Formal model of student competencies in higher education and required skills in the job market

Ena Fuzul, Marko Horvat
Zagreb University of Applied Sciences
Department of Computer Science and Information Technology
Vrbik 8, 10000 Zagreb, Croatia
{ena.fuzul, marko.horvat}@tvz.hr

Abstract. *Higher education is a system that educates, professionally directs and strives to deliver competitive students in the labor market. The study programme determines the set of competencies and knowledge that the student acquires during the course and each course is determined by the curriculum, learning outcomes, achievements, and other data. On the other hand, the real sector is dynamic, defining new jobs, competencies and employment criteria every day. The research presented in this paper aims to define a formal knowledge model for the presentation of student's competencies and match them to the criteria and requirements of the real sector. Using web scraping technologies, the data related to courses and job ads were retrieved. Afterwards they were grouped, categorised and matched using the Web Ontology Language language for further potential comparison of the best candidates and job positions. The results indicate the potential of automatised retrieval and classification of available course data using formal knowledge representation which could lead to a more efficient discovery of employees. The paper concludes with guidelines for further analysis and potential upgrades of the student evaluation process.*

Keywords. Formal knowledge representation, ontologies, web scraping, higher education curriculum, learning analytics

1 Introduction

The goal of higher education is to provide its students with knowledge and experience in order for them to be prepared upon receiving their degrees for today's job market. The process in which students gain the aforementioned knowledge is comprised of a set of interconnected courses which are grouped in a study curriculum. Upon receiving their degrees, in most cases students end up on the job market seeking and competing for employment based on their knowledge. On the other hand, the real sector is different. Employees seek specific knowledge related to their open positions. This knowledge is often stated in a list of requirements for the job and doesn't necessarily correlate with the learning outcomes defined and received from students curriculae. In order to make higher education more competitive and aligned with the current job market, curriculae need to be compared and constantly re-evaluated based on their outcome in order for students receiving the degrees. By creating a formal knowledge model, both the universities and future employers may benefit by easily producing and finding ideal candidates.

Formalization of knowledge about higher education domain in the Croatian education system has already been a subject of intensive research by other authors, per example (Lovrenčić & Čubrilo, 2007) (Mesarić, 2007) (Mesarić & Dukić, 2007) (Konecki & Lovrenčić, 2015) (Grubišić et al., 2016). In addition, there has been an intensive international research in this area, such as (Al-Yahya, George, & Alfaries, 2015) (Chung & Kim, 2016) (Terblanche & Wongthongtham, 2016) (Obeid, 2018), to mention just a few.

In this research we propose a model that expands the domain of formal description of higher education with information about competencies and learning outcomes, and how they relate to specific job ads.

The remainder of the paper is organized as follows; the next section gives and a short introduction to computer ontologies, what is their practical use in the formal reasoning, how they are categorized into upper and domain ontologies depending on the range and scope of knowledge they formally describe. In addition, this section explains ontology computer languages that may be used to implement ontologies with different levels of expressivity. Section 3 explains in the detail the automatic information retrieval process used in this research to extract relevant data about academic programmes and job ads from web-based sources. Section 4 introduces the developed formal model of student competencies, its concepts, their properties and mutual relationships, and how the model was implemented. Section 5 gives insight into many possible directions for future work. Finally, Section 6 concludes the paper.

2 Ontologies as a tool for formal knowledge representation

In a nutshell, formal knowledge representation is concerned with describing knowledge about a particular domain. One of the most common tools for efficient formal knowledge representation are ontologies.

Ontologies, as a tool in computer science, are used for the formal description of high-level content and semantics (Staab & Studer, 2009). Together with a suitable reasoning engine, ontologies have been successfully applied in many areas for knowledge description and automated reasoning. Apart from specialized domain ontologies which as used for a detailed description of specific domains of interest, several upper ontologies have also been developed, such as SUMO (Pease et al., 2002), Cyc (Medelyan & Legg, 2008), and ConceptNet (Speer et al., 2017), which attempt to represent a wide range of general concepts and their relations. Upper ontologies enable the integration of several disjoint domains in a consolidated knowledge base.

