Towards a Methodology for Lipoprotein Ontology

Meifania Chen, Maja Hadzic
Research Lab for Digital Health Ecosystems, Curtin University of Technology, Perth, Australia
m.chen@curtin.edu.au, m.hadzic@curtin.edu.au

Abstract

Abnormal plasma lipoprotein levels have been found to be significantly correlated to cardiovascular disease, the leading cause of mortality and morbidity worldwide. In addition, lipoprotein dysregulation, known as dyslipidemia, is a central feature in disease states such as diabetes and hypertension, which also increases the risk to cardiovascular disease. Despite progress in lipoprotein research, a vast number of the world population suffers from dyslipidemia. One of the major challenges that researchers face is the difficulties in accessing and integrating relevant information amidst massive quantities of heterogeneous data. Ontologies target these problems by providing a semantic framework of the concepts involved in a system of related instances to support systematic querying of information, data mining, as well as form the basis for collaboration between research teams. Lipoprotein Ontology will provide the basis for the design of various applications to enable interoperability between research groups or software agents, as well as the development of tools for the diagnosis and treatment of dyslipidemia. In this paper we present a nine-step methodology for the design of Lipoprotein Ontology. This methodology can be adapted for use in the design of other domain-specific ontologies.

1. Introduction

Lipoproteins are a water-soluble "lipid+protein" complex which serves as a mode of transport for the uptake, storage and metabolism of lipids. The basic lipoprotein structure comprises of a hydrophobic core of triglycerides and cholesteryl esters, surrounded by a hydrophilic outer layer of phospholipids, cholesterol and apolipoproteins. Lipoproteins play a very crucial role in the regulation of biological and cellular functions in humans, and can be impacted by a number of factors, including obesity, diet/nutrition, physical exercise and other factors such as smoking and alcohol consumption. Lipoproteins have been extensively researched since the first isolation of high density lipoproteins (HDL) in 1929 and low density lipoproteins (LDL) in 1950 [1]. In addition, the lipoprotein transport system has been well established since the 1960s [1]. However, in spite of these advances, there is an increasing prevalence of dyslipidemia.

Clinical and epidemiological studies have identified lipoprotein dysregulation as a significant contributor to cardiovascular disease, the leading cause of death in the world today [2]. While a large corpus of literature exists on lipoprotein research, researchers face challenges in navigating through overwhelming amounts of information. The advent of semantic web technologies, specifically ontologies, targets these problems by enabling the assimilation of data and extracting relevant information into effective and efficient problem-solving tools.

Ontologies are a medium of knowledge representation which captures and conceptualises a domain in terms of its associated concepts and instances. They provide a mechanism for sharing a common vocabulary in a domain to facilitate information exchange and are the basis for intelligent retrieval of information. Consequently, ontologies are becoming increasingly relevant in life sciences, as evident from the emergence of a number of biomedical ontologies [3, 4]. Some of these ontologies include the Gene Ontology, Protein Ontology, Lipid Ontology, among others [5, 6, 7].

The application of an ontology to the domain of lipoprotein research is a response to the need to share and reuse the complex and heterogeneous sources of information available. Ontologies provide a semantic framework of the concepts involved in entities and processes in a system of hierarchical and associative relations which supports systematic querying of information. Through a common vocabulary, ontologies present a platform for the integration of data which facilitates collaboration between research groups or software agents. By creating annotations linking primary data to expressions in controlled, structured

978-1-4244-9166-7/10/$26.00 ©2010 IEEE

vocabularies, developing an ontology for lipoprotein pathways will have a positive impact for ongoing lipoprotein research by making the data available to effective searching and algorithmic processing.

In Section 2, we discuss current ontologies in the biomedical domain. Section 3 covers existing tools which use ontologies for intelligent information retrieval and extraction. We then present a nine-step methodology for the design of Lipoprotein Ontology in Section 4. Section 5 provides a brief overview of lipoprotein ontology and Section 6 concludes the paper.

