Linking Data Across Universities: An Integrated Video Lectures Dataset

Miriam Fernandez, Mathieu d'Aquin, Enrico Motta

Knowledge Media Institute, Open University
Walton Hall, Milton Keynes, MK7 6AA, UK
{m.fernandez, m.daquin, e.motta}@open.ac.uk

Abstract. This paper presents our work and experience interlinking educational information across universities through the use of Linked Data principles and technologies. More specifically this paper is focused on selecting, extracting, structuring and interlinking information of video lectures produced by 27 different educational institutions. For this purpose, selected information from several websites and YouTube channels have been scraped and structured according to well-known vocabularies, like FOAF[1], or the W3C Ontology for Media Resources[2]. To integrate this information, the extracted videos have been categorized under a common classification space, the taxonomy defined by the Open Directory Project[3]. An evaluation of this categorization process has been conducted obtaining a 98% degree of coverage and 89% degree of correctness. As a result of this process a new Linked Data dataset has been released containing more than 14,000 video lectures from 27 different institutions and categorized under a common classification scheme.

Keywords: Linked Data, Education, Integration

1 Introduction

Different educational institutions, and even different departments within the same institution, produce yearly large amounts of educational material (videos, slides, documents, etc.). However, when students and educational practitioners have to perform learning and investigation tasks, they generally: (i) not find the best available resources for the topic they aim to investigate or, (ii) spend large amounts of time browsing the websites of different institutions in order to collect and extract the key information about the topic. In this context, we believe that integrating the large amount of educational material produced by different institutions is a key requirement towards educational data sharing and exploitation. The fact that different institutions publish and describe their educational content using different formats, tags, categories and structure, makes this integration process a difficult and challenging problem.

[1] http://xmlns.com/foaf/spec/
[2] http://www.w3.org/TR/mediaont-10/
[3] http://www.dmoz.org/

The emergence of Linked Data (LD)[4] brings to this scenario a new dimension of possibilities under which educational material can be organized, integrated, archived and retrieved. LD refers to a set of principles to put raw data on the Web and making them Web addressable and linkable, so that they can be easily accessed, discovered, connected and reused. The number of universities, research organizations, publishers and funding agencies contributing to the LD cloud is constantly increasing. Universities such as The Open University[5], Southampton[6], Sheffield's Computer Science Department[7], or the University of Münster[8], among others, are embracing the LD principles and releasing educational resources as part of the LD cloud.

In this paper, we aim to expose our experience extracting, structuring and integrating video lectures material from 27 different educational institutions by exploiting LD principles. Since standardized practices for publishing and integrating educational LD across institutions are not yet in place, we expect that this work can contribute to reflect on the evolution of such practices.

The rest of the paper is structured as follows: Section 2 provides an overview of related work. Section 3 describes the processes of selecting, extracting and structuring video lectures from various information sources according to LD principles. Section 4 explains the data integration process, focused on the creation of a common searchable/browsable space for educational material. Section 5 shows the conducted evaluation for the data integration process. Conclusions are shown in Section 6.

2 Related Work

We are currently witnessing a substantial increase in universities adopting the Linked Data initiative. One of the currently strongest activities towards LD production and consumption within the context of education has been carried out by the Open University (OU), in the context of the Lucero project[9] (Linking University Content for Education and Research Online). This project performs OU data extraction, transformation, maintenance and exploitation [14]. At the time of writing, several datasets about publications, podcasts and course descriptions, among others, have been released and are accessible in an open way through online access at `http://data.open.ac.uk`. In addition, several applications[10] have been developed to show the potential of the OU linked data exposure. Although all these applications show several significant advantages of exploiting LD in the educational context, their coverage is currently limited to the OU. There is still no significant integration of educational data across universities that can be exploited by these applications.

Other examples of efforts towards the production and consumption of LD in the educational context are:

[4] `http://linkeddata.org/`
[5] `http://data.open.ac.uk`
[6] `http://data.southampton.ac.uk`
[7] `http://data.dcs.shef.ac.uk/`
[8] `http://lodum.de/about`
[9] `http://lucero-project.info/`
[10] `http://data.open.ac.uk/applications/`

- The University of Sheffield's Department of Computer Science[7], which provides a LD service describing research groups, staff and publications, all semantically linked together [10].
- The University of Southampton, which has recently announced the release of their LD portal[6]. At the time of writing 26 different datasets including information about university buildings, educational videos, or university bus routes have been released.
- The University of Manchester's library catalogue. Its records can now be accessed in RDF format[11].
- The University of Edinburgh, where the university's buildings information is now generated as LD[12].
- The University of Münster, which recently announced LODUM[8], a project with the aim to release the university's research information as LD. This includes information related to people, projects, publications, prizes and patents.

Additionally to the initiatives of publishing educational content as LD, it is important to highlight some of the current works towards integrating library catalogs on a global scale. Some of these works are discussed in [4]. Examples include the American Library of Congress[13], the German National Library of Economics [8], and LIBRIS,[14] the Swedish National Union Catalogue, which publish their subject heading taxonomies as LD. Similarly, the OpenLibrary[15], a collaborative effort to create "one Web page for every book ever published" has published its catalogue in RDF. Scholarly articles from journals and conferences are also well represented through community publishing efforts such as DBLP as LD[16], RKBexplorer[17], and the Semantic Web Dogfood Server [7].