The ontology-based paradigm for annotation and retrieval proposed in the paper consists of terminological and assertional knowledge about high-level education and student skills required in the job market.

Terminological and assertional knowledge are the basic components of a knowledge-based system based on Description Logics (DLs) as a set of structured knowledge formalisms for knowledge representation with decidable reasoning algorithms (Baader et al., 2003). DLs represent important notions about a domain as a concept and role descriptions. To achieve this, DLs use a set of concept and role constructors on the basic elements of a domain-specific alphabet. This alphabet consists of a set of individuals (objects) constituting the domain, a set of atomic concepts describing the individuals and a set of atomic roles that are assigned to the individuals. The concept and role constructors that are employed indicate the expressive power and the name of the specific DL. Here, we use $\mathcal{SHOIN}(\mathcal{D})$ on which Web Ontology Language DL (OWL DL) is based that employs concept negation, intersection, and union; existential and universal quantifiers, transitive and inverse roles, role hierarchy and a number of restrictions. Besides OWL DL, two other variants of OWL exist: Lite and Full. Since OWL Lite is decidable but semantically limited while OWL Full is undecidable but very expressive. OWL DL represents a compromise between adequate expressivity and guaranteed decidability. Importantly, a diverse set of computer tools for knowledge engineering are available which allow construction, management, reuse and reasoning with ontologies in OWL (Musen, 2015). As such OWL DL is an appropriate ontology language for representation and reasoning about high-level semantics about high-level education and required skills in the job market.

3 Automated information retrieval

Data on learning outcomes and jobs requirements had to be collected for the research. The information retrieval procedure was conducted in two parallel processes. One process was designed for retrieving data on study programmes and courses, while the other process was intended for retrieving data on job ads. In both cases, similar retrieval techniques were used although data were acquired from various sources freely available on the web. The former process included API requests and common web scraping techniques to extract data from university and the latter required more complex scraping techniques to extract data from job ad portals. In addition, the whole information retrieval procedure was separated into two phases. Phase 1 focused on accessing various sources of data through different techniques, and phase 2 focused on parsing the aforementioned data for further analysis. The parsed data was then manually copied into the program for further analysis. The information retrieval procedure is shown in Fig. 1.

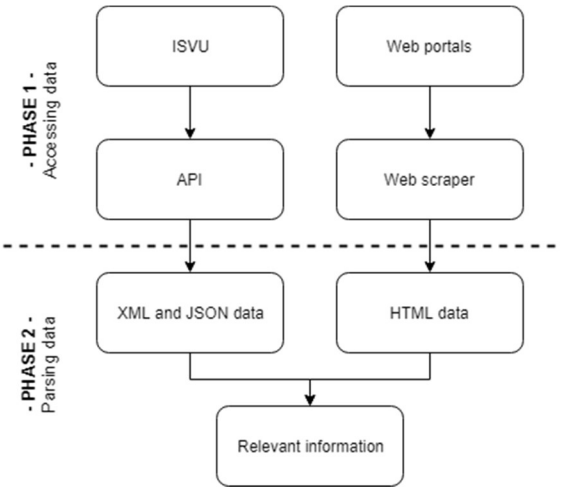

Figure 1. The procedure of automated information retrieval.

To identify a set of competencies and knowledge that student acquires during the study it was necessary to obtain data associated within a university's curriculum and course syllabi (Jokinen, 2015). The focus was on gathering learning outcomes from course syllabi in order to relate and compare them to the needs of the current job market. Additional data collected in the process, such as information about the university and different study programmes was also gathered and used in further research.

