

Ontologies for Social Media Digital Evidence

Edlira Kalemi, Sule Yildirim-Yayilgan

Abstract—Online Social Networks (OSNs) are nowadays being used widely and intensively for crime investigation and prevention activities. As they provide a lot of information they are used by the law enforcement and intelligence. An extensive review on existing solutions and models for collecting intelligence from this source of information and making use of it for solving crimes has been presented in this article. The main focus is on smart solutions and models where ontologies have been used as the main approach for representing criminal domain knowledge. A framework for a prototype ontology named SC-Ont will be described. This defines terms of the criminal domain ontology and the relations between them. The terms and the relations are extracted during both this review and the discussions carried out with domain experts. The development of SC-Ont is still ongoing work, where in this paper, we report mainly on the motivation for using smart ontology models and the possible benefits of using them for solving crimes.

Keywords—Criminal digital evidence, social media, ontologies, reasoning.

I. INTRODUCTION

IN this paper, we present an overview of the use of OSNs and the development of ontologies for gathering intelligence from OSNs to provide it to legal domains.

There are numerous social network data sources, like Facebook, YouTube, LinkedIn, Google Plus, Twitter etc., which are becoming larger every day and hold an abundance of valuable information to use. Different digital crime evidence may be collected from online social networks for crime detection and further analysis. The information made available by social media providers is staggering compared to other sources of information. For example, phone companies can only provide information about phone calls and messages. When a social media company like Facebook responds to a government subpoena, it could provide the user's profile, wall posts, photos that the user uploaded, photos in which the user was tagged, a comprehensive list of the user's friends with their Facebook IDs, and a long table of login and IP data [1]. A vast amount of information from OSNs is publicly available as well and can be used by different parties during a criminal investigation. Despite the need to use social networks, the large data volumes have increased workloads for digital forensic investigators and analysts and the increase in the workloads is becoming unmanageable. Since the growth of this workload is not changing in the near future, there is a need to find new methods on how to facilitate this process. Automated processes are needed in order to help reduce the

amount of workload of investigators without affecting the quality of the results they produce.

Due to what we said in the previous paragraph, the aim of this article is to: 1) present a review on existing smart solutions and models, 2) present a prototype for a criminal ontology, 3) provide the motivations for using ontologies for developing smart solutions and models for solving crimes using OSNs.

To the best of our knowledge, there is a lack of use of ontologies for finding digital criminal evidence in OSNs. OSNs may provide a very efficient approach for solving the crimes in this domain. An ontological representation provides two major benefits: the use of a common language for all the parties in a particular domain and a basis for reasoning [2]. The first benefit of using of a shared ontology across multiple systems creates a consistent language, as the same terms are used to define the same meanings. Google Knowledge Graph, one of the current generations of ontological systems, has the slogan "Things, not strings" [3]. Ontological representation allows a more extensible method of storage than the use of custom databases, as new properties i.e. new classes or relationships append to the existing structures. Different and incompatible knowledge representations will potentially use the same term in a different context, use different terms to mean the same concept, or represent terms with different granularities. The second benefit of ontological representation is its use in the field of symbolic artificial intelligence, and machine learning. Knowledge becomes a graph structure which can be easily navigated, compared and analyzed by machine. Representation of different cases into a single graph allows a more complete, less fractured view for the investigators.

Section II of this article is organized as follows. Part A explains how OSNs are actually used to solve crimes. Part B gives definitions digital evidence and different types of digital criminal evidence. Part C analyses existing ontologies used for collecting and using digital evidence at different stages of criminal investigation and in court. In Section III, an ontology prototype for collecting digital criminal evidence is introduced. Lastly, we give conclusions and future work.

II. BACKGROUND AND REVIEW

A. Using OSNs for Crime Investigation

The European COMPOSITE [4] project has conducted a deep analysis on "Best Practice in Police Social Media Adaptation" where they highlighted that social media is a very good source of information on criminal activities and is also an efficient means of communication with the public. LexisNexis Risk Solutions [5] announced the results of a comprehensive study focused on the use of social media by

Edlira Kalemi, Assistant Professor is with the University of Tirana, Tirana, Albania (corresponding author phone: +355689034567; e-mail: edlirakalemi@gmail.com).