We believe that the increase involvement of the library community in LD[18] will soon enhance the exchange and consumption of educational material, facilitating its search, exploration and comparison across institutions.

3 Generating RDF

Among the different types of educational resources (textual documents, slides, videos, etc.) this paper is focused on: (a) generating and (b) interlinking RDF descriptions from video lectures. In this section we will explain the RDF generation process including: (i) the processes of information selection and extraction, (ii) the process of vocabulary selection and, (iii) the process of information structuring according to the selected vocabularies.

---

[11] http://prism.talis.com/manchester-ac/

[12] http://ldfocus.blogs.edina.ac.uk/2011/03/03/

[13] http://id.loc.gov/authorities/about.html

[14] http://blog.libris.kb.se/semweb/?p=7

[15] http://openlibrary.org/

[16] http://dblp.l3s.de

[17] http://www.rkbexplorer.com/data/

[18] http://www.w3.org/2005/Incubator/lld/

3.1 Selecting and extracting educational information from various sources

When selecting information sources, we have considered two of the currently most popular video lecture containers: YouTube[19] university channels and the videolectures.net website.

YouTube is a video-sharing website on which users can upload, share and view videos. In the context of education, YouTube has been used by several institutions to make their video lectures publicly available on the Web via YouTube channels. For the purpose of this work we have selected the YouTube channels of 25 different universities and research institutions including: Standford, Yale, Harvard, Oxford or Google Talks, among others. The complete list of YouTube channels used for this paper can be found in http://smartproducts1.kmi.open.ac.uk/web-linkeduniversities/index.htm. Video lectures information from YouTube channels is accessed and extracted via the YouTube data API[20]. Among the information that can be accessed through this API we have focused on: (i) video upload feeds and (ii) playlist feeds. Video upload feeds refer to all the videos uploaded by the same university channel. Video playlist feeds are collections of videos available via a particular university channel that may have been uploaded by the university or by other users/institutions. Figure 1 represents a summary of the common properties associated to video uploads and playlist feeds. When querying the YouTube data API, each element (video or playlist) is returned and represented as an entry point with several associated properties. The complete list of properties can be found in the YouTube data API documentation[21].

```
<entry gd:etag='W/"DkADSH47eCp7ImA9WhZWFEg."'>
    <id>tag:youtube.com,2008:video:zZCaHSW88Ts</id>
    <published>2011-02-18T11:41:08.000Z</published>
    <updated>2011-05-15T10:19:39.000Z</updated>
    <category scheme='http://gdata.youtube.com/schemas/2007/categories.cat'
              term='Education' label='Education'/>
    <category scheme='http://gdata.youtube.com/schemas/2007/keywords.cat'
              term='Dr Barry Cooper'/>
    <title>Intro to Professional Practice (Children & Families)</title>
    <author>
      <name>TheOpenUniversity</name>
      <uri>http://gdata.youtube.com/feeds/api/users/theopenuniversity</uri>
    </author>
   <media:description type='plain'>Free learning from The Open University
      http://www.open.ac.uk/openlearn/
      An introduction by Barry Cooper detailing the Postgraduate […]
   </media:description>
   <media:keywords>Dr Barry Cooper, postgraduate qualifications, social work,
      children and families, childcare worker, childcare practitioner, healthcare
      practitioner, flexible pace of study, flexible award, online tutor panel,
      online classroom, ou_k14, ou_e70, open university
   </media:keywords>
   <media:thumbnail url='http://i.ytimg.com/vi/zZCaHSW88Ts/default.jpg'
      height='90' width='120' time='00:03:19.500' yt:name='default'/>
   <yt:duration seconds='399'/>
    <content type='application/x-shockwave-flash'
src='http://www.youtube.com/v/zZCaHSW88Ts?f=user_uploads&app=youtube_gdata'/>
```

[19] http://www.youtube.com/

[20] http://code.google.com/apis/youtube/getting_started.html#data_api

[21] http://code.google.com/apis/youtube/2.0/developers_guide_protocol_understanding_video_feeds.html

```
      <gd:feedLink
href='http://gdata.youtube.com/feeds/api/videos/zZCaHSW88Ts/comments'
countHint='2'/>
```

Fig. 1. Summarized example of a YouTube upload video feed.

Among these properties we have selected for the purpose of this work: the video ID, the publication date, the date at which it was updated, its duration, its title, its description, its authors, the link to the video content, the links to the associated thumbnails and the list of categories and keywords that describe it. Additionally, for videos extracted from a playlist, the playlist identifier is also extracted. When selecting the set of entities and properties on which to apply the LD principles, the goal we had in mind was to have an essential definition of the video lectures which would be reasonably independent from the original source.

