The Role of Taxonomies in Social Media and the Semantic Web for Health Education

A Study of SNOMED CT Terms in YouTube Health Video Tags

S. Konstantinidis¹; L. Fernandez-Luque²; P. Bamidis³; R. Karlsen²,4

¹Aristotle University of Thessaloniki, Lab of Medical Informatics, Medical School, Thessaloniki, Greece;
²Norut, Tromsø, Norway;
³Aristotle University of Thessaloniki, Lab of Medical Informatics, Medical School, Thessaloniki, Greece;
⁴University of Tromsø, Computer Science Department, Tromsø, Norway

Keywords
Health video, tags, tag cloud, SNOMED CT, terms, Medical thesauri, linked data, linked medical data, educational metadata, learning/content management system

Summary
Background: An increasing amount of health education resources for patients and professionals are distributed via social media channels. For example, thousands of health education videos are disseminated via YouTube. Often, tags are assigned by the disseminator. However, the lack of use of standardized terminologies in those tags and the presence of misleading videos make it particularly hard to retrieve relevant videos.

Objectives: i) Identify the use of standardized medical thesauri (SNOMED CT) in YouTube Health videos tags from preselected YouTube Channels and demonstrate an information technology (IT) architecture for treating the tags of these health (video) resources. ii) Investigate the relative percentage of the tags used that relate to SNOMED CT terms. As such resources may play a key role in educating professionals and patients, the use of standardized vocabularies may facilitate the sharing of such resources. iii) Demonstrate how such resources may be properly exploited within the new generation of semantically enriched content or learning management systems that allow for knowledge expansion through the use of linked medical data and numerous literature resources also described through the same vocabularies.

Methods: We implemented a video portal integrating videos from 500 US Hospital channels. The portal integrated 4,307 YouTube videos regarding surgery as described by 64,367 tags. BioPortal REST services were used within our portal to match SNOMED CT terms with YouTube tags by both exact match and non-exact match. The whole architecture was complemented with a mechanism to enrich the retrieved video resources with other educational material residing in other repositories by following contemporary semantic web advances, in the form of Linked Open Data (LOD) principles.

Results: The average percentage of YouTube tags that were expressed using SNOMED CT terms was about 22.5%, while one third of YouTube tags per video contained a SNOMED CT term in a loose search; this analogy became one tenth in the case of exact match. Retrieved videos were then linked further to other resources by using LOD compliant systems. Such results were exemplified in the case of systems and technologies used in the mEducator EC funded project.

Conclusion: YouTube Health videos can be searched for and retrieved using SNOMED CT terms with a high possibility of identifying health videos that users want based on their search criteria. Despite the fact that tagging of this information with SNOMED CT terms may vary, its availability and linked data capacity opens the door to new studies for personalized retrieval of content and linking with other knowledge through linked medical data and semantic advances in (learning) content management systems.
1. Introduction

Over the last decade the World Wide Web has undergone rapid changes. Social media have fostered the interaction, collaboration and sharing of resources. It is clear that today health social media is one of the main vehicles for health communication for both health professionals and patients [1,2].

The retrieval of health social media resources and their description within the mass of public social media has recently become a difficult task due to the enormous explosion of the Social Web. It is becoming increasingly difficult to accurately retrieve quality health resources of educational value [3–5].

Medical terminologies and classification standards have been brought in to simplify the identification and retrieval of health resources. Examples of such terminologies and classification standards include SNOMED CT, Read Clinical Terms, UMLS, GALEN, MEDCIN, CPT-4, LOINC, ICPC-2, ICD-10 [6]. The use of medical vocabularies in patients’ forums, such as PatientsLikeMe [7] has also been explored.

SNOMED CT (Systematised Nomenclature of Medicine – clinical terms) has been re-enforced by the College of American Pathologists (CAP) and, as of April 2007, it is owned, maintained, and distributed by the International Health Terminology Standards Development Organisation (IHTSDO) (http://www.ihtsdo.org/snomed-ct/). SNOMED CT provides a general terminology for electronic health records (EHRs) and contains more than 311,000 active concepts with unique meanings and formal logic-based definitions organized into hierarchies. SNOMED CT is taking a central role in medical informatics core activities in many countries: it is now mandatory in the US government’s incentive program for meaningful use of electronic health records [8], while the UK Department of Health intends it to be the only terminology that will be supported from April 2015 and beyond [9], [10]. SNOMED CT consists of concepts, descriptions and relationships. A concept is a clinical meaning and is identified by a unique number, while each concept is associated with two or more descriptions, which are human readable terms, and information about the term [11]. In this paper we consider as a SNOMED CT term, any term that is human readable and exists in the description of a SNOMED CT concept.

