Traversal and relations discovery among business entities and people using semantic web technologies and trust management

Dejan Lavbič [a, 1], Slavko Žitnik [a], Lovro Šubelj [a], Aleš Kumer [a], Aljaž Zrnec [a] and Marko Bajec [a]
[a] *University of Ljubljana, Faculty of Computer and Information Science, Slovenia*

Abstract. There are several data silos containing information about business entities and people but are not semantically connected. If in integration process of data sources trust management is also employed than we can expect much higher success rate in relations discovery among entities. Majority of current mash-up approaches that deal with integration of information from several data sources omit or don't fully address the aspect of trust. In this paper we discuss semantic integration of personal and business information from various data sources coupled with trust layer. The resulting system has higher and more defined solidity while trust for single entity and also for data source is defined. The case study presented in the paper focuses on integration of personal information from data sources mainly maintained by government authorities who have higher trustability than information from social networks, but we also include other less trusted sources. The developed SocioLeaks system allows users traversal and further relation discovery in a graph based manner.

Keywords. Semantic Web, trust management, semantic integration, ontologies, personal information extraction, SocioLeaks

1. Introduction

System interoperability is an important issue, widely recognized in information technology intensive enterprises and in the research community of information systems. The widely adoption of the World Wide Web to access and distribute information further stressed the need for interoperability. In order to allow software agents to understand and process the web information in a more intelligent way, researchers have created the Semantic Web vision [2], where data has structure and ontologies describe the semantics of the data. When a wide variety of software agents could communicate with each other in a flexible and intelligent way, systems interoperability will be greatly enhanced.

The Semantic Web offers a compelling vision, yet it also raises many difficult challenges. One of the key challenges is the problem of trust as a method of dealing with uncertainty. This aspect is especially significant when dealing with online personal information. There are several online data sources that include information about business entities and people. They provide basic information about business entity, like company name, tax number, business address, legal organization form, share capital, partners and share equity interests, authorized representatives, members of the supervisory board etc. These sources include European Business Register (EBR) for all European based companies and also detailed information for individual countries can be identified, e.g. Public Legal Records and Related Services, Public spending at Public Payments Administration, list of registered researchers, list of doctors performing public health service, people listed in telephone directories etc. In addition to data sources maintained mainly by government authorities there are also others that are community based. First group of sources are nowadays very popular social networks, which include Facebook, LinkedIn etc. Besides semi-structured and structured sources of data, the vast majority of information still can be found in unstructured data sources. One of the largest sources is of course the World Wide Web and for our problem domain the newspaper articles (e.g. BBC, CNN, Delo, Finance etc.) hold very useful information about business entities and people. The structure of the data can be semi-automatically acquired by employing text-mining methods and supported by ontology.

So, interoperability among applications in heterogeneous systems depends critically on the ability to map between their corresponding ontologies and establishing a trust layer. We have a vast amount of information available about people and business entities at our hand that we have to manually digest and traverse to draw some conclusions, add value to raw data at individual data silos and discover some relations which are possible only after integration of several data sources. The integration of this data is not straightforward and requires a lot of tedious tasks which are not automated and usually have to be manually performed by humans. Nevertheless, there is also a problem of trust in individual entity and data source.

This paper tackles the problem of trust management in semantic integration and is structured as follows. First we present related approaches in section 2 with clear definition of the problem and solution proposal. Next, in section 3, we introduce trust in semantic integration of data with the definition of trust used throughout the paper. This section also includes the details of our approach in trust modelling with corresponding algorithm for trust delegation. To support the trust management, different types of trust and several layers of trust are presented. Details of the case study implementation are discussed in section 4. These include the architecture of our approach, data source and common ontologies developed and GUI that enables users searching for patterns across different data sources, traversal and further relation discovery. Finally the last section 5 presents conclusions.