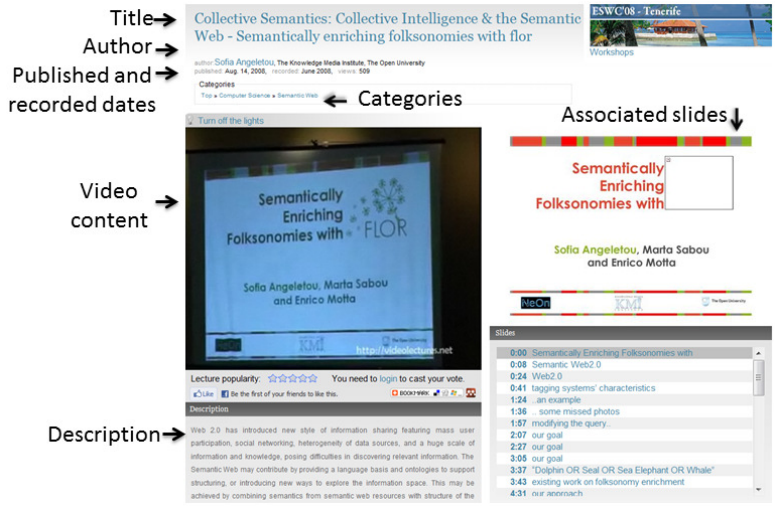

Fig. 2. Screenshot of `videolectures.net` describing one of its videos.

`Videolectures.net` is a website for academic talks launched in 2007. It offers to the scientific, research, business and general public a large collection of video lectures that are enriched with slides. While the vast majority of talks belong to the subject of Computer Science, it also contains videos about Astronomy, Medicine or Philosophy among others. `Videolectures.net` does not provide any API for accessing its data so, for the purpose of this work, a tailor-made HTML scraper has been developed with the aim of extracting a selected set of information. In this case, the properties extracted for each particular video are: the video ID, the publication date, the recording date, its duration, its title, its description, its list of corresponding authors (including authors' names and affiliations), the link to the video content, the link to the associated slides when available, the links to the associated thumbnails and the list of categories used to describe it. A screenshot of the `videolectures.net` website where these properties are displayed for a particular video can be seen in Figure 2.

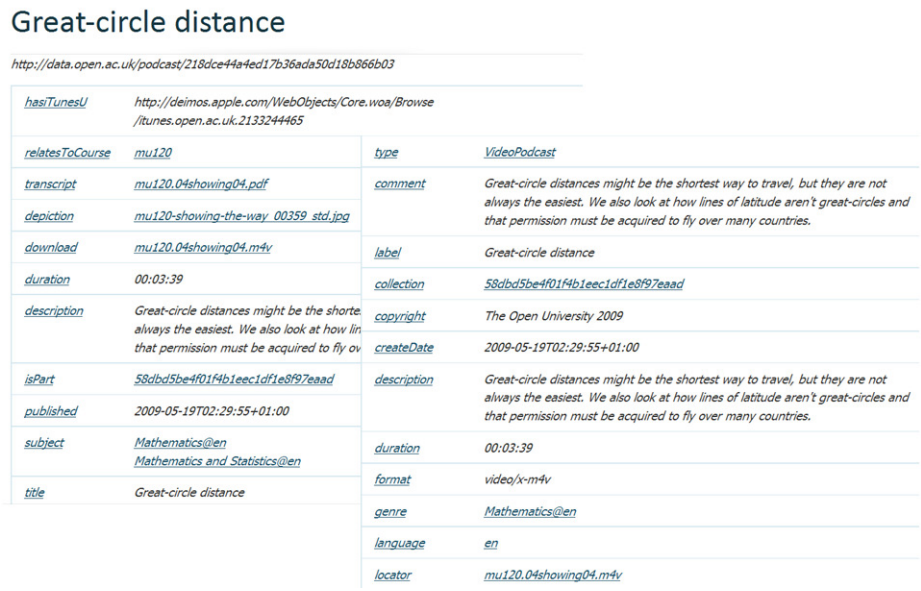

Fig. 3. Example of an OU Podcasts.

Additionally to the two previously mentioned information sources, we have also added to the video lectures linking process an already LD structured video lectures dataset, the OU Podcasts[22]. OU Podcasts is a collection of Audio and Video material related to education and research at the Open University. This video and audio material has been remodeled using LD principles and is currently defined using a variety of ontologies[23]. Figure 3 shows an example of the information associated to the OU Podcast "*Great-circle distance*", including: the video ID, the publication date, the creation date, its duration, its title, its description, its list of corresponding publishers, the link to the video content, the link to the video transcript, the links to the associated thumbnails and the list of categories used to describe it. Note that this information source is already structured according to the LD principles, so the process of RDF generation is not applied to it.

3.2 Reusing vocabularies to describe educational data

As described by Heath and Bizer [4], RDF provides a generic, abstract data model for describing resources using subject, predicate, object triples. However, it does not provide any domain-specific terms for describing classes of things in the world and how they relate to each other. This function is served by taxonomies, vocabularies and ontologies expressed in *SKOS*[24], *RDFS*[25] and *OWL*[26]. In this context, and according to

[22] http://podcast.open.ac.uk/
[23] http://data.open.ac.uk/datasets/
[24] http://www.w3.org/TR/skos-reference/
[25] http://www.w3.org/TR/rdf-schema/

linked data practices [4], "*if suitable terms can be found in existing vocabularies, these should be reused to describe data wherever possible, rather than reinvented*". Reuse of existing terms is highly desirable as it maximises the probability that data can be consumed by applications that may be tuned to well-known vocabularies, without requiring further pre-processing of the data or modification of the application.

