Krassimir Markov, Vitalii Velychko, Oleksy Voloshin
(editors)

# Information Models of Knowledge

I T H E A®
K I E V – S O F I A
2010

Krassimir Markov, Vitalii Velychko, Oleksy Voloshin (ed.)

Information Models of Knowledge

ITHEA®

Kiev, Ukraine – Sofia, Bulgaria, 2010

ISBN 978-954-16-0048-1

First edition

Recommended for publication by The Scientific Concil of the Institute of Information Theories and Applications FOI ITHEA

ITHEA IBS ISC: 19.

This book maintains articles on actual problems of research and application of information technologies, especially the new approaches, models, algorithms and methods fot information modeling of knowledge in: Intelligence metasynthesis and knowledge processing in intelligent systems; Formalisms and methods of knowledge representation; Connectionism and neural nets; System analysis and sintesis; Modelling of the complex artificial systems; Image Processing and Computer Vision; Computer virtual reality; Virtual laboratories for computer-aided design; Decision support systems; Information models of knowledge of and for education; Open social info-educational platforms; Web-based educational information systems; Semantic Web Technologies; Mathematical foundations for information modeling of knowledge; Discrete mathematics; Mathematical methods for research of complex systems.

It is represented that book articles will be interesting for experts in the field of information technologies as well as for practical users.

General Sponsor: Consortium FOI Bulgaria (www.foibg.com).

Printed in Ukraine



**ISBN 978-954-16-0048-1**

C\o Jusautor, Sofia, 2010

# INTEGRATION OF FINANCIAL DOMAIN KNOWLEDGE ON BASE OF SEMANTIC WEB TECHNOLOGIES

**Anatoly Gladun, Julia Rogushina, Rodrigo Martínez-Béjar, Francisco García-Sanchez and Rafael Valencia-García**

***Abstract**: The increasingly huge volume of financial information found in a number of heterogeneous business sources is characterized by unstructured content, disparate data models and implicit knowledge. As Semantic Web Technologies mature, they provide a consistent and reliable basis to summon financial knowledge properly to the end user. In this paper, we present SEFSS, a semantically enhanced financial search engine empowered by semi-structured crawling, inference-driven and ontology population strategies bypassing the present state-of-the-art technology caveats and shortcomings.*

***Keywords**: Semantic Web Technologies, financial knowledge, ontology, Web-services, data crawling*

***ACM Classification Keywords**: I.2.4 Knowledge Representation Formalisms and Methods - Semantic networks, H.3.3 Information Search and Retrieval - Retrieval models*

## Introduction

The increasing availability of financial data collected by banks and other organizations attached to specific applications, pushes the need of real time access very urgent. The optimal solution would allow accessing financial data the way we currently browse unstructured information on the web with popular search engines. Unfortunately, this is not yet possible, since data, residing in thousand incompatible formats, not limited to different data technologies or supporting software, is also incompatible at semantic level. The problem, referred as semantic non-interoperability, is due to differences as the "world-view" level, since related data fields across repository possibly contain data with different meaning, coding scheme or format. Considering this issue on a much broader scale, it applies to thousands of data structures within thousands of databases and messaging formats, spread across the planet. All Semantically-Empowered Financial Search Systems (SEFSS) aims to address these issues by adopting a state-of-the-art semantic knowledge management approach, by extracting meaning from financial data, and unleashing hidden relationships between related but incompatible datasets.

SEFSS will provide access to such relationships by developing a semantically-empowered financial search engine: financial data, crawled from structured and unstructured information sources both publicly available on the Internet and provided by private corporate information sources, will be modeled and semantically annotated. The process will be mostly automated, adopting a supervised information extraction approach based on natural language processing, in order to properly index information fed by harvesting the available sources and extract significant metadata. An improved access level will be provided by exploiting the potential of semantic classification and inference on a domain ontology developed within the project. The lightweight RDF format, used for representing metadata and relationships, will ease this task, especially on the inference and retrieval side. On the language side, although only the handling of English language information will be supported, a particular care will be given to make the system upgradeable to support other EU languages in an easy way.

