Proposed ontology based knowledge acquisition and integration framework for clinical knowledge management

Phanu Waraporn[†*], Phayung Meesad[††], and Gareth Clayton[†††],

[†]Department of Information Technology, Faculty of Information Technology
[††]Department of Teacher Training in Electrical Engineering, Faculty of Technical Education
[†††]Department of Applied Statistics, Faculty of Applied Science
King Mongkut's University of Technology North Bangkok, Bangkok, THAILAND, and
[*]Department of Business Computing, Faculty of Management Science
Suan Sunandha Rajabhat University, Bangkok, THAILAND

Summary
This paper discusses the potential of ontologies as elements in Information Technology (IT) support for Knowledge Management (KM), particularly in the medical domain. It also details fundamental topics of medical ontology based on ontological theory. Finally, it proposes the ontology for clinical knowledge management framework and its implementation in the area of Coronary Heart Disease (CHD).
Key words:
Ontology, Clinical Knowledge Management, Coronary Heart Disease (CHD).

1. Introduction

In Thailand, for the purpose of accurate statistics as well as expense reimbursement, every medical diagnosis in every public hospitals medical records is required to be coded by trained, skilled and experienced coders and later be reviewed by another coder or health and medical related professional as stated in the 2002 Universal Coverage Act enacted 2002 in Thailand [1,2]. At present such a task is handled by the trained and experienced human coders, which often involves the physicians themselves.

Currently the diagnosis situation and procedure coding system in Thailand are as follows [3]. 75% of hospitals in Thailand use ICD10 for coding of diagnosis and ICD-9-CM for coding of procedure. In the view of hospital executives and experts, diagnosis/procedure coding was utilized for statistical report (75%) and for reimbursement (64.19%), 59.87% of the studied hospitals had certificated medical coders in coding practices, but 46.20% of coders studied in hospitals had to work on other jobs. 85% of coders had been trained in diagnosis/procedure coding training course. Even though expertise is required for coding, 43.61% of the coders have between 1-3 years experience and 13.66% have less than 1 year experience.

The important requirements for good coding create a dilemma for the human coders, not only to satisfy themselves but also to cope with the time constraints and huge amount of patient data that need to be thoroughly reviewed. This has inadvertently resulted occasionally in coding errors. Though the medical software and tools are available, they do not fully serve the process well as numerous interactions exist and become the major drawback for the coder to obtain a single plain code [2].
In later sections, we will identify how ontologies, knowledge management principles, particularly the Knowledge Acquisition (KA), and Knowledge Integration (KI), the tools and means by which the foregoing problems can be overcome.

2. Related Work

2.1 Ontology, domain ontology, medical ontology, and medical diagnosis coding ontology

Ontology is a collection of concepts, medical concepts in this context, which are hierarchically linked together according to their underlying relationships within the domain of interest. In the context of computer and information sciences, an **ontology** defines a set of representational primitives with which to model a domain of knowledge or discourse. The representational primitives are typically classes (or sets), attributes (or properties), and relationships (or relations among class members). The definitions of the representational primitives include information about their meaning and constraints on their logically consistent application. In the context of database systems, ontology can be viewed as a

Manuscript received March 5, 2010
Manuscript revised March 20, 2010

level of abstraction of data models, analogous to hierarchical and relational models, but intended for modeling knowledge about individuals, their attributes, and their relationships to other individuals. Ontologies are typically specified in languages that allow abstraction away from data structures and implementation strategies; in practice, the languages of ontologies are closer in expressive power to first-order logic than languages used to model databases. For this reason, ontologies are said to be at the "semantic" level, whereas database schema are models of data at the "logical" or "physical" level. Due to their independence from lower level data models, ontologies are used for integrating heterogeneous databases, enabling interoperability among disparate systems, and specifying interfaces to independent, knowledge-based services. In the technology stack of the Semantic Web standards [4], ontologies are called an explicit layer. There are now standard languages and a variety of commercial and open source tools for creating and working with ontologies [5].

The **domain ontologies** are reusable vocabularies of the concepts within a domain and their relationships, of the activities taking place in that domain, and of the theories and elementary principles governing that domain [6, 7].