2. Related work

2.1. Review of related approaches

Several aspects of Semantic Web and Social Networks are tackled in [10], where author introduces web data and semantics in social network applications, knowledge representation on the Semantic Web, modelling and aggregating social network data, developing social-semantic applications etc. The author also points out the folksonomies that could have also played a key role in breaking down the complexity of building applications with enhanced semantics. Authors in [5] discuss employment of social networking and semantic web technology in software engineering. Their approach is based on using existing Web 2.0 services such as social bookmarking and blogs as the infrastructure to share knowledge artefacts. Similar problems are addressed in [1], where authors highline steps, techniques and technologies for the development of intelligent applications based on Semantic Web Services based on a case study in e-learning systems. They identify several problems with development of this kind of applications. This includes unguided development, performance on loading ontologies, integration mechanisms, fault tolerance etc. An interesting study can be found in [4], where authors discuss the avatar (virtual person on behalf of real users) in the context of intelligent social Semantic Web. They try to identify how people will try to shape and use their avatars with the respect to social data portability. Another study in [13] is giving insight into exploitation of social semantic technology for software development team configuration. They present a system for supporting the design of teams for software development projects, which combines the benefits of semantics and social networks. The authors also propose a full-fledged solution backed with an implementation that has been tested in the scope of small and medium enterprises (SME). Author in [7] discusses about collective knowledge systems and the common points of Social Web and the Semantic Web. His approach proposes a class of applications called collective knowledge systems, which unlock the "collective intelligence" of the Social Web with knowledge representation and reasoning techniques of the Semantic Web.

In the domain of trust management there are several approaches that tackle this problem. Authors in [12] discuss about modelling and evaluation of trust with an extension in Semantic Web. The paper reviews well known methods of trust modelling and trust evaluation and proposes a new method for evaluating trust with greater simplicity and enhanced accuracy. Authors in [9] talk about trust-annotated ontology integration using social modelling. Their approach weights authors of ontologies and the resources they provide. This information is then used to assist integration process for an evolutionary trust model to calculate the level of credibility of resources. Authors in [11] discuss trust management for the Semantic Web. Their approach employs the idea of a web of trust, in which each user maintains trusts for a small number of other users. Each user therefore receives a personalized set of trusts. Authors in [8] talk about trust strategies for the Semantic Web. The paper identifies five common strategies of trust and discusses their envisaged costs and benefits.

There are several web portals that deal with integration of data from several data sources and also variety of approaches that deal with aggregation of personal information about people from different social networks [3]. Majority of methods work with USA based data that can be accessed and consumed in structured and opened manner, while they lack worldwide support, where this information is available mainly in unstructured form. These services mainly offer integration of data from social networks (e.g. FriendFeed, iSearch, Facebook etc.) and also others (e.g. PeopleFinders, Glassdoor, Search Systems etc.) that focus on extraction of personal information from the web, public records databases etc. The basic features can usually be employed for free, while advanced search capabilities are offered for a fee.

2.2. PROBLEM AND PROPOSAL FOR SOLUTION

The related work pointed out that there are several data silos containing information about business entities and people but are not semantically connected. The web of data that would integrate these isolated data sources can be beneficial in several scenarios. One of the problems in modern days that is becoming more and more problematic is the corruption. In crossing the frontiers of legislation some business entities or people can be very innovative by including multiple entities with the purpose of covering the trails of questionable actions. This also includes all sorts of fraud detections with insurance companies, banks and other public institutions. With the availability of such integrated system we can easily traverse the interconnected entities and focus on some specific patterns. If these activities are backed up by integrated data sources with trust management layer we can expect much higher success rate. Majority of current mash-up approaches that deal with integration of information from several data sources omit or don't fully address the aspect of trust. Existing case studies also mainly focus on personal information from social networks which are not very reliable as users for various reasons tend to give false information [6].

In this paper authors propose semantic integration of personal and business information from various data sources coupled with trust layer. The resulting system has higher and more defined solidity while trust for single entity and also for data source is defined. The case study presented in the paper focuses on integration of personal information from data sources mainly maintained by government authorities which have higher trustability than information from social networks.