## Concept and objectives

The need to manage financial data has been coming into increasingly sharp focus for some time. Years ago, these data sat in silos attached to specific applications in banks and financial companies. Then the Web came into the arena, bringing the hurly-burly of data becoming available across applications, departments and entities in general. However, throughout these developments, a particular underlying problem has remained unsolved: data reside in thousands of incompatible formats and cannot be systematically managed, integrated, unified or cleansed. To make matters worse, this incompatibility is not limited to the use of different data technologies or to

the multiple different "flavors'" of each technology (for example, the different relational databases in existence), but also because of its incompatibility in terms of semantics.

The main goal of this work is to develop a semantically empowered financial search engine platform. This SEFSS is designed to gather financial data from very different sources and store them semantically annotated in a repository. This information gathering will occur from both public information sources such as the Internet and from corporate private information sources. Users will then be provided with different services for accessing the data. These services will take advantage of the machine-readable semantic annotations of the financial information in order to provide more sophisticated high-quality functionality to the system's users.

SEFSS includes two functional modules:

1) *Analyst Information Assistant*

The mission of the Analyst Information Assistant Module will be to gather information from such resources about financial information, including financial information web sites, economic opinion sources, financial information aggregators, blogs and any related Web source.

2) *Trader Information Decision*

Searching and integrating data from various sources has become a fundamental issue in financial research, particularly in those fields where massive data gathering is faced. The reason for this is that the need for information integration in such fields is critical, preserving by all means the semantics inherent to the different data sources and formats. In terms of the financial investment and trading domain, such integration would permit to organize properly data fostering the analysis and access of such information to accomplish critical tasks such as investment influence, trading history and analysis of investing trends at a particular time. The mission of this module will be to bring these techniques to its full potential.

It is also important to highlight two major issues regarding the platform functionality. First of all, the system will be capable of accessing both (semi-)structured and natural language-based data sources, thus embracing most of the data available on the Internet. Secondly, although at a first stage SEFSS will focus exclusively on financial data sources in English, the system to be developed will be designed so that it is straightforward to upgrade it to handle other languages like Spanish or German.

Using the vast consortium expertise in semantic, knowledge acquisition and agents technologies, SEFSS concentrates on building an innovative solution for search and management of financial data. To achieve such an ambitious goal a set of functionalities need to be developed starting with *extraction of relevant meaning from structured and unstructured information*, and ending with *information search.*

In development of SEFSS we use a lot of IT and artificial intelligence results:

- Advanced knowledge management systems for information-bound organizations and communities, capable of extracting actionable meaning from structured and unstructured information and social interaction patterns, and of making it available for activities ranging from information search through conceptual mapping to decision making.
- Semantic foundations: probabilistic, temporal and modal modeling and approximate reasoning through objective-driven research moving beyond current formalisms.
- Service architectures, platforms, technologies, methods and tools that enable context-awareness and discovery, advertising, personalization and dynamic composition of services.

## Semantic Web

The project described in this proposal involves the development of a software application based on four cornerstone technologies: the Semantic Web and ontologies, agent technology, logical reasoning and inference, and knowledge acquisition from texts. Next, these technologies state-of-the-art is described and the advances this project aims to bring about are detailed.

***a) Semantic Web and Ontologies (data crawling and storing)***

The World Wide Web (WWW), also called "the Web", was invented in 1989 by Tim Berners-Lee and his colleagues at CERN and it changed the way people gather and access information. Nowadays, the Web is an ever-growing huge data repository. As a consequence, a major bottleneck has emerged when trying to exploit the information represented in the Web, namely, how to find a specific piece of information we may be interested in., ontologies are the backbone technology of Semantic Web [Semantic Web]. Ontologies provide a common vocabulary of an area and define – with different levels of formality - the meaning of the terms and the relations between them [Fensel, 2002]. OWL (Web Ontology Language) (Web Ontology Working Group) is the *de facto* Semantic Web standard ontology language [Maedche, 2001]. A major problem for the success of the Semantic Web vision is the difficulty associated to the semantic annotation of the information already available on the Web. A key contribution of this work is the development of Web crawlers that, in a (semi-)automated way, are able to create RDF [RDF model, 1999] triples from structured, semi-structured and non-structured data on the Web.

