

Original citation:

Al Fayez, Reem Qadan and Joy, Mike. (2017) Using linked data for integrating educational medical web databases based on bioMedical ontologies. The Computer Journal, 60 pp. 369-388.

Permanent WRAP URL:

<http://wrap.warwick.ac.uk/81532>

Copyright and reuse:

The Warwick Research Archive Portal (WRAP) makes this work by researchers of the University of Warwick available open access under the following conditions. Copyright © and all moral rights to the version of the paper presented here belong to the individual author(s) and/or other copyright owners. To the extent reasonable and practicable the material made available in WRAP has been checked for eligibility before being made available.

Copies of full items can be used for personal research or study, educational, or not-for profit purposes without prior permission or charge. Provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

Publisher's statement:

This is a pre-copyedited, author-produced PDF of an article accepted for publication in The Computer Journal following peer review. The version of Al Fayez, Reem Qadan and Joy, Mike. (2017) Using linked data for integrating educational medical web databases based on bioMedical ontologies. The Computer Journal, 60 -388 (3). pp. 369-388. is available online at: <https://doi.org/10.1093/comjnl/bxw096>

A note on versions:

The version presented here may differ from the published version or, version of record, if you wish to cite this item you are advised to consult the publisher's version. Please see the 'permanent WRAP URL' above for details on accessing the published version and note that access may require a subscription.

For more information, please contact the WRAP Team at: wrap@warwick.ac.uk

Using Linked Data for Integrating Educational Medical Web Databases based on BioMedical Ontologies

REEM Q. AL FAYEZ¹ AND MIKE, JOY²

¹*Department of Computer Information Systems, The University of Jordan, Amman 11942, Jordan*

²*Department of Computer Science, University of Warwick, Coventry CV4 7AL, UK
Email: r.alfayez@live.co.uk*

Open data are playing a vital role in different communities, including governments, businesses, and education. This revolution has had a high impact on the education field. Recently, Linked Data are being adopted for publishing and connecting data on the web by exposing and connecting data which were not previously linked. In the context of education, applying Linked Data to the growing amount of open data used for learning is potentially highly beneficial. This paper proposes a system that tackles the challenges of data acquisition and integration from distributed web data sources into one linked dataset. The application domain of this work is medical education, and the focus is on integrating educational content in the form of articles published in online educational libraries and Web 2.0 content that can be used for education. The process of integrating a collection of heterogeneous resources is to create links that connect the resources collected from distributed web data sources based on their semantics. The proposed system harvests metadata from distributed web sources and enriches it with concepts from biomedical ontologies, such as SNOMED CT, that enable its linking. The final result of building this system is a linked dataset of more than 10,000 resources collected from PubMed Library, *YouTube* channels, and Blogging platforms. The final linked dataset is evaluated by developing information retrieval methods that exploit the SNOMED CT hierarchical relations for accessing and querying the dataset. Ontology-based browsing method has been developed for exploring the dataset, and the browsing results have been clustered to evaluate its linkages. Furthermore, ontology-based query searching method has been developed and tested to enhance the discoverability of the data. The results were promising and had shown that using SNOMED CT for integrating distributed resources on the web is beneficial.

Keywords: Linked Data; Biomedical Ontologies; Medical Education; Web Databases; Data Model; Data Integration

Received 00 January 2009; revised 00 Month 2009

1. INTRODUCTION

Recent advances in open education and the growing reliance on the web for acquiring knowledge have had a large impact on education. Nowadays, the web offers students and educators more choices in how to learn and teach. For example, academic institutions have been using the *YouTube* platform for publishing educational videos, and some researchers have adopted blogging for sharing their knowledge with the public. Therefore, searching for educational content on the web is no longer restricted to finding books and articles, and has expanded to include searching for Web 2.0

technologies that support the learning process, such as videos, blogs, wikis, or pictures [1]. This trend applies to all fields of education whether its humanities, scientific, or medical education. The challenge is to be able to search and find free high quality educational materials, since hours spent browsing the web for information is time that could be spent learning. The increased use of public web resources in learning and teaching motivates this research and encourage finding a practical solution that can be applied to ease the search problem. The work presented in this paper is applied to the medical education field as a proof of concept

that can be extended to include different fields if it is proved to be valid. We address the search problem in the field of medical education since various technologies have emerged for enhancing the learning and teaching experience, and have been incorporated for developing medical e-curricula [2]. Besides, research in the field of medical education provided insights into the potential impact of Web 2.0 technologies on enhancing teaching and learning [3].

The process of searching any web data source is thus twofold. Firstly, the search process is made easier for users if the published content is described using representative metadata and thus provides what it needs for matching any search query. Secondly, the user's search query must represent the information the user seeks, and that affects the correctness of the search results. Different metadata models have been implemented by organizations such as IEEE to accommodate the requirements of publishing their content [4]. With various metadata models having been proposed, it has become obvious that there is no ideal standard that accommodates the needs of all publishing organizations. In medical education, the spread of educational libraries providing open content and the massive amount of information published using Web 2.0 technologies emphasize the importance of data integration.

The challenge in integrating distributed web sources is that the datasets are heterogeneous. Databases consist of different attributes for describing their data based on the specific metadata schema applied. Moreover, web content is being published using various techniques, from old traditional relational databases to NoSQL databases. Recently, the term "Web of Data" has emerged after the spread of Linked Data practices for publishing data on the web [5]. It is now widely used in different fields, and have changed how web publishing takes place. Linked Data practice is used for publishing data on the web and connecting related data that were not previously linked, and with the evolution of open data on the web, the adoption of Linked Data is increasingly turning the web into a global data space [6]. In education, as any other field, the use of Linked Data is becoming popular [7]. As a result, large amounts of educational content have been published on the web using Linked Data. A full review of Linked Data proposals in the learning domain is presented in [8] that analyses existing research work in the literature. Despite the fact that applying Linked Data in education can be challenging [9], yet it has been reported about the opportunities it provides for open and distance learning [10, 8].

In this paper, we present a novel system for harvesting and interlink different types of Educational Medical Objects (EMOs) collected from various web data sources into a linked dataset named the Linked Educational Medical Objects (LEMO) dataset. This research illustrates how to tackle the challenges of

data acquisition and integration into appropriate presentation and organization with web data in the context of medical education. In particular, this work focuses on bridging the gap between the content of online educational libraries and Web 2.0 that are both used in learning. Using Linked Data practices, the system exposes the heterogeneous metadata of distributed EMOs and represent them using an RDF/XML metadata schema named the LEMO schema [11], and enriches the dataset collected with content from external ontologies. Using biomedical ontologies, the EMOs are enriched by annotating free-text descriptions provided in their metadata records. Ontology-based annotation allows the system to discover keyword terms in the collected dataset and builds dynamic linkages between its components. As a result, the system successfully establishes linkages between the distributed EMOs building one coherent dataset. The content of the dataset used in this experiment is collected from medical educational libraries and Web 2.0 platforms. High quality educational materials found on *YouTube* and blogs are automatically linked with content of online medical libraries. The data sources involved in this experiment are managed by trustworthy medical educators or organizations. The final linked dataset is stored in an RDF triple store where it consists of the EMOs and the ontology-based enrichments added to their metadata.

The rest of the paper is organized as follows. Section 2 presents background information and related work. Section 3 describes the methodology for building the system and its detailed architecture, and details the RDF triple store used for organizing and enriching the dataset collected. Detailed explanations of the dataset components and the linkages subsequently established are detailed in section 4. Section 5 outlines the evaluation techniques applied to test the system and validate its results. In section 6, we present the experiments conducted and discuss the results of validating the system. Finally, section 7 presents the conclusions and future work.

2. RELATED WORK

The internet has been playing an increasing role in education, and with educational materials being available at no cost and easily accessed with only a computer and internet access, learning and teaching techniques are changing. Recent research has shown the high impact of using the web in education [12]. Educational resources include books, articles, videos, pictures, and any other material that supports the learning process, and such resources are being published freely on the web under the open education movement. New technologies such as Web 2.0 have been incorporated with traditional learning and have proved its efficiency in education [13]. Consequently, the evolution of open data has promising results on

enhancing the quality and availability of education [14]. In the medical education field, researchers have studied the potential use of Web 2.0 for active and collaborative learning [15, 16]. The vast developments in the field of education emphasize the need for better organisation of educational data available on the web.

Students and educators are facing problems of searching and browsing the open data available in order to satisfy their needs, and searching and connecting various educational resources to enhance the knowledge about a topic can be frustrating for them. The search process mainly depends on the user search terms initiated and the description provided with the educational resources available on the web. Organizing and controlling the publishing of data on the web has been well researched in the e-learning field. Metadata for educational materials are defined as “data describing the context, content, and structure of records and their management through time” [17], and with this wide definition of metadata, different organizations have developed their own standards for managing their educational resources [18]. In medical education, several metadata standards have been developed to organize libraries and repositories of medical educational materials such as HealthCare LOM [19], the Health Education Assets Library (HEAL) [20], and National Library of Medicine (NLM) [21]. Each metadata standard differs in the elements used for describing the metadata of the published materials.

Due to the large amount of information on the web and the lack of adopting one standard for publishing all types of educational materials, web-scale integration of educational content is now a widely researched topic. After the emergence of Linked Data as a practice for connecting unstructured data on the web [5], researchers started exploring the potential of applying Linked Data for education, and several educational organizations adopted Linked Data practices for exposing and publishing data. This has created a rich environment for more projects to be developed for supporting web educational data integration [22]. Other projects such as the one presented in [23] have exploited Linked Data for connecting the registries of educational ICT tools in order to help educators in searching for such tools. In medical education, the work presented in [24] focused on the issue of interoperability and reuse of open education materials and proposed a data model for exposing and publishing a dataset of educational medical materials that can be easily reused by students and educators. In other domains such as childhood education, Linked Data have been implemented to enrich the data resulted from using a platform for childhood education and care. Tools has been designed to enhance educational resources recommendation, nutritional monitoring, and health monitoring services based on Linked Data [25]. Other projects targeted the management of learning materials available on the web.

SemUnit project initiated by French higher education institutions exploited Linked Data to integrate French repositories that contain learning materials of high-quality for different domains [26]. Ontologies have been incorporated for the enriching the metadata of learning materials in these repositories such as FOAF for describing persons and organizations and SKOS for describing controlled vocabularies in metadata elements. In another application domain, a work presented in [27] discussed the advantages of exposing the content of the Organic.Edunet portal that is a federation of learning repositories in the domain of organic agriculture. Furthermore, Linked Data have been used in building educational materials from open data on the web, such as the work presented in [28] in which a full curriculum was built from open educational resources for training practitioners how to use Linked Data. Universities and educational institutes have now started to expose and publish their data using Linked Data format, and thus it is clear that the future of education lies in Linked Data. One successful experience of publishing open data is provided by the UK Open University, and it is considered a blueprint for other organisations to open up their data and enables its sharing and reuse [29].

