Using RDF to Represent Concepts in Academic Domain

Dragoş VESPAN
Department of Economic Informatics and Cybernetics,
The Bucharest University of Economic Studies, Bucharest, Romania
dragos.vespan@ie.ase.ro

Semantic metadata describes documents or entities inside a document, providing information about their content or about the entities inside the document. RDF represents a set of WWW specifications used as data model, whose base unit is represented by a resource-attribute-value triple that overcomes the semantic limitations of XML. This paper shows the advantages of using RDF over XML for representing concepts in academic domain.
Keywords: *XML, RDF, Concept Representation, Semantic Web*

1 Introduction

The development of World Wide Web in the last decade led to an information flux so big that it cannot be managed and processed entirely. For example, Facebook, the social network, hosts more than 140 billion photos that take up more than 14 petabytes [1]. Most of the information on the Web, including that automatically generated, is presented in a way that can be easily read and interpreted by humans.

Although the classic search engines based on key words like Yahoo or Google are the most used tools on the Internet, there are still some problems related to how they are used:

- *High relevance and low precision: even if the search returns relevant documents, these documents are hard to be identified if they are found between other tens of thousands documents that are less relevant for the search conducted;*
- *Results dependent on the vocabulary: relevant documents are not returned as they use a different terminology that the one used in search, but with the same semantic search;*
- *Results found in singular documents: if the information to be found is in more web documents, these documents will have to be searched one by one and the information they contain will be manually correlated.*

The main problem that knowledge acquisition system faces nowadays is represented by the fact that the semantic meaning of documents available on the Internet is not accessible to computers. Such a system cannot make the difference between "I am associate professor at the Economic Informatics Department" or "You may think that I am associate professor at the Economic Informatics department. Well… ".

In order to solve this problem, the semantic meaning of the information contained in documents must be associated with descriptors that can be automatically processed. This idea represents the base of Semantic Web, defined and developed by Berners-Lee, the inventor of WWW, HTTP and HTML [2], [3].

Semantic Web and its technologies provide a new approach for the management of information and processes whose fundamental principle is the development and the use of semantic metadata.

Semantic metadata used in Semantic Web describes documents (a Web page or a PDF document) or parts of documents (title, paragraph) or entities inside a document (persons, companies) and provide information about the content of an object (its subject, its relation to other documents) or about an entity inside the document.

The contribution of Semantic Web in the development of knowledge acquisition systems has two main characteristics: firstly, it provides ontologies that act as a common knowledge database all over the web and, secondly, it provides a logic that shows the way terms are correlated in order to build complex concepts and the way these concepts interact with already acquired

knowledge. In this context, ontologies have the role of a universal dictionary in order for all the documents to have the same interpretation of the concepts they contain and of the information they provide.

Knowledge acquisition techniques have the purpose to discover new knowledge by identifying new structures in data analyzed. These techniques may be used to project data sources, like collections of text documents, onto an ontological structure. Document classification may be applied in ontology development when there is a set of predefined categories (e.g. medicine, education, and informatics) and a set of documents which is already categorized into these categories. These categories may be structured in ontologies, like the MeSH ontology for the medical domain or the Yahoo! hierarchy for web documents.

2 Knowledge Representation

Knowledge may be represented through explicit specification of knowledge objects and of relations between these objects. Knowledge representation allows machines to reconfigure and reuse information they store in ways that are not necessarily previously specified. Examples of knowledge representations are: *concept mapping*, *semantic networks* or *conceptual indexing* [4].

Concept mapping is based on educational techniques for improving understanding and memorizing. A map of concepts represents an image of the ideas or subjects found in information, together with the way these ideas and subjects are related one to each other. A document concept mapping represents a visual summary which shows the structure of the information described inside the document.

Semantic networks are deeply associated with the detailed analysis of documents and with networks of ideas. The nodes of the network represent classes or abstract sets of entities grouped on certain common characteristics or properties. These entities are concept instances and the edges inside the semantic network represent relations between concepts. The hypertext may be described as a semantic network whose nodes are represented by documents.

Conceptual indexing maps the ideas and the key objects from a document through indexes of concepts. These indexes represent structured sequences resulted from the deep and complete analysis of a document, which contain links to all the information contained in the text. The structure of the indexes allows the users to locate the information quickly and efficiently by organizing concepts from documents using parent-child, synonym or similar relation types.