3. Trust in semantic integration of data

3.1. The definition of trust

The core concept for the Semantic Web is data integration and use from different sources. The tools for implementing the Semantic Web are designed for encoding data and sharing data from many different sources. In our approach we emphasize the aspect of trust. There are several meanings related to trust [6] when dealing with the process of data integration and trust management. It is viewed diversely in different areas, e.g. in computer networks it is interpreted as security and access control, in distributed systems means reliability and in game theory and policies is viewed as correct decision making under uncertainty. In this paper trust is viewed as a measurable belief that utilizes personal experiences: experiences of others or possibly combined experiences, to make trustworthy decisions about an entity. This view is adapted based on the trust definition given in [12]. In this work a trustworthy decision is assumed to be a transitive process such that there is a web of trust network in which a link between two entities means that a trustworthy decision has been made and the quantitative value of that trust has been evaluated.

3.2. Modelling trust

Our approach is based on RDF language that is a standard model for data interchange on the Web. RDF has features that facilitate data merging even if the underlying schemas differ, and it specifically supports the evolution of schemas over time without requiring all the data consumers to be changed. By supporting RDF, our approach is directly applicable also in the environment of RDFS and OWL which extend RDF.

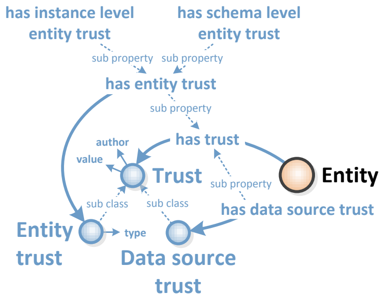


Figure 1. Different types of trust can be defined for each entity.

We define trust for every entity which implies that every RDF triple in our data store has information about trust. Trust management used throughout the paper introduces two independent levels of trust (see figure 1) and trust T of entity e is comprised of:

- **data source trust** $T_{DS}(e)$ and
- **entity trust** $T_E(e)$, which further consists of
  - **schema level entity trust** $T_E^{TBox}(e)$ and
  - **instance level entity trust** $T_E^{ABox}(e)$.

Data source trust $T_{DS}(e)$ defines the level of confidence of data source and is common for all information that is derived from selected data source (e.g. user provided information, collected online have lower level of confidence than information acquired from official databases from public administration sector as depicted in figure 2). Entity trust $T_E(e)$ is defined for every ontological entity (e.g. class, data property, object property, individual etc.) and defines the atomic level of confidence for selected entities.

Trust T of entity e has range $T(e) \in [0,1]$ and is defined as follows:

$$T(e) = T_{DS}(e) \cdot T_E(e) \tag{1}$$

Entity trust $T_E(e)$ is furthermore dependent on schema level entity trust $T_E^{TBox}(e)$ and by crowd sourcing approach of user's votes that result in instance level entity trust $T_E^{ABox}(e)$. Entity trust $T_E(e)$ is defined as follows:

$$T_E(e) = T_E^{TBox}(e) + a \cdot (T_E^{ABox} - T_E^{TBox}) \tag{2}$$

The schema level entity trust $T_E^{TBox}(e)$ is evaluated using the algorithm 1. In calculation the context of entity is included by considering neighbouring entities.

Algorithm 1. Schema level entity trust delegation.

```
$n^{\tau}(e) = 0 \leftarrow$ number of dependent entities of type $\tau$ of entity $e$
$T^{\tau}(e) = 0 \leftarrow$ trust of dependent entities of type $\tau$ of entity $e$

$\boldsymbol{T_E^{TBox}(e)} \vdash$
  for each direct dependent entity $e_i$ of entity $e$ do
    if $type(e_i) = data\ property$ then
      $n^{dp}(e) += 1$
      $T^{dp}(e) += T_E^{TBox}(e_i)$
    else if $type(e_i) = object\ property$ then
      $n^{op}(e) += 1$
      $T^{op}(e) += T_E^{TBox}(range(e_i)) \cdot T_E^{TBox}(e_i)$
    else if $type(e_i) = class$ then
      $n^{C}(e) += 1$
      $T^{C}(e) += T_E^{TBox}(E) \cdot T_E^{TBox}(subClass(e_i))$
    end if
    $T(e) = \frac{T^{dp}(e) \cdot n^{dp}(e) + T^{op}(e) \cdot n^{op}(e) + T^{C}(e) \cdot n^{C}(e) + T_{DS}(e)}{n^{dp}(e) + n^{op}(e) + 1}$
end for
```