2. Ontologies in biomedical science

Biomedical research is increasingly becoming a data-driven endeavour, with large volumes of complex information derived from different sources, with different structures and different semantics. There have been efforts in the recent years in the organising of biological concepts, such as controlled terminologies or ontologies [3, 4]. Terminologies promote a standardised way of naming these concepts. That is, pre-established hierarchies of terms are used to constrain selections made by users in annotating large document corpora. In contrast, ontologies provide an organizational framework of the concepts involved in biological entities and processes in a system of hierarchical and associative relations that allows reasoning about biomedical knowledge. Other systems have also been developed to provide interoperability among different ontologies, such as the Unified Medical Language System [5], a collection of many biomedical vocabularies which provides a common frame of reference among the different research communities. The Open Biomedical Ontologies (OBO) Foundry hosts over 70 open source ontologies associated with phenotypic and biomedical information [6].

Ontologies in biomedical domain may be sorted into three categories:

1. Terminology-based application ontologies, which are systems of controlled vocabularies designed to meet particular needs, such as annotating biological databases or the medical record (e.g., ICD-10, SNOMED).

2. Domain ontologies provide conceptual structure of a particular domain (i.e. Foundational Model of Anatomy). They are general-purpose resources designed to provide a reference and also support a range of different types of research applications.

3. Upper-level ontologies, i.e. DOLCE and Basic Formal Ontology (BFO) (Grenon et al, 2004), describe very general concepts that address a broad range of domain areas by providing domain-independent theories largely through a frame-work of axioms and definitions. Upper level ontologies support very broad semantic interoperability between a large number of ontologies accessible under this upper ontology.

The Gene Ontology (GO) project provides a set of dynamic, controlled vocabularies of gene products that can be applied in different databases to annotate major repositories for plant, animal and microbial genomes [7, 8]. GO is divided into three domains: cellular component, molecular function and biological processes. The use of GO terms by several collaborating databases facilitates uniform queries among them.

Protein Ontology annotates terms and relationships within the protein domain and classifies that knowledge to allow reasoning [9]. Protein Ontology consists of seven generic concepts: *Residues*, *Chains*, *Atoms*, *Family*, *AtomicBind*, *Bind*, and *SiteGroup*. These generic concepts are placed into a class hierarchy of the Protein Ontology (*ProteinComplex*) to derive descriptions for any given protein, including: (1) protein sequence and structure information; (2) protein folding process; (3) cellular functions of proteins; (4) molecular bindings internal and external to proteins and; (5) external factors affecting final protein conformation (Sidhu et al, 2006).

Most recently, Lipid Ontology was developed to provide a structured framework for the effective derivation of lipid-related information [10]. Lipid Ontology mainly serves as a formal annotation for the classification and organisation of information on lipids and to support navigation of text mining results from lipid literature. The ontology has been extended to describe the lipid nomenclature classification explicitly using description logics (OWL-DL) and to support reasoning and inference tasks. Currently, the Lipid ontology has a total of 672 concepts and 75 properties. The ontology is the result of integrating schema components from existing biological database schemas, interviews with laboratory scientists, lipid and text mining experts.

To date, there does not exist a database system of classification specific to lipoproteins. Researchers have identified the role of certain lipoproteins in the Disease Database system (updated 2009), however this is limited to disease implications rather than a structured set of definitions and relations. The LOVD, a database of the LDL receptor (LDLR), contains 1,066 variations of the LDLR gene which encodes the receptor for LDL cholesterol particles [11].

3. Ontology based tools

Ontologies provide the basis for the development of several knowledge-based applications, such as information retrieval, text mining as well as other complex knowledge discovery tools.