To retrieve data about a university programme structure, the university's application programming infrastructure named ISVU and its REST API service was used. ISVU (Information System of Higher Education) is an official information system of higher education institutions in the Republic of Croatia. It implements a great number of functionalities important

for educators and students including editing lesson plans for lectures and exercises, content management of courses and study programmes, review of student exam dates and statuses, long-term storage of data related to the completion of studies, etc. ISVU REST API is a service developed for programmable interaction with the ISVU and like any other API it serves to provide access to resources and methods ("ISVU REST API", 2019). ISVU REST API enables access to all data related to the specific university and it serves as a tool for professors and academic staff. ISVU REST API eases the process of selecting, updating or deleting data related to courses, exams, students and any kind of data related to the higher education system. It is based on the Hypertext Transfer Protocol (HTTP) and it enables resource access through HTTP requests. Using the aforementioned approach, it was possible to access information on numerous structures at the university. The information included types of study programs, elements of each study programme, and the list of courses connected to the study program. It also consisted of many attributes and metadata. Information retrieved through ISVU REST API was the course names, course codes, corresponding semesters in which it is performed, a course being compulsory or elective, etc. (Fig. 2). The data were formatted in XML and JSON. All the data obtained regarding courses and study programme relates to the Zagreb University of Applied Sciences and its undergraduate and graduate studies.

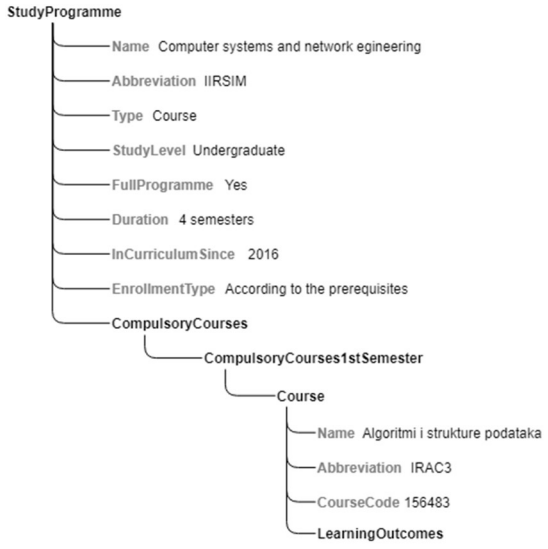

Figure 2. Information retrieved through ISVU REST API.

Additional data retrieval was focused on information related to courses. Despite gathered data through the ISVU REST API essential information regarding learning objectives was still missing. Using web scraping technologies (Baeza-Yates & Ribeiro, 2011) the data were retrieved through the private website of the university. The website, property of Zagreb University of Applied Sciences, holds all additional information on courses, e.g. learning aims, goals, outcomes, assessments, professors, lectures, etc. The website served as an extension with additional data bound to the courses. A web scraper was developed for these purposes. The program was built in Python language using Beautiful Soup and urllib2 libraries (Nair, 2014). Its main objective was to scrape through a website and collect meaningful information. Python's urllib2 module enables opening URLs, basic and digest authentication, redirections, cookies and more and was used to access the university's website.

The second module, Beautiful Soup, was used for navigating, searching through a website and modifying the parse tree. It served as a toolkit for dissecting a document and extracting relevant data which was important in an attempt to collect significant data on courses. In combination, these two modules extracted all the data needed. By handing URLs through urllib2 module it was possible to access the website and forward its source code to Beautiful Soup's parser (Nair, 2014). Beautiful Soup parsed DOMs (Document Object Model) from the source code and extracted relevant data using CSS (Cascading Stylesheet) selectors and IDs (unique identifiers) from the original document. After cleaning the data, removing unnecessary HTML (Hypertext Markup Language) tags and other irrelevant parts the core information was stored. In the case of more advanced websites where data is not automatically available in the DOM, a different approach would be more appropriate. Such cases appear when accessing SPAs (Single Page Applications) where the data is loaded dynamically after the application has been loaded in the user's browser using asynchronous requests. Solution to such problems would be to use a headless approach, i.e. run a fully functional web browser without a GUI (Graphical User Interface), which would allow more control over the application.

The same process was used in collecting job ad information. With a need for data from the labor market, its requirements, demands and technical skills, various data sources were searched. Among various options as potential insight into the real sector, the focus went on Croatia's most popular job ads portals (www.moj-posao.hr and www.posao.hr) which together have more than 11,000 job offerings. Specifically targeting jobs in the field of telecommunication and information technology, more than 600 job offers were retrieved. A large amount of data retrieved by scraping the aforementioned websites resulted in another set of problems. Since there was no mandatory or official standard in writing a job ad it was much harder to obtain the pertinent data. Most job ads do not have a unique written format, they are often designed in the form of a company's job offer template, followed by various descriptions in natural language. To get to the core data, the scraper had to be improved in order to comply with various web scraping throttle rules, CAPTCHA triggers, and complex data structures.