Sule Yildirim Yayilgan, Associate Professor is with Norwegian University of Science and Technology, 2518, Gjøvik, Norway (e-mail: suley@hig.no).

law enforcement for crime investigation and prevention and eight out of 10 law enforcement professionals use it for crime investigations (63 percent) and 51 percent are using social media for crime prevention activities. Through verbal discussion with Albanian Cyber Security department of the State Police, we learned that their staff are using social networks for investigating and collecting information for a certain case after the case is open. Official permission from the OSN provider is obtained for the collection of data from the provider. The procedure of gaining the permission is not a fast and an easy process. The data collected from the OSN provider is not used to prevent crimes. It is neither used to investigate into new point of views for solving a crime nor used as an evidence at court. This is also confirmed by the attorneys in Albania and no data obtained from OSNs is considered as primary or secondary proof. However, they find it of great interest to have the possibility of preventing crimes and providing evidence by using OSN data but they are missing the tools to do this and their staff is overloaded for carrying out this process manually.

One significant example demonstrating the amount of information available to law enforcement from a simple photograph is that of John McAfee, the antivirus company founder who was recently under investigation from law enforcement authorities investigating the murder of his neighbor. McAfee was forced out of hiding when it was found that a photo of him published on a blog was embedded with GPS metadata pinpointing his exact location in Guatemala [6].

Social Media has been successfully used even in cases of Post-Riot Investigations. An example is the Vancouver Canucks Riot where investigators had to process 5,000 hours of raw video in more than 100 formats. A dedicated website was established and spread in social media to allow the public to review photos. The photos were previously collected by investigators and the public would provide names and contact information for suspects whom they recognized. In the U.K. Riots after the Mark Duggan Shooting, investigators used Facebook and other social media sites to gather information and intelligence, synthesizing it into usable evidence [7]. Metropolitan Police Service gathered enough evidence to make more than 4,500 arrests. Numerous types of crimes have also been successfully solved by using information collected from OSNs. A review of using online social networks for investigative activities [8] categorizes the crimes involving OSNs in Classical Crimes (burglary, vandalism, domestic violence etc.) and Digital Crimes (identity theft, cybertalking, child pornography etc.). Some tools for analyzing OSNs have been developed and are being used by investigators as well. Lococitato [9] is one of the analysis tools for OSNs which generates animated click-able maps of the relationships among any users of Facebook, Youtube and Twitter. Other tools that are more recently used are Centrifuge [10], Commatrix [11], and Gephi [12].

B. Digital Criminal Evidence

Digital evidence of an incident is any digital data such as profile pages, chat transcripts, public messages, private e-

mail-type messages, digital photographs, or video [13] that contain reliable information that supports or refutes a hypothesis about the incident [14]. There are many definitions of digital evidence; the definition proposed by the Standard Working Group on Digital Evidence (SWGDE) [15] is any information of probative value that is either stored or transmitted in a digital form. In the transparency report [16] published by Google, it is indicated that the number of requests from government and courts for obtaining user data is growing from year to year. The services from which government require information are Gmail, YouTube, Google Voice and Blogger. Based on the legal process and scope of request, Google is providing different types of data. Examples of types of data from Gmail include subscriber registration information (e.g., name, account creation information, associated email addresses, phone number), sign-in IP addresses and associated time stamps, non-content information (such as non-content email header information), and Email content. From YouTube: Subscriber registration information, sign-in IP addresses and associated time stamps, video upload IP address and associated time stamp, copy of a private video and associated video information, private message content. From Google voice: telephone connection records, billing information, forwarding number, stored text message content, stored voicemail content. From Blogger: Blog registration page, blog owner subscriber information, IP address and associated time stamp related to a specific blog post, IP address and associated time stamp related to a specified post comment, private blog post and comment content. Law enforcements can also require data from Facebook [17] or other OSNs. The OSN user whose data is being inquired should be notified about the inquiry and should be asked for consent for the use of his/her information. Facebook [18] gives the possibility to users to download all their Facebook data for the whole existence of their profile, and all sort of data such as chat, check-ins, deleted friends, friend requests, emails etc. This way the users can provide this information to law enforcement instead of the law enforcement gaining access to the user account on their own.

C. Authenticating OSNs Digital Evidence as Court Proof

Social media evidence is the new frontier of criminal proceedings and it raises unique legal challenges, including issues of admissibility and a defendant's constitutional rights in material that social media companies maintain [19]. There are many cases of defendants who are arrested because of information found or provided from OSNs. Government can seek information from public data or ask for extra data from the social media service providers. Defendants face more significant obstacles than the government when seeking exculpatory evidence from social media companies [20].