Financial Data Crawling and Storage is based on a software component for the crawling, analysis and storage of financial data represented in several formats, extracting RDF metadata records to be used for search and retrieval, according to a Financial Ontology that serves as a unifying data model. This Financial Ontology is found with the requirements of the case studies and a number of alternatives are envisaged to make it consistent and self-contained.

***b) Agent Technology***

The intelligent agents and multiagent systems area has received ever-increasing attention by researchers over the last few years. *Agents* are the computer systems in charge of carrying out this task. Agents can be useful as stand-alone entities that are delegated particular tasks on behalf of a user. However, in the majority of cases agents exist in environments that contain other agents, constituting multiagent systems (MASs) [Gladun, 2005]. A MAS can be seen as a system consisting of a group of agents that can potentially interact with each other [Gladun , 2006].

Nowadays, the agent community is facing the problem of integrating agent technology with Web services. The SEFSS project will lead to the development of a reference model by means of which both intelligent agents and Web Services can seamlessly communicate to each other, so being able to cooperate in a global environment. We use MAS paradigm because financial domain is too complex to be processed by alone agent [Gladun,2006].

***c) Reasoning, Inference Engine***

Ontology reasoning is a research area intensively investigated in the recent years. Most of the techniques and inference engines developed for Semantic Web data are focusing either on reasoning over instances of an ontology with rules support (e.g. Rule based approaches) or on reasoning over ontology schemas (DL reasoning) [Baader, 2003]. The examples of such tools are Pellet [Pellet, 2009], FaCT++ [FaCT++ , 2009] - is the new generation of the well-known FaCT OWL-DL reasoner, KAON2 [KAON2, 2009] - is an infrastructure for managing OWL-DL, SWRL, and F-Logic ontologies.

Reasoning over instances of an ontology, for example, can derive a certain value for an attribute applied to an object. These inference services are the equivalent of SQL query engines for databases, but they provide stronger support (for example, handling of recursive rules). Reasoning over concepts of an ontology, for example, can automatically derive the correct hierarchical location of a new concept in a given concept hierarchy. Nowadays also the integration of rule and DL based reasoning approaches gathered a lot of attention. However most of the approaches are considering rather static information, which does not apply to a dynamic domain such as the financial sector. In SEFSS we plan to build reasoning techniques which deal in a scalable way with the dynamism of the financial sector.

***d) Knowledge Acquisition for Ontologies from Natural Language Texts***

There are different techniques for ontology learning: symbolic techniques, statistical techniques and machine learning techniques. There are two main trends in ontology building: (i) manual ontology building, and (ii) (semi-) automatic ontology building (i.e. ontology learning). Ontology population is the process through which a given

ontology is populated with instances. SEFSS needs to import data from a wide range of channels, being it structured or unstructured [Valencia-Garcia, 2004]. We plan to design component for the crawling, analysis and storage of financial data represented in several formats, extracting RDF metadata records to be used for search and retrieval [Rogushina, 2006].

This tool gets the reference ontology and natural language free texts as inputs and, using a set of NLP tools, it obtains a populated ontology. The ontology population process is comprised of two main phases: NLP phase and Population Phase. The scenario could be as follows (see Fig. 1):