SNOMED CT and other reference terminologies provide formal concept definitions that allow for an easier and computer readable coding and storage of clinical data [12–14]. SNOMED CT is regarded as the most comprehensive terminological system for coding clinical information [12], while it provides some codes for a few score details in the form of individual concepts [15].

Over the years SNOMED CT has been used in multiple research efforts to conceptualize health related information such as clinical records [16] or health interventions [17–19]. Legacy interface terminologies [20], or domain-specific interface terminologies have been developed based on SNOMED CT [12]. Mapping biomedical text to metathesaurus [21] or mapping terminologies (e.g. International Classification for Nursing Practice -ICNP) with SNOMED CT is an ongoing research activity to support the interoperability of different health care concepts [22]. However, only a few research endeavors have attempted to identify the use of SNOMED CT and other thesauri for annotating health social media, either in an automatic or human intervened way, or through the use of semantic technologies [23], in an effort to achieve best practice in medical education (cf. the m Educator project [24]).

In parallel developments, numerous standards have been initiated in recent years in an attempt to describe educational processes or educational resources that exist on the Web [25]. Web2.0 has changed dramatically the way users collaborate and share content in the health domain [26], [27]. Wikis, blogs, forums and in general social media hold a new type of knowledge that is more difficult to retrieve using the traditional techniques of content retrieval. However, standardizing the description of social media resources and/or information is itself a difficult procedure.

One third of American Internet users have tagged online content [28], and 6% of health information seekers have tagged health content [29]. When users tag health social media they are not just making it easier for themselves to use this information at a later stage, they are also making it easier for others to search for and retrieve these resources. In addition, the tags that describe that information can be seen in some cases as additional content to the information/resource [30]. Moreover, in some other cases, tags form special terminology cases, usually linked to a theme or a practice community; these have recently been given the name of folksonomy, which could essentially act as a user-generated terminology for the specific community, while also providing the keywords necessary for successful content and information searches.

In contrast, terminologies and thesauri in the field of healthcare information technology are certainly required in the process of fixing/stabilizing the language within a specific domain. To this extent, terminologies and thesauri foster unambiguous intra-(and extra)-community communications, providing a common language. Taxonomies facilitate semantic indexing, information retrieval and exchange, as well as the linking between heterogeneous systems. Furthermore taxonomies could serve as the basis for documentation through coding and classification [6]. Another area of research is the creation of lay-person vocabularies that are mapped into medical standardized vocabularies to bridge the gap between patients and professionals [31]. Tagging can also be used to personalize content [32].

Moreover, advances on the semantic web front (or Web3.0), have led to several ontologies and ontology formats that have also capitalized on the semantics of metadata created by tagging activities. The Simple Knowledge Organization System (SKOS) (http://www.w3.org/2004/02/skos/); schemes like the Description of a Project (DOAP) or the Friend of A Friend (FOAF) and other ontology formats enable the structure of tagging activities among other advancements to facilitate modeling and the integration of tagging data [33]. To this extent, the definition of metadata and ontologies are necessary steps to create the “Web of Data” (aka semantic web or Web 3.0) [34]. However, another necessary step
is the integration of different semantically-enhanced data sources that can be machine-readable. An approach to solve that integration problem is the “Linked Data” initiative [35]. Tim Berners-Lee has suggested that Linked Data will be a key element in the creation of the future smart web [36].

In this piece of research, we focus on identifying the use of predefined thesauri (SNOMED CT) while annotating YouTube Health videos with tags that are uploaded from more than 500 American hospitals, and constitute a web resource pool of some thousands of health videos referring to surgery. The value of these resources is enormous since many of them can easily be used to educate health professionals and patients. However, finding this kind of resource in the current information overload situation affecting the web is not an easy task.