Following these guidelines, several vocabularies have been selected to describe the information extracted from the YouTube university channels and `videolectures.net`. The following list represents the chosen vocabularies:

- Dublin Core[27]: is a widely deployed vocabulary for representing provenance, particularly by the use of the `dcterms:creator` and `dcterms:published` predicates. It also provides descriptive predicates such as `dcterms:title`, `dcterms:description` or `dcterms:subject`.
- FOAF[28], the Friend Of A Friend vocabulary: defines terms for describing persons, their activities and their relations to other people and objects. The class `foaf:Person` and the predicates `foaf:name` and `foaf:homepage` are examples of reused elements to describe authors of video lectures.
- The W3C Ontology for Media Resources[29]: is both a core vocabulary (a set of properties describing media resources) and mappings to a set of metadata formats currently describing media resources published on the Web. Examples of reused elements include: `ma:publisher`, `ma:createData` or `ma:description`.
- The Media Vocabulary[30]: defines a minimal scheme for media content. Classes like `media:Recording`, to instanciate video lectures, as well as predicates like `media:download` or `media:depiction` have been reused to describe the video content and its associated thumbnails.
- The Nice Tag Ontology[31]: describes tags as generally as possible. Tags associated to videos from `videolectures.net` and YouTube channels have been modeled using the `nt:isRelatedTo` predicate.

  Note that `dcterms`, `foaf`, `ma`, `media`, and `nt` are the corresponding prefixes for the vocabularies' associated namespaces.

3.3 Structuring information according to the previously selected vocabularies

When structuring information in RDF, one of the main discussions raised in the Linked Data community are the best practices to generate Unified Resource Identifiers (URIs). URIs should be representative as names for things (real-world entities or abstract concepts) and should be designed to be simple. The W3C Interest Group has generated "*Cool URIs for the Semantic Web*[32]", a guideline about good

---

[26] `http://www.w3.org/TR/owl-features/`
[27] `http://dublincore.org/documents/dcmi-terms/`
[28] `http://xmlns.com/foaf/spec/`
[29] `http://www.w3.org/TR/mediaont-10/`
[30] `http://payswarm.com/vocabs/media`
[31] `http://ns.inria.fr/nicetag/2010/09/09/voc.html`
[32] `http://www.w3.org/TR/cooluris/`

practices for URI generation. This guideline has been followed, when applicable, during the development of the present work.

The base URI, common to all elements of the dataset, is `http://linkeduniversities.org`.

Individuals of the class `media:Recording` have been generated to represent video lectures objects. The URIs for this type of objects are identified by five main elements: the base URI, the type of educational material (video, audio, text, etc.), the educational institution producing this material (Carnegie Mellon University, Open University, etc.), the storage/communication source used by the institution (YouTube, Podcasts, etc.) and the primary key, or video identifier within the storage source. The properties of each video are structured according to the set of vocabularies described in Section 3.2. An example of how to structure a video lecture, including its assigned URI and its list of associated properties is described in Table 1.

Table 1. Structure of a video lecture.

Example of the structure associated to a video lecture	
`http://linkeduniversities.org/video/CarnegieMellonU/youtube/B135229F3706D215/9949817F2FB77F0C`	
`rdf:type`	`media:Recording`
`media:download`	`http://www.youtube.com/watch?v=TOTuStPIeFc&feature=youtube_gdata_player`
`dcterms:title`	`CMU Football Engineering Summer 2008 Video`
`rdfs:label`	`CMU Football Engineering Summer 2008 Video`
`dcterms:description`	`Football […]Summer 2008 Video`
`foaf:thumbnail`	`http://i.ytimg.com/vi/TOTuStPIeFc/3.jpg`
`foaf:thumbnail`	`http://i.ytimg.com/vi/TOTuStPIeFc/1.jpg`
`foaf:thumbnail`	`http://i.ytimg.com/vi/TOTuStPIeFc/2.jpg`
`foaf:thumbnail`	`http://i.ytimg.com/vi/TOTuStPIeFc/0.jpg`
`media:duration`	`155`
`dcterms:isPart`	`http://linkeduniversities.org/video/CarnegieMellonU/youtube/playlist/B135229F3706D215`
`ma:publisher`	`http://linkeduniversities.org/video/CarnegieMellonU/youtube/user/footballtracking`
`dcterms:published`	`2011-06-03T23:23:53.262Z`
`nt:isRelatedTo`	`http://linkeduniversities.org/video/CarnegieMellonU/tag/cmu`
`nt:isRelatedTo`	`http://linkeduniversities.org/video/CarnegieMellonU/tag/sports`
`nt:isRelatedTo`	`http://linkeduniversities.org/video/CarnegieMellonU/tag/football`
`nt:isRelatedTo`	`http://linkeduniversities.org/video/CarnegieMellonU/tag/engineering`
`dcterms:subject`	`http://dmoz.org/Society/People`
`dcterms:subject`	`http://linkeduniversities.org/video/CarnegieMellonU/dmoz/Society/People`
`dcterms:subject`	`http://dmoz.org/Sports/Football/Rugby_Union`
`dcterms:subject`	`http://linkeduniversities.org/video/CarnegieMellonU/dmoz/Sports/Football/Rugby_Union`