For this research, the collection of data were reduced to the set of the most fitting jobs ads for further analysis. Jobs ads in the final dataset had the most standardized written form, and as such were presumably the most suitable for extraction of significant information. After scraping the websites and manual selection of job ads the final dataset contained 30 of the most fitting job ads. These ads were used the standardised jobs ads web portal template allowing for easier data extraction.

Each ad had a variety of attributes such as the job title, job description, employee and company name, working location, obligations, requirements, benefits information, etc. There was an issue related to the description of the required skills and knowledge. Even though ads were nearly identical, each of them differed in describing desired requirements. This was a common problem and there was currently no adequate programmatic way to tackle this NLP (Natural Language Processing) problem automatically, therefore ads were manually reviewed and the corresponding requirements categorized.

4 Formal model of student competencies

After the data had been collected, cleaned and organized it was transformed into an ontology. An ontology represents knowledge on a subject and it is described by different objects and relations between them. Ontologies provide classes, properties, individuals, and data values to describe data and may be regarded as documents in the Semantic Web paradigm (Hall et al., 2009). The ontology was built using OWL which is a XML-based language that can be used to describe formal associations among resources. A resource can be anything with a Uniform Resource Identifier (URI).

An ontology named Learning Outcomes Course Objectives Ontology (LOCOOnto) (Fig. 3) was created for this project and it represents all the data gathered as formal concepts and relations between them. By definition, ontologies are a representation of a shared understanding about a specific domain and as such enables the derivation of implicit knowledge through automated inference with reasoning engines. Similarly, the LOCOOnto enables a common understanding about the content of student competencies in higher education. The ontology was built using Protegé, an open-source ontology editor and framework for building intelligent systems developed by Standford University (Musen, 2015).

The scraped data describes an organized and formal system of knowledge received from educational systems and the current job market. First, the data were deeply analyzed, divided into logical segments, then later combined and designed into a structure using Protegé. Defining this logical structure and sorting data into classes, attributes, individuals and their relations was the biggest part of the research.

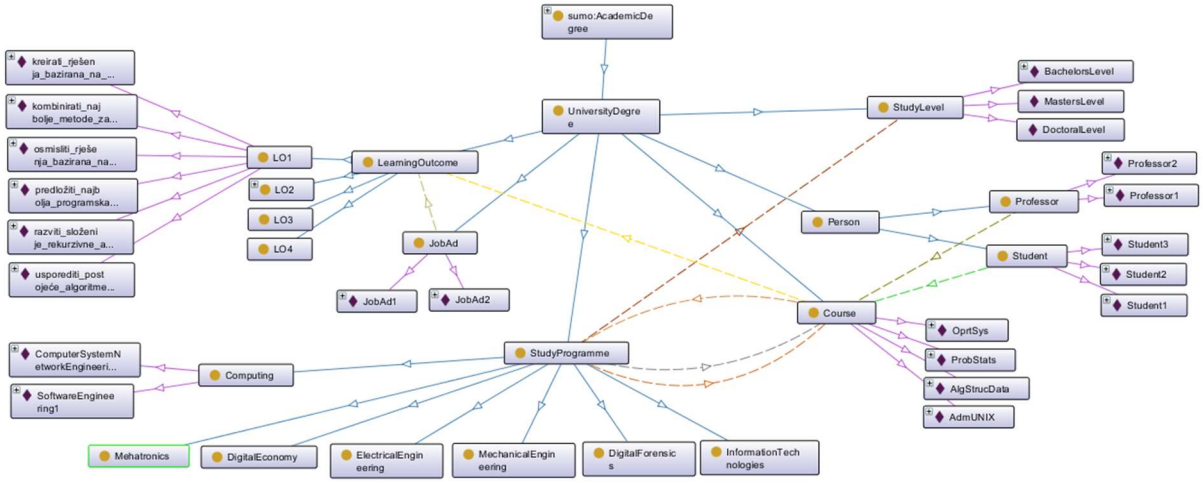

Figure 3. LOCOOnto ontology structure.