In recent years, the scene in health education has been changed by the use of social media, while YouTube Videos have been used as educational resources [37–39]. The purpose of this paper is not to check the validity of YouTube videos as learning resources, or identify an algorithm that investigates their quality. Instead, our objective is to identify whether the tags of the videos collected through a preselected list of Hospital channels by making use of the HealthTrust algorithm [40, 41], contain SNOMED CT terms in their user assigned tags that would facilitate their retrieval, together with the process for mapping and correlating them with other resources by means of semantic web and linked data principles. For this reason, this paper aims to numerically check the matches between the tags of these videos and SNOMED CT terms, so as to investigate the feasibility of video enrichment and its correlation/interlinking with other existing information and knowledge by using the “same as” property.

Thus, the objective of this paper is threefold. First, demonstrate an information technology (IT) architecture for retrieving the tags (i.e. the folksonomy) of these health (video) resources. Second, investigate the relative percentage of the used tags that relate to SNOMED CT terms. These results add value and knowledge to a third endeavour: demonstrate how such resources may be properly exploited within the new generation of semantically enriched content or learning management systems.

The remainder of this paper is as follows. In the methods section, detailed information is provided regarding: i) the architecture of the IT infrastructure that enables the subsequent study of the resources; and, ii) the material at hand. In the following section, numerous diagrams and figures depict the results obtained from the study of the material and its handling in an exemplar semantically enriched content/learning management system. The importance of these achievements together with their current limitations and future prospects are discussed in the final section.

2. Methods

The identification of SNOMED CT terms in the tags of a total of 4,307 YouTube health videos was a time-consuming and difficult procedure. The videos were collected from channels published in the Ed Bennett List with US-based hospitals (http://ebennett.org/hsnl/). To achieve a valid and independent comparison between SNOMED CT terms and YouTube tags, we selected 4,307 videos that were automatically retrieved using a framework whose development was based on the results from mEducator and MyHealth-Videos projects [42]. This framework includes a semi-automated translation of YouTube metadata of health educational videos into mEducator metadata [43] and their subsequent publication as linked data. The mEducator scheme/ontology constitutes of 10 mandatory fields (identifier, title, creator, IPR license, language of resource, language of metadata, creation date, metadata creation data, keywords and description) and a number of optional fields (educational objectives, educational outcomes, assessment methods, educational context, technical description, discipline, etc.). An optional description of a resource that is repurposed can also be included [43]. In addition to the web and mobile interfaces, the framework consists of four basic components (Figure 1) a web crawler that extracts YouTube metadata and stores them in a content management system (i.e., Drupal v.7.0); 2) a module that automatically translates YouTube metadata (e.g. tags, description, title) into the standardized semantic annotation of the mEducator scheme[43]; this module is used for pre-filling the description form that may be used later by medical educators to describe the content; 3) a wizard facilitating the expert evaluation and completion of the mEducator metadata for each video[43]; 4) a module for the automatic publication of these educational videos as linked data; this module uses a customized Drupal SPARQL endpoint through which the mEducator metadata are exported as Linked Data in the cloud.

We developed a prototype portal for surgery videos. The described framework parts were developed as modules for the Drupal v7.0 content management system. The crawler was developed in MyHealth-Videos project [40, 41].

We also developed procedures for the automatic identification of SNOMED CT terms existing in the tags of the identified educational health videos. These procedures consisted of i) comparison functions that correlated the terms stored in the aforementioned portal database with SNOMED CT terms; ii) web-services of BioPortal REST services from the National Center for Biomedical Ontology. This REST API made it possible to search for terms in different ontologies, allowing the narrowing down of a search into only one ontology (e.g. SNOMED CT). Each query within the REST API returned an XML file containing the ‘concept id’ of the term, if it existed, or the terms that included the specific word that search was performed with.

The comparison of SNOMED CT terms and YouTube tags was divided into two major categories. This was done using the BioPortal REST service optional argument “exactmatch” which enabled the match of the entire concept name. The specific description of the BioPortal REST service for the aforementioned search functionality is that “… the search attempts to match both partial and exact queries, giving more...” [35].
weight to exact matches. In single-word searches, the wildcard character (•) is automatically appended to the end of the word. In phase searches (multiple words), the wildcard character is appended to the end of each word.