The use of ontologies and connecting them with open linked data is also a promising research topic in the field of open data. Ontologies are used for semantically enriching unstructured data on the web, and then using Linked Data, the enriched content can be exposed and connected with external sources on the web. Enriching data on the web can be performed either at the client side of a system or at the server side, and the work presented in [30] investigated the advantages and disadvantages of both techniques. Another use of ontologies is presented in [31] where server side enrichment of queries was tested by enriching medical images stored in a library using the MeSH ontology and enriching the queries with the same ontology for enhanced search results. Applying semantic enrichment on User Generated Content (UGC) can also be beneficial, but usually the metadata describing the UGC content is not of a high quality. For example, users’ tags and folksonomies provided for describing *YouTube* content is not representative of the content. One research proposed a solution for enhancing the discoverability of videos that lack sufficient metadata and proposed a semantic video search engine named “yovisto” [32]. It presented an exploratory search engine that expands the search query with terms extracted from DBpedia. Furthermore, in another research both the metadata and the query can be extended and enriched with Linked Data [33] and that builds a platform for browsing video annotation. This work is applied on a repository of videos provided for the history course at the Open University. Using ontology enrichment, tag expansion has been applied in [34] for enhancing the metadata describing *YouTube*

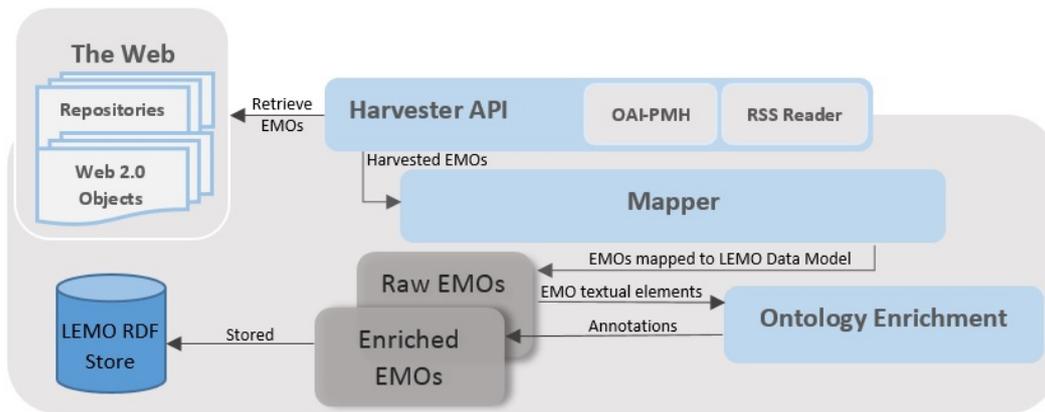


FIGURE 1: The LEMO System architecture

videos. In the field of medicine, having well established biomedical ontologies [35], such as SNOMED CT ontology and MeSH ontology, has enabled ontology enrichment of medical data in general. Hence, our proposed system presented in this research is applied on the field of medical education. It takes advantage of the well-established ontologies available for medicine and apply it for integrating educational medical materials from distributed sources.

3. METHODOLOGY

The amount of educational content on the web is increasing and the type of the content continues to vary. IEEE has launched a metadata IEEE LOM standard in 2002 which defined a Learning Object as “any entity, digital or non-digital, that may be used for learning, education or training” [36]. Recently, different types of educational content have been integrated with the curriculum taught in schools and universities. In this paper, we refer to the educational content harvested from different sources as Educational Medical Objects (EMOs). Linked Data practice is used for exposing and linking the EMOs in the system developed; hence the system is called the *Linked Educational Medical Objects (LEMO) system* and the resulting dataset is the *LEMO dataset*.

The system developed consists of various components for collecting data, mapping, and linking them into one coherent dataset. Figure 1 illustrates the architecture of the LEMO system.

3.1. Harvesting

The system is designed for collecting and linking diverse educational objects from distributed sources on the web. Therefore, we conducted a survey on medical students and educators to get feedback about their habits when searching for online content. Other than the normal search engines such as *Google* and *Bing*, students and educators have the tendency to follow the publications of specific journals of interest, key

researchers in the field, and specific academic institutes’ publications. Also, they search in some dedicated libraries which provide open content for their users.

Furthermore, we investigated the methods used for importing content from the web. Some journals, academic institutes, and well known researchers started to have their own Blogs or *YouTube* channels, and in some cases both, for sharing their content. The need for having such delivery channels arise from the increasing use of different types of educational objects in the learning process. These delivery channels, *YouTube* and Blogs, are bundled with RSS feeds which can be read easily using an RSS reader. In the LEMO system, we have developed an endpoint for reading RSS feeds, which is used for collecting EMOs published on journal blogs and *YouTube* channels. Examples of such sources are *Khan academy medicine*³ channel, the blog of *emergency medicine cases*⁴, and the blog of *The New England Journal of Medicine (NEJM)*⁵. The size of the data retrieved using this endpoint is detailed in later sections. Figure 2 illustrates an example of an article published in the New England Journal of Medicine (NEJM). The RSS feeds button highlighted in the orange rectangle at the top right corner of the figure is used to identify the RSS feeds URL that can be used as an input for the RSS feeds harvesting endpoint.

As for the traditional libraries where books and articles are stored, few of them provide open content to their users, and even fewer provide the ability for others to harvest and store their data. One of the popular repositories in medicine which provides open content is *PubMed library*. The library is set up with a service that provide access to its metadata. This service is an implementation of the Open Archives Initiative for Metadata Harvesting (OAI-PMH) which is a protocol for retrieving metadata from digital repositories [37]. Since the PubMed library is popularly used and provides access to its content, we used this

³<https://www.youtube.com/user/khanacademymedicine>

⁴<http://emergencymedicines.com/>

⁵<http://www.nejm.org/>



The screenshot shows the homepage of The New England Journal of Medicine. At the top left is the journal's logo, a circular seal with the text 'THE NEW ENGLAND JOURNAL OF MEDICINE' and the years '1812', '1823', and '1928'. To the right of the logo is the title 'The NEW ENGLAND JOURNAL of MEDICINE'. Below the title is a navigation bar with links: HOME, ARTICLES & MULTIMEDIA, ISSUES, SPECIALTIES & TOPICS, FOR AUTHORS, and CME. The main heading is 'Browse Figures & Multimedia' with an RSS icon to its right. Below this, it says 'Showing 1 to 20 of 62 Articles'. A 'FILTER RESULTS' sidebar on the left lists 'By Specialty' with options: 'All Specialties' (62), 'Medical Practice, Training, and Education' (62), 'Primary Care/Hospitalist/Clinical Practice' (42). The main content area shows 'SORT BY: Newest | Oldest | Most Viewed | Most Cited'. A featured article is titled 'VIDEOS IN CLINICAL MEDICINE' with the sub-title 'Examination of the Retina'. It includes a small image of a person's eye being examined and the text: 'August 20, 2015 | Liu Y.Wu F.Lu L.Lin D.Zhang K. | N Engl J Med 2015; 373:e9'.

FIGURE 2: An example of blog article published in NEJM that can be harvested using RSS feeds

library for collecting books and articles to complement our dataset. In the LEMO system, we developed a second endpoint based on OAI-PMH protocol for harvesting the library content.

After harvesting data using the developed endpoints, the LEMO dataset consists of EMOs of different types collected from various sources, hence we developed a mapper which maps the various heterogeneous formats of metadata files harvested into one proposed metadata schema that accommodates their differences.

3.2. Mapping

All the data harvested by the LEMO system, whether collected using the RSS reader endpoint or the OAI-PMH endpoint, are retrieved in XML format. These files describe the metadata attributes of the EMOs in the dataset, and a metadata schema that accommodates the different attributes of all the heterogeneous metadata files harvested is essential. Therefore, the LEMO metadata schema was proposed in [11] after conducting comparative studies between existing educational metadata schemas used in the field and analysing the elements required for describing the different EMOs harvested in the system. The detailed process of developing the LEMO metadata schema is beyond the scope of this paper. The work presented in [11] details experiments of applying the proposed LEMO metadata for describing different types of EMOs.

The EMOs harvested from the PubMed are described using a metadata schema that is based on the DCMI data model, while the RSS feed records are described using a simple metadata element set due to the summarized information provided in when describing *YouTube* videos and Blog articles. One limitation we faced in this work is the incompleteness of the metadata records for videos and blogs. Searching such metadata will not be successful and enriching its content will be less effective. The proposed LEMO metadata schema is based on Dublin Core Metadata Initiative (DCMI) standard [38] which consists of a flat structure of 15 elements. This element set covers all the attributes used for describing the metadata of the EMOs harvested. The flat structure of DCMI enables further refinements to be added in the LEMO metadata schema without compromising the metadata interoperability. The new refinements added in the LEMO schema focus on adding attributes which enrich the textual elements of the EMOs using external ontologies. The LEMO metadata schema is developed in RDF/XML format which enables automatic linking of EMOs using URIs and RDF. The new elements proposed in the LEMO metadata schema are defined using the prefix “lemo” while the original DCMI elements are defined using the prefix “dc”. The new elements proposed in the LEMO metadata are used for storing the results of the enrichment process.

3.3. Enrichment and linking

After harvesting and mapping the data from the web, the EMOs' textual elements are enriched by annotating it with biomedical ontologies. In this process, the input is the *titles* and the *descriptions* of all EMOs harvested. We specify the ontology to be used for enriching these textual elements with semantic annotations and the output of this process is an annotated dataset of EMOs with keyword terms discovered in its textual elements.

The annotation process is based on existing ontologies used in the medical field. Ontologies have been used in libraries for indexing entries and ease the search process, such as the use of MeSH ontology for indexing PubMed library entries [39]. In LEMO system, we adopted different ontologies and tested them for preliminary results. We experimented with different ontologies in order to test the feasibility of the annotation process in connecting EMOs. The work presented in [40] details the experiments conducted on a small dataset of around 2000 EMOs. The experiments compared the results of using two popular ontologies — MeSH and SNOMED CT for enriching harvested content.

The enrichment process consists of two main steps: first, the keyword terms are annotated in the title and the description of EMOs using SNOMED CT ontology in order to enrich the metadata of EMOs. The second step is responsible of weighting and ordering the keywords annotated in each EMO according to its importance for that EMO in order to select a more representative sample of keywords as the subject keywords categorizing that EMO. These subject keywords will be the entries of the subject property in the LEMO metadata schema for each EMO. Algorithm 1 and algorithm 2 explain the steps of terms discovery and terms filtering respectively.

