Metadata Extraction from Open edX Online Courses Using Dynamic Mapping of NoSQL Queries

Dmitry Mouromtsev
ITMO University
St. Petersburg, Russia
d.muromtsev@gmail.com

Aleksei Romanov
ITMO University
St. Petersburg, Russia
gloomspb@gmail.com

Dmitry Volchek
ITMO University
St. Petersburg, Russia
dvolchekspb@gmail.com

Fedor Kozlov
ITMO University
St. Petersburg, Russia
kozlovfedor@gmail.com

ABSTRACT

The paper describes use cases and architecture of the course extraction plugin for the Open edX platform build upon Linked Open Data. The issue of frequent repetitions of educational materials within the MOOC and relativity of recommendation tools for course developers is considered. Comprehensive review of the designed ontology and mapping, as well as evaluation using test courses are given. The last part of the paper discusses new possibilities and opportunities for the future work.

Categories and Subject Descriptors

I.2.4 [**Knowledge Representation Formalisms and Methods**]: Semantic networks

General Terms

eLearning system; edX; semantic web technologies in education

Keywords

Semantic Web; Linked Learning; edX; education; metadata; Linked data in education; educational ontology population

1. INTRODUCTION

The educational technology market is growing rapidly[1]. According to some of the researches, its capitalization rate approached 100 billion by the end of 2015. Furthermore, the number of people involved in this process is also increasing. For example, the audience of the Coursera[2] that had been launched in 2012 exceeded 10 million users in 2014 and increased up to 15 million in September 2015[3]. The edX has reached approximately 5 million users[4]. The number of enrolling people increases quickly due to the fact that the edX platform is opened, and many other institutions independently deploy it.

The obvious advantages of e-learning include the following: simple 24/7 access to the educational materials of the world's top universities, relevance and completeness of the information, the possibility to use external sources. All these benefits could be extended even more if powered by the semantic web technologies[1]. Moreover, these technologies allow to link together, process, refine and reuse already existing information while applying the Linked Data principles[2].

1.1 Motivation

Taking into account rapid growth of MOOC and educational platforms, a new challenge of revealing relationships between existing courses and building tools for their creation arises. Overcoming this challenge will allow to link materials in different domains, reduce duplication of resources and develop search capabilities.

The main purpose of this paper is to propose means of data extraction in the edX platform. According to the official statistics, more than 100 institutions and universities are using the platform[5]. This differentiates the edX from the other educational platforms. Thus, the data extraction using semantic web technology will allow to find relevant information and reuse it many times and integrate content from different open educational resources developed for the Open edX platform[6] and it could be interlinked with other educational resources[3].

The Open edX platform is an open-source project that boasts more than 200 contributors and more than 100 pull requests on the GitHub[7]. Therefore, it is an emerging and rapidly growing platform. The most important issues are the

WWW'16 Companion, April 11–15, 2016, Montréal, Québec, Canada.
ACM 978-1-4503-4144-8/16/04
http://dx.doi.org/10.1145/2872518.2890462.

[1] http://elearningindustry.com/elearning-statistics-and-facts-for-2015
[2] https://coursera.org
[3] https://www.edsurge.com/news/2015-09-08-udacity-coursera-and-edx-now-claim-over-24-million-students
[4] https://twitter.com/edxonline/status/631844606964035588
[5] https://github.com/edx/edx-platform/wiki/Sites-powered-by-Open-edX
[6] https://open.edx.org/
[7] https://github.com/edx/edx-platform

complexity of the current API and the platform modification for specific purposes.

To create a solution for data extraction from the Open edX platform and demonstrate results of its work, the following tasks were set up:

1. To define classes and attributes of study courses by analyzing:
   - external structure of courses based on the study of the already developed Learning Management System (LMS) courses in the edX;
   - stages of course design in the edX Studio;
   - structure of the Open edX database segment directly related to the course content.
2. To develop an ontology based on the defined classes.
3. To write a plugin for processing and conversion of data into triplets by:
   - deploying a server that is available on the Internet in order to use the Open edX;
   - constructing the database queries to obtain the information about the courses;
   - processing and mapping data in accordance with the ontology.
4. To use the proof of concept to assess fullness and quality.

1.2 Related work

Semantic technologies based on machine-interpretable representation formalism give good grounds for describing objects, sharing and integrating information, and inferring new knowledge together with other intelligent processing techniques[4].