The Web is the most important and the most used environment for many-to-many knowledge exchange. Thus, a knowledge representation language for the Web must accomplish a series of requests related to the exchange format:

- *universal description capacity: because not all of the potential uses of knowledge may be anticipated, a Web based interchange format must describe any type of knowledge;*
- *syntactic interoperability: software applications should be able to analyze the data carrying knowledge and should be able to get a representation of these data that can be exploited;*
- *semantic interoperability: concepts found in data should potentially be associated which requires a context analysis.*

XML (eXtensible Markup Language) [5] has the capacity of universal description and can represent anything that has a defined grammar. XML was created to overcome HTML drawbacks, especially related to the lack of efficiency in complex client-server communication management and to the impossibility to define new tags in order to solve users' needs.

XML is a content oriented language, with a very simple syntactical structure. Its tagging elements are identified by opening tags (<professor>) and closing tags (</professor>) and may be specialized by adding attributes and values to the opening tag. By using this structure, XML meets also the property of syntactic interoperability as any XML parser

may analyze any XML data being, usually, a reusable component.
With respect to the semantic interoperability, XML has some disadvantages: there is no way to recognize a semantic unit in a certain domain as XML deals only with the structure of the document and does not provide a common interpretation of data contained in a document.
The main idea of Semantic Web is to provide support for creating a Web distributed at data level and not at information presentation level. In order to implement Semantic Web, it is required to have a data model that allows knowledge distribution over the Internet.
The classical model for data representation is represented by tables, where each line represents a described element, each column represents an attribute of the described element and each cell represents an attribute value of the element. Table 1 presents an example of data regarding the curricula for the third year, first semester, of Economic Informatics specialization.

Table 1. Data table that represents the curricula for third year, first semester, of Economic Informatics specialization

No.	Name of discipline	Type of discipline	Form of evaluation	Credit points	Working group managing the discipline
1	Economy of information	A	E	2.00	Economic Cybernetics II
2	Multimedia	O	V	2.00	Advanced Programming Languages
3	Object oriented programming	O	E	2.00	Advanced Programming Languages
4	Economic informational systems	O	V	2.00	Computer Programming
5	Data structures	O	E	2.00	Advanced Programming Languages
6	Decision theory	A	E	2.00	Economic Cybernetics II

Modelling the strategy of data distribution over the Internet, data in Table 1 can be represented on different machines using multiple strategies. Such a strategy is represented by distributing data on rows to different machines (Fig. 1). Each machine in the network is responsible to keep the information contained in one or more rows of the table. Any query related to an entity represented by a certain discipline can be solved by returning the appropriate row. In this strategy, the server should indicate at global level which column describes a certain property.