Information Retrieval. The GoPubMed search engine was created to allow users to explore PubMed search results with the GO and Medical Subject Headings (MeSH). GoPubMed retrieves PubMed abstracts for a search query, detects terms from the GO and MeSH in the abstracts, displays the subset of GO and MeSH relevant to the keywords and displays only articles containing GO and MeSH [12]. The search engine is developed in a way that any ontology (for example Lipoprotein Ontology) can be integrated and used for a domain-specific literature search. This provides an overview of the literature and enables users to retrieve relevant information efficiently. Some other examples are: Textpresso [13] which performs semantic searches through *Caenorhabditis elegans* literature using an ontology of 14,500 terms based on Gene Ontology; BioIE [14] which is a rule-based system which extracts information pertaining to biological interactions and annotates the results using Gene Ontology terms; GenIE [15] which uses both simple processing techniques as well as syntactic and semantic analysis based on domain ontology to extract information about biochemical pathways, sequences, structures and functions of genomes and proteins. GENIA [16] covers information about biological reactions concerning transcription factors in human blood cells. All the MEDLINE abstracts on this topic and their titles have been marked-up for biologically meaningful terms. These terms have been semantically annotated using the GENIA ontology, which is a taxonomy of 47 biologically relevant nominal categories.

Text mining. A number of approaches have combined text mining with ontologies to annotate database entries with segments of biomedical literature, enabling targeted abstract document delivery [12, 17, 18]. A number of text-mining tools have been reviewed; the most promising tool that can be applied to Lipoprotein Ontology is TerMine. TerMine is a term extraction tool which extracts candidate terms and provides an interface for importing these terms to an OWL ontology. TerMine takes into account several statistical characteristics of the candidate term, such as the total frequency of occurrence in the corpus, the frequency of the term as part of other longer candidate terms and the length of the candidate term. TerMine has been found to be successful with regards to predicting long terms, but not short ones [19]. This is a crippling disadvantage within the lipoprotein domain as important key words such as LDL, HDL, will be overlooked. A tool which has been found to predict short terms effectively is TF-IDF (Term Frequency-Inverse Document Frequency), which is a weighting scheme used to evaluate the importance of a word in a collection of documents [19]. These two tools could therefore be used in conjunction with each other to achieve maximal results.

Knowledge discovery. The system architecture of Lipid Ontology is shown in Fig 1[10]. It consists of a content acquisition engine which takes user keywords and retrieves full-text research papers from public databases and converts them to a custom format ready for text mining. A workflow of Natural Language Processing algorithms identifies target concepts or keywords and tags individual sentences according to the terms they contain. Sentences are instantiated (as A-boxes), using a custom designed Java program, to the lipid ontology's literature specification (sentence concept) and relations to instances of each target concept are added into the ontology. The fully instantiated ontology is reasoned over using the reasoning engine RACER [20] and its A-box query language nRQL [21]. A custom built visual query interface facilitates query navigation over instantiated concept hierarchies, object properties and the visualisation of datatype properties in the ontology.

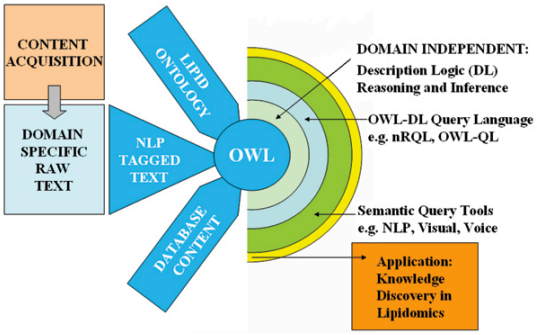

Fig. 1. Ontology-centric knowledge-delivery system architecture [10]

Knowledge sharing. Ontologies can also be employed as robust knowledge bases which will facilitate the sharing of experimental data between different research groups. One such example of the use of ontologies towards this goal is the Reference Information Model (RIM) developed by HL7 [22]. The RIM is a standard explicitly targeted to enable "consistent sharing and usage of data across multiple 'local' contexts" [22]. Although the RIM has been widely accepted by many health care organizations, a major challenge lies in the complexity of applying clinical data to the various classes in the RIM [23].

Similarly, there have been efforts in bridging the gap between medical and biological information through large multi-granular datasource built by using ontological principles [24].