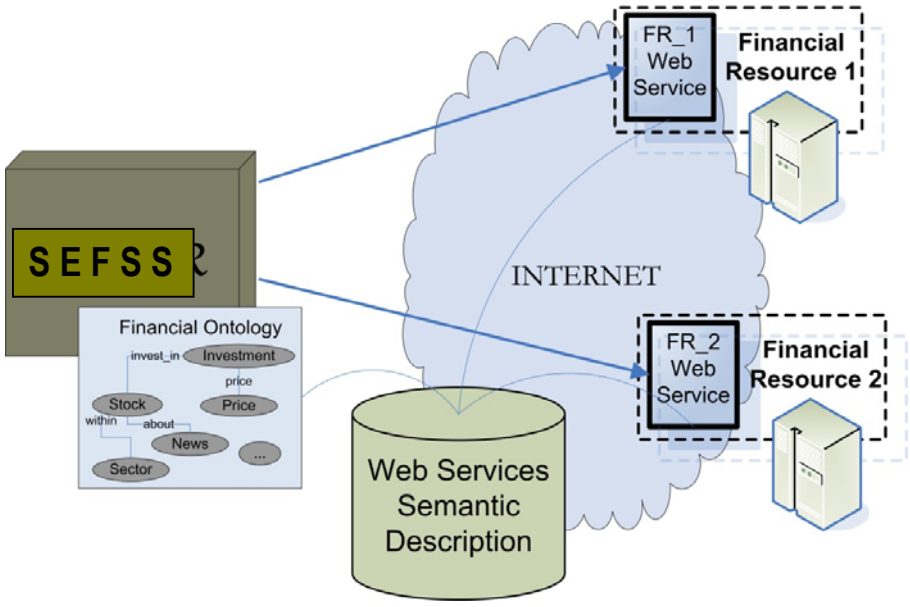


Fig. 1. Crawling the Web for financial information.

The ontology population methodology, based on existing research results from the University of Murcia, will be used for implementing an interface for assisted or supervised processing of free text information, and significant effort will be put to make the process as automated as possible. This methodology gets a reference ontology and natural language free texts as inputs and using a set of NLP tools obtain a populated ontology (see Fig. 2). This ontology must be verified by an expert.

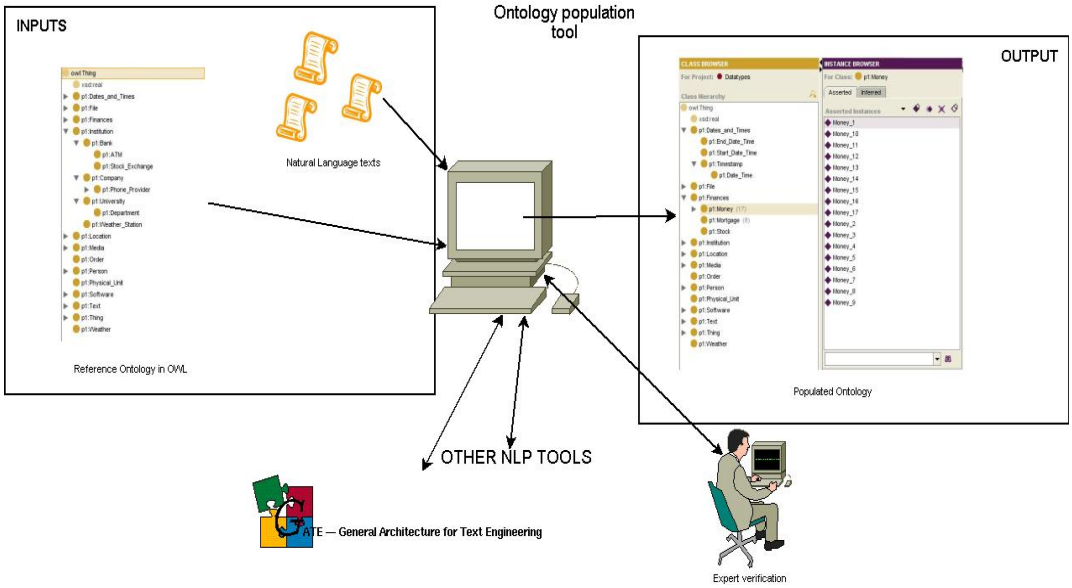


Fig. 2. Ontology Population Tool.