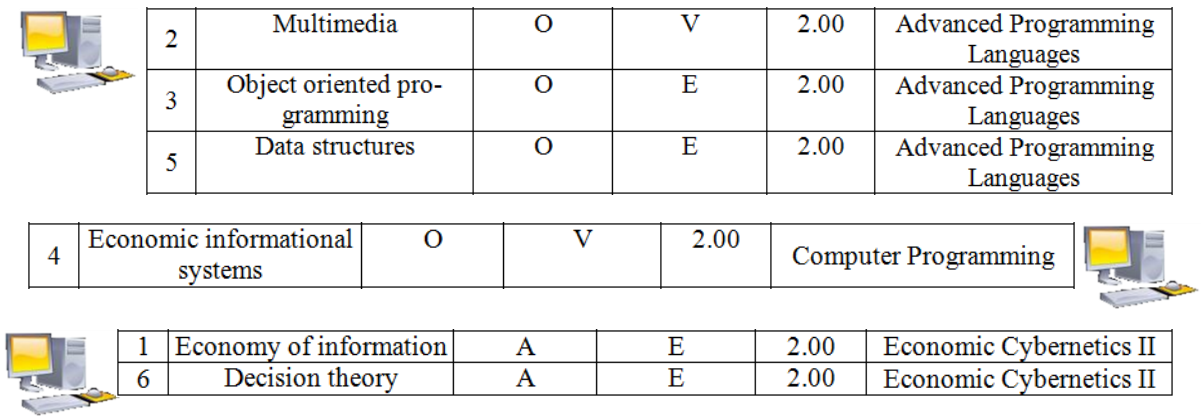
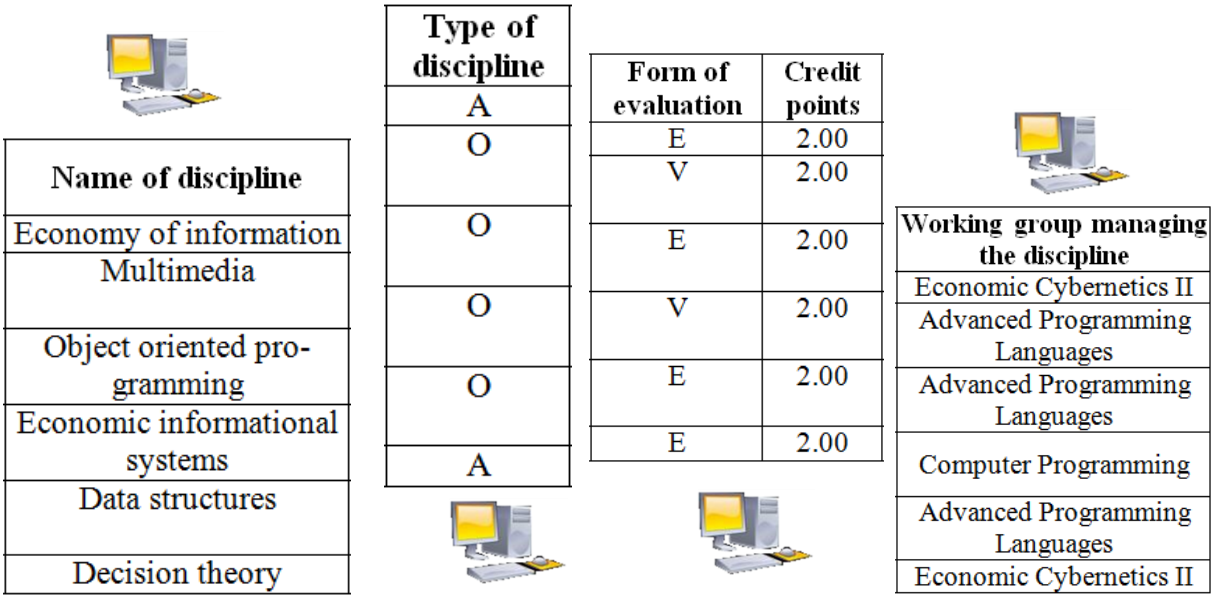


Fig. 1. Data distribution on rows

Fig. 2 represents another strategy for data distribution, where each machine is responsible for one or more columns of the original table. This solution allows each machine to be responsible of a certain type of information. If we are not interested in the type of discipline, then we will not consider the information on the machine that stores this information. Also, if we want to add new information about entities (like number of course hours and laboratory hours for each discipline) then we can add a new machine to the network without affecting the existing ones. This strategy is based on the description of an entity identifier at global level.



Name of discipline
Economy of information
Multimedia
Object oriented pro-gramming
Economic informational systems
Data structures
Decision theory

Type of discipline
A
O
O
O
O
A

Form of evaluation	Credit points
E	2.00
V	2.00
E	2.00
V	2.00
E	2.00
E	2.00

Working group managing the discipline
Economic Cybernetics II
Advanced Programming Languages
Advanced Programming Languages
Computer Programming
Advanced Programming Languages
Economic Cybernetics II

Fig. 2. Data distribution on columns

A combination of the two strategies, in which data is distributed neither on rows nor on columns, is the cells data distribution (Figure 3). Each machine is responsible for a certain number of cells in the table. Two machines may share both the description of a single entity and the use of a certain property. This strategy combines the flexibility of the strategies described above but also the costs associated. Global references are required for each column header and also for each row.