The algorithms are applied to the LEMO dataset of size $|LEMO| = n$ where $LEMO = \{emo_1, emo_2, emo_3, \dots, emo_n\}$. After the enrichment process, each emo_i can be represented by a set of the keywords discovered in its textual content having $emo_i = K_i$ and $K_i = KT_i \cup KD_i$ where KT_i and KD_i denote the keywords discovered in the title and the description of emo_i respectively. Annotating the LEMO dataset results in having a large number of keyword terms K representing the whole LEMO dataset, as explained in algorithm 1.

The annotation process is applied using the BioPortal annotator endpoint⁶ which is provided by the BioPortal repository as a web service [41]. Using this endpoint, we send the raw text as an input along with the name of the ontology used in the enrichment process. The endpoint examines the text together with the ontology classes and returns the relevant annotations. Ontologies are formal representations of knowledge with definition of concepts and their relations [42], hence the LEMO

⁶<http://bioportal.bioontology.org/annotator>

Algorithm 1 Terms Discovery of EMOs

Input : Textual content of EMO elements *titles*, *descriptions*

Output : The set of keywords for all EMO elements K

```

for  $emo_i$  in  $LEMO$  do
  title  $\leftarrow$  getTitle( $emo_i$ )
  desc  $\leftarrow$  getDescription( $emo_i$ )
   $KT_i \leftarrow$  annotateText(title)
   $KD_i \leftarrow$  annotateText(desc)
   $K_i = KT_i \cup KD_i$ 
  add  $K_i$  to the set  $K$ 
end for

```

dataset linkages are based on the ontology classes' relations. The ontology used in enriching the LEMO dataset in this paper is the SNOMED CT ontology [43], which can be represented in a graph structure G_{snomed} . The keyword terms K discovered in the LEMO dataset is a subset of the G_{snomed} vertices. As a result, the relations between LEMO dataset annotations can be represented in the graph G_{lemo} based on the original ontology graph. The graph structure representation of the LEMO dataset keywords K is stored in the LEMO system and used as the ground reference for further processing of the LEMO dataset.

The next step of the enrichment is to filter the terms discovered in each EMO. Using the algorithm proposed in 2, we filter the terms into a smaller set used as the entries for the subject property of the EMOs. First, for each emo_i , each keyword term in the set K_i is weighted based on its number of occurrences in the title or the description of the EMOs. Higher weights are given to terms discovered in the title since they give a good indicator about the subject of the EMOs [44]. Next, weights are updated based on their location in the G_{lemo} graph. Since each emo_i is represented by a set of K_i , then based on the G_{lemo} structure, we can represent each EMO using a graph structure G_{emo_i} . The weights of the keyword terms of each EMO are updated to the accumulated weights of its descendent terms. Updating the weights of the terms to consider their hierarchical positions in the ontology is beneficial for categorizing the EMOs into subject keywords. Doing so, terms which are leaf nodes in the ontology will have lower weights compared with terms in higher levels of the ontology. The final step for filtering the terms is to normalize the weights of each EMO keyword terms K_i and specify a threshold for each EMO's set of keywords. Based on the threshold, the keywords are split and terms with the highest weights are named KS_i and stored in the *dc:subject* property for each emo_i .

The goal of annotating EMOs using ontologies is to build relations between these EMOs. Based on the keywords discovered in each EMO, linkages can be built between EMOs having the same keyword terms

Algorithm 2 Terms Filtering of EMOs

Input : The sets of K_i for each emo_i
Output : The sets KS_i , keywords chosen as subjects for each emo_i

for emo_i in *LEMO* **do**
 for $k_{i_j} \in K_i$ **do**
 weight \leftarrow calculateOccurrence(k_{i_j})
 weight \leftarrow weightBasedonHierarchy(k_{i_j})
 Assign weight $\rightarrow k_{i_j}$
 end for

normalizeWeights(K_i)
 $KS_i \leftarrow$ sortAndSplit(K_i)
end for

annotated in their textual content, but this is not efficient for large datasets. In order to accomplish more accurate linkages in the LEMO dataset, a link is considered to be valid between two EMOs emo_i and emo_j if they have at least one term in common in their subject keyword terms stored in their subject property, i.e. $|KS_i \cap KS_j| \geq 1$. Links between EMOs based on the keywords discovered in its titles or its descriptions are still established but not considered as valid links in this work.

At this stage, the LEMO dataset is enriched and links are established between its EMOs based on similar annotations. Storing and organizing the LEMO dataset is based on the proposed LEMO metadata schema extending DCMI metadata schema with new properties introduced describe the enrichments of the dataset. Since the LEMO schema is implemented in RDF/XML format, the full dataset of LEMO and its annotations are organized in triple store called the LEMO RDF triple store explained in the next section.

3.4. Storage and organization

The LEMO dataset layered design, illustrated in figure 3, explains the organization of the LEMO dataset components. The LEMO dataset is stored in RDF/XML format using Linked Data practice, and is beneficial for interlinking the EMOs with external ontologies. The layered design of the LEMO triple store in figure 3 demonstrates the components of each layer and its relation with other layers. This layered design eases the understanding of the LEMO metadata properties and their relations. The figure also provides snippets of XML describing the LEMO metadata properties used for representing each component of the LEMO dataset, where the top layer represents the largest components of the dataset which are the EMOs themselves, while lower layers represent smaller components which are added to enrich the content of the dataset.

The **first layer** consists of the metadata of the

EMO resources, described using the original DCMI elements. The values of the *title* and *description* properties are not stored as textual values — instead, they point to new resources representing the text and its annotations in the second layer. The collection of *title* and *description* resources of all EMOs in the dataset are stored in the **second layer** using new URIs that link them to the original EMOs. It also links the titles and the descriptions with their annotations using the new proposed LEMO metadata properties such as *lemo:lemoTitleAnnotation* property. The **third layer** stores the annotations discovered in the LEMO dataset. The collection of resources are described using LEMO properties which are illustrated in the XML snippet. The properties are used to store the indices of the annotated text (*lemo:lemoFrom* and *lemo:lemoTo*) and the class it maps to in the ontology (*lemo:lemoClassID*, *lemo:lemoClassTerm*, \dots , etc.). Finally, the set of SNOMED CT classes used to annotate the keyword terms form the **fourth layer**. In this layer, the ontology class relations are retrieved based on their original structure in the SNOMED CT ontolog and stored using the new LEMO proposed metadata property *lemo:adjacentTo*. The results of testing the LEMO system are stored in the LEMO RDF triple store, and details about the dataset, its annotations, and linkages are presented in the next section.

4. LEMO RDF TRIPLE STORE

The final LEMO dataset consists of EMOs of different types and sizes, described using the LEMO data Model. The enrichment process annotates the content of the dataset using the SNOMED CT ontology. The keyword terms discovered in the *title* and the *description* of the EMOs are then filtered into a smaller number of terms to categorize the EMOs into subjects. In this section, the statistics of the resulted LEMO dataset are detailed. We also compare the results of the linkages established in the dataset based on different properties of the LEMO metadata schema: the title, description, and subject properties.

4.1. Dataset

The components of the LEMO dataset are detailed in table 1. The table details the number of resources harvested grouped by type, and the number of keywords terms resulted from enriching the EMOs using the SNOMED CT ontology. The majority of the dataset content is harvested from the PubMed library using OAI-PMH harvesting endpoint. Also, videos and blogs were harvested from a list of *YouTube* channels and blogging platforms managed by well-known medical institutes using RSS feeds reader endpoint.

Applying the enrichment process to the harvested dataset using the SNOMED CT ontology resulted in large amounts of annotations. The metadata provided

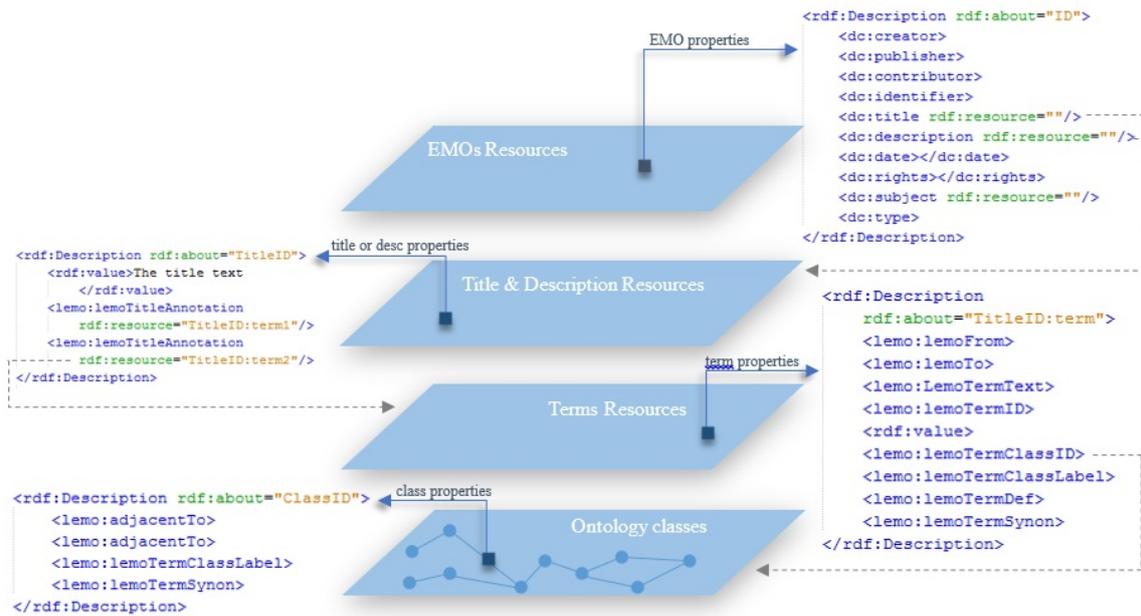


FIGURE 3: The LEMO triple store layered design

TABLE 1: The LEMO dataset components

Type of EMOs	Number of EMOs	SNOMED Keyword Terms		
		Title	Description	Total
Article	8742	56708	307431	364139
Video	1259	3297	5348	8645
Blog	461	1494	9766	11260
Total	10462	61499	322545	384044

for videos and blogs is not well documented since they are user generated content. For example, the description field of the video might be missing, and the title field is not descriptive enough for the content of the video. Thus, the annotations resulted from enriching videos and blogs are less than the articles annotations, as shown in table 1. In the LEMO dataset, EMOs of type video are the least enriched EMOs in the dataset compared to their size.

The keyword terms discovered are weighted and filtered into a smaller set of terms representing the EMOs subject properties. Keyword terms discovered in the titles are generally repeated in the description text of any published content on the web. Also, titles tend to be less detailed than the description in order to indicate the general topic of the content published. Table 2 shows that the results of applying term filtering to the set of keyword terms discovered in the LEMO dataset, compliant with general practice.