Semantic technologies hold a significant promise in enhancing learning experience and teaching process in Higher Education (HE). This promise is based on the potential of semantic technologies to express meaning for learning resources, teaching resources, people, and learning objectives with the help of ontologies and annotation. Given semantic annotations, more efficient discovery and matching among learners, teachers, and learning resources can be achieved. The affordances of semantic technologies are increasingly significant and potentially transformative for the HE sector[5] considering the volume of learning resources online and the growing number of learners with access to collaboration tools and online repositories. It is important to point out that semantic technologies are widely implemented in educational resources.

Linked University and Open University with more than 250,000 enrolled students are the most popular projects. The main purpose thereof is exposing the public data as Linked Data. In general, many large universities and institutions are trying to describe the different ways of interaction using the semantic web technology. For example, BBC Curriculum Ontology provides a data model and vocabularies for describing the National Curricula within the UK.

Furthermore, the research of Zablith F.[6] describes the way of semantic technology integration into education by creating a linked data layer that serves as a conceptual connection between higher education courses. In the research, the author sets out the applications that reflect flows of the learning materials between different courses. It should be noted that these flows are based on interlinked concepts in e-learning environments.

The ECOLE[7] system collects educational content from different sources and shares it with the university learning systems. The implemented ECOLE system allows to exchange the educational content between universities and other institutions.

The most recent investigations relating to the AFEL[8] (Analytics for Everyday Learning) are noteworthy. AFEL provide developing tools for informal and collective learning by understanding the needs of the persons involved. The needs are determined by analysis of online social data on the basis of retrieving, extracting and enriching information from Web environments. mEducator that applies the principles of linked data[8] and standardizes medical information provides users with access to medical education resources. mEducator has plugins for Moodle and Drupal CMS as well as for standalone version[9].

Another example is the SlideWiki that opens up new opportunities for working with presentations[10] and changes the process of creation, dissemination and use of educational materials making them reusable and flexible.

It is worth to note the work of some major universities in this direction: the Open University, Southampton University and Oxford University. They provide access to open data while using and developing their own ontology. All of this stimulates new researches[9] and promotes development of educational ontologies.

2. ONTOLOGY DEVELOPMENT

An important question related to the educational semantic web is how a course can be represented in a formal, semantic way to be interpreted and manipulated by both computers and humans[11]. This problem can be solved by the means of the ontology development. The developed ontology (Fig.1) consists of 18 Classes, 14 Object Properties and 23 Data Properties. It is based on Top-level ontologies[12] such as: AIISO[10] that provides classes and properties to describe the internal organizational structure of an academic institution; BIBO[11] that provides main concepts and properties for describing citations and bibliographic references; FOAF[12] (an acronym of Friend of a friend), the ontology describing persons, their activities and their relations to other people and objects and TEACH[13], the Teaching Core Vocabulary, that is a lightweight vocabulary providing terms to enable teachers to relate things in their courses together.

Open edX platform courses have a very specific and strictly organized structure, so they cannot be fully and correctly described by using existing ontologies. In order to take into account all the features of the structure of the course, presented ontology has been developed. After that, it will be possible to map existing courses on the ontology and to obtain unified data model.

[8] http://projects.kmi.open.ac.uk/afel/

[9] http://blogs.pjjk.net/phil/a-short-project-on-linking-course-data

[10] http://purl.org/vocab/aiiso/schema.