## Related Work

Web 2.0 is unleashing a number of possibilities that, combined with the Semantic Technologies, could result in significant success . Since the work on improving search results spans over and binds together a number of research initiatives, in this section we briefly describe related work.

Searching has been subject of intensive research but a more concrete survey on filtering search results and optimizing results yields also a remarkable amount of efforts. Following research successfully implemented in the Google search engine, a number of search variants related to the work presented have been explored such as using faceted search, including its application to multimedia faceted metadata for image search and browsing or navigating RDF data.

Wherever possible, research results will be exploited for the internal development and support of new products and services. These products and services will lead to a competitive advantage of the participating organizations and will create a substantial benefit for the targeted users. In order for the exploitation to be effective, an integrated approach will be necessary, combining experience and expertise from the development department and solution management, and the involvement of a user base [represented by the consortium partners - if there are no user partners: which will be involved through dedicated surveys]. Integral part of the exploitation approach is the identification of a [one or more] use case which will serve as the validation point throughout the project - from customer requirements, to building the demonstrator up to the evaluation of the results.

The project will create exploitation plans for the results of individual participants and for the consortium as a whole in order to ensure that the developed technologies have a significant impact in the market and that they do not remain as theoretical developments. The SEFSS project will include a specific component and a work package devoted to exploitation issues. The major results of the exploitation work package are the Market Analysis and the Exploitation Plan. As part of the management structure, an Exploitation Management Board will ensure the successful execution of the exploitation work package and will ensure that the other work packages take into account results from the exploitation work package.

The industrial partners within the project have clear exploitation routes for the technology in a number of ways, and their individual exploitation intentions will be stated in the exploitation plan. The industrial partners are all experienced in generating and protecting intellectual property. In some cases software, notably certain software modules generated by the academic partners, may be open-source, which will further encourage the take-up of the technology.

The SEFSS exploitation strategy consists in outline of:

- Tracking important commercial and technical developments in the Semantic Web.
- Analyzing more deeply and in detail of the financial sector.
- Identifying the results of SEFSS that are exploitable: semantically-empowered financial search engine platform.
- Identifying the appropriate distribution channels to exploit the different results of the project.
- Analyzing the impact of the financial search engine platform in terms of quantitative and qualitative impact, when targeting the deployment and integration of the platform in financial entities. This evaluation will be based on real-world implementations.

Different activities will be sought to represent the project actively and establish liaisons with other relevant projects, standard organizations, and institutions that can be of benefit for the project. The different partners will establish contacts with other companies outside the consortium and preparing the market for the technology adoption. These derived products, contacts, potential users and exploitation plans will be documented in a final exploitation plan report.

Coming up, details on the intentions of the project participants in the dissemination activities are:

Metaware has a clear exploitation path for foreseen project results, as a way to add more powerful and insightful technologies and approaches into the existing Business Intelligence products and solutions. The direct peering with top financial and pubic bodies, thanks to a strong background in the financial services, may also allow providing interesting opportunities for further extension of the results, and technological transfer.

## Conclusions and discussion

Conventional wisdom holds that new Semantic Technologies promise a powerful paradigm for solving integration problems with current state-of-the-art IT infrastructure such as financial search engines and applications. However, there has SEFSS not been significant progress in terms of real developments, particularly because of the Semantic Technologies problem: Web information or data source providers would always request for a good excuse or reason, a good application or benefit from providing metadata. However, if the metadata is not generated, no application or value-added functionality can be achieved. But the explosive growth of a number of structured metadata formats in blogs, wikis, social networking sites and online communities has transformed the Web in recent years. Mainstream media has taken notice of the so-called Web 2.0 revolution and business success stories have gathered stream.

These technologies are blooming overnight and providing “metadata farms” whose potential can be unleashed by Semantic applications, which will gain momentum towards a Web generation in which Semantic and Web 2.0 technologies will end up meeting.

