Semantic Interoperability in Multi-Disciplinary Domain. Applications in Petroleum Industry

Jon Atle Gulla[1] and **Darijus Strasunskas**[1] and **Stein L. Tomassen**[1]

Abstract. The petroleum industry is a technically challenging business with high investments, complex projects and operational structures. There are numerous companies and public offices involved in the exploitation of a new oil field, and there is a high degree of specialization among them. Even though standardization has been considered important in this industry for many years, there is still very little integration across phases and across disciplines. An industrially driven consortium launched the Integrated Information Platform project in 2004, in which semantic standards based on OWL and Semantic Web technologies were to be developed for the subsea petroleum industry. In this paper, we present the IIP project in more detail and discuss applications for semantic information interoperability and retrieval.

1 INTRODUCTION

The petroleum industry in Norway is technically challenging with subsea installations and difficult climatic conditions. It is industrially still quite fragmented, in the sense that there is little collaboration between phases and disciplines in large petroleum projects. There are many specialized companies involved, though their databases and applications are not necessarily well integrated with each other. Research done by the Norwegian Oil Industry Association (OLF) shows that there is a need for more collaboration and integration across phases, disciplines and companies [1]. The existing standards do not provide the necessary support for this, and the result is costly and risky projects and decisions based on wrong or outdated data.

This paper presents the Integration Information Platform (IIP) project [2] and preliminary results. The project's goal is to extend and formalize an existing terminology standard for the petroleum industry, ISO 15926. Using Semantic Web technologies, we turn this standard into a real ontology that provides a consistent unambiguous terminology for subsea petroleum production systems. However, creating and maintaining ontologies is both time-consuming and costly. Consequently, ontologies are applied for many different tasks to increase return on investment (ROI). Therefore, the IIP project focuses on reuse of ontologies in traditional vector-space information retrieval (IR) systems, in addition to rules-based notification. Considering multi-disciplinary domain and a big variation of terminology used one of the challenges is adoption of the created ontology to the document space. Finally, it is necessary to consider how ontologies will be used in those applications, i.e. application specific ontology value is an important concern in IIP.

The paper is structured as follows. In Section 2 we go through the structures and challenges in the subsea petroleum industry, explaining the status of current standards and the vision of future integrated operations. In Section 3 the IIP project is briefly introduced. Whereas in Section 4, we discuss chosen approaches. Finally, the conclusions are drawn in Section 5.

[1] Dept. of Computer & Information Science, Norwegian University of Science & Technology (NTNU), NO-7491 Trondheim, Norway;
email: Jon.Atle.Gulla@idi.ntnu.no, Darijus.Strasunskas@idi.ntnu.no, Stein.L.Tomassen@idi.ntnu.no

2 THE SUBSEA PETROLEUM INDUSTRY

The Norwegian subsea petroleum industry is a technically challenging business. Sophisticated equipment and highly competent companies are needed, and the projects tend to be both large and expensive. Many disciplines and competences need to come together in these projects, and their success is highly affected by the way people and systems are able to collaborate and coordinate their work. On the Norwegian Continental Shelf (NCS) there are traditional oil companies, specialized service companies and smaller ICT service companies. The multidiciplinarity of the industry causes in various perspectives towards the domain, and contextual usage of different terminologies. One of the challenges is to deal with contextual information and multi-perspective data integration in the multidisciplinary industry.

Both the projects and the subsequent production systems are information-intensive. When a well is put into operation, the production has to be monitored closely to detect any deviation or problems. The next generation subsea systems include numerous sensors that measure the status of the systems and send real-time production data back to certain operation centers. For these centers to be effective, they need tools that allow them to understand this data, relate it to other relevant information, and help them deal with the situation at hand. There is a challenge in dealing with all this information, but also in interpreting information that is deeply rooted in various technical terminologies.