4. Design of lipoprotein ontology

Numerous methodologies have been proposed for ontology development, such as Knowledge Engineering Methodology [25], DOGMA Methodology [26], TOVE Methodology [27], METHONTOLOGY [28], OnToKnowledge (OTK) Methodology [29] and SENSUS Methodology [30]. Knowledge Engineering Methodology enables the construction of ontologies at the knowledge level by defining classes using natural language before formalizing the conceptualisation. DOGMA methodology separates the conceptualisation of a domain (ontology base) from their application (commitment layer), which allows for reusability and scalability in reasoning about formal semantics. The TOVE methodology use motivating scenarios and a set of competency questions to determine the scope of the ontology to be modeled. METHONTOLOGY proposes a set of activities to develop ontologies based on its life cycle and prototype refinement. OnToKnowledge is a process-oriented methodology that focuses on knowledge management and maintenance in enterprises based on an analysis of usage scenarios. The SENSUS methodology derives domain-specific ontologies from large ontologies and enables reusability of knowledge since they have a common underlying structure.

The methodology we use to develop Lipoprotein Ontology integrates aspects of various existing methodologies. Some features of this ontology include:

- Openness
- Reusability
- Evidence-based
- Potential to evolve
- Compatibility with neighbouring ontologies such as Lipid Ontology and Protein Ontology

This methodology mainly covers three broad processes: specification, conceptualisation and implementation. Knowledge acquisition occurs throughout the three stages. Detailed steps for these processes will be described below and shown in Fig 2.

4.1 Specification

1. Identify purpose and scope

This is to ensure that the ontology created is purpose-driven and contain the right level of granularity for knowledge-based queries. The scope of Lipoprotein Ontology is within the lipoprotein domain, and the main source of knowledge is PubMed, a public online repository of peer-reviewed journal articles in the biomedical domain.

2. Literature review

A broad literature survey on lipoproteins will be conducted to define the basic concepts for the formulation of competency questions.

3. Formulate competency questions

The formulation of competency questions supports knowledge acquisition and also serves as a validation technique for completeness and consistency of the ontology.

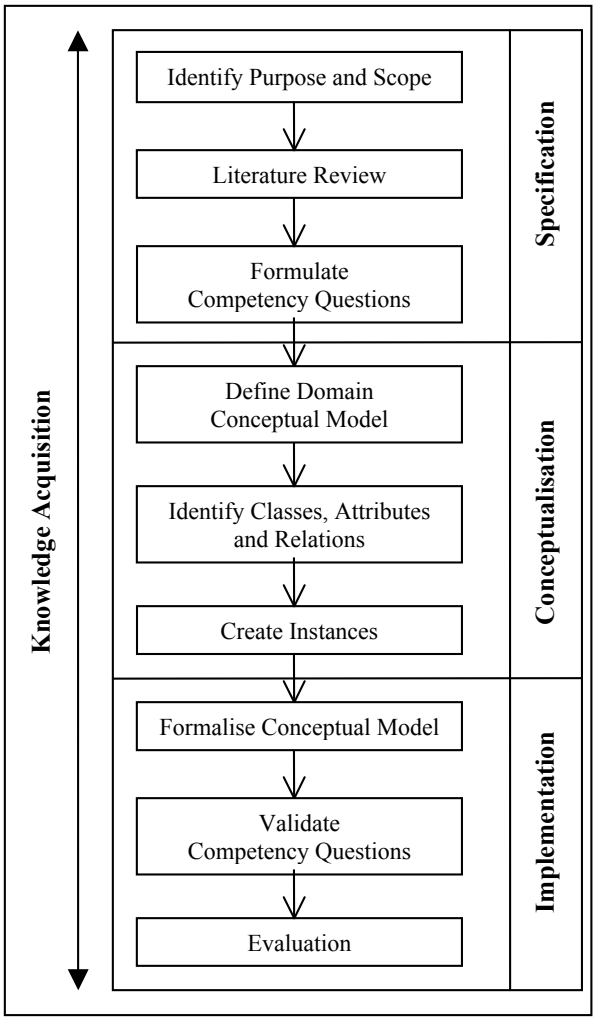

Fig. 2. Methodology for Lipoprotein Ontology development

4.2 Conceptualisation

1. Define domain conceptual model

The primary aim of the conceptualisation phase is to identify core concepts and their definitions, discern secondary concepts and describe the main relationships between these concepts. Based on the literature review,

the conceptual framework for Lipoprotein Ontology will cover all areas of lipoprotein research, namely classification, metabolism, pathophysiology, etiology and treatment.