Our future work focuses on tuning and optimizing that crossway and enabling the transition from research and academic prototypes to practice and from standards to deployment. Tool vendors and manufacturers are reluctant to implement products until they see a market forming, but we envisage the market needs as a powerful driving force to make these solutions mature and business oriented.

## Bibliography

[Semantic Web] Semantic Web Challenge. - http://challenge.semanticweb.org/.

[Fensel , 2002] Fensel, D.: Ontologies: A Silver Bullet for Knowledge Management and Electronic Commerce. Springer, Heidelberg (2002).

[Maedche , 2001] Maedche, A., Staab, S.: Ontology learning for the semantic web. IEEE Intelligent Systems 16(2), 72–79 (2001).

[RDF Model, 1999] RDF Model and Syntax Specification. W3C Proposed Recommendation. - January 1999. - http://www.w3.org/TR/PR-rdf-syntax.

[Gladun, 2005] Anatoly Gladun and Julia Rogushina: Ontologies as a Perspective Direction of Intellectualization of Informational Retrieval in Multiagent Systems of E-commerce // The Proceedings of XI-th Intern. Conf. "KDS’2005”, Varna, Bulgaria, pp.112-120.

[Gladun , 2006] Gladun A. , Rogushina J. Ontological Approach to Domain Knowledge Representation for Informational Retrieval in Multiagent Systems // International Jornal "Information Theories & Applications", V.13, N.4, 2006. – P.354-362.

[Baader , 2003] Baader, F., Calvanese, D., McGuinness, D.L., Nardi, D., Patel-Schneider, P.F. (eds.): The Description Logic Handbook. Cambridge University Press, Cambridge (2003).

[Pellet:, 2009] Pellet: OWL 2 Reasoner for Java.- http://clarkparsia.com/pellet .

[FaCT++ , 2009] FaCT++ - http://owl.man.ac.uk/factplusplus/ .

[KAON2, 2009] KAON2 - http://kaon2.semanticweb.org/

[Valencia-Garcia, 2004] Valencia-Garcia, R., Ruiz-Sanchez, J.M., Vicente, P.J.V., Fernandez-Breis, J.T., Martinez-Bejar, R.: An incremental approach for discovering medical knowledge from texts. Expert Syst. Appl. 26(3), 291–299 (2004)

[Rogushina , 2006] Rogushina J., Gladun A.: "Semantic Search of Internet Information Resources on Base of Ontologies and Multilinguistic Thesauruses" // International Journal «Information Theories and Applications», vol.14, 2006.-P.117-129.

## Authors' Information

***Dr. Anatoly Gladun*** *– senior scientific researcher, Associate Professor, International Research and Training Centre of Information Technologies and Systems, National Academy of Sciences and Ministry of Education of Ukraine, 44 Glushkov Pr., Kiev, 03680, Ukraine ; e-mail: glanat@yahoo.com .*

*Major Fields of Scientific Research: Intellectualization of computer networks, Intelligent Software Agents and multiagent systems, Semantic Web technologies, Information Retrieval.*

***Dr. Julia Rogushina*** *– senior scientific researcher, Associate Professor, Institute of Software Systems of National Acedemy of Sciences of Ukraine, 44 Glushkov Pr., Kiev, 03680, Ukraine; e-mail: _jjj_@ukr.net*

*Fields of Scientific Research: Semantic Web technologies, knowledge management, data mining. knowledge acquisition.*

***Dr. Rodrigo Martínez-Béjar*** *- Full Professor, Departamento de Informatica y Sistemas, Univeridad de Murcia, Spain, rodrigo@um.es .*

*Major Fields of Scientific Research: General theoretical information research, Multi-dimensional information systems.*

***Francisco Garcia-Sanchez - ,*** *Departamento de Informatica y Sistemas, Univeridad de Murcia, Spain, fgarcia@um.es .*

*Major Fields of Scientific Research:*

***Rafael Valencia-Garcia -,*** *Departamento de Informatica y Sistemas, Univeridad de Murcia, Spain, valencia@um.es .*

*Major Fields of Scientific Research:*