The instance level entity trust $T_E^{ABox}(e)$ is furthermore defined as follows:

$$T_E^{ABox}(e) = \frac{n^+}{n^+ + n^-} \tag{3}$$

where n^+ is the number of positive user's votes and n^- is the number of negative user's votes. The parameter a defines to which degree the user's votes are incorporated into entity trust calculation. Parameter a is furthermore defined as follows:

$$a = \begin{cases} 0 & ; \quad n^+ + n^- < t \\ \frac{max(n^+, n^-)}{n^+ + n^-} & ; \quad n^+ + n^- \geq t \end{cases} \tag{4}$$

where t is a threshold that indicates the minimum number of votes so that the instance level trust is relevant.

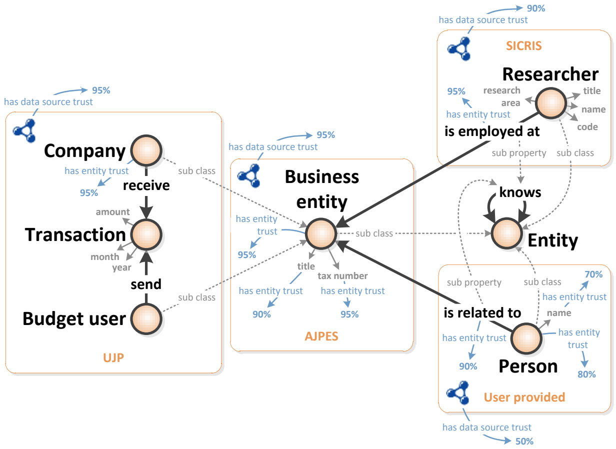


Figure 2. Excerpt of SocioLeaks ontology.

Figure 2 depicts an example from SocioLeaks system, developed within this study. The default values of trust for data source and entities are set to 100%, but users can further define any deviations. The trust for “Business entity” is calculated as $T(Business\ entity) = 92\%$, which considers the data source trust of AJPES and the context of “Business entity” with all corresponding entities within its context. $T(Business\ entity)$ depicts that we can trust the instances of class “Business entity” with the degree of 92%. This information represents the fact described with attributes title and tax number and furthermore a sub concept of entity class.

The trust for “Researcher” is calculated as $T(Researcher) = 90\%$, which considers the data source trust of SICRIS and the context of “Researcher” with all corresponding entities within its context. $T(Researcher)$ depicts that we can trust the instances of class Researcher with the degree of 90%. This information represents the fact described with attributes research area, title, name and code, relation to business entity and furthermore a sub concept of entity class.

The trust for “Person” is calculated as $T(Person) = 64\%$, which considers the data source trust of user provided information and the context of “Person” with all corresponding entities within its context. The lower value of trust is influenced by the fact that user provided data source has a priori trust of $T_{DS}(Person) = 50\%$ and also entity trusts for attributes are set to lower values. Users can elevate the value of trust for any selected entity by using crowd sourcing approach of voting and therefore influencing the trust value in the range of $[0\%, 100\%]$.

4. SocioLeaks case study implementation

4.1. Technology

The SocioLeaks system was implemented using open source technologies of Apache Jena framework that support recent W3C standards in Semantic Web and linked-data applications. The system architecture is defined in several independent layers (see figure 3) that can be further optimized or extended for performance or semantic aspects. The schematic part of knowledge base (TBox component) is stored in-memory, while ontology instances (ABox component) are persisted in native tuple store TDB. The selected graph based triple store supports SPARQL queries with additional support for property functions, aggregates and arbitrary length property paths. For reasoning capabilities Apache Jena Inference layer is included. We used built-in rule reasoner that supports transitive closure with some custom defined capabilities. This chosen setting is a compromise between semantic expressiveness and end user performance.