[11] http://purl.org/ontology/bibo/

[12] http://xmlns.com/foaf/spec/

[13] http://linkedscience.org/teach/ns/teach.rdf

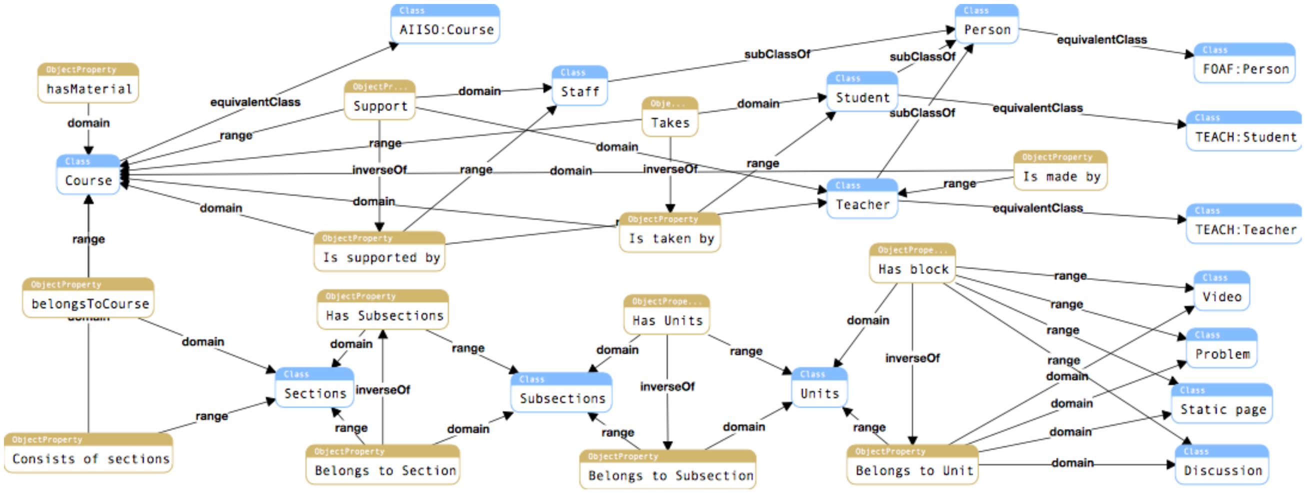

Figure 1: edX ontology

The main classes of the ontology that describe course structure are as follows:

- **Course** (equivalent to AIISO:Course) is the main class of the entire ontology, which has Data properties: Course image, start and end date, number of hours per week (estimated time for successful completion of the course), title, overview.

- **Sections** is a class describing main sections of a course. It has the following data properties: title, start date and visibility. This derives from the fact that course sections appear gradually (usually every week).

- **Subsections** is a class containing a description of the main elements of a section. It has such data properties as title, start date, visibility and deadline (it is necessary for a student to finish the assignments by the date in order to receive points).

- **Units** is a class describing Subsection elements. The class data properties include title and visibility. It combines the following classes:

  – *Static page* is a class containing a description of content of the study course in HTML format;

  – *Video* is a class for description of educational materials in the video format;

  – *Problem* is a class for different tasks and quizzes;

  – *Discussion* is a class for discussion. This class is considered as one of the greatest communication channels.

The following classes describe the key persons involved in the educational process within the course.

- **Person** is class that is equivalent to the FOAF [13] ontology class with the same title. It has subclasses of persons involved in the interaction as part of work with the platform:

  – *Student* (equivalent to TEACH:Student) is a class describing a student;

  – *Teacher* (equivalent to TEACH:Teacher) is a class to be populated by instances of teachers who participated in the course development and support educational process;

  – *Staff* is a class for storage of information about employees. These employees might not had been the developers but their contribution to the course support was significant. Such people help to organize prompt and competent interaction with students.

The ontology development was divided into two stages. At the first stage, the layout was designed based on analysis of the user interface available for end users of the Open edX platform, i.e. students and teachers. This allowed us to assess the total size of the developed ontology, identify the main classes, relationships, properties, facets, instances. At the same time, the technical implementation of the platform was considered.

After the above mentioned preparation and getting access to the direct implementation of the data storage model of the course structure, the second stage began. This stage was aimed at clarification of developed ontology details and consideration of the ontological model peculiarities. For example, all the components of the hierarchical structure of the course (Sections, Subsection, Units, etc.) were considered as subclasses of each other with a common parent class *Course* at the first stage. However, at the second stage, the solution was revised because all of these elements had been sufficiently independent and each of them had had full and detailed description as well as a unique structure. Otherwise, the child classes would have cumbersome description lacking information because they had been inherited from the parent class.

In a real model of the structure, all of these elements are individual objects that form a hierarchy using *block_type* parameter. This parameter can be *Course*, *Chapter*, *Sequential*, *Vertical* corresponding to *Course*, *Section*, *Subsection*, *Unit* classes of the described ontology and *children* parameter. The objects of the lower level within the element concerned are specified in *children* parameter. In this respect, it was decided to implement the components as individual classes linked by Object Property, for example, *Course hasSection Section*.

To demonstrate the ontology mapping, we used one branch of the first week of the course available on the edX.org, created by MIT. The course structure is clearly mapped on the ontology. Units with ”- - -” inscription mean other subsection branches that are not included in figure 2 to avoid cluttering.