Current medical information systems contain ambiguous communication of complex and detailed medical concepts whereby several agents must interact between each other to share the results and hence use a common and standardized medical terminology with precise and unequivocal meanings. The **medical ontologies** are therefore developed to enable the reusing and sharing of medical data [8].

Medical diagnosis coding ontology is an ontological engineering development to construct components based distributed ontologies of patient data, patient laboratory information, patient drug prescription and consumable using reference standards with one dedicated goal of automating the medical coding process presently limited to the skilled and experienced medical personnel.

2.2 Leading medical ontologies

GALEN [8] OpenGALEN, is a clinical terminology represented in the formal and medically oriented language GRAIL.[9] This language was specially developed for specifying restrictions used in medical domains. GALEN was intended to be used with different natural languages and integrated with different coding schemata. It is based on a semantically sound model of clinical terminology known as the GALEN COding REference (CORE) model. The GALEN CORE top-level ontology establishes four general categories (which are subclasses of DomainCategory):

1. Structures (GeneralisedStructure), which are abstract or physical things with parts that are time-independent (such as microorganism, protocol or heart).
2. Substances (GeneralisedSubstance), which are continuous abstract or physical things that are time-independent, such as bile, drugs or radiation.
3. Processes (GeneralisedProcess), which are changes that occur over time, such as irradiation, clinical act or breathing.
4. Modifiers (ModifierConcept), which refine or modify the meaning of the other three categories, such as severe diabetes. In this ontology the following types of modifiers are considered: modifiers of aspect (classified in feature, state, selector, and status), unit, modality, role, general level of specification, and collection.

The category ValueType is also a subclass of DomainCategory. It defines value types such as Integer, Ordinal, etc. Besides, the category Phenomenon is included in this top-level ontology. It gathers the medical intuitions of Disease and Disorder and it is associated with domain categories, more specifically with structures, processes, and substances, and with the modifiers of feature, state, and collection. The GALEN CORE top-level ontology defines relationships between concepts that belong to general categories. These relationships are called attributes (DomainAttribute) and are divided into two types: constructive attributes, which link processes, structures and substances together; and modifier attributes, which link processes, structures and substances to modifiers.

UMLS [8] (Unified Medical Language System), developed by the United States National Library of Medicine, is a large database designed to integrate a great number of biomedical terms collected from various sources (over 123 sources in the 2008 edition) such as clinical vocabularies or classifications (MeSH, SNOMED, RCD, etc.). UMLS is structured in three parts: Metathesaurus, Semantic Network and Specialist Lexicon.

1. The Metathesaurus contains biomedical information about each of the terms included in UMLS. If a term appears in several sources, which is usual, a concept will be created in UMLS with a preferred term name associated with it. The original source information about the terms (such as, definition, source, etc.) is attached

to the concept and some semantic properties are also specified, such as concept synonyms, siblings and parents, or the relationships between terms.

2. The Semantic Network is a top-level ontology of biomedical concepts and relations among these concepts. The Semantic Network was not derived from the biomedical sources integrated in UMLS but created as a part of UMLS with the aim of providing a consistent structure or categorization in which the Metathesaurus concepts are included. Each Metathesaurus concept is attached to a concept or concepts of the Semantic Network. Thus, the Semantic Network was introduced in order to solve the heterogeneity among the UMLS sources, and it could be considered the result of the integration of the UMLS sources.
3. The Specialist Lexicon contains syntactic information about biomedical terms to be used in natural language processing applications.

ON9 [8,10] is a set of medical ontologies that includes some terminology systems, like UMLS.. The ontologies are represented by boxes. Thick dashed boxes are sets of ontologies (some show the elements explicitly). Continuous arrows mean included in, and dashed arrows mean integrated in. The ontologies at the top of the hierarchy are the Frame Ontology and the set of KIF ontologies. To link these ontologies with the generic ontology library several ontologies have been defined: Structuring-Concepts, Meta-Level-Concepts and the Semantic-Field-Ontology. The sets of Structural ontologies and of Structuring ontologies contain generic ontologies. Particularly, the Integrated-Medical-Ontology includes all the generic ontologies used to gather the terminological ontologies of the five terminological systems.