OSN evidences can be used more easily for criminal investigation but it is much more difficult to use them as admissible social media evidence in court. Social media is subject to the same rules of evidence as paper documents, but creates hurdles in admissibility as it can be more easily manipulated by its nature [21]. Consequently, the methods of

authentication include 1) presenting a witness with the personal knowledge of the digital evidence published in the OSN profile (they wrote it, they received it, or they copied it) 2) professionals control the computer directly to see if it was used to post or create the information 3) attempting to obtain the same information from the actual social media company that maintained it in the ordinary course of their business [22]. In order for the evidence to be accepted as valid from the court, chain of custody for digital evidence should be kept, or it must be known who exactly, when and where came into contact with evidence in each stage of the investigation [23].

There are two distinct types of authentication that must occur for collecting evidence from social networking sites 1) authenticating the authorship of the evidence on the website and 2) authenticating that the proof used at a trial which may typically be a printout of a web page is a fair and an accurate representation of what was provided earlier, for example a print screen of a webpage [24]. In the case of Google for the records to be admissible, they do not provide any expert as testimony. Instead they provide a written certificate of authentication and they consider it sufficient proof [25]. Facebook also does not provide expert testimony support. The records are self-authenticating in the eyes of the law [26]. If a case requires any special certificate other than the certificate of authentication, the court should request it from Facebook.

Admissibility of digital evidence is one challenge. Another challenge of social media is the communication between attorneys and jurors, or jurors and defendants. Sometimes jurors conduct their own investigation into a case, post their opinions about the case on social media websites, or attempt to "friend" parties, lawyers, witnesses, or judges. The inappropriate use of social media has led to penalties for both jurors and attorneys. One case was Jacob Jock, and he was jailed after being found guilty on a criminal contempt of court charge after he Facebook-friended a defendant in a trial for which he had just been selected as a juror [27].

Given the expansion of social media and the potential challenges relating to the reliability or authentication of social media, the authentication and admissibility will continue to be a hard process. Smart solutions to support the actors of this process in proving facts and avoiding mistakes would be of great help and be in main focus of the researchers from the field.

D.Existing Ontologies

According to Noy and McGuinness, ontology creates a common definition among a domain of information within a certain area. By doing this, common information structures can be formed, knowledge can be reused, assumptions within a domain can be made, and every piece can be analyzed [28]. In this review, we analyze existing ontologies and proposed ontological models for the legal and forensics domain. A simple ontology was developed in NISLAB [29] named Criminal Ontology [30] and a successful model has been developed and tested for identifying if some individual is suspect or not based on the data gathered from Facebook. However, this ontology is not a robust one and further

development of it is needed. Research in the field is also not much developed. Some research and resulting ontologies have appeared in the related fields. We give a review of published ontologies and research for the legal and cyber domain.

In [31], a Cyber ontology has been developed starting from an initial malware ontology and then extended with utility ontologies that are focused on time, geospatial information, person, events, and network operations. Malware is one of the most prevalent threats to cyber security, and the Malware Attribute Enumeration and Characterization (MAEC) [32] language provides a store of knowledge (i.e. use cases, idioms, suggested practices etc.) that can be readily leveraged and it allows for the faster development of countermeasures by enabling the ability to leverage responses to previously observed malware instances.

In [33], an ontological model has been developed for finding the correct specialization, certification, and education within the cyber forensics domain. This model may also be used to develop curriculum and educational materials.

Digital Chain of Custody Digital Evidence Ontology (DCoDeOn) is an ontology which can be used for forming a common understanding of the structure of digital forensics. This will benefit forensic investigators, those sharing digital evidence, and software agents [34].

Bezzazi presents a small formal cybercrime ontology to demonstrate how law articles and legal cases could be defined, so that the problem of case resolution is reduced to a classification problem, and that counterfactual reasoning may be held over it [35].

The LKIF [36] core ontology of basic legal concepts is part of a generic architecture for legal knowledge systems, which will enable the interchange of knowledge between existing legal knowledge systems. LKIF is developed in the Estrella [37] project and has two main roles: 1) the translation of legal knowledge bases into different representation formats and formalisms and 2) a knowledge representation formalism that is part of a larger architecture for developing legal knowledge systems. LKIF core consists of 15 modules, each of which describes a set of closely related concepts from both legal and commonsense domains [38].