Looking at the table, it is important to highlight certain design decisions:

- The content of the title has been duplicated within the properties `dcterms:title` and `rdfs:label`. Since most current Semantic Web applications exploit the `rdfs:label` predicate as the main descriptive property of the object.
- The association of a video with a playlist is reflected using the `dcterms:isPart` predicate.
- The set of tags and categories describing the video are associated using the `nt:isRelatedTo` predicate. In addition, these tags are mapped to the base URI, i.e., the `linkeduniversities.org` domain.
- The use of the property `dcterms:subject` is extensively described in section 4. Basically it reflects the categorization of the video lecture within the unified searchable/browsable space. As we can see, the value of this property is also

duplicated to maintain the URI of its original source, but also to add it as part of our base URI.

Individuals of the class `foaf:Person` have been generated to represent authors. The same URI elements used to represent individuals of the class `media:Recording` are used to represent this class (with the exception of the primary key, or author identifier, which is generated taking into account the author's name). An example of the structure of a video lecture, including its assigned URI and its list of associated properties, is described in Table 2.

Table 2. Structure of a Person/author.

Example of a structure associated to an author		
`http://linkeduniversities.org/video/videolectures/michel_dumontier`		
`rdf:type`	`foaf:Person`	
`foaf:name`	`Michel Dumontier`	
`foaf:homepage`	`http://videolectures.net/michel_dumontier`	
`vcard:organization-name`	`Carleton University`	
`<http://linkeduniversities.org/video/videolectures/6593>`	`dcterms:contributor`	`<http://linkeduniversities.org/video/videolectures/michel_dumontier>`

As we can see, every identified author has at least an associated name, homepage, and organization. The last row of Table 2 describes how a video lecture is associated to its corresponding authors by the property `dcterms:contributor`.

4 Integrating Educational Information

YouTube channels, `videlectures.net` and OU Podcasts, use different classification schemes and systems. To unify the search and exploration tasks of all these educational material it is necessary to integrate the extracted videos under a common searchable/browsable space, in this case under a common topic hierarchy. To address this problem, three key issues should be tackled:

i) Select the most appropriate categorization scheme under which these video materials should be classified.
ii) Analyze the classifications assigned by each particular information source (YouTube channels, `videolectures.net`, OUPodcast) to determine how this information can be mapped to the common categorization scheme.
iii) Propose a categorization approach to classify every video lecture to the common categorization scheme.

4.1 Selecting a common categorization scheme

When selecting the potential categorization schemes, three main requirements have been considered: (i) to be general, i.e., aiming to cover all subjects in "the universe of information", (ii) to be fully public and, (ii) to be available in RDF. Following these requirements, four potential categorization schemes have been selected:

- DMOZ[33], the Open Directory Project (ODP) topic hierarchy: the ODP is the largest, most comprehensive human-edited directory of the Web. It is constructed and maintained by a vast, global community of volunteer editors and is 100% free. RDF dumps of this topic hierarchy and its content are available for download[34]. ODP data powers the core directory services[35] for many of the Web's largest search engines and portals, including Netscape Search, AOL Search and Google. In addition, more than seventy-five languages are currently represented in this topic hierarchy and, at the time of writing, it claims to have over 1,007,233 categories.
- DBpedia categories[36]: The DBpedia project extracts various kinds of structured information from Wikipedia editions in 97 languages and combines this information into a large cross-domain knowledge base. The DBpedia knowledge base currently claims to describe more than 3.5 million things. It provides three different classification schemes for things: (i) Wikipedia categories, (ii) the YAGO Classification, derived from the Wikipedia category system using WordNet and, (iii) WordNet Synset Links, generated by manually relating Wikipedia infobox templates and Word Net synsets. Although DBPedia currently constitutes one of the main cores of the Web of Data, erroneous Wikipedia categories also cause the derivation of false facts [13].
- Library of Congress Subject Headings (LCSH)[37]: the LCSH comprises a thesaurus of subject headings, maintained by the United States Library of Congress, for use in bibliographic records.
- The International Press Telecommunications Council (IPTC) News Codes[38]: The IPTC creates and maintains sets of concepts to be assigned as metadata values to news objects like text, photographs, graphics, audio and video files and streams. Among this metadata they provide several taxonomies to describe the content of news items.

Although LCSH and IPTC are high quality classification schemes, developed and maintained by the library and journalism communities, to the best of our knowledge, they only support the English language. Considering that educational resources may be accessed and described in different languages we have opted for selecting a multilingual classification scheme, i.e., either DBPedia or ODP. DBPedia is currently considered the core of the LD cloud and there is a high level of activity towards its development. On the other hand, the ODP classification scheme has been longer established and there is a wide range of sites that are successfully exploiting it. Although both classification schemes seemed suitable for the task at a hand we have selected ODP because of its maturity and the availability of tools to exploit it in the context of classification tasks.