The above selected outstanding medical ontologies are capable of assisting the human in the medical coding process but are not designed to fully function and automate the translational works.

2.3 Selected medical standards and references

In this study and future in-depth research works, the standards and references in use will mainly involve, but not be limited to, the following three standards and references.

World Health Organization (WHO) **International Classification of Disease** (ICD) - Version 10 (ICD-10): The ICD-10 is the international standard diagnostic classification for all general epidemiological and many health management purposes. These include the analysis of the general health situation of population groups and monitoring of the incidence and prevalence of diseases and other health problems in relation to other variables such as the characteristics and circumstances of the individuals affected. It is used to classify diseases and other health problems recorded on many types of health and vital records including death certificates and hospital records. In addition to enabling the storage and retrieval of diagnostic information for clinical and epidemiological purposes, these records also provide the basis for the compilation of national mortality and morbidity statistics by WHO Member States. [11]

The features of ICD 10 in graphical presentation linking other features are shown in Figure 1. The categories and concepts are grouped into connected categories and presented in Figure 2.

Unified Medical Language System (UMLS): This is a U.S. National Library of Medicine (NLM) project to develop and distribute a multi-purpose, electronic "Knowledge Source" and associated lexical tool for system developers. UMLS products are useful in investigating knowledge representation and retrieval questions. [13]

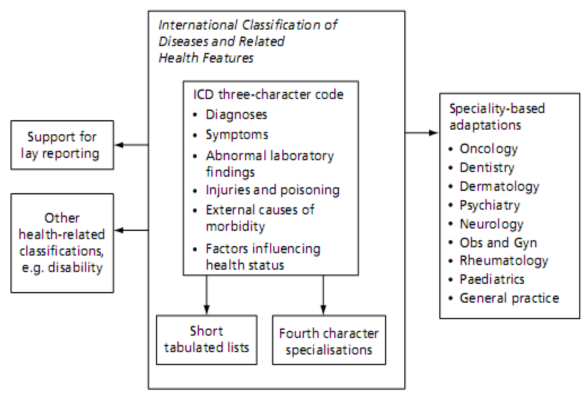

Figure 1 Overview of International Classification of Diseases [12]

Chapter	Codes	Description
I	A00-B99	Certain infectious and parasitic diseases
II	C00-D48	Neoplasms
III	D50-D89	Diseases of the blood, blood-forming organs and certain diseases of the immune system
IV	E00-E90	Endocrine, nutritional and metabolic disorders
V	F00-F99	Mental and behavioural disorders
VI	G00-G99	Diseases of the nervous system
VII	H00-H59	Diseases of the eye and adnexa
VIII	H60-H95	Diseases of the ear and mastoid process
IX	I00-I99	Diseases of the circulatory system
X	J00-J99	Diseases of the respiratory system
XI	K00-K93	Diseases of the digestive system
XII	L00-L99	Diseases of the skin and subcutaneous system
XIII	M00-M99	Diseases of the musculoskeletal system and connective tissue
XIV	N00-N99	Diseases of the genitourinary system
XV	O00-O99	Diseases of pregnancy, childbirth and puerperium
XVI	P00-P96	Certain conditions originating in the perinatal period
XVII	Q00-Q99	Congenital malformations, deformations and chromosomal abnormalities
XVIII	R00-R99	Symptoms, signs and abnormal clinical and laboratory findings not elsewhere classified
XIX	S00-T98	Injury, poisoning and certain other consequences of external causes
XX	V01-Y98	External causes of morbidity and mortality
XXI	Z00-Z99	Factors influencing health status and contact with services

Figure 2 The division of ICD 10 code into chapters [12]

Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT): This is a comprehensive list of clinical terminology, originally created by the College of American Pathologists and, as of April 2007, owned, maintained, and distributed by the International Health Terminology Standards Development Organisation (IHTSDO). The NLM is the U.S. Member of the IHTSDO. The license terms have been incorporated into the License for Use of the UMLS Metathesaurus. Licensees of the UMLS Metathesaurus have access to SNOMED CT in multiple formats - as part of the UMLS Metathesaurus, where it is linked to many other biomedical terminologies and natural language processing tools [14,15]. SNOMED CT is considered to be the most cited and exhaustive worldwide referencing sources are tested. The system did generate useful and relevant concepts and categories but not in compliance with ICD 10 for the coder to perform his work effectively. Also, it has a non uniform structure for context and non negation but with a proliferation of tools. These drawbacks can be solved by description logic (DL) using expressive language which exists in ontology design [18]. Furthermore, terminologies are increasingly based on ontologies developed in description logics and related languages such as the new Web Ontology Language, OWL. The use of description logic has been expected to reduce ambiguity and make it easier to determine logical equivalence, deal with negation, and specify electronic health records (EHRs). However, this promise has not been fully realized in part because early description logics were relatively inexpressive and the relation between coding systems, EHRs, and ontologies expressed in description logics has not been fully understood [18]. The DL model has many advantages. It establishes a formal semantics for assertions and suggests a syntax, provides a basis for understanding expressiveness and computational complexity, through correspondence with known results from DL, and finally helps to clarify the relationships among existing concept representation methods.

Health Level 7 (HL7) Health Level Seven is one of several American National Standards Institute (ANSI) - accredited Standards Developing Organizations (SDOs) operating in the healthcare arena. Most SDOs produce standards (sometimes called specifications or protocols) for a particular healthcare domain such as pharmacy, medical devices, imaging or insurance (claims processing) transactions. Health Level Seven's domain is clinical and administrative data. Health Level Seven is like most of the other SDOs in that it is a not-for-profit volunteer organization. Its members-- providers, vendors, payers, consultants, government groups and others who have an interest in the development and advancement of clinical and administrative standards for healthcare—develop the standards. Like all ANSI-accredited SDOs, Health Level Seven adheres to a strict and well-defined set of operating procedures that ensures consensus, openness and balance of interest. A frequent misconception about Health Level Seven (and presumably about the other SDOs) is that it develops software. In reality, Health Level Seven develops specifications, the most widely used being a messaging standard that enables disparate healthcare applications to exchange keys sets of clinical and administrative data [16].

2.4 Ontological engineering

This refers to the set of activities that concern the ontology development process, the ontology life cycle and the methodologies, tools and languages for building ontologies [8]. This is also related to knowledge engineering, a general process of knowledge base construction [17].

2.5 Selected ontological building tools and editors

There are currently wide varieties and numbers of tools for deployment. Each serves different purposes and therefore any selection may have to consider the nature of domain and characteristic of the relevant data. We deploy and implement our framework using HOZO.

2.6 Information Technology (IT) supported Knowledge Management (KM)

Knowledge Management is an interdisciplinary topic with roots in Economics, Information Technology, Pedagogy, Psychology, and Organization Theory [19,20]. The Knowledge Management can be defined according to [21] as a (1) systematically managed organizational activity; (2) which views implicit and explicit knowledge as a key strategic resource of an organization, and thus; (3) aims at improving the handling of knowledge at the individual, team, organization, and inter-organizational level; (4) in order to achieve organizational goals such as better innovation, higher quality, more cost-effectiveness, or shorter time-to-market; (5) by employing tools, techniques, and theories from manifold areas such as IT, strategic planning, change management, business process management, innovation management, human resource management, and others; and (6) in order to achieve a planned impact on people, processes, technology, and culture in an organization [21]

Table 1: Knowledge-based KM Technology, adapted from [22]

Knowledge Sharing	**Knowledge Dissemination**	**Knowledge Capturing and Codification**	**Knowledge Creation**
Ontology-based portal / Information Portal	**Expert Systems**, Lesson-learned and Best Practice Systems	**Knowledge Acquistion and Coding tools**	**Knowledge Discovery and Data Mining Systems** and Creativity Systems

Table 1 illustrates how information technology supported knowledge management can conform to our work, particularly with regard to the use of ontologies.

3. The Proposed Framework

In our case, an Ischaemic Heart Disease ontology is manually built by reusing the framework, high level, relations and concepts defined in accordance with [23]. Figures 3 and 4 illustrate the property inheritance and top level, respectively.