The majority of the keywords, selected after filtering and weighting as subjects categorizing EMOs, are keywords annotated in the title. The terms filtering process was able to filter the large number of terms annotated and reduce the subjects selected to only 27%

TABLE 2: Terms filtering results

	Number of terms	Number of subjects	Percentage
Title	61499	47586	77.3%
Description	322545	59010	18.2%
Total	384044	106596	27.7%

of the full terms set, 77% of them being title annotated keyword terms.

4.2. The ontology

The ontology incorporated in the LEMO system for enriching the EMOs in the LEMO dataset is the SNOMED CT ontology. It is an acronym for Systematized Nomenclature of Medicine - Clinical Terms (SNOMED CT). The ontology provides a comprehensive healthcare terminology that contains interrelated concepts, supported by synonyms and definitions [45]. The SNOMED CT ontology classes are used to represent the Term resources annotating the EMOs are represented as Class resources as detailed in figure 3. The collection of Class resources and their relations are denoted by a graph G_{lemo} that represent the bottom layer of the LEMO triple store in figure 3. Hence, the G_{lemo} graph is considered a subset of the SNOMED CT ontology that can be denoted as a graph named G_{snomed} . Details about the number of the concepts or classes in each graph and the depth of that graph are detailed in table 3. The table presents a comparison between the SNOMED CT ontology and the subset of classes stored in the LEMO triple store.

The number of the ontology concepts that are stored

TABLE 3: SNOMED CT vs. LEMO store classes

Metrics	G_{snomed}	G_{lemo}
Number of Classes	316031	29283
Maximum Depth	28	25

as Class resources in the RDF store is a small subset of the SNOMED CT classes as detailed in the table. The number of EMOs described in the LEMO RDF store is more than 10,000 EMOs which are annotated with more than 29,000 concepts from the SNOMED CT ontology. Harvesting more data from the web and enriching it might increase the number of ontology concepts used and increase the size of the subgraph G_{emo} representing the relations between the ontology concepts. The maximum depth represents the length of the deepest branch in the SNOMED CT taxonomy. It is organized based on the is-a hierarchical relations between its concepts. The maximum depth of the G_{emo} graph is 25 levels. That means that some concepts annotating the EMOs are in the lower levels of the ontology hierarchy.

4.3. Link results

After the subject selection process, links are generated between EMOs based on their subject property. A link exists between two EMOs (e.g. emo_i and emo_j), if they have at least one similar annotated class in their subject property list of keywords (i.e. $|KS_i \cap KS_j| \geq 1$). The links in the LEMO dataset are considered direct links, therefore, if there is a link from node a to node b the link will be counted twice instead of once. Links can be generated based on the keyword terms discovered in the titles and descriptions of EMOs. However, since the number of annotations made is very large, the smaller set of keyword terms, representing the subjects of EMOs, are considered the basis of the linkage process in the LEMO system. Figure 4 illustrates the percentage of the links made based on the keyword terms found in different properties of EMOs: title, description, and subject.

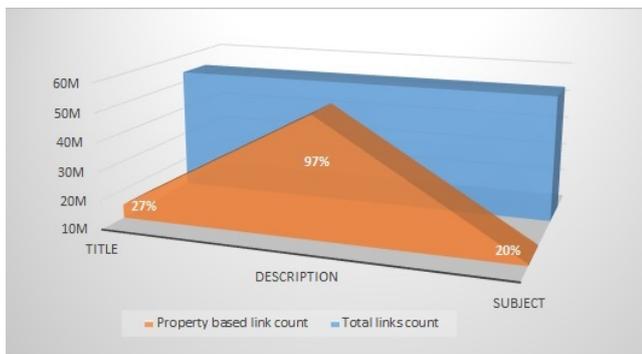


FIGURE 4: Percentages of links made based on the properties of the EMOs

The total number of links made based on all the keywords annotated is large (more than 50 million). Keyword terms annotated in the title or description properties of an EMO can link it to any other EMO annotated with the same keyword. Hence, most of the links are generated based on the description annotated terms since there are more of them than title terms. Generally, most of the keywords mentioned in the title will be repeated in the description and annotated in both. Such keywords will have higher weights and will be selected as subject keywords for the EMO. The links considered in this evaluation are the subject based links only, and links based on the subject property form 20% of the overall number of links (around 10 million). This indicates that the average number of links for each EMO in the dataset is around 100. Thus, the number of links for each EMO varies based on its annotation. The quality of the links based on the subject property is stronger than those based on other fields due to the terms filtering process.

5. EVALUATION

We evaluated the performance of the LEMO system by validating the established linkages in the LEMO dataset. The main goal of the LEMO system was to build a coherent linked dataset from distributed sources on the web. Hence, in this section, we explain the techniques used in validating the linkages built within the LEMO dataset.

Since the LEMO dataset is large, it is hard to validate its content using experts' judgements. The dataset consists of more than 10,000 EMOs with more than 10 million links connecting these EMOs, as detailed in section 4. Therefore, we simulated two techniques of information retrieval for accessing the LEMO dataset: browsing and querying searching. First, we evaluated the general behaviour of browsing the LEMO dataset based on the SNOMED CT ontology classes. We also conducted clustering experiments in order to emphasize the strongly linked and weakly linked communities within the LEMO dataset discovered while browsing. Secondly, we proposed an ontology based query technique and performed random queries to test it. We compared the results of this technique with text based query results. Next, we describe the details of each evaluation technique.

5.1. Preliminary ontology based browsing

Browsing is a basic kind of information seeking behaviour used to satisfy users' information needs [46]. Browsing the LEMO data set is based on the SNOMED CT ontology, which is a hierarchical structure of classes [47] rooted under the node **SNOMED CT concept** and can be easily browsed in the BioPortal repository⁷. The ontology consists of 19 classes at

⁷<http://bioportal.bioontology.org/ontologies/SNOMEDCT>

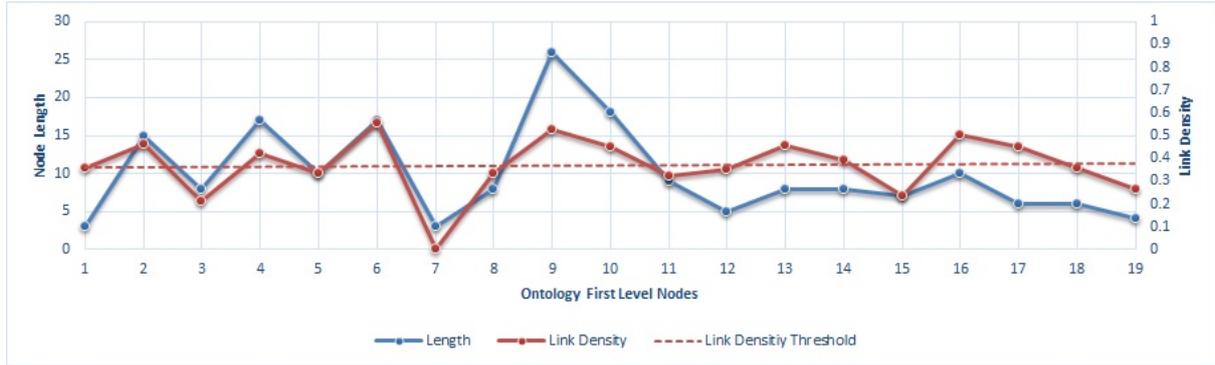


FIGURE 5: The link density resulted from browsing first level ontology nodes compared to their length

the first level descending from the root node. The SNOMED CT ontology demonstrates the hierarchical relations between the ontology classes from generalized classes to more specialized ones.

The first step to evaluate the LEMO system is to compare the results of browsing different levels of the SNOMED CT ontology. Since the SNOMED CT ontology is not physically stored the LEMO system, the subset of its classes which are annotated in the LEMO dataset forms the LEMO graph G_{lemo} . This subgraph is stored at the lowest level of the LEMO RDF triple store as explained in section 3.4. The graph structure of G_{lemo} is transformed into a tree structure to ease its navigation for browsing.

Browsing any information system starts at a general level, then the user navigates into deeper levels of specialized concepts. Browsing LEMO dataset follows the same procedure, where top level classes of the ontology are more generalized than the lower levels. We simulated the process of browsing the dataset content. Selecting an ontology class will retrieve all EMOs in the LEMO dataset which are annotated in this class or any of its descendent classes as stored in G_{lemo} graph. The annotations considered in these experiments are those in the subject property only and we did not consider the title and description annotations in this evaluation.

In order to get an overview of LEMO dataset density and distribution over the classes of the SNOMED CT ontology, we experimented with browsing the various levels of the G_{lemo} graph. Selecting a class or a node in the graph will retrieve a subset of linked EMOs from the LEMO dataset. Assume that the LEMO dataset is denoted by graph $G = (V, E)$, where V is the EMOs composing the full dataset, and E is the edges representing the linkages between these EMOs based on its subject annotations. Selecting a class while browsing retrieves a subset of the LEMO dataset denoted by V_i . This set of EMOs is a subset of V and E_i denotes its edges, thus it can be represented by a graph $G_i = (V_i, E_i)$. The density of the retrieved dataset $D(G_i)$ is calculated as shown in equation 1.

$$D(G_i) = \frac{|E_i|}{|V_i| \cdot (|V_i| - 1)} \quad (1)$$

The density $D(G_i)$ measures the links density between the retrieved dataset of EMOs V_i where $D(G_i) \in [0, 1]$. The larger the density is, the more related the EMOs retrieved are [48].

In order to get an overview of the LEMO dataset distribution over the SNOMED CT ontology classes, we calculated the density of the EMOs retrieved after browsing the first level nodes (19 classes) of the ontology. Figure 5 illustrates the density of the results retrieved when browsing the first level nodes compared to the length of its descendent tree of classes. As can be seen in the figure, there is a correlation between the node length and its link density. The density results have the tendency to increase when the node length increase. The relation between the two variables is measured using the Pearson correlation coefficient statistical measure [49]. The value of this coefficient for the values presented in figure 5 is 0.633. This value indicates a moderate positive correlation between the length and the density values.

The nodes with longer trees descendent from them contain larger numbers of classes than other shorter trees. This indicates that the browsing results are affected by the number of classes related to the node chosen. In the next step we conduct further evaluations on highly dense sets of EMOs. Hence, we took the average density of all first level nodes as a threshold, and then narrowed the experiments to include only the nodes which are above that threshold. Only 9 out of the 19 first level nodes of the LEMO graph are above the threshold line. The focus now moves to validating the browsing of these 9 branches only. The dense content retrieved will be more efficient for showing significant results for identifying communities.

