Malaisé, V., Miles, A. and Schreiber, G. (2006), "A method to convert thesauri to SKOS", *The Semantic Web: Research and Applications: 3rd European Semantic Web Conference, ESWC 2006, Budva, Montenegro*, June 11-14, 2006, Proceedings, *Springer*, Vol. 4011, p. 95.

Wenige, L., Berger, G. and Ruhland, J. (2018), "SKOS-based concept expansion for LOD-enabled recommender systems", *Proceedings of the 12th International Conference on Metadata and Semantics Research (MTSR 2018)*, *Springer*, pp. 101-112.

Corresponding author
Dimitrios A. Koutsomitropoulos can be contacted at: koutsomi@ceid.upatras.gr