---

[33] http://www.dmoz.org/docs/en/about.html

[34] http://www.dmoz.org/rdf.html

[35] http://en.wikipedia.org/wiki/Open_Directory_Project

[36] http://dbpedia.org/About

[37] http://id.loc.gov/authorities/about.html

[38] http://www.iptc.org/site/NewsCodes/NewsCodes_Retrieval_in_Different_Formats

4.2 Analyzing the categorization schemes of each information source

Mapping video lectures categorized under different schemes, to a common searchable/browsable space of topics is a challenging problem. YouTube videos, for example, are categorized by YouTube categories, as well as by user's and developer's tags[39]:

- Each video can be associated with one predefined YouTube category, such as Comedy, News or Sports. A video's category is identified by the `<media:category>` and the `<category>` tags for which the value of the scheme attribute is `http://gdata.youtube.com/schemas/2007/categories.cat`.
- Each video can be associated with an arbitrary number of keywords, which are also known as tags. A video's tags are identified using the `<media:keywords>` tag in API requests and responses. Keyword tags are also identified by `<category>` tags for which the value of the scheme attribute is `http://gdata.youtube.com/schemas/2007/keywords.cat`.
- Each video can also be associated with an arbitrary number of developer tags. Video developer tags are identified in `<media:category>` and `<category>` tags for which the value of the scheme attribute is `http://gdata.youtube.com/schemas/2007/developertags.cat`.

`Videolectures.net` uses its own categorization scheme that contains 23 main root elements including: Architecture, Arts, Biology, Business, Chemistry or Computer Science among others. The categorization scheme is available through their website.
OU Podcasts are classified under three different categorization systems: OU specific subject headings, iTunes categories[40] and iTunes U categories[41].

In addition to the categorization information used by each individual source, properties such as title and description, available in all three sources, can be used as additional information to generate the corresponding mappings to the ODP categorization scheme.

4.3 The categorization approach

As mentioned in the previous section, we have three main different types of information to extract the most accurate ODP categories for each particular video lecture: (i) its source-dependent categories, (ii) its associated tags, and (iii) the text extracted from its title and description.

When mirroring this problem to current state of the art approaches, we found several interesting works that have attempt to: (i) generate mappings between category hierarchies [6, 9, 11], (ii) generate mappings from tag information spaces to category hierarchies [1, 12] and (iii) classify textual documents under category hierarchies [2, 5]. While the previously mentioned works are focused on using only one type of information, our purpose is to exploit simultaneously, tags, source-

[39] `http://code.google.com/apis/youtube/1.0/reference.html`

[40] `http://itunes.apple.com/us/genre`

[41] `http://deimos.apple.com/rsrc/doc/iTunesUAdministrationGuide/iTunesUintheiTunesStore/chapter_13_section_3.html`

dependent categories and associated textual descriptions, to extract the most accurate ODP categorization for each video lecture.

For this purpose, among the available systems and techniques for information classification we have decided to reuse TextWise[42] software and services. The categorization service provided by Textwise identifies the main topic categories for an input text or URI using the ODP 2010 categorization[43]. According to Textwise, the categorization of content is performed by analyzing the dimensions and weights of the content's Semantic Signature[44], where a Semantic Signature represents the concepts in a text through a weighted vector entry of typically several thousand semantic dimensions.

Our proposal is therefore to exploit the three different types of information to generate the input text that the TextWise service needs to perform the categorization process. When generating this text, it is important to keep in mind that tags, and domain-dependent categories, are the key video classification properties. Although properties like title and description provide a more extended and coherent characterization of the video content, we have empirically observed that they tend to become ambiguous information elements when they are used to extract the key video topics. Based on these facts, the categorization approach is formulated as follows:

Let $S = \{s_1, s_2 ..., s_n\}$ be the list of different educational institutions, or information sources. Let $V_i = \{v_{i_1}, v_{i_2}, ..., v_{i_m}\}$ be the complete list of video lectures extracted from the educational institution S_i where each video lecture v_{ij} has associated: a set of tags, $T_{v_{ij}}$, a set of categories, $C_{v_{ij}}$, a title $Title_{v_{ij}}$ and a description, $Desc_{v_{ij}}$. Following these definitions, the pseudo-code of the proposed categorization approach is described in Table 3.

Table 3. Pseudo-code of the proposed video categorization approach.

Categorization approach
For $i = 1$ to n
select information source s_i
For $j = 1$ to m
select the video lecture v_{ij}
extract $T_{v_{ij}}$, extract $C_{v_{ij}}$, extract $Title_{v_{ij}}$, extract $Desc_{v_{ij}}$
$HTMLD_{v_{ij}}$ = HTMLDoc ($T_{v_{ij}}$, $C_{v_{ij}}$, $Title_{v_{ij}}$, $Desc_{v_{ij}}$)
TextWise ($D_{v_{ij}}$, 2)

Basically, for each video lecture, v_{ij}, the approach extracts its set of tags, $T_{v_{ij}}$ and source-dependent categories, $C_{v_{ij}}$, its title, $Title_{v_{ij}}$ and its description, $Desc_{v_{ij}}$. With