LRI-core upper ontology [39] provides anchors and interpretation to the various legal domain ontologies. It can be used for tagging and annotating the hearing documents, searching these documents, and structuring the set of retrieved documents. A real case where LRI core has been used and tested is provided in the ontology about Dutch criminal law developed under the E-Court project.

FOLaw is another core ontology developed by authors in [40] and specifies functional dependencies between different types of knowledge involved in legal reasoning and the authors define it as an epistemological framework.

For the automatic evidence analysis and for representing digital evidence in a semantic manner, authors in [41] have proposed an ontological approach. They have concluded that specialized forensic and security tools significantly contribute to and (semi-)automate parts of the analysis of large volumes of digital data that may be collected during digital

investigation process.

Digital Evidence Exchange (DEX) [42] is a standard of digital evidence provenance that explicitly defines the set of tools and transformations that led from acquired raw data to the resulting product. It is an XML format and is independent of the forensic tool that discovered the evidence. Digital Forensics XML (DFXML) [43] is another more recent XML language for digital forensics research and is designed to be an interchange format between forensic tools, to represent a wide range of forensic information and forensic processing results.

In [44], the author presents a prototype for analyzing email content using ontologies and inferencing in order to give answers to some forensics questions in a faster way e.g. are there any deviations in the email account history during the period of crime incident, etc.

From this review and to our best knowledge, no ontology has been developed so far for intelligence gathering from OSNs; for classifying digital evidence found in OSNs or for any other purpose that facilitates the process of solving crimes using OSN data. Existing ontologies for legal domain do mainly focus on helping the process in court in the context of law-case matching and the forensic ontologies deal with cybercrimes, not with ordinary crimes. The digital evidence these ontologies provide is not gathered from OSNs. Digital evidence gathered with existing tools as the ones listed in Section A are not smart models. The amount of data is big, where investigators and analysts have to spend ever increasing resources on locating and analyzing criminal data on social networks.

Ontologies for supporting digital evidence collection, classification, and intelligence gathering from OSNs are needed and should be developed in order to provide to the police a means to accelerate and facilitate crime prevention, investigation, and solving.

III. PROPOSED PROTOTYPE ONTOLOGY

There is a lack of scientific research about developing and using domain ontology in the digital forensics field. Additionally, ontologies that use OSN profiles to support crime solving are missing. Reason for this is that digital forensics is a multidisciplinary field and only the knowledge of technical aspects is not enough.

To fill the gap of developing and using ontologies to keep OSN profiles to support crime solving, we are in the process of designing and developing a smart ontology. The aim is to support solving digital and ordinary crimes faster and to make the job of the investigators and the other law institution actors easier. The Albanian Police and the other law enforcement agencies (LEAs) in Albania do not have the specialized tools for preventing crimes and collecting evidence using OSNs which leads to having the whole process done manually. The existing tools for dealing with the social criminal activity analysis are not intelligent and they mainly rely on the contextual analysis of data. An example of this is using only the co-occurrence of natural language terms appearing in a document to find the relationship between criminal activities in a network.

This review and discussions with the police and attorneys in Albania have given us the knowledge and the basis to design the initial version of the SC-Ont ontology. We have perceived the ontology as having three main pillars: People, Crimes and Crime Solving as illustrated in Fig. 1.