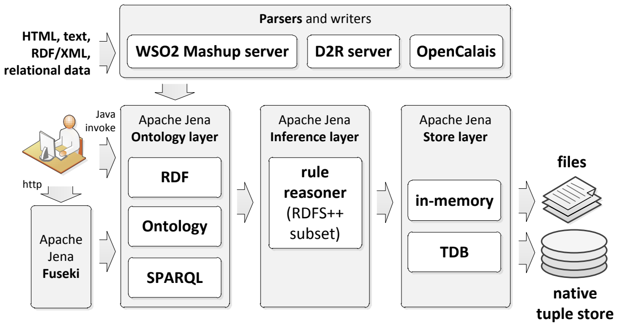


Figure 2. SocioLeaks architecture.

Ontology layer within our architecture supports RDF triples at its ground level and extends it with RDFS and some limited features of OWL (owl:sameAs, owl:inverseOf, owl:TransitiveProperty). Again, this is a compromise to enable greater performance but still regain high level of semantic expressiveness. To import information from external data sources custom parsers were developed for scraping information from mainly HTML presentation of data by using XQuery and regular expression technology. For data publishing, Fuseki, a data publishing server, was used for presenting and updating RDF models over the web using SPARQL and HTTP.

4.2. Ontologies

To support main functionalities of SocioLeaks system, several external data sources were integrated to enable users cross data source search and pattern discovery to find hidden relations between business entities and people. Figure 4 depicts a sub set of data sources used in our system. Every data source has defined data source trust and entity trust for each comprising entity.

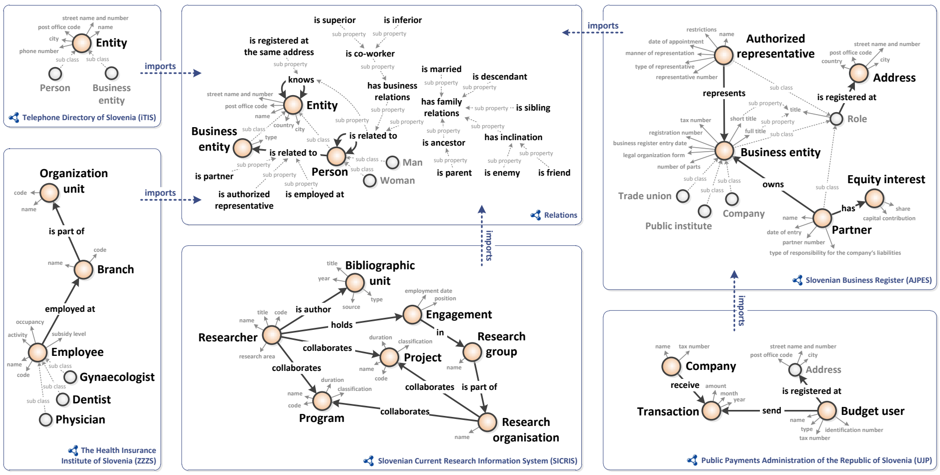
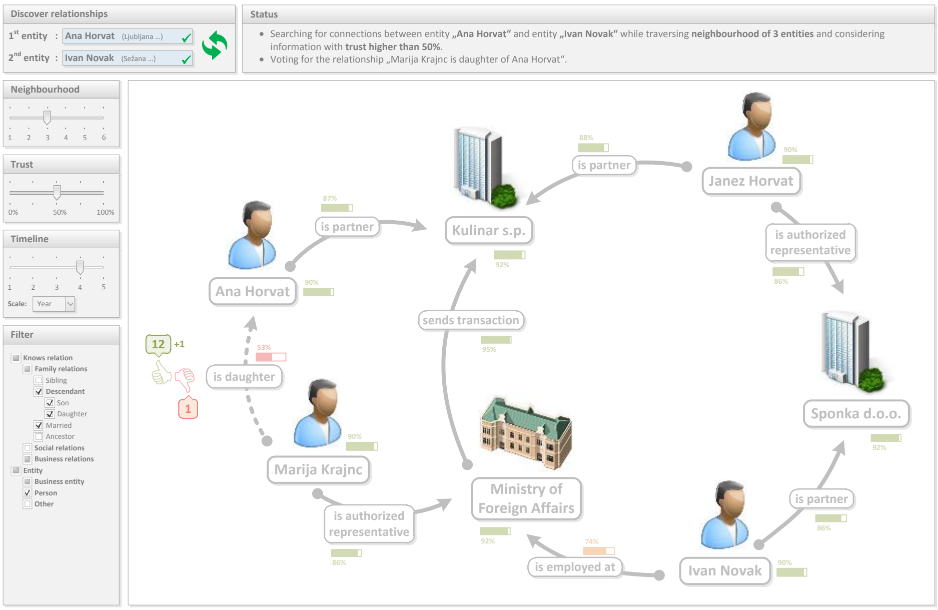