The multitude of companies involved, with their own applications and databases, makes coordination and collaboration more important than in the past. For the industry as a whole, this severely hampers the integration of applications and organizations as well as the decision making processes in general:

- **Integration**. Even though there is some cooperation between companies in the petroleum sector, this cooperation tends to be set up on an ad-hoc basis for a particular purpose and supported by specifically designed mappings between applications and databases. There is little collaboration across disciplines and phases, as they usually have separate databases rooted in different goals, structures and terminologies. It is of course possible to map data from one database to another, but with the complexity of data and the multitude of companies and applications in the business this is not a viable approach for the industry as a whole.
- **Decision making**. A current problem is the lack of relevant high-quality information in decision making processes. Some data is available too late or not at all because of lack of integration of databases. In other cases relevant data is not found due to differences in terminology or format. And even when information is available, it is often difficult to interpret its real content and understand its limitations and premises. This is for example the case when companies report production figures to the government using slightly different terminologies and structures, making it very hard to compare figures from one company to another.

XML is already used extensively in the petroleum industry as a syntactic format for exchanging data. Over the last few years, there

have been several initiatives for defining semantic standards to achieve semantic interoperability and information sharing in the business.

2.1 ISO 15926 Integration of Life-Cycle Data

ISO 15926 is a standard for integrating life-cycle data across phases (e.g. concept, design, construction, operation, decommissioning) and across disciplines (e.g. geology, reservoir, process, automation). It consists of 7 parts, of which part 1, 2 and 4 are the most relevant to this work. Whereas part 1 gives a general introduction to the principles and purpose of the standard, part 2 specifies the modeling language for defining application-specific terminologies. Part 2 comes in the form of a data model and includes 201 entities that are related in a specialization hierarchy of types and sub-types. It is intended to provide the basic types necessary for defining any kind of industrial data. Being specified in EXPRESS [3], it has a formal definition based on set theory and first order logic.

Part 4 of ISO 15926 is comprised of application or discipline-specific terminologies, and is usually referred to as the Reference Data Library (RDL). These terminologies, described as RDL classes, are instances of the data types from part 2, are related to each other in a specialization hierarchy of classes and sub-classes as well as through memberships and relationships. If part 2 defines the language for describing standardized terminologies, part 4 describes the semantics of these terminologies. There is ongoing work in the Norwegian offshore industry to provide a comprehensive standardized terminology for the petroleum industry in part 4. Part 4 today contains approximately 50.000 general concepts like motor, turbine, pump, pipes and valves.

ISO 15926 is still under development, and only Part 1 and 2 have so far become ISO standards. In addition to adding more RDL classes for new applications and disciplines in Part 4, there is also a discussion about standards for geometry and topology (Part 3), procedures for adding and maintaining reference data (Part 5 and 6), and methods for integrating distributed systems (Part 7). Neither ISO 15926 nor other standards have the scope and formality to enable proper integration of data across phases and disciplines in the petroleum industry.

2.2 The Vision of Integrated Operations

The Norwegian Oil Industry Association proposed the Integrated Operations program in 2004. The fundamental idea is to integrate processes and people onshore and offshore using new information and communication technologies. Facilities to improve onshore's abilities to support offshore operationally are considered vital in this program. Personnel onshore and offshore should have access to the same information in real-time and their work processes should be redefined to allow more collaboration and be less constrained by time and space. OLF has estimated that the implementation of integrated operations on the NCS can increase oil recovery by 3-4%, accelerate production by 5-10% and lower operational costs by 20-30% [1].

Central in this program is the semantic and uniform manipulation of heterogeneous data. Decisions often depend on real-time production data, visualization data, and background documents and policies, and the data range from highly structured database tables to unstructured textual documents. This necessitates intelligent facilities for capturing, tracking, retrieving and reasoning about data.

Figure 1 illustrates the objectives of the integrated operations initiative. Whereas we in the current situation have numerous databases that need to be mapped to each other on an ad hoc basis, we envision a semantic standard in the future that supports integration and interoperability between data from all phases and disciplines. Suppliers's applications interact with the operators' data through standardized semantic interfaces, making sure that a unified terminology is used and data is consistent and unambiguous. The implementation requirements for integrated operations include the introduction of proper standards for efficient sharing and exchange of information.