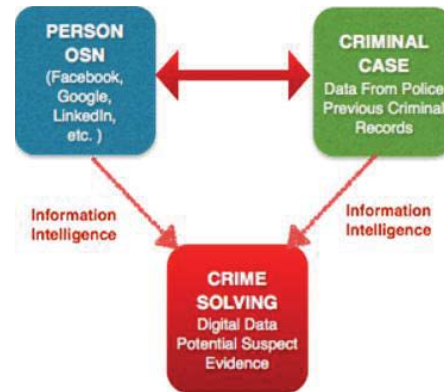
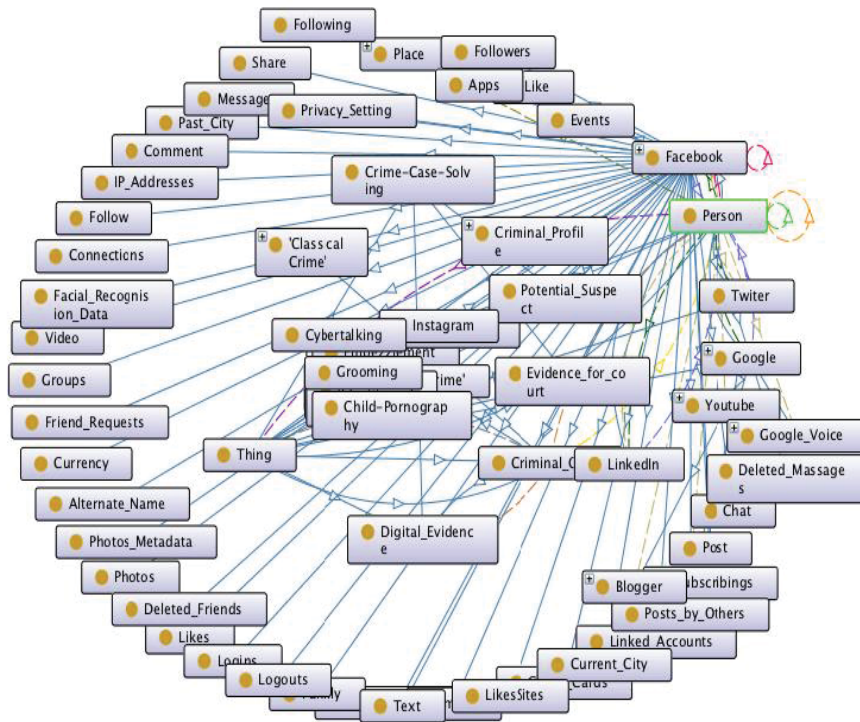


Fig. 1 Top layers of SC-Ont Ontology

Information about people found in OSNs profiles, the specific crime case that the police is working on to solve and the police database on previous precedents will lead to some strong hints for the crime investigator. The classes and the relations between classes that compose our ontology are illustrated in Fig. 2. At this stage, we have focused on information of user profiles that can be obtained from Google and Facebook when law enforcement requires data for someone's profile. We focus on both forensics crimes and crimes that occur in daily life. We are developing the ontology using Protégé [45] framework and OWL 2 Web Ontology Language. OWL 2 ontologies provide classes, properties, individuals, and data values and are stored as Semantic Web documents [46].

Some SWRL [47] rules have been written for categorizing instances in ontology; either as a part of their natural belonging or they may become part of another category in order to fulfill the defined rules. An example is that the police is working on an un-solved criminal case and the case is defined by having occurred at a location, on a date and at a certain time. If several individuals have "checked-in" in this location around the same time or if they have been tagged by their friends to have been in that place around the same time (in the OSN), then these individuals will be categorized under the "Potential Suspect" class (Fig. 2).

Some of the object properties or relations between classes are: 'for case', 'has account in', 'author-of', 'check_in', 'co-author', 'follow', 'followed', 'geolocation', 'has-author', 'is_private', 'is_public', 'foaf:knows' (this object property is imported from the Friend Of A Friend ontology [48]), 'shareVia', 'subscribe', 'vality_url', 'visited' etc.