Functions were created for each of these two categories to group and compare: i) unique tags of all videos (a tag that existed in more than one video was checked only once); ii) all tags of all videos (a tag existed many times); iii) tags that existed per one video only. The aim of YouTube health video tag comparison was to identify whether these videos could be used to semantically enrich their content with predefined terminologies (SNOMED CT terms) in the sets of YouTube health video tags. The distribution of percentages and the cumulative frequencies of SNOMED CT terms existing in YouTube Health video tags in different cases (i), (ii) and (iii) mentioned above) would allow us to estimate the average possibility of SNOMED CT terms existing in YouTube Health video tags and, moreover, how many semantic annotations could be made with predefined terminologies (SNOMED CT in our case) or if there was a need for semantic annotation through folksonomies.

3. Results

The 4,307 videos that were automatically extracted from YouTube to which in total 64,367 tags were assigned by their providers; the overall number of unique tags was 7,798. Our results showed that the percentage of correspondence of 7,798 unique tags with SNOMED CT terms was 20.6% with the exact match set to false, while this percentage was limited to 4.7% with the exact match option set to true (Figure 2).

The percentage of existing SNOMED CT terms in overall tags (aka all tags regardless of how many times they existed in these videos) in our case was 7,798. The percentage of existing SNOMED CT terms in overall tags was 20.6% with the exact match option set to true (Figure 2).
video tags) of all videos (64,367 tags) was 37.6% with the “exactmatch” argument set to false, while the corresponding percentage with the “exactmatch” argument enabled was 7.4%.

Figure 3 depicts the percentage of SNOMED CT terms existing in YouTube tags per video with the “exactmatch” argument set to false while the mean value of the percentage was 34.3%. The number of videos that did not contain any SNOMED CT term (17.7% of the total videos), while the number of videos that contained more SNOMED CT terms in their tags than the average of 34.3% was 1973 (45.8% of the overall videos). If we exclude the videos that did not contain any SNOMED CT term in their tags, the average percentage of SNOMED CT terms existing in YouTube videos tags per video increased to 41.7%.

Figure 2 Average Percentage of SNOMED CT terms in YouTube health video tags. This figure shows the difference between unique and overlapped tags with the parameter “exactmatch” enabled or disabled. The comparison of average percentages revealed the existence of SNOMED CT terms in both cases; the case of “exactmatch” set to “off” dominated clearly. The existence of SNOMED CT terms in YouTube health videos tags could become promising when their connection/interlinking with other resources was considered (e.g. through semantic annotation; see text).

Figure 3 Percentage of SNOMED CT terms existing in YouTube tags per video with the “exactmatch” argument set to false; their mean value is shown in red. Each YouTube Health video has a unique percentage of the SNOMED CT terms that it contains in its tags; almost half of the depicted percentages were above the mean value of 34.34%, which raises the possibility of connecting/interlinking them with other related resources through semantic enrichment, as the latter could exploit SNOMED CT.
The following diagram (Figure 4) shows the distribution of percentages of existing SNOMED CT terms per YouTube Health video tags with the “exactmatch” parameter set to false. The standard deviation of this percentage was quite high (25.1), thereby revealing that there was no standard/usual number of SNOMED CT terms per video’s tags, but the distribution of SNOMED CT terms per video’s tags could vary greatly.

As the following diagram (Figure 5) shows, the distribution of the percentages of SNOMED CT terms existing in YouTube tags per video with the “exactmatch” argument set to false excluding videos with no SNOMED CT term was not normal. The standard deviation of this condition was still high enough (21.4) with the average percentage being 41.7%.

From the Diagram of Cumulative Frequencies (Figure 6) of percentage of SNOMED CT terms existing in YouTube tags per video with the “exactmatch” argument set to false, it is clear that half of the videos’ tags (51.57%) contained more than one SNOMED CT term out of three (33.3% to 100% of tags were SNOMED CT terms), while the number of videos decreased (14.56%) when they were described with two SNOMED CT terms out of three (62.5% to 100% of tags were SNOMED CT terms). However, there was a small percentage (2.3%) of total video numbers that were fully described (87.5% to 100% of tags were SNOMED CT terms).