2. Identify classes, attributes and relations

The first step in developing an ontology includes defining classes and arranging them in a hierarchy. As a result, relations between classes are described in an unambiguous way and the possible values are filled.

3. Formalise conceptual framework

The conceptual framework for Lipoprotein Ontology will be formalised using the software Protégé [31]. The Protégé OWL plug-in will be used to build the ontology and Jambalaya plug-in allows for visualisation of the ontology. The ontology will contain concepts, rules, restrictions and constraints.

4.3 Implementation

1. Create instances

Populate the ontology by creating instances for their given classes. This can be done semi-automatically by using TerMine plug-in or other similar tools.

2. Validate competency questions

Competency questions that were defined at the initial specification stage of ontology development must be tested against the complete model. Competency questions can be encoded into OWL Probe classes, and the ontology can be queried to check that the ontology can answer all expected competency questions.

3. Evaluation

After the ontology has been developed, it must be tested for:

- Completeness, that the concepts and relationships are explicitly stated and each definition is complete
- Consistency, that the definitions are consistent and do not include contradictory information
- Conciseness, that the ontology does not store any unnecessary definitions
- Expandability, that the user can add new definitions to an ontology and more knowledge to its definitions without altering the set of well-defined properties

This can be done by evaluating the complete and consistent model against the initial source documents to check that the model has successfully represented relationships present in the initial documents or definitions.

Some preliminary work has been done on Lipoprotein Ontology, which was presented at the 22nd IEEE International Symposium on Computer-Based Medical Systems in August, 2009 [32]. Lipoprotein Ontology covers the classification of lipoproteins, pathways of lipoprotein metabolism, pathophysiology of lipoproteins, causes of lipoprotein dysregulation as well as treatment of dyslipidemia.

5. Lipoprotein ontology

The lipoprotein transport system is critical for the supply, exchange and clearance of essential lipids in the body. Various lipoproteins, apolipoproteins, enzymes, transporters, and receptors in this system constitute a delicate physiologic balance; disruption of one or more components of this system results in abnormal lipoprotein levels and increases the risk to cardiovascular disease. There have been considerable efforts by a multitude of different research groups working disparately to investigate different aspects of lipoprotein research. Thus, the need for a common information repository is warranted in order to fully appreciate the implications of lipoprotein dysregulation. By incorporating specific aspects of lipoprotein research in Lipoprotein Ontology, not only in terms of the classification of lipoproteins, but also understanding the metabolic pathways, pathophysiology, causes and treatment of abnormal lipoprotein levels, this impacts not only on identifying the risks, but also provides effective preventative measures. In this paper, we present a model for Lipoprotein Ontology using Protégé, which consists of five sub-ontologies:

1. Classification
2. Metabolism
3. Pathophysiology
4. Etiology
5. Treatment

An overview of Lipoprotein Ontology is shown in Fig 3.

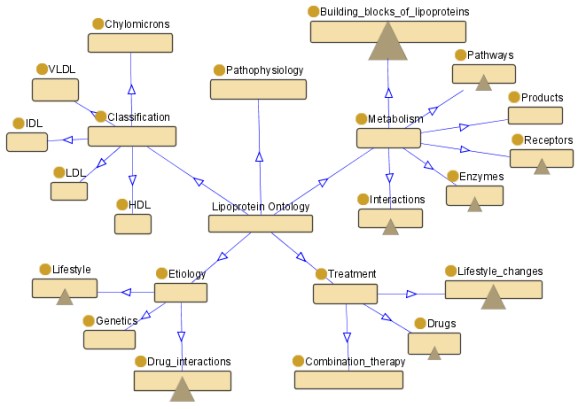

Fig. 3. Lipoprotein Ontology model consisting of five sub-ontologies and their subclasses

In our future work, Lipoprotein Ontology will be populated with concepts and instances from the literature. The database from which knowledge will be

elicited is PubMed, which is a public online repository of peer-reviewed journal articles in the biomedical domain.