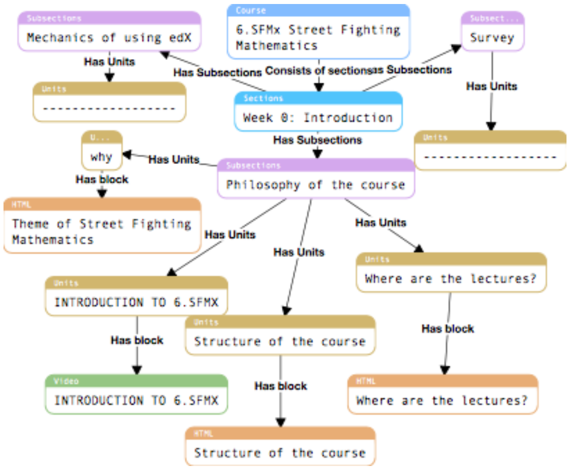

Figure 2: Example of the real course mapping

3. METHOD

Despite the fact that the Open edX is widely used open-source platform, its data structure, data processing, requesting and usage are difficult. The Course Structure API[14] provides the information only about the structure, but not about the content itself. XBlock API provides the information on content, but it is in pre-alpha[15]. The data about courses, their content and all the course changes are stored in the document-oriented database MongoDB. That’s why SQL database rewriters as SPARQLify[16] and similar ones are no use. The main task is writing the plugin to process the data obtained from MongoDB of the edX into the triple store followed by the SPARQL Endpoint deployment.

The edX course database consists of three collections:

1. ***active_versions*** collection stores brief information about the course and current published version (called *published-branch*) containing *ObjectId* which is used to create a relationship with *structures* collection.

2. The objects that contain the structure of each course are located in ***structures*** collection using *published-branch:ObjectId* from the *active_versions* collection. Owing to the features of the document-oriented database model, the course description seems to be cumbersome and lacking information at first. It is also explained by the fact that one block contains only written information about the whole course structure. The main parameter of data sampling is *block_type* parameter that allows to define format of content. Moreover, it is important to take into account the hierarchy of the course units shown in figure 3. Furthermore, the hierarchy is defined by the ”children” parameter from the parent block. The main course content is located in the lowest-level blocks, i.e. static page, problem, discussion and video. However, this content is stored in *definitions* collection.

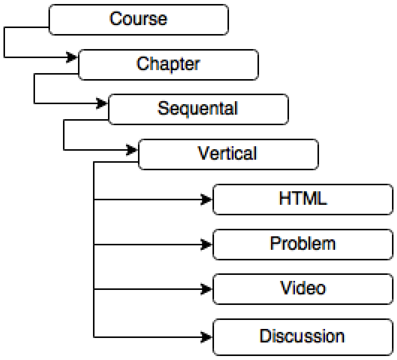

Figure 3: Course structure hierarchy

3. ***definitions*** collection is developed for storing the main content of the course. Objects of this collection create a relationship with the *structures* collection by the *block_id* parameter.

Based on the storage structure analysis, all the relationships were determined, and data about the courses and their content were collected with subsequent export into the triple store, which is necessary for mapping with the developed ontology.

The suggested method consists of the following steps shown on figure 4:

1. ”Mongo-Parse Plugin” collects course data from the Open edX MongoDB storage with MongoDB queries and exports them into the local triple store.

2. Triples are imported into RDF storage system (AllegroGraph, Sesame, etc.).

3. User can execute SPARQL queries through the SPARQL endpoint.

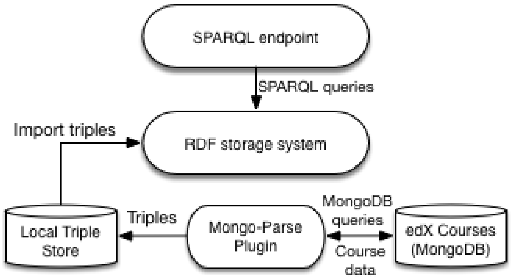

Figure 4: Concept of the method described

4. IMPLEMENTATION

For the purposes of implementation and performance of the task, Ubuntu 12.04 LTS server was deployed and the Open edX Cypress platform installed. To load course data from the Open edX MongoDB, to convert the data to N-Triples[17], and to perform ontology mapping, server-side scripting language PHP was used. PHP was chosen because the

[14] http://edx.readthedocs.org/projects/edx-platform-api
[15] https://github.com/edx/XBlock
[16] https://github.com/AKSW/Sparqlify
[17] http://www.w3.org/TR/n-triples/

authors are experienced in PHP programming and because PHP is suitable for the architecture concerned. For working with triple stores and for SPARQL Endpoint deployment, AllegroGraph was chosen[18]. It is designed for maximum load and query speed.