5.2. Identifying communities based linkages

Browsing the LEMO dataset retrieves sets of EMOs annotated with one or more classes of the ontology nodes selected. To test the validity of the annotations

of the EMOs and the resulting linkages, clustering the retrieved datasets must result in groups of related EMOs. Normally, high level browsing will result in large datasets with fewer connections, while deep level browsing will result in smaller and more related datasets. Furthermore, clusters of EMOs retrieved from the same branch of the ontology hierarchy must be highly connected compared to clusters retrieved from different branches of the ontology.

In order to perform validation via clustering, the EMO linkages can be represented in graph $G = (V, E)$ where V are the EMOs and E are the edges which are the links between these EMOs. The graph is transformed into a similarity matrix of EMOs and their relations. Since the LEMO dataset is of size $|LEMO| = n$ where $LEMO = \{emo_1, emo_2, emo_3, \dots, emo_n\}$, and the linkages between the EMOs are based on the subject terms similarity, the LEMO dataset linkages can be represented in a similarity matrix S of size $n \times n$ given as

$$S = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \dots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nn} \end{pmatrix} \quad (2)$$

where x_{ij} represents the similarity between the subject property terms of the two EMOs emo_i and emo_j . Hence, the value of $x_{ij} = |KS_i \cap KS_j|$.

In this paper, we applied the agglomerative hierarchical clustering method in order to analyse the linkages established. Since the similarity matrix is based on the ontology classes used for annotations, hierarchical clustering methods define the hierarchies of nested clusters [50]. In this method clusters are created by merging the most similar points in the dataset into clusters based on a distance matrix, and stops when all the clusters have merged into one big cluster. The results of merging the clusters are represented in a dendrogram, which represents the detailed merging process composing a hierarchical tree of clusters. At some level in the tree, some meaningful clusters might be found. Several measures can be used to decide on the best number of clusters. The silhouette coefficient is used in this experiment since it assesses both the separation and cohesion of clusters [51]. We calculate the average silhouette coefficient for clusters resulting from agglomerative hierarchical clustering, and for each experiment we changed the parameter determining the number of clusters and compared the results of the silhouette coefficient. The detailed experiments are explained in the next section.

The clustering is applied on the datasets retrieved while browsing the nodes of the G_{lemo} graph. For each node, we simulated the browsing process for different levels from 1 to 5. We conducted a comparison between clusters resulted from different branches of the ontology. The clusters were compared and evaluated using the

following evaluation measures.

5.2.1. Evaluation measures for cluster quality

In order to measure and compare the efficiency of the clustering experiments, internal measurements of the cluster quality were calculated [52]. As noted above, since the LEMO dataset is large, it is not practical to validate its linkages by experts in the field of medical education. The LEMO dataset is represented by the graph $G = (V, E)$, and based on the similarity matrix S which represents the LEMO dataset linkages, all the pairwise distances among the EMOs in the LEMO dataset are calculated. The internal measures are all based on the $n \times n$ distance matrix \mathbf{W} given as

$$\mathbf{W} = \left\{ \delta(x_i, x_j) \right\}_{i,j=1}^n \quad (3)$$

where

$$\delta(x_i, x_j) = \|x_i - x_j\| \quad (4)$$

is the Euclidean distance between x_i, x_j . That is, the distance weight $w_{ij} = \mathbf{W}(i, j)$ for all $x_i, x_j \in V$.

Given a clustering $C = \{C_1, C_2, \dots, C_k\}$ where k is the number of clusters, and given any subsets $S, R \subset V$, define $W(S, R)$ as the sum of the weights w on all edges with one vertex in S and the other in R , given as

$$W(S, R) = \sum_{x_i \in S} \sum_{x_j \in R} w_{ij} \quad (5)$$

Also, given $S \subseteq V$, we denote by \bar{S} the complementary set of vertices, that is, $\bar{S} = V - S$.

The internal measures are based on various functions over the intracluster and intercluster weights. In particular, the sum of all the intracluster and intercluster weights over all clusters are given by W_{in} and W_{out} respectively. Also, the number of distinct intracluster edges are denoted by N_{in} , and intercluster edges by N_{out} . Then the total number of distinct pairs of points N is $N = N_{in} + N_{out}$.

The following is an explanation for the internal measures used for validating the clustering experiments, followed by the results and a visualization of the results.

1. BetaCV Measure:

The BetaCV measures the quality of the clusters generated based on the ratio between the intracluster and intercluster distances. Equation 6 shows how the BetaCV is measured.

$$BetaCV = \frac{W_{in}/N_{in}}{W_{out}/N_{out}} = \frac{N_{out} \sum_{i=1}^k W(C_i, C_i)}{N_{in} \sum_{i=1}^k W(C_i, \bar{C}_i)} \quad (6)$$

It evaluates the mean intracluster distances to the mean intercluster distances. The smaller the BetaCV ratio, the better the clustering. It indicates that, on average, the distances between points in the same cluster are smaller than distances between points in different clusters.

2. Normalized Cut (NC) Measure:

The normalized cut measure can be used in the clustering process to determine the best cut for cluster partitioning. It can also be used as an internal measure of cluster quality. As for all the measures, we apply equation 7 on the distance matrix \mathbf{W} . The value of NC is maximized when the intracluster distances are much smaller compared to the intercluster distances.

$$NC = \sum_{i=1}^k \frac{W(C_i, \overline{C_i})}{W(C_i, V)} \quad (7)$$

where the volume of cluster C_i , denoted as $W(C_i, V) = W(C_i, C_i) + W(C_i, \overline{C_i})$, so that

$$NC = \sum_{i=1}^k \frac{W(C_i, \overline{C_i})}{W(C_i, V)} = \sum_{i=1}^k \frac{1}{\frac{W(C_i, C_i)}{W(C_i, \overline{C_i})} + 1} \quad (8)$$

The higher the normalized cut value the better. The NC value is maximized when the ratio between the intracluster distances and the volume of the cluster are as small as possible across all the k clusters.

3. Modularity:

The modularity objective for graph clustering is used as the third internal measure calculated using equation 9. The modularity measures the difference between the actual and expected distances within the clusters.

$$Q = \sum_{i=1}^k \left(\frac{W(C_i, C_i)}{W(V, V)} - \left(\frac{W(C_i, V)}{W(V, V)} \right)^2 \right) \quad (9)$$

As the equations are based on the distances matrix, the smaller the modularity measure the better the clustering. It indicates that the intracluster distances are low compared to the expected distances to the other clusters.

4. Davies-Bouldin (DB) Index:

This measure is based on the cluster mean and variance values, and measures the quality of cluster separation. Let μ_i denote the cluster C_i mean, given by

$$\mu_i = \frac{1}{n_i} \sum_{x_j \in C_i} x_j \quad (10)$$

Further, let the variance σ_{μ_i} denote the spread of the points around the cluster mean defined in equation 11.

$$\sigma_{\mu_i} = \sqrt{\text{var}(C_i)} \quad (11)$$

The Davies-Bouldin measure for pair of clusters C_i and C_j is defined in equation 12.

$$DB_{ij} = \frac{\sigma_{\mu_i} + \sigma_{\mu_j}}{\delta(\mu_i, \mu_j)} \quad (12)$$

DB_{ij} indicates how compact the clusters are compared to the distance between their means. Based on the DB_{ij} values for all pairs of clusters, the Davies-Bouldin Index is defined in equation 13

$$DB = \frac{1}{K} \sum_{i=1}^k \max_{j \neq i} \{DB_{ij}\} \quad (13)$$

The smaller the DB value the better the clustering. The index is calculated based on the largest DB_{ij} ratio for each cluster C_i . Hence, it will give a good indication about how well the clusters are separated from each other.

5.3. Ontology based querying

Another method for validating the LEMO system is query searching. We developed a prototype for an interface for querying the LEMO triple store based on the SNOMED CT ontology. Since the LEMO triple store is not published yet for users access, the query interface is developed on the local server to experiment with the algorithm proposed. The search field on the interface binds the user entries to the SNOMED CT ontology classes and allows the user to choose a class from the ontology to search instead of typing free-text values as illustrated in figure 6. This auto-complete text box is bound using a SPARQL query that read the ontology classes from the LEMO triple store. The results of performing the search process illustrated in the text box is shown in figure 7. The results retrieved are combination of EMOs of different types retrieved from the LEMO triple store via the prototype web interface developed. A full LEMO website should be developed for browsing and querying the LEMO dataset in the future.

An ontology-based query searching algorithm is proposed in algorithm 3. The method exploits the Term resources that describe the EMOs annotations and the ontology concepts used to annotate these Terms in order to expand the search query and retrieve wider range of results.

Starting with a query class Q , we build a query vector $QVector$ based on the class adjacency properties stored in the G_{lemo} graph, following which EMOs annotated with any of the classes in the query vector are retrieved. The classes in the vector query are then weighted based on the number of co-occurrences with the main class Q entered by the user. Normalizing the weights of the query vector classes is performed based on the size of the search results retrieved. The class Q weight is equal 1, and other class weights, depend on its co-occurrence with Q . The result of running this query are ranked according to its distance from $QVector$. Each EMO is represented using a vector of the same length as $QVector$ and weighted according to the annotation weights stored in the **rdf:value** property. The search results are ranked based on their Euclidean distance from the query vector $QVector$.

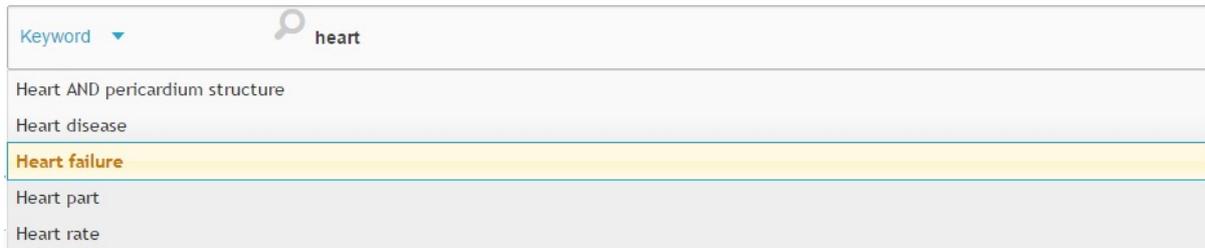


FIGURE 6: The ontology-based LEMO search user interface

Search Results for: "Heart failure" is 45

Episode 4: Acute Congestive Heart Failure

Dr. Eric Letovsky and Dr. Brian Steinhart describe a practical way to approach patients with undifferentiated SOB and acute congestive heart failure, the utility of various symptoms and signs in the diagnosis of CHF, as well as the controversies surrounding the best use of BNP and Troponin in the ED. A discussion of the use of ultrasound for patients with SOB as well as the indications for formal Echo are reviewed. In the second part of the episode they discuss the management of acute congestive heart failure based on a practical EM model, as well as the difficulties surrounding disposition of patients with CHF.

The post [Episode 4: Acute Congestive Heart Failure](#) appeared first on [Emergency Medicine Cases](#).