The core ontology concept, i.e. the topmost class in LOCOOnto, is UniversityDegree that encompasses six other concepts: StudyProgramme, LearningOutcome, Course, StudyLevel, JobAd, and Person. Each of these subsumed concepts defines one independent information domain that represents particular knowledge about the content of a university degree. All scraped and collected data were categorized into one of these classes. The data were automatically retrieved through the Python program and subsequently manually migrated to the OWL format and copied to Protegé tool as concept instances. Therefore, the whole process of building the research dataset involved automatic procedures and manual intervention.

Concept StudyProgramme describes the structure of undergraduate and graduate studies with attributes

such as name, education level, full-time programme, number of semesters, university programme, enrollment model, etc. Each instance of the class StudyProgramme is one programme held at the Zagreb University of Applied Sciences which results in a total of 7 study programmes. Each programme is connected to the StudyLevel class in order to separate degrees of education. Therefore, class StudyLevel consists of three subclasses named BachelorsLevel, MastersLevel, and DoctoralLevel. They are related through the property hasStudyLevel which specifies the relationship between these two concepts and joins the programme to a specific education level. StudyProgramme class also has another connection to Course class. This relation separates compulsory courses from elective ones. Further, StudyLevel has two additional object properties: hasCompulsoryCourse and hasElectiveCourse. In this way courses are connected to the study programme in one-to many relationships. Courses belong to study the programme and the study programme is comprised of courses. Next concept was LearningOutcome that represents a list of all individual learning outcomes necessary for a particular university degree. StudyProgramme is a subsuming concept for Computing, Mechatronics, DigitalForenscics, DigitalEconomy, MechanicalEngineering, ElectricalEngineering and InformationTechnologies, that all represent one specific study programme taught at Zagreb University of Applied Sciences. Finally, the concept Person is subsumed by two different concepts: Student and Professor.

Importantly, as shown in Fig. 4 we decided to derive UniversityDegree from sumo:AcademicDegree concept thus integrating LOCOOnto with an existing SUMO upper ontology. This is essential in facilitating knowledge reuse and by providing a common attachment point between various domain ontologies.

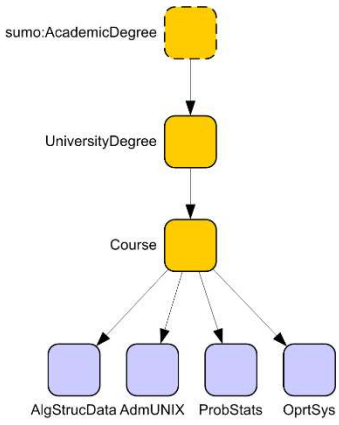

Figure 4. Relationship between concepts sumo:AcademicDegree, UniversityDegree, and Course with corresponding individuals (AlgStrucData, AdmUNIX, ProbStats, OprtSys).

One of the concepts that UniversityDegree subsumes is Course which, depending on a particular curriculum, has different instances specifically taught courses. The instances represent courses Algorithms and data structures (AlgStrucData), UNIX systems administration (AdmUNIX), Probability and statistics (ProbStats) and Operating Systems (OprtSys).

Concepts Course and JobAd are related to LearningOutcome (Fig. 5). This relationship is vital to define a unitary relationship between job ads posted by employers and courses which students have passed to achieve their university degree. Student competencies are defined by the courses a student has passed. Therefore by linking Course and JobAd student competencies are also connected to the needs of the real sector. By using this relationship it is possible to extract only students that have specific skills and job vacancies that suit their distinct academic degrees.

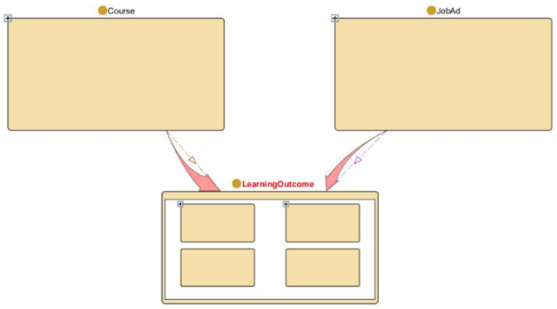

Figure 5. The link between concepts Course and JobAd relating student competencies to the skills required by companies.

