Using Linked Open Data to Enrich a Corporate Memory of Universities

Cristal Karina Galindo Durán¹, R. Carolina Medina-Ramírez² and Mihaela Joganaru-Mathieu³

¹, ² Departamento de Ingeniería Eléctrica, Universidad Autónoma Metropolitana, Unidad Iztapalapa, D.F. México.
³ Institut H. Fayol, École Nationale Supérieure des Mines de Saint Étienne, Saint Étienne, France.

Abstract - In this work we describe a proposal for managing educational resources produced by three high level educational institutions. Our approach takes into account not only the meta informations inside educational resources but also the meta informations extracted from linked open data in the web associated to massive open online courses (MOOCs). We visualize educational resources as a semantic corporate memory where semantic technologies and Linked Open Data approaches can improve the information integration. Text mining takes an important part in our approach for generating links between concepts and detecting some important associations.

Keywords: Link Open Data, corporate memory, MOOC, text mining, Semantic web

1 Introduction

A corporate memory (CM) is an explicit and persistent representation of knowledge inside an organization, which can be embodied in persons or documents both with heterogeneous content and formats (doc, xls, ppt, etc.) [1]. In particular, these documents may contain data from different areas of knowledge, so document management requires in a broad sense an intelligent storage and management in order to draw on the knowledge inside them and links with large pools of domain data, personal or social data. We consider as Semantic Resource Integration (SIR) the search process and significant data retrieval existing inside information resources (documents, people) available in a digital format. This integration is based on the use of semantic technologies and linked open data approaches in order to retrieve pertinent documents or navigate between linked documents. Our approach considers an educational corporate memory (ECM) case study: the semantic integration of educational resources produced by three high level educational institutions: École Nationale Supérieure des Mines de Saint Étienne (FR), Universidad Autónoma Metropolitana, Unidad Iztapalapa (Mx) and Escuela Militar de Ingenieros (Mx), see the Figure 1 to look the main functions of this proposal.

We also take into account: the explosive growth of educational resources, heterogeneity in format, content and structure, concepts ambiguity, among others. The main users of the ECM are students, lecturers, professors and people who are related to the educational domain. These educational resources can be consulted in order to retrieve specific documents or people according to a concept, or something that is well known, finding information that is not explicit in educational resources.

Although document characterization and document access can be simple to implement in the case of a centralized or distributed corporate cloud system, information search is not immediate, some text mining functionalities are needed in order to automate the educational resource characterization process: for example classification to detect the main domain, the level and the language of a MOOC or other external document, named entity recognition to detect the keywords, mining links to detect the area and the level of an educational resource. On the other hand, this organizational memory can be enriched with documents and external information to organizations from other web sources such as: funding.
bodies, enterprises, customers, press documents, documents from partners or descriptive semantic information.

The ability to easily access the full corporate memory and the presence of semantic search indexes and resources are conditions that greatly simplify the entire process of mining. In Figure 2, we show a representation of a variety educational resources and indexes associated. Automatic generation of educational resource descriptions helps the automatic creation of indexes.

Figure 2. Indexes representing the retrieval information

Linked Open Data has a vital role in open education where MOOCs are the main protagonists. The open education is emerging as a collective web of linked educational data about courses, resources, platforms, instructional models, experts, professors, and so on [2]. However, all these initiatives do not converge yet to a cohesive and truly connected global view [2]. Universities are joining the movement by deploying their own Linked Data platforms, and many other organizations and initiatives are also exploring how reuse, integrate and interoperate isolated Open Educational Resources (OER) repositories using Linked Data [3]. Our proposal tries to contribute to the Open Educational Resource integration in order to be a truly connected global view of OER repositories and MOOCs. Metadata enables a level of interoperability between different learning platforms [4], there are currently no standards or metadata exchange in massive open online courses (MOOCs), we consider it as a corporate memory and thus to enable resource sharing across different platforms. Indexing is a crucial phase in Open Educational Resources retrieval, so text mining techniques could help to automate this phase. After obtaining indices we can apply semantic technologies and linked open data, which have shown the feasibility of formal representation of knowledge in a specific domain for seeking information considering the content of the resource. The use of ontologies, axioms, reasoner and semantic descriptions of educational corporate memory resources are significantly increase the quality of response of information search systems and open educational resource integration.

The rest of the paper is structured as follows: Section II introduces some concepts that will be used throughout the document. Section III describes an overview of our proposal a
complete system for managing an Open Educational Corporate Memory using Semantic Web technologies and Linked Open Data (LODCMS). In section IV, we describe the process to be followed for obtaining educational resource descriptions and then propose semantic index which could guide the educational resources storage and retrieval. Finally, in Section IV some conclusions and future work are presented.

2 Theoretical and explanatory framework

The MOOCs are distinguished from other online courses, because they provide academic support as well as support and guidance to students throughout the course. The main features are: a) Mass: We do not known exactly how many students will be and may be more than can be met in person; b) Open: materials are available for free; c) Online: Profit of different communications channels and tools offered by Internet.

There are two main streams on MOOCs [5] considering educational and organizational models, these categories are: cMooC and xMOOC. cMOOC are developed and conducted by academics through open courses and web platforms. They are based on a participatory and collaborative approach. xMOOC: is a kind of of MOOC where the teacher is still the center of instruction, she/he prepares lectures, makes online discussions and others activities.

Many initiatives have been emerged around MOOCs and produced some platforms such as: Coursera [6], Udacity [7] EDx [8] MiriadaX [9], among others.