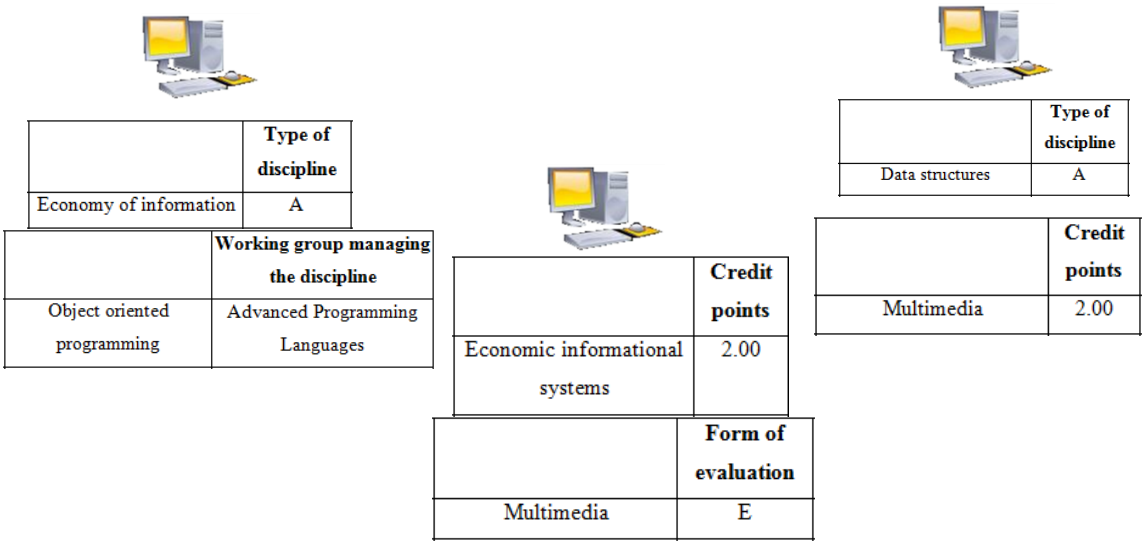


Fig. 3. Data distribution on cells

Thus, each cell is represented by three values: a global reference for the row, a global reference for the column and the actual value of the cell. This third strategy is used by Resource Description Framework (RDF) for representing data on the internet.

4 Resource Description Framework

RDF represents a set of WWW specifications used as data model, whose base unit is represented by a resource-attribute-value tuple. Resources are represented by row identifiers, attributes are represented by column identifiers and values are stored in the cells of the table.

RDF is a general-purpose language for representing information on the Web [6], is domain independent and allows defining a specific terminology through RDF Schema (RDFS). This schema defines the vocabulary used in RDF data models by specifying properties that may be applied to objects and by describing relations between objects [7]. For instance, we may write:

memberOfWorkingGroup is a subclass of memberOfDepartment.

This sentence states that all members of working groups are members of departments so there is an associated meaning for the statement "is a subclass of". This statement must be the same for all RDF based applications and should not be differently interpreted by each application.

The importance of RDF Schema is emphasized in Fig. 4, where automated search of members of department returns only the name Dragoş V. This result is correct from the XML point of view but is not sufficient from the semantic point of view. A person would have included, implicitly, Paul P and Marian D in the category of members of department, because:

- all members of working groups are members of department (memberOfWorkingGroup is a subclass of memberOfDepartment);
- courses are taught only by members of working groups.

```
<memberOfWorkingGroup> Dragos V </memberOfWorkingGroup>
<memberOfDepartment> Paul P </memberOfDepartment>
<course name="Multimedia">
        <isTaughtBy> Marian D </isTaughtBy>
</course>
```

Fig. 4. XML elements for defining entities in educational organization

This type of knowledge uses the semantic model of the academic domain and cannot be represented in XML or RDF, but is representative for the knowledge described in RDF Schema. This way, RDF Schema allows the semantic information to be accessible to machines, in concordance with the vision of Semantic Web [8].

The main concepts of RDF are resources, properties and statements. Resources are objects (authors, disciplines, institutions, and places) associated to an identifier unique on the Internet, called Uniform Resource Identifier (URI). An URI may be represented by any unique type of identifier (URL, web Address) and does not automatically provide the access to the resource it refers to. URI schemas may be defined both for Web locations and for other objects, like phone numbers, ISBN numbers or geographical locations.

Properties describe relations between the resources (i.e. "taught by", "written by", "title", and "testing method") and are also identified through URI. The idea of using URI to identify resources and relations between them is very important since it provides a unique global scheme for naming them.

Statements associate properties and resources. A statement is represented by a resource-attribute-value triple, where values may be other resources, numbers or strings. This way, resources are described through attributes which relate objects (represented by resources) to values of the attributes which may include links to other resources [9].

An example of statement is:

Marian D is teaching the Multimedia course

The easiest way to interpret this information is the use of the tuple (http://www.dice.ase.ro/courses/multimedia, isProfessorOf , http://www.dice.ase.ro/members/ MarianD). This tuple (x,P,y) may be represented through a logical formula P(x,y) where the binary predicate P connects the subject x and the object y [10].

Table 2 describes a set of tuples through subject, predicate and object:

Table 2. Representation of statements using tuples

Subject	Predicate	Object
Economy of information	Type of discipline	A
Object oriented programming	Working group managing the discipline	Advanced Programming Languages
Economic informational systems	Credit points	2.00
Multimedia	Form of evaluation	E
Data structures	Type of discipline	A
Multimedia	Credit points	2.00

When more tuples refer the same entity (Table 3), they can be represented through a directed graph where each tuple represents an edge from its subject towards its object and the predicate is the label of the edge.