5 Future work

Creating a rudimental system for collection of acquired knowledge and basic comparison to the current job market requirements has proved there is a large potential for improvement and collaboration. There are several key directions which could allow for a better final outcome. The system would greatly improve by expanding the amount of data that is collected. This could be done by scraping more job portals and ads which could then be integrated into the system. Another large problem that could be addressed is the NLP problem of freeform job ad descriptions. The system could benefit by employing advanced NLP algorithms to better understand and group similar job requirements, either based on simple string similarity and word processing (stemming, etc.) or more advanced techniques of identifying, classifying and clustering (Ding & He, 2004) similar job requirements. Further improvement could be accomplished by upgrading the current ontology into a more modern OWL version in order to define relations with a higher specificity. By mapping the developed ontology to other upper ontologies, apart from the SUMO, the described system could allow for a wider context of understanding and help improve the overall Semantic web (Horvat, Grbin, & Gledec, 2013). However, fragmentation of existing ontologies still represents a

significant challenge for the practical integration of knowledge within the Semantic web paradigm (Horvat, Dunđer & Lugović, 2016). Expanding the scope of the system would allow for the inclusion of other categories contributing to the total of knowledge, e.g. job recruitment and work experience (Fazel-Zarandi & Fox, 2009), online courses (Zaletelj & Košir, 2017), sentiments and emotional states (Horvat et al., 2009) (Horvat, Popović & Ćosić, 2012) (Horvat, Bogunović & Ćosić, 2014), and computer-aided medicine and well-being (Ćosić et al., 2013). Formal knowledge representation about emotional states and how they influence mental processes such as cognition, attention, memory, and learning is extremely important for the construction of a more robust decision-making system in the areas of education and pedagogy.

Finally, in the future, we intend to validate the developed ontology model in cooperation with the student alumni organization and interested employers.

6 Conclusion

Higher education is a complex system that brings out competitive students to the job market. Students are represented by a degree of their university which indicates their field of knowledge. To gain insight into achieved knowledge throughout the study a formal model was created. The process of collecting data using web scraping followed by a logical organization of data segments into an ontology could serve as a tool for easier selection of potential employees. Such a system would make knowledge more definable to the company and to the employee. The method developed in this research represents just a starting point in the automatization of the hiring process. Tools used in this paper serve as an indicator that the hiring process can be automated with the use of smart data extraction in combination with a classification of data using formal knowledge representation. An intersection between the demands of the job market and university learning outcomes indicate the possibility of an effective and systematical finding of a potential employer. This method can be further improved by implementing NLP solutions and more advanced knowledge representation. In that case, an overview of a student's knowledge could be extended with more categories, e.g. work experience and various informal learning activities.

References

Al-Yahya, M., George, R., & Alfaries, A. (2015). Ontologies in E-learning: review of the literature. International Journal of Software Engineering and Its Applications, 9(2), 67-84.

Baader, F., Calvanese, D., McGuinness, D., Patel-Schneider, P., & Nardi, D. (Eds.). (2003). The description logic handbook: Theory, implementation and applications. Cambridge university press.

Baeza-Yates, R., & Ribeiro, B. D. A. N. (2011). Modern information retrieval. New York: ACM Press; Harlow, England: Addison-Wesley,

Chung, H., & Kim, J. (2016). An ontological approach for semantic modeling of curriculum and syllabus in higher education. International Journal of Information and Education Technology, 6(5), 365.

Ćosić, K., Popović, S., Horvat, M., Kukolja, D., Dropuljić, B., Kovač, B., & Jakovljević, M. (2013). Computer-aided psychotherapy based on multimodal elicitation, estimation and regulation of emotion. Psychiatria Danubina, 25(3), 0-346.

Ding, C., & He, X. (2004, July). K-means clustering via principal component analysis. In Proceedings of the 21st international conference on Machine learning (p. 29). ACM.