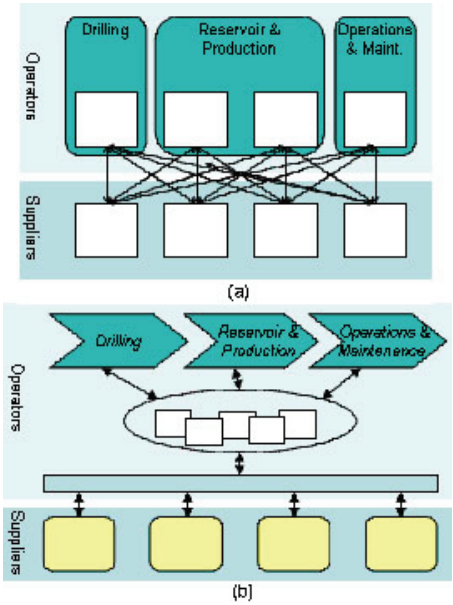

Figure 1. (a) Current situation; (b) The vision of integrated operations

3 THE INTEGRATED INFORMATION PLATFORM PROJECT

The Integrated Information Platform (IIP) project is a collaboration project between companies active on NCS and academic institutions, supported by the Norwegian Research Council (NFR). Its long-term target is to provide high quality real-time information for decision making at onshore operation centers.

The IIP project addresses the need for a common understanding of terms and structures in the subsea petroleum industry. The objective is to ease the integration of data and processes across phases and disciplines by providing a comprehensive unambiguous and well accepted terminology standard that lends itself to machine-processable interpretation and reasoning. This should reduce risks and costs in petroleum projects and indirectly lead to faster, better and cheaper decisions.

The project is identifying an optimal set of real-time data from reservoirs, wells and subsea production facilities. The OWL web ontology language is chosen as the markup language for describing these terms semantically in an ontology. The entire standard is thus rooted in the formal properties of OWL, which has a model-theoretic interpretation and to some extent support formal reasoning. A major part of the project is to convert and formalize the terms already defined in ISO 15926 Part 2 (Data Model) and Part 4 (Reference Data Library), which we will come back to in the next Section. Since the ISO standard addresses rather generic concepts, though, the ontology must also include more specialized terminologies for the oil and gas segment. Detailed terminologies for standard products and services are included from other dictionaries and initiatives (DISKOS, WITSML, ISO 13628/14224, SAS), and the project also opens for the inclusion of terms from particular proc-

esses and products at the bottom level. In sum, the ontology being built in IIP has a structure as shown in Figure 2.

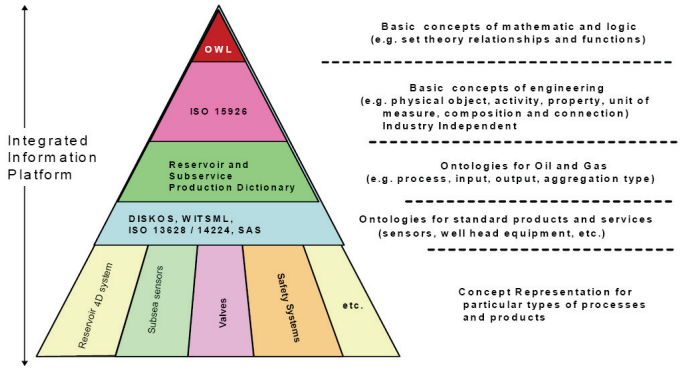

Figure 2. The standardization approach in IIP

4 APPROACH AND DISCUSSION

The success of the new ontology, and standardization work in general, depends on the users' willingness to commit to the standard and devote the necessary resources. If people do not find it worthwhile to take the effort to follow the new terminology, it will be difficult to build up the necessary support. This means that it is important to provide environments and tools that demonstrate the value of using the ontology. Intelligent ontology-driven applications must demonstrate the benefits of the new technology and convince the users that the additional sophistication pays off.