Table 3. Subject-predicate-object tuples that represent resources

Subject	**Predicate**	**Object**
D.I.C.E.	Teaches	General Informatics
C.S.I.E.	Includes	D.I.C.E.
D.I.C.E.	Publishes	Journal of Economy Informatics
Dragoş V	Member of	D.I.C.E.
General Informatics	Taught by	Dragoş V

General Informatics	Managed by	Informatics Working Group
Dragoş V	Director of	Knowledge acquisition through text mining
Knowledge acquisition through text mining	Financed by	C.N.C.S.I.S
Dragoş V	Member of	Informatics Working Group
Dragoş V	Publishes in	Journal of Economy Informatics
Informatics Working Group	Part of	D.I.C.E.

The graph in **Error! Reference source not found.** presents the same information as Table 3, everything that we know about *Dragoş V* (either as subject or as object) being displayed in a single node. This graph is the semantic network.

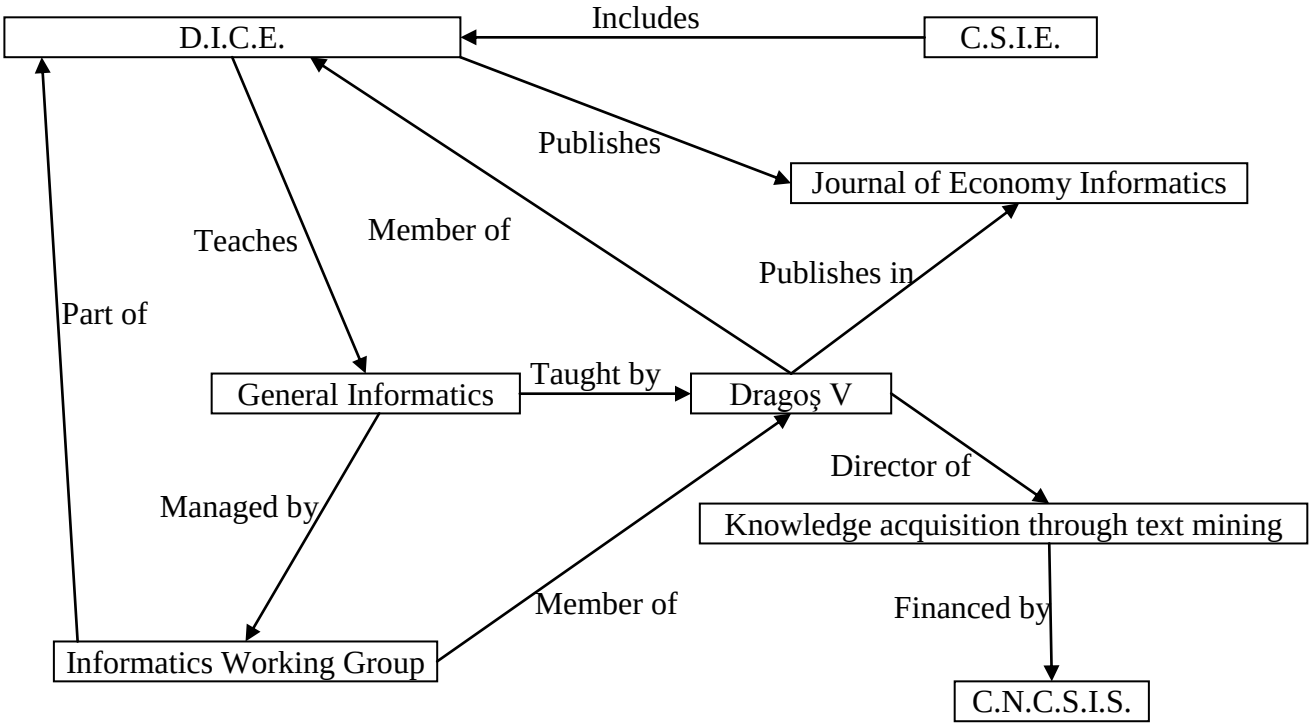


Fig. 5. Representation through a directed graph of tuples in Table 3

The Semantic Web provides representations of knowledge which may be accessed and processed by machines. In order to accomplish this, RDF tuples may be represented in XML. A RDF document may be represented trough a XML element with the tag *rdf:RDF*. The content of this element is represented by a number of descriptors which use *rdf:Description* tags. Each such description represents a statement about a resource.

```
<?xml version="1.0" encoding="UTF-16"?>
<rdf:RDF
        xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
        xmlns:ent="http://info.ase.ro/onto/entity"
        xmlns:edu="http://info.ase.ro/onto/educational-organization">
        xmlns:geo="http://www.w3.org/2003/01/geo/wgs84_pos#"
        <rdf:Description rdf:about="http://www.csie.ase.ro">
                <geo:lat>44.4477</geo:lat>
                <geo.long>26.0990</geo:long>
                <edu:includes rdf:resource="http://www.dice.ase.ro"
                        ent:name="D.I.C.E."/>
        </rdf:Description>
</rdf:RDF>
```

Fig. 5. XML RDF code to describe

Inside an RDF XML document there are two types of nodes: resource nodes and property nodes. Resource nodes are objects or subjects of statements and have an *rdf:about* attribute holding the URI of the resource it represents. Resource nodes may contain only property nodes representing statements. In the code in Fig. 5, there is a single resource node *rdf:Description* and three statements with the subject <http://www.csie.ase.ro> and the predicates *geo:lat, geo:long* şi *edu:includes*. Property nodes contain alphanumeric values („44.4477", „26.0990", „D.I.C.E.").

5 Conclusions

The main advantage of using Semantic Web is that anyone can create a language by simply publishing the information using RDF in order to describe a set of URIs, their purpose and how they should be used. Semantic Web is based on the lowest power principle: as fewer rules as better. Thus, Semantic Web is very less restrictive in expression.