Figure 4. Set of data source ontologies.

Data in SocioLeaks is mainly from government based registers and other community based. There is information about entities from Telephone Directory of Slovenia (iTIS) where name, address and phone numbers of people and business entities can be found. The Health Insurance Institute of Slovenia (ZZZS) is an extensive source of information about physicians, dentists and gynaecologists and bring into system information about their names, occupancy, activity, employment in branches of organization units. Another valuable source of information is Slovenian Business Register (AJPES), where registered business entities can be found with several additional information – tax number, business register entry date, legal organization form, type, address, list of partners and authorized representatives. The AJPES data is further enriched by employment of Public Payments Administration (UJP) where information about money transaction between companies and budget users can be found. The information about researchers, research organizations and their bibliographic units is extracted from Slovenian Current Research Information System (SICRIS). All external data sources are extended by relations ontology, which is also the user entry point to SocioLeaks system. The relations ontology basically defines class “entity” and relation “knows” that relates two entities. Entity is then furthermore partitioned into business entity and person, while knows relations is specialised in several sub relations (has business relations, has family relations, is employed at, is registered at the same address etc.).

4.3. Case study

Currently SocioLeaks system holds almost 8M triples from several data sources as presented in section 4.2. Its main purpose is to enable users’ traversal and relations discovery among business entities and people employing Semantic Web technologies. This process is supported by HTML5 front end where users can search for entities; while results are displayed in graph based manner to enable them the traversal and further relation discovery (see figure 5).



Discover relationships

1st entity : Ana Horvat (Ljubljana ...) ✓

2nd entity : Ivan Novak (Sežana ...) ✓

Status

- Searching for connections between entity **„Ana Horvat“** and entity **„Ivan Novak“** while traversing **neighbourhood of 3 entities** and considering information with **trust higher than 50%**.
- Voting for the relationship „Marija Krajnc is daughter of Ana Horvat“.

Neighbourhood: 3

Trust: 50%

Timeline: 4

Scale: Year

Filter

- ▣ Knows relation
  - ▣ Family relations
    - □ Sibling
    - ☑ Descendant
      - ☑ Son
      - ☑ Daughter
    - ☑ Married
    - □ Ancestor
  - □ Social relations
  - ▣ Business relations
- ▣ Entity
  - ▣ Business entity
  - ☑ Person
  - □ Other

Figure 5. User front end of SocioLeaks system.

Users can either enter one selected entity and display all related associations to some level (set by neighbourhood setting) or enter two entities by finding all existing routes between those two entities and all other that are part of the path. User has the ability to set the filtering with selection of considered entities (business entity, person etc.) and selection of knows relations (son, daughter, married, employed etc.).

There is also additional feature of timeline filtering with selection of timeframes for information we are interested in (e.g. display all relations that were present between selected company and its legal representative during some period of time).

The novelty of our approach is the ability to incorporate trust in visualizing relations between people and business entities. We enable users to select the acceptable level of trust (see section 3) which is then considered in searching the relations and displaying them in a graph based manner. Every entity is accompanied by an indicator (value of trust displayed as progress bar and percentage value). By including the trust information users have more complete information about displayed relations among entities and can make more educated decisions.