[42] http://textwise.com/

[43] http://textwise.com/api_docs/labels/2010-ODP-Topic-Category-Mapping.txt

[44] http://textwise.com/technology-0

this information it generates an HTML document, $HTMLD_{v_{ij}}$, from which the Textwise service extracts up to two ODP classifications of v_{ij} . The generated HTML document contains the following structure:

```
<html>
      <head>
         <title> Title_{v_ij} </title>
         <meta name="keywords" content=" T_{v_ij} , C_{v_ij} ">
      </head>
       <body> <p> Desc_{v_ij} </p></body>
</html>
```

As we can see, the HTML page title and body correspond to the title and description of the video lecture respectively. Each tag and source-dependent category is added as a meta keyword element of the HTML document. Following this approach, tags and categories are emphasized within the HTML page. This emphasis is expected to produce a positive impact when using the TextWise categorization service, because it decreases the relevance of more ambiguous properties, such as title and description. To visualize the HTML page generation process, let's consider the video lecture presented in Figure 2:

Table 4. HTML page associated to the v_{ij} video lecture presented in Figure 2.

v_{ij} **Information**	**HTMLDoc** ($T_{v_{ij}}$, $C_{v_{ij}}$, $Title_{v_{ij}}$, $Desc_{v_{ij}}$)
s = videolectures.net, $T_{v_{ij}} = \varnothing$, there are no tags associated to this video. $C_{v_{ij}}$ = {Computer Science, Semantic Web} $Title_{v_{ij}}$ = "Collective Intelligence [...] enriching folksonomies with Flor" $Desc_{v_{ij}}$ = "Web 2.0 has introduced […] with help of the Semantic Web"	<html> <head> <title> Collective Intelligence [...] enriching folksonomies with Flor</title> <meta name = keywords content= "Computer Science", "Semantic Web"> </head> <body> Web 2.0 has introduced […] with help of the Semantic Web </body> </html>

For this generated HTML page, the TextWise service produces as response:

- Reference/Knowledge_Management (id=495), w=0.71
- Reference/Libraries/Library_and_Information_Science (id=497), w= 0.53

The TextWise service provides not only the ODP categorization label and its corresponding id, but also a weight which reflects the confidence of the service in the proposed classification. As we can see in the example, for the two proposed answers, the first one may be considered correct by most evaluators, but the correctness of the second one is more arguable. Following some empirical tests, we have decided to set up a threshold of 0.5 to accept the proposed categorization as valid.

5 Evaluating the Categorization Problem

Achieving a high degree of correctly categorized video lectures is a key requirement towards the generation of a high quality interlinked dataset of educational material. In this section we describe the evaluation conducted to assess the quality of the video lectures data integration process. The evaluation pursues three key goals: (i) measuring the *coverage* of the categorization process; i.e., how many video lectures have been assigned at least to one ODP category; (ii) measuring the *correctness* of the categorization process; i.e., which percentage of the assigned categories are considered correct and; (iii) measuring the *specialization* of the categorization process; i.e., are the assigned ODP classifications the most specialized ones or is it possible to find a more refined ODP category to describe the same video content?

Coverage: To evaluate the coverage of the categorization process we have analyzed the number and percentage of video lectures for which no ODP categories were assigned. From the total of 14,311 videos lectures extracted from the 27 different educational institutions, a total of 14,037 (**98%**) were successfully categorized using the approach presented in section 4. Additionally 55% of the videos were assigned a second ODP category. Over the remaining 274 video lectures we have performed an empirical analysis to find out the different reasons for their lack of classification. The most significant one is the use of different languages to represent the properties of the video lecture. As an illustration, consider the video lecture defined by the URI `http://videolectures.net/inovativna_slovenija2010_golobic_kis/`. This video lecture has its title described in Slovenian language "*Kdaj inovativna Slovenija?*" and its classification described in English language " *Top » Technology » Innovation*". Other reasons include the simultaneous lack of video description, tags and categories.

Correctness and Specialization: To evaluate the correctness and specialization of the categorization approach, we have engaged 3 different evaluators in the campaign. Each of them has evaluated the categories assigned to 675 video lectures (25 randomly selected video lectures for each of the 27 information sources). Considering that 252 of the 675 selected videos where assigned two different ODP categories, the total number of video categorizations judged by each evaluator was 927. Note that, when randomly selecting the 25 video lectures for each information source, we have previously discarded those ones for which no ODP category was assigned. To judge the correctness and specialization of each video categorization, the evaluators were provided with: (i) all the available video information, (ii) its assigned categories and, (iii) the complete ODP hierarchical classification. Each video categorization was judged using a value from 0 to 2, where each number implies: (0) the classification is incorrect, (1), the classification is correct but a more specialized category could have been assigned, and (2), the classification is correct and the evaluator has not found any more specialized category in the ODP.

For each video categorization, given the three user's evaluations, the categorization was considered correct if at least two evaluators were rating it with values higher than 0, and it was considered specific, if at least two evaluators were rating it at level 2 and the remaining evaluation was not 0. There was a substantial agreement among users. Fleiss' kappa statistic [3] measuring user's agreement was k=0.71 (a value k=1 means complete agreement). Once the agreement results were established we found that over

the 927 video categorizations 831 (**89%**) were considered correctly classified and 475 (**51%**) were considered specialized.