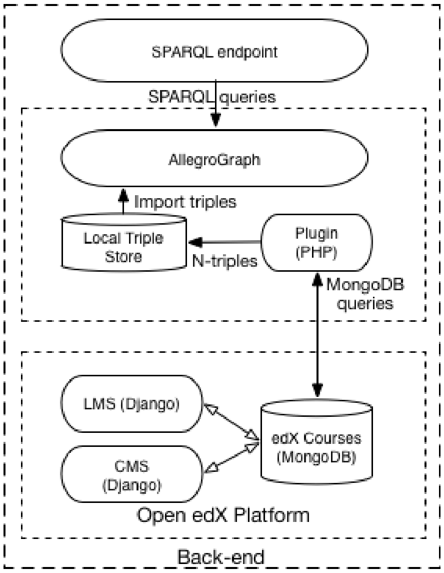

Figure 5: The overall architecture of data extraction

As shown in figure 5 of overall architecture, PHP Plugin connects to the Open edX storage using MongoDB queries returning courses with their structure and lectures in a text string with HTML tags. Plugin strips out all HTML tags and removes invalid characters for the further correct import of the string into triple store. Mapping of data to ontology classes is performed in the N-Triples Language and results are saved in a local triple store. The latter is updated at specified intervals. After that, AllegroGraph loads data from this triple store into its own graph database and users have access to this Linked Data through integrated SPARQL Endpoint.

5. EVALUATION

The proof of concept was used to assess the achievement of the following goals:

1. A check of developed plugin correctness in order to find critical errors. It also included assessments of labor intensity and plugin time characteristics.

2. An assessment of conformity and completeness of the results was used to determine the quality and fullness of mapping with the developed ontology.

3. Identification of discrepancies and inaccuracies for further elimination thereof.

4. Based on values and assessments obtained, it is necessary to give an opinion on the prospects of this project.

The plugin runs with no critical errors, the quantitative metrics given below are proposed for the assessment: analyzing structure of the database of the Open edX and Plugin development: 10 man-hours; total Courses: 6, Sections: 24, Subsections: 97, Units: 446, Static pages (lectures): 841; script time (MongoDB queries and creating local triple store): average of 0.76s after 1,000 executions; number of triples after ontology mapping: 9,367.

At the second step of evaluation, the following courses were created and uploaded in the edX platform: three MIT courses distributed under Creative Commons License[19]: Multivariable Calculus(I), Explore Engineering(II), Introduction to Computer Science and Programming(III); one default edX Demonstration Course(IV); four other courses with the structure only and no lectures. These courses provide data necessary for verification of the mapping completeness as well as for plugin and subsequent test SPARQL queries execution.

As the examples, the queries to count the total number of the sections, subsections, units and static page lectures were considered. The obtained data was recorded in table 1.

Table 1: Course materials mapping

Course	Sections	Subsections	Units	Static pages
I	6/6	26/26	265/265	609/611
II	9/9	14/10	37/33	44/44
III	6/6	35/35	98/102	96/106
IV	6/6	14/14	39/39	84/84

The table values reflect the number of elements found in the ”MongoDB query/SPARQL query” formats. Execution time of SPARQL queries is less than 0.002s. Based on the obtained data analysis, it can be concluded that the considered data set mapping is executed at a satisfactory high level. At SPARQL queries stage it became clear that minor errors (like 98/102 units for III course) were due to the fact that the duplicate triplets had appeared during the mapping. The differences relating to II course were predictable since it had been exported from Moodle CMS[20].

Query example that shows all Static page names of edX Demonstration Course with 84 results: ”EdX Exams”, ”Introduction: Video and Sequences”, etc. is given below.

```
select  ?n where {
?m exo:belongsToUnit ?s.
?s exo:belongsToSubsection ?h.
?h exo:belongsToSection ?o .
?o exo:belongsToCourse
<http://www.semanticweb.org/EdxOntology/
Main#ObjectId('56703c17457ebc4e4d8e595c')> .
?m rdfs:label ?n}
```

It should be noted that triple store does not store all of course data, for example, problems, video subtitles, different external materials that could be used in the course, not to mention data which is not directly related to the educational information including start date, deadlines, weight of an assignment, etc. From the standpoint of knowledge base creation, the specified data is not as important as learning materials. However, the data can be successfully used in the analysis of the relevance and quality of the course structure. The data can be used to improve or develop the course, or to review the learning materials, leading to positive changes in the knowledge base. It is worth to note that the developed ontology includes all of these elements, and their mapping is considered as the prospect for further research.