Open

blog

Medical School - Heart Failure with Preserved Ejection Fraction (Diastolic Heart Failure)

Brief discussion of heart failure with preserved ejection fraction, otherwise known as diastolic heart failure. Heart failure has become one of the most comm...

Open

video

Recompensation of Heart and Kidney Function after Treatment with Peritoneal Dialysis in a Case of Congestive Heart Failure

We report the case of a 57-year-old woman suffering from congestive heart failure. Due to refractory congestions despite optimised medical treatment, the patient was listed for heart transplantation and peritoneal dialysis was initiated. Peritoneal dialysis led to a significant weight loss, reduction of hyperhydration and extracellular water obtained by bioimpedance measurement, and a significant improvement in clinical and echocardiographic examination. Furthermore, residual kidney function increased during the long-term followup, and subsequently peritoneal dialysis was ceased. Pulmonary artery pressure and left ventricular ejection fraction remained stable and the patient did well. This case demonstrates the possibility of treating hyperhydration due to congestive heart failure with peritoneal dialysis resulting in recompensation of both heart and kidney functions.

Open

Article

Adjuvant Use of Ivabradine in Acute Heart Failure due to Myocarditis

We report two cases of young men in whom acute heart failure due to myocarditis was diagnosed. The patients had been transferred to the intensive care unit (ICU) with commencing symptoms of acute heart failure and consecutive multiorgan failure for further treatment and to evaluate the indication for implantation of a ventricular assist device or for high urgent orthotopic heart transplantation. In both patients, the If-channel inhibitor ivabradine was administered off-label to provide selective heart rate reduction, and thus support hemodynamic stabilization. Though currently considered off-label use in patients suffering from severe hypotension and acute heart failure, the use of ivabradine may beneficially influence outcome by allowing optimization of the patient's heart rate concomitant to initial measures of clinical stabilization.

Open

Article

FIGURE 7: Query results for "Heart failure"

Algorithm 3 Ontology-based Query

Input : Ontology class to be queried Q , LEMO dataset $LEMO$

Output : Ranked Search Result set of EMOs R

$RelClasses \leftarrow getRelatedClasses(Q)$ \triangleright Stores adjacent classes to Q

for $c \in RelClasses$ **do**

$qResults \leftarrow getEMOsAnnotatedWith(c)$

add $qResults$ to $ResSet$ $\triangleright ResSet$ is the final search results

end for

$QVector \leftarrow weightQVector(RelClasses)$ \triangleright Weight related classes to Q

for $d \in ResSet$ **do**

$dVector \leftarrow weightDVector(d)$ \triangleright Weight d annotations based on $QVector$

end for

for $d \in ResSet$ **do**

calculatedEuclideanDist($dVector, QVector$)

end for

$R \leftarrow Sort(ResSet)$ \triangleright Sort results ascendingly

In this experiment, we compared the results of query searching using the ontological based approach against the text based approach. The set of queries tested are restricted to one ontology class only, and is compared against one word text based queries, giving more direct and easy to compare results. In order to evaluate the proposed ontology based query algorithm, we use the overlap coefficient and the Jaccard similarity coefficient for comparing the similarity of the search results [53].

6. RESULTS AND DISCUSSION

We discuss the results of the evaluation techniques explained in the previous section, and detail the results of the experiments conducted.

6.1. Browsing the LEMO dataset

As explained in section 5.1, we simulated the behaviour of browsing the LEMO dataset based on its annotations stored in G_{lemo} graph. In this experiment, we focused on browsing 9 dense nodes and its descendent branches as shown in figure 5. In any well-established information retrieval system, the deeper we browse into the dataset the more related results we should find. We tested the behaviour of browsing the LEMO dataset at different levels deep. Selecting classes of the ontology at different levels results in datasets retrieved which are of different sizes and density. We calculated the average size of the datasets G_i retrieved and their density $D(G_i)$ when selecting all the nodes at the same level. Figure 8 illustrates the results of browsing different levels of

the ontology classes. The results show that browsing the first level classes of the ontology results in large datasets retrieved which have low density. A deep level of browsing results in significantly smaller datasets retrieved and with higher numbers of links between their components.

We explained in previous sections the LEMO dataset distribution over the ontology classes. We illustrated that browsing specific nodes results in retrieving more dense datasets. The LEMO dataset was harvested from distributed sources with no specific topics defined when harvesting. Hence, small part of the SNOMED CT ontology concepts were used to annotate the EMOs in the dataset and that resulted in a distribution of EMOs over the branches of the ontology. The results of this experiment are based on browsing the most dense branch descending from the 9 nodes selected as the most dense nodes in the ontology represented in G_{lemo} graph. Browsing the dense branches can allow us to extend the browsing into deeper levels compared to less dense branches. We calculated the link density of the datasets retrieved when browsing these branches. Doing so, we can illustrate the changes of browsing deeper into the LEMO dataset. Figure 9 illustrates the density of the datasets retrieved when browsing 9 out of the 19 nodes at different levels.

The graph illustrates the link density values between the results retrieved when browsing each node at different levels of depth from 1 to 7. From this experiment, we notice that, for most of the nodes, the link density values start to decrease after level 5. This can be due to smaller number of EMOs retrieved at those levels which are not highly connected to each other. Other branch results such as “clinical findings” and “body structure” nodes have link density values higher than 0.5 at level 7. This indicates that the EMOs retrieved at that level for such nodes are more connected and related to each other. It also indicates that large numbers of EMOs are annotated with the classes of these nodes since browsing deep levels is still possible with high density values.

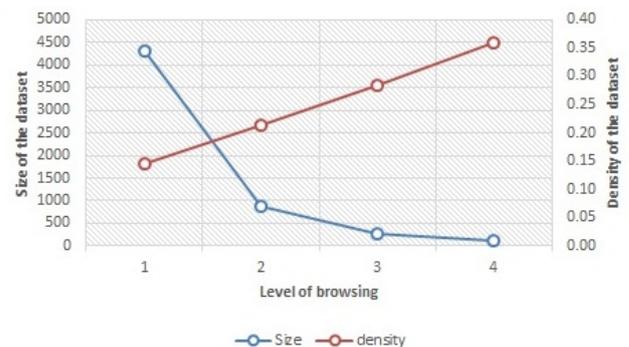


FIGURE 8: Results of browsing different levels

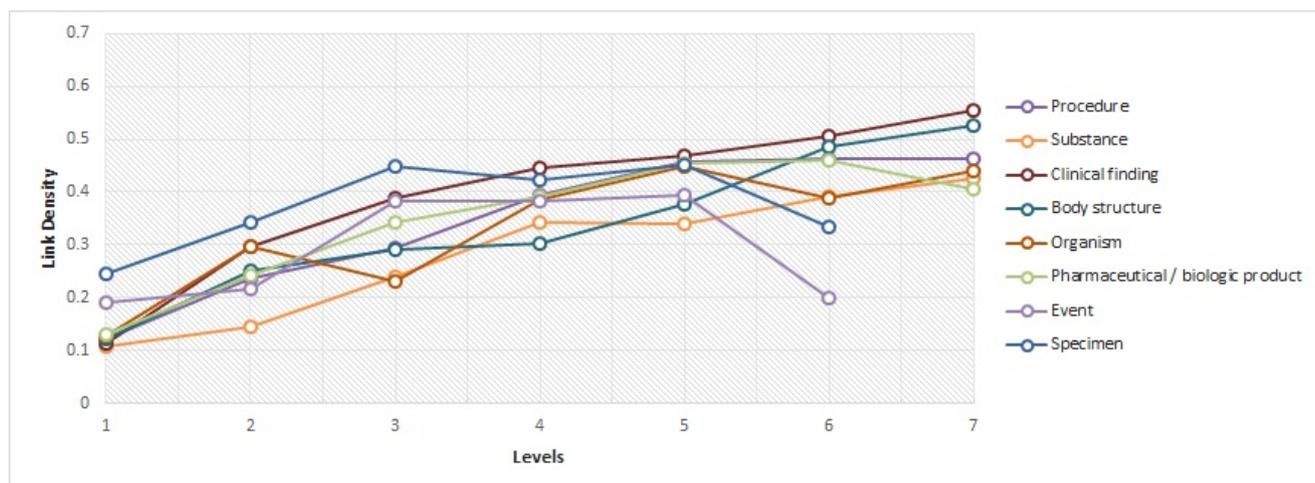


FIGURE 9: The link density score variation at different levels of browsing

6.2. Comparison of clustering experiments

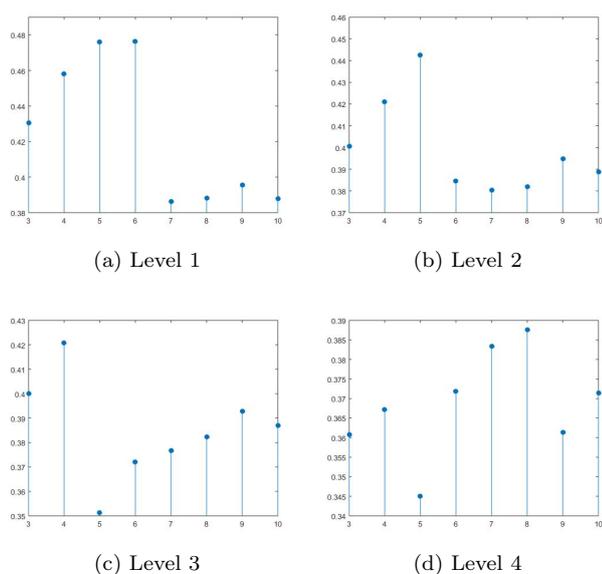


FIGURE 10: The silhouette plots for node 2 (Substance branch)

The full experiment of clustering the result data when browsing LEMO is detailed in table 4. The table details the results of applying clustering on the retrieved EMOs datasets of size (m) at each trial. At different levels, the highest silhouette value (s) and its associated number of clusters (k) are listed. The average silhouette value and link density of each branch are also detailed. The silhouette values for each clustering experiment applied for different levels of browsing for node 2 (Substance branch) are illustrated in figure 10. The number of clusters k with the highest silhouette value is chosen as the best number of clusters for the dataset tested.

We notice that clustering is more efficient with larger datasets retrieved compared to smaller ones.