The most important benefit brought by publishing information using RDF is represented by the fact that, once available on a public domain access, the scope of the information may be easily changed. Using URIs for describing resources results in a decentralization of the way the information is published: there should not be any central authority to publish all data and its description language.

The imbricated structure resource-attribute-value satisfies the request for universal description that a knowledge representation language must fulfil. Also, RDF accomplishes the request of interoperability because there are a lot of application independent RDF parsers available. Regarding the semantic interoperability, RDF has a significant advantage over XML: the structure resource-attribute-value provides natural semantic units because all objects are independent entities.

References

[1] C. Boja, A. Pocovnicu and L. Batagan, „Distributed Parallel Architecture for "Big Data"," *Informatica Economica*, vol. 16, nr. 2, pp. 116-127, 2012.

[2] T. Berners-Lee, J. Hendler and O. Lassila, „The Semantic Web," *The Semantic Web*, http://www.sciam.com/article.cfm?id=the-semantic-web, 2001.

[3] W3, „Tim Berners-Lee," 2008. [Online]. Available: http://www.w3.org/People/Berners-Lee/.

[4] J. N. D. Gupta and S. K. Sharma, „An Overview of Knowledge Management," in *Knowledge Management: Concepts, Methodologies, Tools, and Applications*, IGI Global, 2008, pp. 1-22.

[5] T. Bray, J. Paoli, C. Sperberg-McQueen, E. Maler and F. Yergeau, „Extensible markup language (XML) 1.1," 15 04 2004. [Online]. Available: http://www.w3.org/TR/2004/REC-xml11-20040204/.

[6] L. van Ruijven, „Ontology for Systems Engineering," *Procedia Computer Science*, vol. 16, pp. 383-392, 2013.

[7] A. Doan, A. Halevy and Z. Ives, „12 - Ontologies and Knowledge Representation," in *Principles of Data Integration*, Boston, Morgan Kaufmann, 2012, pp. 325-344.

[8] G. Antoniou and F. van Harmelen, Semantic Web Primer, second edition,

Boston: Massachusetts Institute of Technology, 2008.

[9] W. Hesse, "Ontologies in the Software Engineering process," *Proceedings of the Workshop on Enterprise Application Integration EAI*, 2005.

[10] A. Zimmermann, N. Lopes, A. Polleres and U. Straccia, „A general framework for representing, reasoning and querying with annotated Semantic Web data," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 11, pp. 72-95, 2012.

Dragos VESPAN is Lecturer at Department of Economic Informatics and Cybernetics of the Bucharest University of Economic Studies from Bucharest, Romania. In 2002 he graduated the Faculty of Cybernetics, Statistics and Economic Informatics at the Bucharest University of Economic Studies of Bucharest and since 2008 he has a PhD degree on artificial intelligence. Also, he is certified as IPMA level C in Project Management by Romanian Project Management Association. His work focuses on artificial intelligence, text mining and internet technologies.