Users have also the ability to influence the trust management by voting on selected facts. Figure 5 depicts user action of voting for a fact that Marija Krajnc is a daughter on Ana Horvat. The user vote is taken into account when calculating instance level entity trust (see section 3 for details). By considering the schema level entity trust the overall level of trust for selected entity can increase or decrease, based on the user agreement.

5. Conclusions and future work

In this paper we proposed the use of Semantic Web technologies for semantic integration of data about business entities and people coupled with trust layer. By following this approach we enable users to seamlessly integrate data into web of data and introduce a single point of access by allowing users to perform queries across several data sources considering trust information about every entity. By defining data source ontologies that represent formal view of individual data source we set common ground for integration of unlimited data sources that we might want to include in the future. Data source ontologies are then linked to upper ontology for the domain of business entities and people which can be queried by using SPARQL language and therefore providing users a panoramic and integrated view on data sources. The added value of our approach is introduction of trust management layer that enables filtering the data based on the user preference. Several layers of trust are introduced – data source trust, schema level entity trust and instance level entity trust. By following this approach presented in the paper a priori trust is calculated based on the schema and data source while users are still allowed to influence the trust by employing crowd sourcing approach of voting for individual instances. The future work will include extending the set of integrated data sources and increasing the support for richer semantics by extending RDFS++ subset of supported OWL features. Further work will also focus on performance issues by optimization of triple store layer and improving querying capabilities.

References

[1] H. Barros, A. Silva, E. Costa, Bittencourt, II, O. Holanda, and L. Sales, Steps, techniques, and technologies for the development of intelligent applications based on Semantic Web Services: A case study in e-learning systems, *Engineering Applications of Artificial Intelligence* **24** (2011), 1355-1367.

[2] T. Berners-Lee and O. Lassila, The Semantic Web, *Scientific American* **284** (2001), 34-43.

[3] R. Bodle, REGIMES OF SHARING Open APIs, interoperability, and Facebook, *Information Communication & Society* **14** (2011), 320-337.

[4] A. Brasoveanu, M. Nagy, O. Mateut-Petrisor, and R. Urziceanu, The Avatar in the Context of Intelligent Social Semantic Web, *International Journal of Computers Communications & Control* **5** (2010), 477-482.

[5] J. Dietrich, N. Jones, and J. Wright, Using social networking and semantic web technology in software engineering - Use cases, patterns, and a case study, *Journal of Systems and Software* **81** (2008), 2183-2193.

[6] J. Golbeck, *Trust on the World Wide Web: A Survey*, 2006.

[7] T. Gruber, Collective knowledge systems: Where the Social Web meets the Semantic Web, *Journal of Web Semantics* **6** (2008), 4-13.

[8] A. Harith, Y. Kalfoglou, and N. Shadbolt, Trust strategies for the semantic web, in: *Trust, Security and Reputation Workshop at the ISWC04*, Hiroshima, Japan, 2004, pp. 78-85.

[9] D. Hooijmaijers and M. Stumptner, Trust-annotated ontology integration using social modelling, *Expert Systems* **25** (2008), 237-252.

[10] P. Mika, *Social Networks and the Semantic Web*, Springer, 2007.

[11] M. Richardson, R. Agrawal, and P. Domingos, Trust management for the semantic web, in: *Semantic Web - Iswc 2003*, D. Fensel, K. Sycara, and J. Mylopoulos, eds., Springer-Verlag Berlin, Berlin, 2003, pp. 351-368.

[12] S. Shekarpour and S.D. Katebi, Modeling and evaluation of trust with an extension in semantic web, *Journal of Web Semantics* **8** (2010), 26-36.

[13] R. Valencia-Garcia, F. Garcia-Sanchez, D. Castellanos-Nieves, J.T. Fernandez-Breis, and A. Toval, Exploitation of social semantic technology for software development team configuration, *Iet Software* **4** (2010), 373-385.