6. Conclusion

The growth in biomedical data being generated through new theories and experimental techniques has led to an overwhelming increase in information. This is relevant in the context of lipoprotein research, where new discoveries are added to the literature pool constantly. To effectively merge the wealth of information that currently exists in the literature and the new findings generated in lipoprotein research, we propose a methodology for the design of Lipoprotein Ontology. This methodology covers three broad processes: (1) specification, (2) conceptualisation and (3) implementation. The methodology proposed for Lipoprotein Ontology can be adapted for use in the design of other domain-specific ontologies. Lipoprotein Ontology will provide the basis for the design of various applications to enable interoperability between research groups or software agents, as well as the development of tools for the diagnosis and treatment of dyslipidemia.

7. References

[1] R.E. Olson, "Discovery of the Lipoproteins, Their Role in Fat Transport and Their Significance as Risk Factors", in *The Journal of Nutrition*, vol. 128, pp. 439S-443S

[2] [AHA] American Heart Association. 2009. Heart disease and stroke statistics – 2009 update [online] http://www.americanheart.org/presenter.jhtml?identifier=3000090

[3] Y.A. Lussier, O. Bodenreider O, "Clinical Ontologies for discovery Applications", in *Semantic Web: Revolutionizing knowledge discovery in the life sciences,* Springer, pp. 101-119, 2007

[4] P. Lambrix, H. Tan, V. Jakoniene, L. Strömbäck, "Biological Ontologies", in *Semantic Web: Revolutionizing knowledge discovery in the life sciences.* Springer, pp. 85-99, 2007

[5] O. Bodenreider, "The Unified Medical Language System (UMLS): integrating Biomedical terminology", in *Nucleic Acids Research*, vol. 32, no. 1, pp. 267-270, 2004

[6] B. Smith, M. Ashburner, C. Rosse, C. Bard, W. Bug, W. Ceusters, et al., "The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration", in Nature Biotechnology, vol. 25, pp. 1251 – 1255, 2007

[7] Gene Ontology, http://www.geneontology.org

[8] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, et al., "Gene ontology: tool for the unification of biology", in Nature Genetics, vol. 25, pp. 25-29, 2000

[9] A.S. Sidhu, T.S. Dillon, E. Chang, "Integration of Protein Data Sources through PO", in *Proceedings of the 17th International Conference on Database and Expert Systems Applications* (DEXA 2006), Poland, pp. 519-527, 2006

[10] C.O. Baker, R. Kanagasabai, W.T. Ang, A. Veeramani, H.S. Low, M.R. Wenk, "Towards ontology-driven navigation of the lipid bibliosphere", in *BMC Bioinformatics*, vol. 9, pp. S5, 2008

[11] S.E. Leigh, A.H. Foster, R.A. Whittall, C.S. Hubbart, S.E. Humphreys, "Update and analysis of the University College London low density lipoprotein receptor familial hypercholesterolemia database", in Annals of Human Genetics, vol. 72, pp. 485-495, 2008

[12] A. Doms, M. Schroeder, "GOPubMed: exploring PubMed with the GO", in *Nucleic Acid Research* vol. 33, pp. 783-786, 2005

[13] H. Muller, E. Kenny, P. Sternberg, "Textpresso: An ontology-based information retrieval and extraction system for biological literature", in *PLoS Biology*, vol. 2, no. 11, p.e309, 2004

[14] J. Kim, J. Park, "BioIE: Retargetable information extraction and ontological annotation of biological interactions from the literature", in *Journal of Bioinformatics and Computational Biology*, vol. 2, no. 3, pp. 551-568, 2004