Recall, the multidisciplinary settings of the petroleum industry. The multidisciplinarity results in different views on the domain followed by vast terminology variation between disciplines, e.g. oil companies, specialized service and ICT service companies. Non-consistent usage of terminology causes the problems in documents exchange among the industrial partners (see illustration in left part of Figure 2). Furthermore, the variation in terminology may prohibit successful commitment to the ontology and its adoption in daily work routines. Therefore, we propose an approach to bridge the gap among terminologies by constructing a feature vector for each of the concepts in the ontology (see right part of Figure 3).

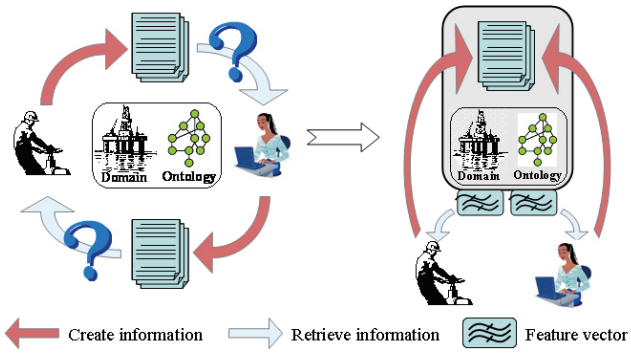

Figure 3. Multidisciplinary interpretation of domain and ontology

Development of the approach is inspired by a linguistics method for describing the meaning of objects – the semiotic triangle (known as triangle of meaning or Ogden's triangle, as well) [4]. In our approach, a feature vector connects a concept and a document collection (Figure 4), i.e., the feature vector is tailored to the terminology used in a particular collection of the documents (that is company or discipline specific). The construction of feature vector is further explained in section 4.3 and [5].

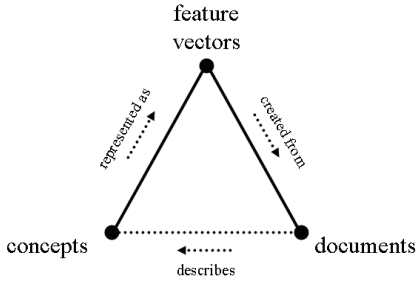

Figure 4. Explanation of a feature vector by adapted semiotic triangle

4.1 Semantic Web Technology and Interoperability

The general idea in the Semantic Web is to annotate each piece of data with machine-processable semantic descriptions. These descriptions must be specified according to a certain grammar and with reference to a standardized domain vocabulary. The domain vocabulary is referred to as an ontology and is meant to represent a common conceptualization of some domain. The grammar is a semantic markup language, as for example the OWL web ontology language recommended by W3C. With these semantic annotations in place, intelligent applications can retrieve and combine documents and services at a semantic level, they can share, understand and reason about each other's data, and they can operate more independently and adapt to a changing environment by consulting a shared ontology.

Interoperability can be defined as a state in which two application entities can accept and understand data from the other and perform a given task in a satisfactory manner without human intervention. We often distinguish between syntactic, structural and semantic interoperability [6, 7]:

- *Syntactic interoperability* denotes the ability of two or more systems to exchange and share information by marking up data in a similar fashion (e.g. using XML).
- *Structural interoperability* means that the systems share semantic schemas (data models) that enable them to exchange and structure information (e.g. using RDF).
- *Semantic interoperability* is the ability of systems to share and understand information at the level of formally defined and mutually accepted domain concepts, enabling machine-processable interpretation and reasoning.

For the Semantic Web technology to enable semantic interoperability in the petroleum industry, it needs to tackle the problem of *semantic conflicts,* also called *semantic heterogeneity*. Since the databases are developed by different companies and for different phases and/or disciplines, it is often difficult to relate information that is found in different applications. Even if they represent the same type of information, they may use formats or structures that prevent the computers from detecting the correspondence between data.