[18] http://franz.com/agraph/allegrograph/

[19] https://github.com/mitocw

[20] https://github.com/mitocw/moodle2edx

6. CONCLUSIONS

The identified tasks related to ontology development, validation of method of interaction with the Open edX platform and implementation of plugin were completed, and the main goal of data extraction was achieved.

The developed ontology and plugin for e-learning systems based on the edX platform allow users (teachers and course developers) to download data through the SPARQL endpoint. In the meantime, the Linked Open Data e-learning system uses learning materials in developing and updating e-learning courses by refining and reusing already existing information.

Thus, further work will evolve in the following ways: documentation and implementation of the described method as a component of Open edX; documentation and publication of the developed ontology; full ontology mapping with course information on any language. As a result, recommendation service that can analyze terms, learning materials and offer already existing in other courses or knowledge bases parts of the course will be created for course developers.

Recent links of the LMS of edX system, AllegroGraph, SPARQL Endpoint, ontology documentation, developed plugin and source code can be found at the Laboratory of Information Science and Semantic technologies GitHub:
`https://github.com/ailabitmo/edx-ontology`

This work was partially financially supported by the Government of Russian Federation, Grant 074-U01.

7. REFERENCES

[1] M. d'Aquin, "Linked data for open and distance learning," 2012.

[2] C. Bizer, T. Heath, and T. Berners-Lee, "Linked data-the story so far," *Semantic Services, Interoperability and Web Applications: Emerging Concepts*, pp. 205–227, 2009.

[3] S. Dietze, H. Q. Yu, D. Giordano, E. Kaldoudi, N. Dovrolis, and D. Taibi, "Linked education: interlinking educational resources and the web of data," in *Proceedings of the 27th annual ACM symposium on applied computing*, pp. 366–371, ACM, 2012.

[4] P. Barnaghi, W. Wang, C. Henson, and K. Taylor, "Semantics for the internet of things: early progress and back to the future," *International Journal on Semantic Web and Information Systems (IJSWIS)*, vol. 8, no. 1, pp. 1–21, 2012.

[5] T. Tiropanis, D. Millard, and H. C. Davis, "Guest editorial: Special section on semantic technologies for learning and teaching support in higher education," *IEEE Transactions on Learning Technologies*, no. 2, pp. 102–103, 2012.

[6] F. Zablith, "Interconnecting and enriching higher education programs using linked data," in *Proceedings of the 24th International Conference on World Wide Web Companion*, pp. 711–716, International World Wide Web Conferences Steering Committee, 2015.

[7] D. Mouromtsev, F. Kozlov, L. Kovriguina, and O. Parkhimovich, "Ecole: Student knowledge assessment in the education process," in *Proceedings of the 24th International Conference on World Wide Web Companion*, pp. 695–700, International World Wide Web Conferences Steering Committee, 2015.

[8] P. D. Bamidis, E. Kaldoudi, and C. Pattichis, "meducator: A best practice network for repurposing and sharing medical educational multi-type content," in *Leveraging Knowledge for Innovation in Collaborative Networks*, pp. 769–776, Springer, 2009.

[9] M. Hendrix, A. Protopsaltis, I. Dunwell, S. de Freitas, P. Petridis, S. Arnab, N. Dovrolis, E. Kaldoudi, D. Taibi, E. Mitsopoulou, *et al.*, "Technical evaluation of the meducator 3.0 linked data-based environment for sharing medical educational resources," in *2nd International Workshop on Learning and Education with the Web of Data, Lyon, France*, vol. 4, 2012.

[10] A. Khalili, S. Auer, D. Tarasowa, and I. Ermilov, "Slidewiki: elicitation and sharing of corporate knowledge using presentations," in *Knowledge Engineering and Knowledge Management*, pp. 302–316, Springer, 2012.

[11] R. Koper, "Use of the semantic web to solve some basic problems in education: Increase flexible, distributed lifelong learning; decrease teacher's workload," *Journal of Interactive Media in Education*, vol. 2004, no. 1, pp. Art–5, 2010.

[12] C. Keßler, M. d'Aquin, and S. Dietze, "Linked data for science and education.," *Semantic Web*, vol. 4, no. 1, pp. 1–2, 2013.

[13] D. Brickley and L. Miller, "Foaf vocabulary specification 0.98," *Namespace document*, vol. 9, 2012.