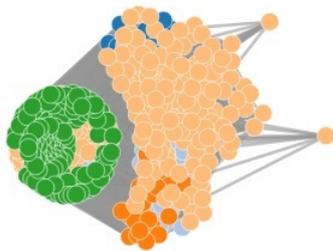
The largest dataset was retrieved when browsing node 3 (Clinical finding), and resulted in more than 8000 EMOs retrieved which are annotated with its descendent classes. The clustering consistently works well at deeper levels too. Node 3 has the highest average silhouette value of all the clustering performed at different levels, while the lowest average silhouette value is related to node 9 (Specimen) which has the lowest number of EMOs retrieved. For further discussion, the node with the highest average silhouette value is selected as a case study. The clustering experiment gives a good indication for the distribution of the LEMO dataset over the SNOMED CT ontology resulting from the enrichment process. At some branches, the nodes did not extend to more than 4 levels as in the cases of nodes 7 and node 9. This experiment does not give any validation of the correctness of the linkages made in LEMO based on its subject annotations. It only indicates that at different levels of browsing, although the data retrieved in smaller levels is a subset of those in the higher levels, the results of the clustering will change and in some cases improve.

A comparison between two datasets of clusters was conducted in order to validate the efficiency of the annotation process. The first dataset contains the clusters resulting from clustering the most dense node in the ontology (node 3) at level 6 named **node dataset** which consists of 5 clusters. The second dataset is combined from 5 clusters resulting when browsing deep levels of different branches named **branches dataset**. Both datasets consist of 5 clusters in order to compare their internal quality measures. The results retrieved from node 3 at level 6, **node dataset**, were clustered using agglomerative hierarchical clustering. The results retrieved were 568 EMOs clustered into 5 clusters. Figure 11a illustrates the results. The clusters are not well separated. The links between EMOs from different clusters makes visualization of the distinct clusters hard. Although the data are highly linked, the

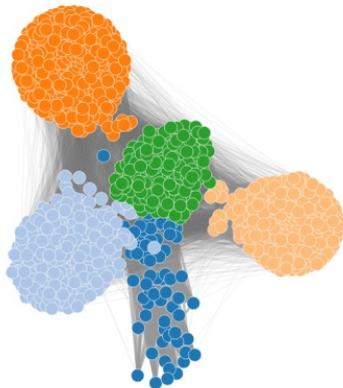
TABLE 4: Experiment of clustering tree branches at different levels

nodes \ levels	level 1			level 2			level 3			level 4			level 5			Avg(s)	Link Density
	<i>m</i>	<i>s</i>	<i>k</i>														
node 1	6950	0.475	4	6134	0.468	5	2784	0.462	4	801	0.427	3	568	0.462	3	0.459	0.464
node 2	6732	0.478	6	4932	0.442	5	4776	0.420	4	3881	0.385	8	2493	0.408	9	0.427	0.424
node 3	8495	0.529	5	7055	0.552	8	4820	0.548	3	4585	0.554	3	1297	0.549	3	0.546	0.556
node 4	7977	0.477	5	6919	0.466	5	6907	0.465	5	5405	0.443	5	1760	0.361	9	0.442	0.527
node 5	3748	0.5	3	2551	0.508	3	2051	0.509	3	2051	0.509	3	1934	0.506	3	0.506	0.449
node 6	2512	0.44	4	513	0.342	7	245	0.37	5	224	0.4	3	215	0.385	3	0.387	0.459
node 7	1049	0.499	8	341	0.449	6	24	0.387	6	24	0.387	6	-	-	-	0.431	0.394
node 8	2305	0.536	5	1541	0.481	5	1368	0.47	5	990	0.457	5	457	0.518	6	0.492	0.502
node 9	135	0.342	7	59	0.41	6	59	0.41	6	59	0.41	6	-	-	-	0.393	0.452

clusters are compact but not well separated. The results are logical since the dataset contains EMOs retrieved at a deep level of one node. The second dataset, **branches dataset**, is illustrated in figure 11b. It consists of 1322 EMOs combined from 5 clusters of different sizes from different branches, hence the clusters are well separated from each other and each cluster is compact and has weaker linkages to other clusters.



(a) clusters of one branch



(b) clusters of different branches

FIGURE 11: Comparison of visualizing the clusters resulted from the validation process

Evaluating the internal measures of the two experiments resulted in the following results detailed in

TABLE 5: Comparison of the evaluation measures

Datasets	internal evaluation measures			
	BetaCV	NC	Modularity	DB
Node dataset	0.6115	0.8567	-0.0144	0.1116
Branches dataset	0.4339	0.8938	-0.0216	0.0621

table 5, which indicate that the **branches dataset** is better clustered than the **node dataset**. The lower the values of the BetaCV, Modularity, and Davies-Bouldin indices the better the clustering is, while higher values for NC indicates better clustering, as explained in the previous section. The results are fairly close to each other, suggesting that both clustering results are good, but in comparison the **branches dataset** is better as measured by the results of all the criteria. This indicates that the ontology-based browsing was able to retrieve EMOs that are related to each other and have similar topic based on their Term resources annotation. The comparison results support this conclusion since the clusters of different branches are considered of a better quality than clusters of one branch as detailed in the table.

6.3. Querying the LEMO dataset

In this section, we present the results of searching the LEMO dataset. We compared two approaches of query searching: ontology based searching and text based searching. As explained in section 5.3, we proposed an algorithm for query searching based on the SNOMED CT ontology classes used for annotations. The algorithm explains the steps of performing a query search on a local server as the RDF store is not published for user access. Hence, testing the query searching interface is not possible with real users. therefore, to test the validity of this proposed algorithm, we ran automatic random queries on the LEMO dataset and compared the results with text based searching. The set of random queries were restricted to selecting one ontology concept from the set of ontology classes stored in the RDF store. The proposed algorithm

TABLE 6: Ontological-based vs. Text-based query searching results

Query Class	Ontology based results (O)	Text based results (T)	Overlap Coefficient ($O \cap T$)	Jaccard Similarity Coef.
Hepatitis	27	21	100%	0.78
Influenza	30	25	92%	0.71
Muscle	66	65	95%	0.89
Brain	61	49	100%	0.80
Renal Disease	36	4	100%	0.11
Hypoglycemia	49	40	100%	0.82
Vasculitis	87	53	98%	0.59
Leukemia	78	63	95%	0.741
Appendicitis	68	68	100%	1.00
Glomerulonephritis	97	75	100%	0.77

utilizes the relations between the ontology classes used in the annotation process to enrich the query sent from the user. A user can query the LEMO dataset using an auto-complete field which restricts the user to choosing a single ontology class. The algorithm then expands the query class into a query vector of related classes weighted according to its importance. We ran 10 random queries as a simulation of users querying the LEMO dataset, then compared the search results against simple text based searching. The sample queries are detailed in table 6. All the queries are composed of one syllable class ontology in order to compare with text based searching. The goal from these random queries was to evaluate the validity of the proposed algorithm. Hence, one word syllable classes are chosen in order to avoid the complexity of text based searching and perform exact text matching for one word.

The results presented in table 6 detail the size of the search results when performing ontology based searching and text based searching. The results are compared using the overlap coefficient and the Jaccard similarity coefficient. These measures compare the coverage of the ontological based approach and the similarity of the search results in both approaches. In most of the queries, the ontology based approach resulted in a larger number of search results retrieved, due to enriching the query sent with other related classes in the ontology. Also, the overlap coefficient results indicate that the ontology based search covers most of the search results from the text based approach. In some cases, the overlap coefficient shown is not 100% which indicates that some of the text based search results are not retrieved using the ontology based approach. Also, the similarity coefficient indicates that the ontology based approach can fail to cover all the results of the text based approach. This might indicate that the annotation process failed to annotate some terms in the textual metadata.

The Jaccard similarity coefficient measures pairwise similarity of the two search results retrieved. The value of this measure ranges from 0 to 1 where 1 means that the two compared sets are exactly similar. A low Jaccard coefficient value results when the two dataset are not similar in their size or content. In this

experiment, low Jaccard coefficient values indicate that the ontology based approach covers a larger number of search results than the text based approach. For example, in the case of “Renal Disease”, although the overlap coefficient is 100%, the Jaccard similarity value is low. Notice that the size of the ontology based search results is larger than that of the text based approach results, and resulted in expanding the class “Renal Disease” into a query vector of 24 other related classes (renal vascular disorder, nephritis, nephrosis and nephrotic syndrome, renal impairment, *cdots*). In this case, the search results retrieved include any EMO annotated with any of the related classes ranked by relevance to the original query sent. In another example, searching for “Vasculitis” will expand the query to include more general classes including hypersensitivity angiitis, arteritis and phlebitis. The goal of the proposed algorithm was to utilize the annotations enriching LEMO in the query searching process, and the results of this experiment indicate the effectiveness of the proposed algorithm in discovering more EMOs about the topic despite some limitations that can occur in the annotation process.

7. CONCLUSIONS

The wide range of open data available on the web has made searching for content, that can be used to learn a particular topic, a time-consuming task. The work presented in this paper has proposed a practical solution to the problem identified and applied it the field of medical education as a proof of concept. The proposed solution adopts Linked Data practices for exposing and connecting Educational Medical Objects (EMOs) of various types. Using SNOMED CT ontology we annotated and enriched the data collected and enabled linkages between data items. The main goal was to build a linked dataset of various types of EMOs collected from distributed web data sources. The system proposed in this paper has implemented methods that create, update, and store Linked Data in an RDF store that manages all the harvested EMOs. This resulted in a linked dataset of more than 10,000 educational materials varying in types including articles, videos, and blogs. The work can be

extended to include other educational material types such as virtual patients, and pictures. The RDF store content has been evaluated via techniques developed for accessing and retrieving EMOs from the RDF store. These techniques have exploited the ontology-based annotations enriching the metadata in order to enhance browsing and querying the RDF store. Clustering the browsing results has identified communities in the dataset based on the annotations similarities and that indicates the success of the ontology-based browsing techniques. Also, ontology based searching has enabled larger numbers of results to be retrieved compared to text based searching. The work can be extended to include larger datasets and can be tested using different ontologies. At this point, the dataset has not been published for users' access, and experiments were conducted to simulate the user behaviour. Future work will focus on providing an advanced user interface for accessing and using the system proposed. The evaluation methods have been useful for validating the new techniques and methods proposed in this work for aggregating and integrating distributed educational objects from the web. One limitation in the evaluation of this work is not involving expert users. Further experiments can be conducted with smaller clusters of datasets where linkages between the EMOs can be evaluated with expert users to validate these linkages.