4.2 Industrial Ontologies

In recent years a number of powerful new ontologies have been constructed and applied in selected domains. This is particularly true in medicine and biology, where Semantic Web technologies and web mining have been exploited in new intelligent applications [6, 8, 9]. However, these disciplines are heavily influenced by government support and are not as commercially fragmented as the petroleum industry. Creating an industry-wide standard in a

fragmented industry is a huge undertaking that should not be underestimated. In this particular case, we have been able to build on an existing standard, ISO 15926. This has ensured sufficient support from companies and public institutions. There is still an open question, though, what the coverage of such an ontology should be. There are other smaller standards out there, and many companies use their own internal terminologies for particular areas. The scope of this standard has been discussed throughout the project as the ontology grew and new companies signalled their interest. For any standard of this complexity, it is important also to decide where the ontology stops and to what extent hierarchical or complementing ontologies are to be encouraged. Techniques for handling ontology hierarchies and ontology alignment and enrichment must be considered in a broader perspective.

4.3 Ontology-driven Information Retrieval

For an Information Retrieval tool developed in IIP, we are adding a mechanism to adopt the ontology with the words used in particular discipline (i.e. by particular company) [5]. Figure 5 illustrates the overall architecture of the ontology-based information retrieval system. The individual components of the system will be given a brief account.

Feature vector miner: This component associates concept from the ontology with relevant terms from the document space. An ontology concept is a class defined in the ontology being used. These concepts are extended into *feature vectors* with a set of relevant terms extracted from the document collection using text-mining techniques. The *feature vectors* provide interpretations of concepts with respect to the document collection and needs to be updated as the document collection changes. This allows us to relate the concepts defined in the ontology to the terms actually used in the document collection.

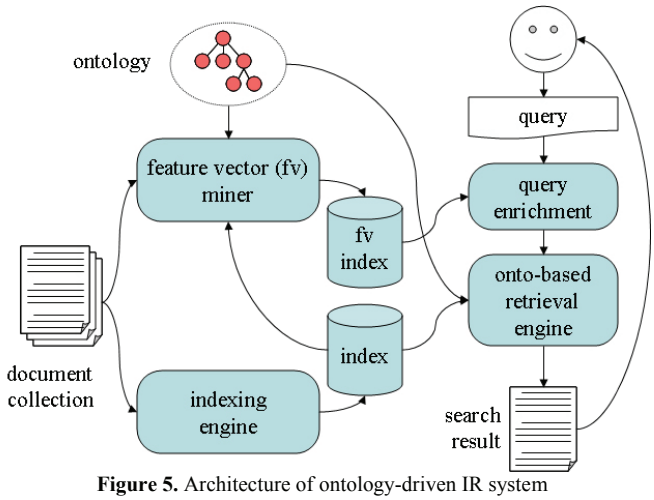

Figure 5. Architecture of ontology-driven IR system

Indexing engine: The main task of this component is to index the document collection. The indexing system is built on top of Lucene, which is a freely available and fully featured text search engine from Apache. Lucene is using the traditional vector space approach, counting term frequencies, and using *tf.idf* scores to calculate term weights in the index.

Query enrichment: This component handles the query specified by the user. The query can initially consist of concepts and/or ordinary terms (keywords). Each concept or term can be individually weighted. The concepts are replaced by corresponding feature vectors.

Onto-based retrieval engine: This component performs the search and post-processing of the retrieved results.

4.4 Rule-based Notification

Since the Semantic Web is still a rather immature technology, there are still open issues that need to be addressed in the future. One problem in the IIP project is that we need the full expressive power of OWL (OWL Full) to represent the structures of ISO 15926-2/4. Reasoning with OWL specifications is then incomplete and inference becomes undecidable [10]. Here we consider investigate the limits of inference using the ontology implemented in OWL Full. This will allow identifying possible scenarios and restrictions in using OWL Full for a such scale project. This is important, since one of the application areas is specification of rules that will be used to analyze anomalies in real-time data from subsea sensors. At that point we will need to exploit the logical properties of OWL and start experimenting with the next generation rule-based notification systems.