- [17] <https://www.facebook.com/safety/groups/law/guidelines/>. Accessed: (17.10.2015).
- [18] <https://www.facebook.com/help/405183566203254>. Accessed: (17.10.2015).
- [19] Justin P. Murphy & Adrian Fontecilla, Social Media Evidence in Government Investigations and Criminal Proceedings: A Frontier of New Legal Issues, 19 RICH. J.L. & TECH 11 (2013), available at <http://jolt.richmond.edu/v19i3/article11.pdf>.
- [20] Daniel K. Gelb, Defending a Criminal Case from the Ground to the Cloud, 27 CRIM. JUST. 28, 29 (2012).
- [21] Griffin v. State, 19 A.3d 415, 424 (Md. 2011) (recognizing “the potential for abuse and manipulation of a social networking site by someone other than its purported creator”).
- [22] Justin P. Murphy & Adrian Fontecilla, Social Media Evidence in Government Investigations and Criminal Proceedings: A Frontier of New Legal Issues, 19 RICH. J.L. & TECH 11 (2013), available at <http://jolt.richmond.edu/v19i3/article11.pdf>.
- [23] Cosic, J., & Baca, M. (2010). Do we have full control over integrity in digital evidence life cycle? Information Technology Interfaces (ITI), 2010 32nd International Conference on, 429–434.
- [24] 7 WASH. J.L. TECH. & ARTS 209 (2012) <http://digital.law.washington.edu/dspace-law/handle/1773.1/11111>.
- [25] <https://www.google.com/transparencyreport/userdatarequests/legalproceedings/>. Accessed: (18.10.2015).
- [26] <https://www.facebook.com/safety/groups/law/guidelines/>. Accessed [19.10.2015].
- [27] <http://www.heraldtribune.com/article/20120216/ARTICLE/120219626>. Accessed [14.10.2015].
- [28] Noy N, McGuinness D. Ontology development 101: a guide to creating your first ontology. Available from: http://protege.stanford.edu/publications/ontology_development/ontology101-noy-mcguinness.html; 2001 [retrieved 13.10.15].
- [29] <https://www.nislab.no/>. Accessed: (22.10.2015).
- [30] B, Z. K., Imran, A. S., & Yildirim-Yayilgan, S. (2015). Social Computing and Social Media, 9182, 148–157. <http://doi.org/10.1007/978-3-319-20367-6>.
- [31] Obrst, L., Chase, P., & Markeloff, R. (2012). Developing an Ontology of the Cyber Security Domain. *Seventh International Conference on Semantic Technologies For. Intelligence, Defense, and Security – STIDS 2012.*, 49–56. Retrieved from http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-966/STIDS2012_T06_ObrstEtAl_CyberOntology.pdf.
- [32] MAEC - Malware Attribute Enumeration and Characterization. [Online] <http://maec.mitre.org/>. Accessed: (21.10.2015).
- [33] Brinson, A., Robinson, A., & Rogers, M. (2006). A cyber forensics ontology: Creating a new approach to studying cyber forensics. *Digital Investigation*, 3(SUPPL.), 37-43.
- [34] Čosić, J., Čosić, Z., & Bača, M. (2011). An ontological approach to study and manage digital chain of custody of digital evidence. *Journal of Information and Organizational Sciences*, 35(1), 1–13.
- [35] Bezzazi, E. H. (2007). Building an ontology that helps identify criminal law articles that apply to a cybercrime case. *Icsoft 2007: Proceedings of the Second International Conference on Software and Data Technologies, Vol P/Dps/Ke/Muse*, 179–185.
- [36] Hoekstra, R., Breuker, J., Di Bello, M., & Boer, A. (2007). The LKIF core ontology of basic legal concepts. *CEUR Workshop Proceedings*, 321, 43–63.
- [37] <http://www.estrellaproject.org/lkif-core/>. Accessed: (14.09.2015).
- [38] <https://github.com/RinkeHoekstra/lkif-core>. Accessed: (12.10.2015).
- [39] Breuker, J., Elhag, A., Petkov, E., & Winkels, R. (2002). Ontologies for Legal Information Serving and Knowledge Management. *Legal Knowledge and Information Systems. Jurix 2002: The Fifteenth Annual Conference*, (July 2015), 73–82.
- [40] Breuker, J., & Hoekstra, R. (2004). Epistemology and ontology in core ontologies: FOLaw and LRI-Core, two core ontologies for law. *Proceedings of the EKAW04 Workshop on Core Ontologies in Ontology Engineering*, 15–27.
- [41] Dosis, S., Homem, L., & Popov, O. (2013). Semantic Representation and Integration of Digital Evidence. *Procedia Computer Science*, 22, 1266–1275.
- [42] Levine, B. N., & Liberatore, M. (2009). DEX: Digital evidence provenance supporting reproducibility and comparison. *Digital Investigation*, 6(SUPPL.), 48–56.
- [43] Garfinkel, S. (2012). Digital forensics XML and the DFXML toolset. *Digital Investigation*, 8(3-4), 161–174.
- [44] Kota, V. K. (2012). An Ontological Approach for Digital Evidence Search. *International Journal of Scientific and Research Publications*, 2(1), 2250–3153. Retrieved from www.ijssrp.org. (18.10.2015)
- [45] <http://protege.stanford.edu/>. Accessed (15.09.2015)
- [46] <http://www.w3.org/TR/owl2-syntax/>. Accessed (20.09.2015)
- [47] Semantic Web Rule Language, <http://www.w3.org/Submission/SWRL/>, Accessed (15.10.2015)
- [48] <http://xmlns.com/foaf/spec/> Accessed (15.11.2015)