REFERENCES

- [1] Newland, B. and Byles, L. (2014) Changing academic teaching with web 2.0 technologies. *Innovations in Education and Teaching International*, **51**, 315–325.
- [2] Fleischer, D. M., Posel, N. H., and Steacy, S. P. (2004) New directions in medical e-curricula and the use of digital repositories. *Academic Medicine*, **79**, 229–235.
- [3] Popoiu, M. C., Grosseck, G., and Holotescu, C. (2012) What do we know about the use of social media in medical education? *Procedia-Social and Behavioral Sciences*, **46**, 2262–2266.
- [4] Sampson, D. (2004) The evolution of educational metadata: From standards to application profiles. *Proceedings of the IEEE International Conference on Advanced Learning Technologies*, pp. 1072–1073. IEEE Computer Society.
- [5] Bizer, C., Heath, T., and Berners-Lee, T. (2009) Linked data-the story so far. *International Journal on Semantic Web and Information Systems*, **5**, 1–22.
- [6] Heath, T. and Bizer, C. (2011) Linked data: Evolving the web into a global data space. *Synthesis Lectures on the Semantic Web: Theory and Technology*, **1**, 1–136.
- [7] 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. *Proceedings of the 27th Annual ACM Symposium on Applied Computing*, pp. 366–371. ACM.
- [8] Vega-Gorgojo, G., Asensio-Pérez, J. I., Gómez-Sánchez, E., Bote-Lorenzo, M. L., Muñoz-Cristobal, J. A., and Ruiz-Calleja, A. (2015) A review of linked data proposals in the learning domain. *Journal of Universal Computer Science*, **21**, 326–364.
- [9] Dietze, S., Sanchez-Alonso, S., Ebner, H., Qing Yu, H., Giordano, D., Marenzi, L., and Pereira Nunes, B. (2013) Interlinking educational resources and the web of data: A survey of challenges and approaches. *Emerald Program: Electronic Library and Information Systems*, **47**, 60–91.
- [10] d'Áquin, M. (2012) Putting linked data to use in a large higher-education organisation. *Proceedings of the Interacting with Linked Data (ILD) Workshop at Extended Semantic Web Conference (ESWC)*, pp. 9–21.
- [11] Al Fayez, R. Q. and Joy, M. (2014) A framework for linking educational medical objects: Connecting web2.0 and traditional education. *Web Information Systems Engineering-WISE 2014*, pp. 158–167.
- [12] Rajiv, M. L. and Sridhar, M. (2011) Enhancing teaching & learning with web tools. *An International Journal of Engineering Sciences ISSN*, **1**, 2229–6913.
- [13] Yuen, S. C.-Y. and Yuen, P. (2008) Web 2.0 in education. *Society for Information Technology & Teacher Education International Conference*, pp. 3227–3228.
- [14] Smith, M. S. and Casserly, C. M. (2006) The promise of open educational resources. *Change: The Magazine of Higher Learning*, **38**, 8–17.
- [15] Kaldoudi, E., Konstantinidis, S., and Bamidis, P. D. (2010) Web 2.0 approaches for active, collaborative learning in medicine and health. In Fiaidhi, J. and Mohammed, S. (eds.), *Ubiquitous Health and Medical Informatics: Advancements in Web 2.0, Health 2.0 and Medicine 2.0*, pp. 127–149. IGI Global, Hershey, PA, USA.
- [16] Sandars, J. and Schroter, S. (2007) Web 2.0 technologies for undergraduate and postgraduate medical education: an online survey. *Postgraduate Medical Journal*, **83**, 759–762.
- [17] Franks, P. and Kunde, N. (2006) Why metadata matters. *Information Management*, **40**, 55–58.
- [18] Devedžić, V. (2006) *Semantic web and education*. Springer Science & Business Media.
- [19] Smothers, V. (2004). Healthcare learning object metadata: Specifications and description document. http://www.medbiq.org/sites/default/files/files/HealthcareLOMSpecifications_pointrelease.pdf. Online. Accessed on January 13, 2014.
- [20] Candler, C. S., Uijtdehaage, S. H., and Dennis, S. E. (2003) Introducing HEAL: The health education assets library. *Academic Medicine*, **78**, 249–253.
- [21] Lipscomb, C. E. (2000) Medical Subject Headings (MeSH). *Bulletin of the Medical Library Association*, **88**, 265–266.
- [22] Domingue, J., Mikroyannidis, A., and Dietze, S. (2014) Online learning and linked data: lessons learned and best practices. *Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion*, pp. 191–192.
- [23] Ruiz-Calleja, A., Vega-Gorgojo, G., Asensio-Perez, J. I., Bote-Lorenzo, M. L., Gomez-Sanchez, E., and Alario-Hoyos, C. (2012) A Linked Data approach for the discovery of educational ICT tools in the Web of Data. *Computers & Education*, **59**, 952–962.

- [24] Mitsopoulou, E., Taibi, D., Giordano, D., Dietze, S., Yu, H. Q., Bamidis, P., Bratsas, C., and Woodham, L. (2011) Connecting medical educational resources to the Linked Data cloud: the mEducator RDF Schema, store and API. *Proceedings of Linked Learning 2011: the 1st International Workshop on eLearning Approaches for the Linked Data Age, 8th Extended Semantic Web Conference (ESWC2011)*, pp. 1–15.
- [25] Alonso-Roris, V. M., Míguez-Pérez, R., Santos-Gago, J. M., and Álvarez-Sabucedo, L. (2012) A semantic enrichment experience in the early childhood context. *Frontiers in Education Conference (FIE), 2012*, pp. 1–6. IEEE.
- [26] Isaac, Y., Bourda, Y., and Grandbastien, M. (2012) Semunit-french unt and linked data. *LiLe-2012 at WWW-2012*, pp. 6–12. CEUR workshop proceedings.
- [27] Sicilia, M. A., Ebner, H., Sánchez-Alonso, S., Álvarez, F., Abián, A., and García-Barriocanal, E. (2011) Navigating learning resources through linked data: a preliminary report on the re-design of Organic. Edunet. *Proceedings of Linked Learning, 2011*, 1–8.
- [28] Mikroyannidis, A., Domingue, J., Maleshkova, M., Norton, B., and Simperl, E. (2014) Developing a curriculum of open educational resources for linked data. *Proceedings of 10th Annual OpenCourseWare Consortium Global Conference (OCWC)*, pp. 1–8.
- [29] Zablith, F., Fernandez, M., and Rowe, M. (2015) Production and consumption of university linked data. *Interactive Learning Environments, 23*, 55–78.
- [30] Ritze, D. and Eckert, K. (2014) Data enrichment in discovery systems using linked data. *Data Analysis, Machine Learning and Knowledge Discovery*, pp. 455–462. Springer.
- [31] Díaz-Galiano, M. C., García-Cumbreras, M., Martín-Valdivia, M. T., Montejo-Ráez, A., and Urena-López, L. (2007) Integrating MeSH ontology to improve medical information retrieval. *Workshop of the Cross-Language Evaluation Forum for European Languages*, pp. 601–606. Springer.
- [32] Waitelonis, J., Sack, H., Hercher, J., and Kramer, Z. (2010) Semantically enabled exploratory video search. *Proceedings of the 3rd International Semantic Search Workshop at the 19th International World Wide Web Conference (WWW)*, pp. 1–8. ACM.
- [33] Yu, H. Q., Pedrinaci, C., Dietze, S., and Domingue, J. (2012) Using linked data to annotate and search educational video resources for supporting distance learning. *IEEE Transactions on Learning Technologies, 5*, 130–142.
- [34] Choudhury, S., Breslin, J. G., and Passant, A. (2009) Enrichment and ranking of the YouTube tag space and integration with the linked data cloud. *International Semantic Web Conference*, pp. 747–762. Springer.
- [35] Hoehndorf, R., Dumontier, M., and Gkoutos, G. V. Evaluation of research in biomedical ontologies. *Briefings in Bioinformatics, 14*, 696–712.
- [36] Learning Technology Standards Committee (2002) IEEE Standard for learning object metadata. *IEEE Standard, 1484*.
- [37] Lagoze, C. and Van de Sompel, H. (2003) The making of the open archives initiative protocol for metadata harvesting. *Library Hi Tech, 21*, 118–128.
- [38] Powell, A., Nilsson, M., Naeve, A., Johnston, P., and Baker, T. (2007). DCMI abstract model. <http://travesia.mcu.es/portalnb/jspui/handle/10421/3481>. Online. Accessed on July 2, 2015.
- [39] Jonquet, C., LePendu, P., Falconer, S., Coulet, A., Noy, N. F., Musen, M. A., and Shah, N. H. (2011) NCBO Resource Index: Ontology-based search and mining of biomedical resources. *Web Semantics: Science, Services and Agents on the World Wide Web, 9*, 316–324.
- [40] Al Fayez, R. Q. and Joy, M. (2015) Applying NoSQL databases for integrating web educational stores- An Ontology-Based approach. *Data Science*, pp. 29–40. Springer.
- [41] Noy, N. F., Shah, N. H., Whetzel, P. L., Dai, B., Dorf, M., Griffith, N., Jonquet, C., Rubin, D. L., Storey, M. A., Chute, C. G., et al. (2009) Biportal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Research, 1*, 1–4.
- [42] Rubin, D. L., Shah, N. H., and Noy, N. F. (2008) Biomedical ontologies: a functional perspective. *Briefings in Bioinformatics, 9*, 75–90.
- [43] Elevitch, F. R. (2005) SNOMED CT: electronic health record enhances anesthesia patient safety. *AANA Journal, 73*, 361–366.
- [44] Van Hage, W. R., De Rijke, M., and Marx, M. (2004) Information retrieval support for ontology construction and use. *The Semantic Web-ISWC 2004*, pp. 518–533. Springer.
- [45] Stearns, M. Q., Price, C., Spackman, K. A., and Wang, A. Y. (2001) SNOMED clinical terms: overview of the development process and project status. *Proceedings of the AMIA Symposium*, pp. 662–666. American Medical Informatics Association.
- [46] Zhang, J. (2008) *Visualization for information retrieval*. Springer.
- [47] Lee, D., de Keizer, N., Lau, F., and Cornet, R. (2014) Literature review of SNOMED CT use. *JAMIA Journal of the American Medical Informatics Association, 21*, 11–19.
- [48] Lieberam-Schmidt, S. (2010) Web structure. *Analyzing and Influencing Search Engine Results*, pp. 49–103. Springer.
- [49] Benesty, J., Chen, J., Huang, Y., and Cohen, I. (2009) Pearson correlation coefficient. *Noise Reduction in Speech Processing*, pp. 1–4. Springer.
- [50] Ceri, S., Bozzon, A., Brambilla, M., Della Valle, E., Fraternali, P., and Quarteroni, S. (2013) Classification and clustering. *Web Information Retrieval Data-Centric Systems and Applications*, pp. 39–56. Springer Berlin Heidelberg.
- [51] Granichin, O., Volkovich, Z., and Toledano-Kitai, D. (2015) Cluster validation. *Randomized Algorithms in Automatic Control and Data Mining*, Intelligent Systems Reference Library, **67**, pp. 163–228. Springer Berlin Heidelberg.
- [52] Zaki, M. J. and Wagner Meira, J. (2014) *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press.
- [53] Zhang, J. (2008) Information retrieval preliminaries. In Zhang, J. (ed.), *Visualization for Information Retrieval*, The Information Retrieval Series, **23**, pp. 21–46. Springer Berlin Heidelberg.