In the case of the “exactmatch” argument set to true, the distribution of percentage of SNOMED CT terms existing in YouTube tags per video (Figure 7) was slightly different to the videos that had over one third of SNOMED CT terms in their tags, which was negligible (3.9% of total videos), while the standard deviation was lower (11.1) than for videos queried with “exactmatch” disabled. The mean value of the percentages of SNOMED CT terms in YouTube tags per video with the “exactmatch” argument set to true was 7.1%, which was close to the percentage of SNOMED CT terms on all overlapped tags queried with “exactmatch” enabled.

The Cumulative Frequencies Diagram (Figure 8) reveals that there were a huge number of videos that did not have a SNOMED CT term in their tags with the “exactmatch” parameter set to ‘on’. The percentage of videos with absolutely no terms was as high as 58.1%.

If one wanted to compare the distribution of percentages of SNOMED CT terms existing in YouTube tags per video, excluding videos with no SNOMED CT term at all, and while the “exactmatch” argument was set to true or false, the values around the mean percentage (17.1% in the case of “exactmatch” enabled) had a standard deviation of 11.1 which was higher than one with 0 percentage. This is because most of the values were close to zero percentage, so the distribution did not vary greatly (Figure 9).

YouTube Health video tags that are mapped to SNOMED CT terms facilitated the procedure of linking these further with
other resources as well as the Linked Open Data cloud. This could make the annotation and enrichment of metadata much easier since many health datasets contain or are linked with other medical resources that are described by medical taxonomies such as SNOMED CT. Figure 10 depicts such an example for the SNOMED CT term “leukemia” which in turn is related with three different datasets.

Thus, the work described above presents a robust and highly promising way of correlating health resources in heterogeneous systems in which linking is performed directly through SPARQL endpoints which follow the mEducator scheme/ontology [43]. As SNOMED CT is proposed as a best practice point in the mEducator scheme/ontology for describing medical education resources [43], the whole mechanism appears promising regarding the correct subsequent use of these social media resources in content or learning management systems. So, for example, searching in other SPARQL endpoints which also use SNOMED CT terms is now enabled, which greatly increases the probability of retrieving related medical educational resources.

Figure 11 depicts the current architectural set-up in the mEducator project in which different (content or learning management) systems enabled with SPARQL endpoints that follow the mEducator scheme/ontology [43], are interlinked with the aforementioned portal to exploit powerful uses of such videos in medical education. The depicted architecture allows the health media resources to be connected not only with resources from the Linked Open Data cloud, but also with other educational resources provided by the mEducator project.

4. Discussion

In this paper, we have presented a web based architecture capable of retrieving, sharing, annotating and studying health education videos from YouTube. As educational resources contained on social media platforms such as YouTube have undergone exponential growth in recent years, studying the correlation of the tagging mechanism of this content with taxonomies traditionally developed within the field of Health Informatics and Information Technology appears to have great potential. The architecture outlined in this paper, is, to the best of our knowledge, the first one not only to be presented in such a study, but also the first to provide the mechanisms for their simultaneous exploitation within content and learning management systems which follow principles of both the social and the semantic web (i.e. both Web2.0 and Web3.0).

The increased percentage of SNOMED CT terms existing in all the 64,367 tags in comparison with 7,798 unique tags, in both searches with the “exactmatch” option set on ‘on’ and ‘off’, reveal that health YouTube videos contained one SNOMED CT term out of five in their tags. This was high enough to expect useful interlinking with other resources through some semantic annotation.
videos tags tend to contain repeated SNOMED CT terms; the latter, in the case of no “exactmatch” can be up to one SNOMED CT term out of three (1/3) health YouTube video tags, while in the case of ‘exactmatch’ it can be one out of ten (1/10). All of the selected YouTube health videos come from preselected channels that usually contain the name and location of the hospital in their tags.