4.5 Application-specific Ontology Value

The quality of ontologies is a delicate topic. It is important to choose an appropriate level of granularity. In this project we have been fortunate to have an existing standard to start with. What was considered satisfactory in ISO 15926 may however not be optimal for the ontology-driven applications that will make use of the future ontology. Ultimately, we need to consider how the ontology will be used in these applications.

The ontology value quadrant [11] in Figure 6 is used to evaluate an ontology's usefulness in a particular application. The ontology's ability to capture the content of the universe of discourse at the appropriate level of granularity and precision and offer the application understandable correct information are important features that are addressed in many ontology/model quality frameworks (e.g. [12, 13, 14, 15]). But the construction of the ontology also needs to take into account dynamic aspects of the domain as well as the behavior of the application. For Ontology-driven Information Retrieval this means that we need to consider the following issues about content and dynamics [11]:

Concept familiarity. Terminologies are used to subcategorize phenomena and make semantic distinctions about reality. Ideally the concepts preferred by the user in his queries correspond to the concepts found in the ontology.

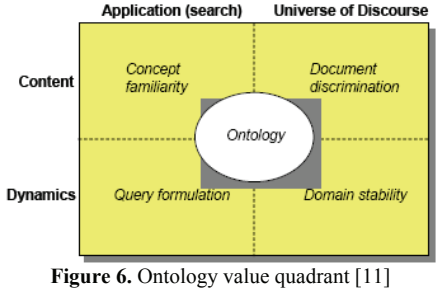

Figure 6. Ontology value quadrant [11]

Document discrimination. The structure of concepts in the ontology decides which groups of documents in the collection can theoretically be singled out and returned as result sets. Similarly, the concepts preferred by the user indicate which groups of documents she might be interested in and which distinctions between documents she considers irrelevant. If the granularity of the user's preferred concepts and the ontology concepts are perfectly compatible,

combinations of these terms can single out the same result sets from the document collection.

Query formulation. The user queries are usually very short, like 2-3 words, and specialized or generalized terms tend to be added to refine a query [16]. This economy of expression seems more important to users than being allowed to specify detailed and precise user needs, as very few use advanced features to detail their query.

Domain stability. The search domain may be constantly changing, and parts of the domain may be badly described in documents compared to others. The ontology needs regular and frequent maintenance, making it difficult to depend on the availability of domain experts.

5 CONCLUSIONS

The Integrated Information Platform project is one of the first attempts at applying state-of-the-art Semantic Web technologies in an industrial setting. Existing standards are now being converted and extended into a comprehensive OWL ontology for reservoir and subsea production systems. The intention is that this ontology will later be approved as an ISO standard and form a basis for developing interoperable applications in the industry.

With the new ontology at hand, the industry will have taken the first step towards integrated operations on the Norwegian Continental Shelf. Data can then be related across phases and disciplines, helping people collaborate and reducing costs and risks. However, there are costs associated with building and maintaining such an ambitious ontology. It remains to be seen if the industry is able to take full advantage of the additional expressive power and formality of the new ontology. The work in IIP indicates that both information retrieval systems and sensor monitoring systems can benefit from having access to an underlying ontology for analyzing data and interpreting user needs.

One of the main applications developed in IIP is an ontology-driven information retrieval system [5]. Here, the concepts in the ontology are associated with contextual definitions in terms of weighted feature vectors tailoring the ontology to the content of the document collection. Further, the feature vector is used to enrich a provided query. Query enrichment by feature vectors provides means to bridge the gap between query terms and terminology used in a document set, and still employing the knowledge encoded in ontology.

Also, we can build more complete semantic descriptions of documents and add more reasoning capabilities to our information retrieval tools. We will then see if a strong semantic foundation makes it easier for us to handle and interpret the vast amount of data that are so typical to the petroleum industry.