Linked Data refers to data published and reusable on the Web and readable by a machine. Its meaning defines that are linked to external data sets [10] [11]. Open Data refers to the legal interoperability of data [12]. Keep in mind that the use of linked data approach does not require the use of open data, however in order to get a potential profit, it is necessary to publish data as Linked Open Data good practices.

The open data format for publication is the Resource Description Framework (RDF) [13]. In 2006 Berners-Lee established four rules for publishing data on the Web; these rules are known as "Open Data Principles" [14] and consist of:

1. Use URIs to identify resources published on the Web.
2. Use HTTP URIs and so people can find those names.
3. When someone looks for a URI, provide useful information, using the standards (RDF, SPARQL, etc.).
4. Include links to other URIs, so that they can discover more things.

The process for looking for a Web URI in order to obtain information on a referenced resource is called dereferenced [15]. URIs can identify resources which could be [16]: a) a resource that can be transmitted through the Web, b) concrete or abstract entities that cannot be transmitted through the Web.

Once Open Data is published, we can perform various operations such as: storage, visualization, query,
linking data, reasoning on data to name a few. Figure 3 shows graphically the operations on Linked Open Data.

![Figure 3. Operations on Linked Open Data](image)

### 3 Overview LODCMS

The interest of this work is focus on the knowledge and resources in three educational organizations (teaching and research), in which there are complete knowledge about their members, internal documents, etc., and incomplete information from external sources (e.g. funding bodies, corporate web sites). The problem is studied from the perspective of data integration (expressed in RDF standard), where we have a domain ontology (representing in OWL format) as a central knowledge model for validating data. We can use reasoning on domain ontology and data (instances) to infer relationships between resources in the memory.

The data mining techniques [17] or text [18] help directly for classifying or clustering, supervised and unsupervised techniques respectively. Also other text mining techniques will be very useful to find general patterns or features in collections or groups of documents and serve to confirm or discover new semantic links, for this reason the links mining techniques [19] could be applied as well.

All mining methods (text, data or links) are based on document content (structured or unstructured). It is essential to generate indexes related to educational resources available in the corporate memory considering the availability, integrity and privacy. The use of a corporate cloud seems a good approach to storage and manage this indexes. Others approaches like platform (PaaS model) [20] or type service (SaaS) [21] are also considered.

Our goal is to propose a complete system for managing an Educational Corporate Memory using Semantic Web technologies (LODCMS). In particular, open linked data and ontologies. Running LODCMS imposes the existence of modules and the analysis of the dialogue with the user, collections management, generation and management of resources corresponding to the indices corporate memory and motor resolution of questions which in turn interacts with engines: information retrieval, reasoning and data mining. Another module would be able to decide if the result delivered to the user (the answer) is worth or not worth to be kept permanently. This decision will depend on the feedback from the same user.

In Figure 4 we show user interactions with LODCMS and other interactions with the outside. The user will ask the system and have an answer (like a document, a list of documents, a collection, a link or an ontology entry, and also the user could send her/his feed-back and indicate, if the response is correct or meaningful. The administrator naturally manages the whole system and is able to integrate new semantical resource. The system can automatically detect the new internal documents and ontologies and also the changes of the extra-resources that are known and can do its self-alimentation.

![Figure 4.- The interaction with the outside of LODCMS: users and Semantic Web.](image)

The resources contain semantic ontologies, thesauri, dictionaries, wikis, layout databases (relational or not). The data contained in the documents refer to internal resources (own institution) or outside (other web resources). Ontologies are much more complete and general resources, and are easy to handle, for this reason in the system there is in a gap between ontologies and other semantic resources that are more heterogeneous.

The auto-feed functionality would treat external data (global web) and internal data (semantic resources themselves, information resources, and ontologies) in a unified way. Indexing latter to access the first one, through links, reaching the globality of interesting information (global web). The indices denote therefore necessary to access the data (documents and ontologies) as links refer to a very fine way to a part of the collections or documents and the ability indices equally treat all data.

In figure 5 we indicate the whole data mining process which contains the mainly phases of a classic mining process: preprocess, choosing parameters, methods, interpretation and evaluation of the results, feedback, and also our mining process contains a feedback grown of the semantic resources we used. The phases as preprocessing, more exactly
representing a textual document or a structured document, can be better done using semantic resources as thesaurus or bag of named entity.

Not only one step of mining is done, but also successive steps in order to detect some important information like language, domain (as biology, mathematics, etc.), areas of domain (as trigonometry, cellular biology, etc.), topics and levels of domain (as college, high education). Automatically, if language and domain are missing, a classification step can be done to detect this information. To detect areas and subareas of a domain some various mining process can be done as named entity recognition, classification inside a well-known resource, topic modeling.

In the mining process we can also detect some patterns after a topic modeling process, a clustering between a small collection and an association rules algorithm or detect some interesting link using link analysis (see [22] for more details), this new knowledge after an expert validation step can be integrate inside our indexes and semantic resources.

![Diagram](image.png)

Figure 5. Proposed methodology for the whole mining process by generating also indexes and links between documents and terms.

4 Conclusion and future work

A proposal for semantic resource management of an organizational memory (LODCMS) is described in this article. The contribution of semantic technologies and linked open data was emphasized. We can also mention that our LODCMS system may be able to generate data that would be opened and linked to others, both to increase the visibility of the organization, to support the growing development of the Semantic Web.

A methodology for obtaining descriptions and index generation, based on the concept of clustering is presented, and the importance of these for the inclusion of semantic Web technologies showed.

5 References