The framework we have proposed constitutes one of the first attempts to fuse health social media (i.e. Web2.0) content with the “Web of data” [34] (i.e. Web3.0); the widely used and accepted taxonomy of SNOMED CT has been used here as an example that propitiates this fusion. The existence of SNOMED CT terms in the YouTube Health videos tags (Figures 4 and 7) makes it possible to connect these videos to other medical resources through annotations and metadata enrichments. Medical resources are often described by taxonomies (e.g. PubMed papers keywords, mEducator Resources subject, Diseases label, etc), a significant number of which use SNOMED CT terms. Recent advances on the Linked Data front provide the unprecedented opportunity and challenge of linking YouTube health videos with these resources using proper metadata enrichments. For example, a video containing the tag “Leukemia” (Figure 10) that has been automatically recognized by the system as a SNOMED CT term may be automatically enriched with further related educational resources. The system annotator easily maps this term and enriches resource metadata with resources that are already known to contain SNOMED CT terms in their metadata and which appear in different datasets, such as clinical drug administration forms, that are mapped to SNOMED CT and FDA; LinkedCT open data, which is a results registry/database of federally and privately supported clinical trials conducted in the United States and around the world; and Linked Life Data, which provides access to more than 20 datasets including PubMed, DrugBank, Gene Ontology, etc. and many more. These enriched metadata could add value to the health education experience with the advantageous (automatic to the user) proposal of relative resources and knowledge that exists in the Linked Open Data (LOD) cloud. This means that while the user browses a YouTube health video there will be no further need to login to different system/search engines to search and retrieve resources or other related resources which have already been brought in automatically by the system. Thus, this work represents a significant breakthrough, because ‘linked knowledge’ with YouTube Health videos is not always easy to discover using traditional search engines and approaches.

While noting that metadata of YouTube health videos with no SNOMED CT terms at all can also be annotated and enriched, videos already containing tags that belong to a known terminology greatly enhance the connectivity prospects with other resources. This mechanism makes YouTube Health videos with SNOMED CT terms

![Figure 8](cumulative_frequencies.png) Cumulative Frequencies of percentage of SNOMED CT terms existing in YouTube tags per video with the “exactmatch” argument set to true; the existence of SNOMED CT terms with the “exactmatch” argument set to true was low, and the semantic annotations with other resources became more promising using folksonomies rather than SNOMED CT.

![Figure 9](distribution.png) Distribution of percentages of SNOMED CT terms existing in YouTube tags per video with the “exactmatch” argument set to true, excluding YouTube health videos with no SNOMED CT term; videos without SNOMED CT terms did not affect the distribution of percentages, but they did foster the following notion: it was possible to connect/interlink videos with at least one SNOMED CT term in their tags (with the “exactmatch” argument set to true) with other related resources using some semantic annotation.
The connectivity of YouTube health videos with other health resources could also be achieved in a different way. A search module can search and retrieve medical educational resources from other systems that share their resources as RDF triples, not just retrieve resources from the LOD cloud (Figure 11). mEducator, an EU funded best practice network (funded by the European Commission under the eContentPlus2008 programme) [24], elaborated on pedagogical, technical, standardization, cultural, social and legal issues towards a standard-based infrastructure that enables the sharing of state-of-the-art digital medical educational content among medical educators and students in European higher academic institutions. mEducator has compared two contemporary ways of achieving this content sharing, namely, mEducator 2.0 – a solution based on Web2.0 (mash-up) technologies, and mEducator3.0 – a solution based on semantic web services and Linked Data. Within the mEducator project three different systems were developed (namely MILES+ [45], MELINA+ and LL+ [46]) that share the metadata of their high-quality medical educational resources through SPARQL endpoints. The search module of YouTube Health Video Portal could bring to the educational processes state-of-the-art educational resources that come from different instantiations of those systems, which are spread around the world, or by other YouTube Health Video Portal instantiations. The educational experience could be enhanced either by the automatic provision of the above resources, or through the users themselves by giving them the chance to access such high-quality medical educational resources. SNOMED CT terms foster this educational experience since it is one of the two

Figure 10 An example of “further” linking (enriching) the tag “Leukemia” (contained in numerous YouTube Health videos, but also in SNOMED CT as a term) with Pubmed articles, Clinical Data (in the Linked Data Cloud) and the UMLS Definitions space.
The difference in the existence of SNOMED CT terms in YouTube health video tags between the "exactmatch" "false" and "true" options is significant. The absence of SNOMED CT terms rose from 17.7% to 58.1% when "exactmatch" was set to "off" and "on" respectively. To this extent, the average percentage of SNOMED CT terms existing in YouTube health video tags was 34.3% with "exactmatch" disabled, while it was only 7.1% when the "exactmatch" parameter was set to "on". On the other hand, the standard deviation of SNOMED CT terms existing in YouTube health video tags was 25.1% and 11.1% when "exactmatch" was set to "off" and "on" respectively. The standard deviation of SNOMED CT terms existing in YouTube health video tags was high enough in both cases ("exactmatch" parameter set to "off" and "on"), but the latter was lower due to the high number of YouTube health video tags that did not contain any SNOMED CT terms. These differences reveal that SNOMED CT terms do exist in YouTube health video tags, but their relative presence in the case of SNOMED CT searches would facilitate the correlation with resources in either of the aforementioned cases.