[15] P. Cimiano, U. Reyle, J. Saric, "Ontology-driven discourse analysis for information extraction", in *Data and Knowledge Engineering*, vol. 55, pp. 59-83, 2005

[16] J. Kim, T. Ohta, Y. Tateisi, J. Tsujii, "Genia corpus – semantically annotated corpus for bio-textmining", in *Bioinformatics*, vol. 19, pp. i180–182, 2003

[17] E.B. Camon, D.G. Barrell, E.C. Dimmer, V. Lee, M. Magrane, J. Maslen, et al., "An evaluation of GO annotation retrieval for BioCreAtIvE and GOA", in BMC Bioinformatics, vol. 6,S17, 2005

[18] H.M. Müller, E.E. Kenny, P.W. Sternberg, "Textpresso: An ontology-based information retrieval and extraction system for biological literature", in PLoS Biology, vol. 2, pp. 1984-1998, 2004

[19] D. Alexopoulou, T. Wächter, L. Pickersgill, C. Eyre, M. Schroeder, "Terminologies for text-mining; an experiment in the lipoprotein metabolism domain", BMC Bioinformatics, vol. 9, 2008

[20] F.I. Baader, I. Horrocks, U. Sattler. "Description logics as ontology languages for the semantic web", in Lecture Notes in Artificial Intelligence, Springer; 2003

[21] K. Wolstencroft, R. Stevens, V. Haarslev, "Applying OWL Reasoning to Genomic Data", in Semantic Web: Revolutionizing knowledge discovery in the life sciences, pp. 225-248, Springer; 2007

[22] L. Vizenor, "Actions in Health Care Organizations: An Ontological Analysis", in *MEDINFO*, Amsterdam: IOS Press, pp. 1403-1407, 200

[23] J.A. Lyman, K. Scully, S. Tropello, J. Boyd, J. Dalton, S. Pelletier, C. Eyhazy, "Mapping From a Clinical Data Warehouse to the HL7 Reference Information Model", in *AMIA Annual Symposium Proceedings*, vol. 920, 2003

[24] A. Kumar, Y.L. Yip, B. Smith, P. Grenon P, "Bridging the gap between medical and bioinformatics: an ontological case study in colon carcinoma", in *Computers in Biology and Medicine*, vol. 36, pp. 694-711, 2006

[25] M. Uschold, M. Gruninger, "Ontologies: Principles, methods and applications", in Knowledge Engineering Review, vol. 11, 1996.

[26] J. De Bo, P. Spyns, R. Meersman, "Creating a "DOGMAtic" multilingual ontology infrastructure to support a semantic portal", in R. Meersman et al. (eds), On the Move to Meaningful Internet Systems 2003: OTM 2003 Workshops, LNCS (Vol. 2889, pp. 253-266). Springer-Verlag, 2003

[27] M. Gruninger, M. Fox, "Methodology for the design and evaluation of ontologies", in D. Skuce (ed.), IJCAI-95 Workshop on Basic Ontological Issues in Knowledge Sharing, 1995

[28] M. Fernández, A. Gómez-Pérez, N. Juristo, "METHONTOLOGY: From Ontological Art Towards Ontological Engineering", Spring Symposium Series, pp. 33-40, 1997

[29] Y. Sure, S. Staab, R. Studer, "On-To-Knowledge Methodology (OTKM)", in Handbook on Ontologies, pp. 117-132, 2004

[30] W. Swartout, R. Patil, K. Knight, T. Russ, "T.: Towards Distributed Use of Large-Scale Ontologies", in AAAI 1998 Spring Symposium on Ontologicial Engineering, USA, 1997

[31] The Protégé Ontology Editor and Knowledge Acquisition System. http://protege.stanford.edu

[32] M. Chen and M. Hadzic, "Lipoprotein ontology as a functional knowledge base", in Proceedings of the IEEE international symposium on computer-based medical systems (CBMS 2009), 2009