Fazel-Zarandi, M., & Fox, M. S. (2009). Semantic matchmaking for job recruitment: an ontology-based hybrid approach. In Proceedings of the 8th International Semantic Web Conference (Vol. 525).

Grubišić, A., Stankov, S., Žitko, B., Tomaš, S., Brajković, E., Volarić, T., ... & Šarić, I. (2016, January). Empirical evaluation of intelligent tutoring systems with ontological domain knowledge representation: a case study with online courses in higher education. In Proceedings of the 13th International Conference Intelligent Tutoring Systems, ITS (pp. 469-470).

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. ACM SIGKDD explorations newsletter, 11(1), 10-18.

Horvat, M., Dunđer, I., & Lugović, S. (2016). Ontological heterogeneity as an obstacle for knowledge integration in the Semantic Web. Polytechnic and Design, 4(2), 85.

Horvat, M., Popović, S., Bogunović, N., & Ćosić, K. (2009). Tagging multimedia stimuli with ontologies. In MIPRO, 2009 32nd International Convention (pp. 203-208). IEEE.

Horvat, M., Popović, S., & Ćosić, K. (2012). Towards semantic and affective coupling in emotionally annotated databases. In MIPRO, 35th International Convention (pp. 1003-1008). IEEE.

Horvat, M., Grbin, A., & Gledec, G. (2013). WNtags: A web-based tool for image labeling and retrieval with lexical ontologies. Frontiers in artificial intelligence and applications, 243, 585–594.

Horvat, M., Bogunović, N., & Ćosić, K. (2014). STIMONT: a core ontology for multimedia

stimuli description. Multimedia tools and applications, 73(3), 1103-1127.

ISVU REST API, Informacijski sustav visokih učilišta RH, Sveučilišni računski centar, https://www.isvu.hr/api/, retrieved on 19 February 2019.

Jokinen, J. P. (2015). Emotional user experience: Traits, events, and states☆. International Journal of Human-Computer Studies, 76, 67-77.

Lovrenčić, S., & Čubrilo, M. (2007, January). Ontologies in the Higher Education Domain. In Proceedings of the 18th International Conference on Information and Intelligent Systems, Varaždin.

Konecki, M., & Lovrenčić, S. (2015, January). Ontology-based approach in education of programming. In International Academic Conference on Social Sciences and Humanities in Prague 2015.

Medelyan, O., & Legg, C. (2008). Integrating Cyc and Wikipedia: Folksonomy meets rigorously defined common-sense.

Mesarić, J. (2007). Integracija znanja u obrazovanju. Informatologia, 40(3), 216-222.

Mesaric, J., & Dukic, B. (2007, June). An approach to creating domain ontologies for higher education in economics. In 2007 29th International Conference on Information Technology Interfaces (pp. 75-80). IEEE.

Musen, M. A. (2015). The protégé project: a look back and a look forward. AI matters, 1(4), 4.

Nair, V. G. (2014). Getting Started with Beautiful Soup. Packt Publishing Ltd.

Obeid, C., Lahoud, I., El Khoury, H., & Champin, P. A. (2018). Ontology-based recommender system in higher education. In Companion Proceedings of the The Web Conference 2018 (pp. 1031-1034). International World Wide Web Conferences Steering Committee.

Pease, A., Niles, I., & Li, J. (2002). The suggested upper merged ontology: A large ontology for the semantic web and its applications. In Working notes of the AAAI-2002 workshop on ontologies and the semantic web (Vol. 28, pp. 7-10).

Speer, R., Chin, J., & Havasi, C. (2017). Conceptnet 5.5: An open multilingual graph of general knowledge. In Thirty-First AAAI Conference on Artificial Intelligence.

Staab, S., & Studer, R. (2009). Handbook on ontologies. Springer.

Terblanche, C., & Wongthongtham, P. (2016). Ontology-based employer demand management. Software: Practice and Experience, 46(4), 469-492.

Zaletelj, J., & Košir, A. (2017). Predicting students' attention in the classroom from Kinect facial and body features. EURASIP journal on image and video processing, 2017(1), 80.