YouTube health videos that are not in English are probably going to contain non-English tags. At the moment, a similar analysis could be conducted in some other languages, since SNOMED CT is available in US/UK English, Spanish, Danish and Swedish, while translation to other languages is currently under development. However, content retrieval and content correlation is still feasible due to semantic web developments. For instance, currently running EU funded projects like MORMED (http://www.mormed.eu/) [47] could provide a solution for non-English tagged content to correlate it with related resources or translate it for end users. However, retrieving user-generated multilingual content is still, largely, an open issue.

However, there are many YouTube health video tags that are not matched with SNOMED CT terms (▶ Figures 6 and 8). Here, the correlation with other medical educational resources could be achieved through folksonomies that are also supported by the mEducator scheme/ontology. In this case, the public must adopt the same user-generated terminology to search or automatically retrieve the resources they want based on their search criteria.

Record keeping of user search behaviors, either by taxonomies or by folksonomies would therefore enhance the personalized educational experience and in the case of SNOMED CT searches would facilitate the correlation with resources in either of the aforementioned cases.

The difference in the existence of SNOMED CT terms in YouTube health video tags between the "exactmatch" "false" and "true" options is significant. The absence of SNOMED CT terms rose from 17.7% to 58.1% when "exactmatch" was set to "off" and "on" respectively. To this extent, the average percentage of SNOMED CT terms existing in YouTube health video tags was 34.3% with "exactmatch" disabled, while it was only 7.1% when the "exactmatch" parameter was set to "on". On the other hand, the standard deviation of SNOMED CT terms existing in YouTube health video tags was 25.1% and 11.1% when "exactmatch" was set to "off" and "on" respectively. The standard deviation of SNOMED CT terms existing in YouTube health video tags was high enough in both cases ("exactmatch" parameter set to "off" and "on"), but the latter was lower due to the high number of YouTube health video tags that did not contain any SNOMED CT terms. These differences reveal that SNOMED CT terms do exist in YouTube health video tags, but their relative per-
of health social media resources by taking into consideration the user type (e.g. healthcare professional or patient).

The proposed framework [42] makes it possible for YouTube health videos to be brought into the era of semantic web and linked data. They can be shared and “enriched” more easily by annotating SNOMED CT terms, facilitating their connection with the rest of resources existing in the Linked Open Data cloud. Moreover, videos that are semi-automatically described by the mEducator ontology and include SNOMED CT terms as tags (or as ‘subject in mEducator terminology) constitute semantically-enhanced descriptions of the medical educational materials. In the context of health education, it will be possible to search across different repositories for educational content (including YouTube health videos) and provide additional (and hopefully effective and therefore powerful) recommendations for a personalized learning experience through the provision of additional content/knowledge from the enriched terms of the resource in question.

On the other hand, user behaviors in social media cannot be recorded easily. The type of information that a user shares could be very heterogeneous [49] and the tagging of this information with SNOMED CT terms could also vary (Figures 4 and 7). Having said that, it is at least clear that, if such tagging exists, the framework we have outlined in this paper opens up new avenues for studying personalised retrieval of educational content and sharing through taxonomies, thesauri and linked data for social media platforms.

**Acknowledgment**

This work has been co-funded by the project “mEducator: Multi-type Content Repurposing and Sharing in Medical Education”, supported by the eContentplus program, Information Society and Media Directorate-General, European Commission (ECP 2008 EDU 418006). This work has received further support from the Norwegian Research with an Yggdrasil grant (211158/F11). This project is part of the Tromsø Telemedicine Laboratory co-funded by the Norwegian Research Council, project 174934.

**References**