Main future work is an inclusion of rules to be used to analyze anomalies in the real-time data from the subsea sensors. Then we will need to evaluate and investigate the logical properties of OWL and start experimenting with the next generation rule-based notification systems.

Acknowledgements

This research work is funded by the Integrated Information Platform for reservoir and subsea production systems (IIP) project, which is supported by the Norwegian Research Council (NFR). NFR project number 163457/S30.

REFERENCES

[1] OLF. Integrated Work Processes: Future work processes on the Norwegian Continental Shelf. The Norwegian Oil Industry Association. URL: http://www.olf.no/?28867.pdf (Accessed: 2006 03 05).

[2] Sandsmark, N.; Mehta, S. Integrated Information Platform for Reservoir and Subsea Production Systems. In Proceedings of the 13th Product Data Technology Europe Symposium (PDT 2004), Stockholm, (2004).

[3] International Standards Association. Industrial automation systems and integration – Product data representation and exchange. Part 11: Description methods: The EXPRESS language reference manual. URL: http://www.iso.org/iso/en/CatalogueDetailPage.CatalogoueDetail?CSNUMBER=18348. (Accessed: 2006 03 05)

[4] Ogden, C.K., and Richards, I.A. *The Meaning of Meaning*. 8th Edition, New York, Harcourt, Brace & World, Inc., (1923).

[5] Tomassen, S.L.; Gulla, J.A.; Strasunskas, D. Document Space Adapted Ontology: Application in Query Enrichment. Proceedings of 11th International Conference on Applications of Natural Language to Information Systems (NLDB'2006), Klagenfurt, Austria, Springer-Verlag, *LNCS 3999*, 46-57, (2006).

[6] Aguilar, A. Semantic Interoperability in the Context of e-Health. 2005. CDH Seminar. URL: http://m3pe.org/seminar/aguilar.pdf (Accessed: 2006 03 05).

[7] Dublin Core. Dublin Core Metadata Glossary. http://library.csun.edu/mwoodley/dublincoreglossary.html. (2004).

[8] Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. Nature Genet. **25**, 25-29, (2000).

[9] Pisanelli, D. M. (Ed.). Ontologies in Medicine. Volume 102 Studies in Health Technology and Informatics. IOS Press. (2004).

[10] Horrocks, I., Patel-Schneider, P. and F. van Harmelen. From SHIQ and RDF to OWL: The making of a web ontology language. *Journal of Web Semantics* **1**(1), 7–26, (2003).

[11] Gulla, J. A.; Borch, H.O.; Ingvaldsen, J.E. Unsupervised Keyphrase Extraction for Search Ontologies. . Proceedings of 11th International Conference on Applications of Natural Language to Information Systems (NLDB'2006), Klagenfurt, Austria, Springer-Verlag, *LNCS 3999*, (2006).

[12] Burton-Jones, A.; Storey, V.C.;Sugumaran, V.; Ahluwalia, P. A Semiotic Metrics Suite for Assessing the Quality of Ontologies. *Data and Knowledge Engineering*, **55**(1), 84-102, (2005).

[13] Gangemi, A.; Catenacci, C.; Ciaramita, M.; Lehmann, J. Ontology evaluation and validation. An integrated formal model for the quality diagnostic task. Technical Report, ISTC-CNR, Trento, Italy, (2005). URL: http://www.loa-cnr.it/Files/OntoEval4OntoDev_Final.pdf (Accessed: 2006 04 02)

[14] Lindland, O. I., Sindre, G., Solvberg, A.: Understanding Quality in Conceptual Modeling. *IEEE Software*, **11** (2), 42-49, (1994).

[15] Lozano-Tello, A.; Gomez-Perez, A. Ontometric: A method to choose appropriate ontology. *Journal of Database Management*, **15**(2), 1-18, (2004)

[16] Gulla, J.A., Auran, P.G., Risvik, K.M.: Linguistic Techniques in Large-Scale Search Engines. Fast Search & Transfer, (2002) 15 p.