6 Conclusions and Discussion

This paper presents our work and experience interlinking educational information across universities through the use of LD principles and technologies. More specifically, this paper is focused on selecting, extracting, structuring and interlinking information of video lectures produced by 27 different educational institutions. For this purpose, selected information from several websites and YouTube channels have been scraped and structured according to several existing vocabularies. To integrate this information, the extracted videos have been categorized under a common searchable/browsable space, the taxonomy defined by the Open Directory Project. As a result of this process a new LD educational dataset has been released containing more than 14,000 video lectures from 27 different institutions. These videos have been categorized under a total of 569 different ODP categories. Among the most popular ones we can highlight: Science/Math, Science/Physics and Computers/Artificial_Intelligence.

High levels of coverage (**98%**) and accuracy (**89**%) have been achieved during the integration process. The complete dataset is available under `http://smartproducts1.kmi.open.ac.uk/web-linkeduniversities/index.htm`. Here, the reader can find a complete description of the dataset, including the RDF dumps for each institution, a SPARQL endpoint, and several SPARQL query examples.

Regarding our lessons learned we propose five main ingredients for a successful production and integration of educational content through the use of LD principles.

1. LD principles are simple. However, identifying available data, obtaining access to it and remodeling it is a high-cost process. Making educational institutions understand that it is worth doing it is a critical factor.
2. There is a need to agree on a set of collective vocabularies to model and structure educational information. Following a bottom up approach, those vocabularies should initially focus on modeling common elements across educational institutions like: educational material, courses or research staff.
3. There is a need to agree on common searchable/browsable spaces under which educational information can be explored and retrieve. Establishing a common space of topics under which educational material and courses can be classified could be a good starting point.
4. Establishing qualitative criteria and quantitative evaluation measures to assess these criteria are key requirements for the development of high quality educational LD.
5. Educational LD is not about a killer application, but is about multiple small things that are made easier (integrating information across university departments, enriching information with external sources, sharing educational content across institutions, etc.) Proposals should emerge about how to integrate the benefits of LD within the universities' practices and workflows.

To be truly effective, many of these improvements should be the results of community-wide efforts rather than advances at the level of individual research groups. We believe that this is an important time for the development of the education's Web of LD. Collaborative efforts to produce and integrate educational information are needed to achieve the envisioned data space from which, information across educational institutions will be search, explored, compared and retrieved in an homogenous way.

References

1. Cantador, I., Konstas, I., Jose, J. M. (2011) *Categorising Social Tags to Improve Folksonomy-based Recommendations*. Journal of Web Semantics 9(1), pp. 1-15.
2. Cesa-Bianchi, N., Conconi, A., Gentile, C. (2004) *Regret Bounds for Hierarchical Classification with Linear-Threshold Functions*. Lecture Notes in Computer Science, Vol. 3120, pp. 93-108.
3. Fleiss, J. L., Cohen, J. (1973) *The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability*. Educational and Psychological Measurement, Vol. 33, pp. 613-619.
4. Heath, T., Bizer, C. (2011) *Linked Data: Evolving the Web into a Global Data Space (1st edition)*. Synthesis Lectures on the Semantic Web: Theory and Technology, 1:1, pp. 1-136. Morgan & Claypool.
5. Hofmann, T., Cai, L., Ciaramita, M. (2003) *Learning with Taxonomies: Classifying Documents and Words*. Syntax, Semantics and Statistics Workshop (NIPS).
6. Kalfoglou, Y., Schorlemmer, M. (2003). *Ontology mapping: the state of the art*. The Knowledge Engineering Review, 18 (1), pp. 1-31.
7. Möller, K., Heath, T., Handschuh, S., Domingue, J. (2007) *Recipes for semantic web dog food - the ESWC and ISWC metadata projects*. 6th International Semantic Web Conference and 2nd Asian Semantic Web Conference, Busan, Korea.
8. Neubert, J. (2009) *Bringing the "thesaurus for economics" on to the web of linked data*. Workshop on Linked Data on the Web, WWW2009.
9. Noy, N. (2004) *Semantic integration: a survey of ontology-based approaches*. *SIGMOD Rec.* 33, 4, pp. 65-70.
10. Rowe, M. (2010) *Data.dcs: Converting legacy data into linked data*. Workshop on Linked Data on the Web, WWW2010.
11. Shvaiko, P., Euzenat, J. (2005) *A Survey of Schema-Based Matching Approaches*. Journal on Data Semantics IV. Lecture Notes in Computer Science, 2005, Vol. 3730, pp. 146-171.
12. Specia, L., Motta, E. (2007) *Integrating Folksonomies with the Semantic Web*. The Semantic Web: Research and Applications. Lecture Notes in Computer Science, 2007, Vol. 4519, pp. 624-639.
13. Suchanek, F.M., Kasneci, G., Weikum, G. (2007) *Yago: a core of semantic knowledge*. 16th international conference on World Wide Web, WWW2007. ACM, New York, NY, USA, pp. 697-706.
14. Zablith, F., Fernandez, M., Rowe, M. (2011) *The OU Linked Open Data: Production and Consumption*. eLearning Approaches for the Linked Data Age. Extended Semantic Web Conference 2011.