3.1 Medical Knowledge Integration Framework and Process

In a complex domain such medicine, knowledge integration requires consolidation of the heterogeneous knowledge and reconfirmation of it with the induced models since it originates from sources with different levels of certainty and completeness. Therefore, new models are collectively learned and comprehensively evaluated based on existing knowledge. Figure 4 illustrates the overview of medical knowledge integration.

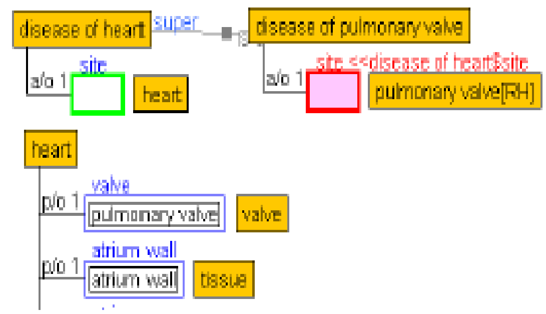

Figure 3 Property Inheritances in HOZO Heart Disease [23].

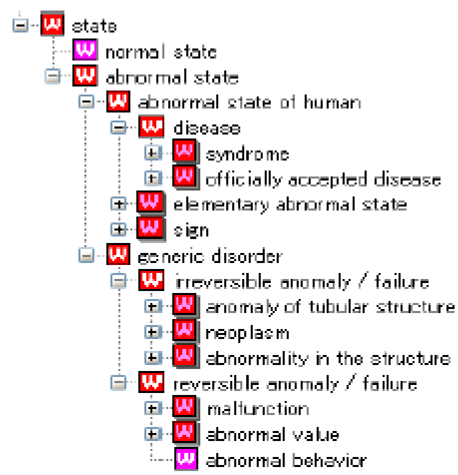

Figure 4 High Level category in HOZO [23].

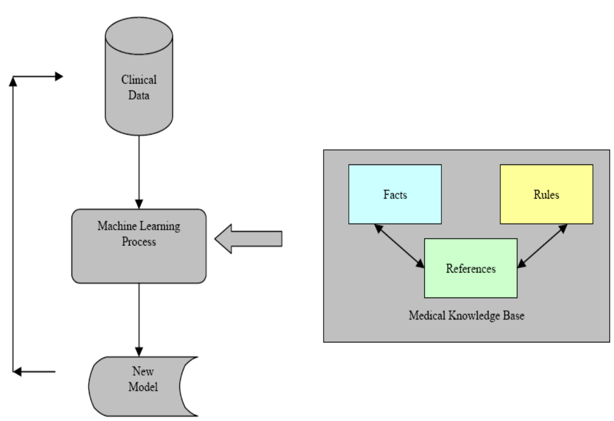

Figure 4 Overview of medical knowledge integration architecture [24].

Patient records in use are the discharge summary records, which generally contain the principal diagnosis (PDx) information. The particulars of secondary diagnosis (SDx) may appear provided that there is an additional ailment requiring treatment in addition to the PDx. SDx are the details about the comorbidity, complication, status of the disorder; chronic or acute and external cause of the ailment.

Additionally, depending on the necessity, the treatment may involve other medical procedures (PROC) and such details are itemized. In the case where only PDx is reported, the coding process is fairly simple and fast. Generally, however, the patient records contain many SDxs and multiple PROC. The coder duty is to translate this PDx(s), SDx(s) and PROC information into corresponding ICD 10 code(s). Other patient data aside from the discharge summary includes but is not limited to prescription information sheet, laboratory results, drug prescribed nurse's notes, progress notes, and other records [2].

From Figure 5, an ICD 10 for Coronary Heart Disease (CHD) is presented in a hierarchical form. The top first level is the high/concept level class where the second and third levels are major and minor classes, respectively. The minor class gives more specific information yet not comprehensively enough when compared with other major medical concepts and ontologies like UMLS, GALEN, SNOMED CT, etc. This is due to the fact that information is served differently.

Key words/terms extracted from patient discharge summaries will undergo the coding process and provided that all relevant information are in agreement, this summary will finally be assigned the relevant ICD 10 code(s) as specified in the ICD 10.

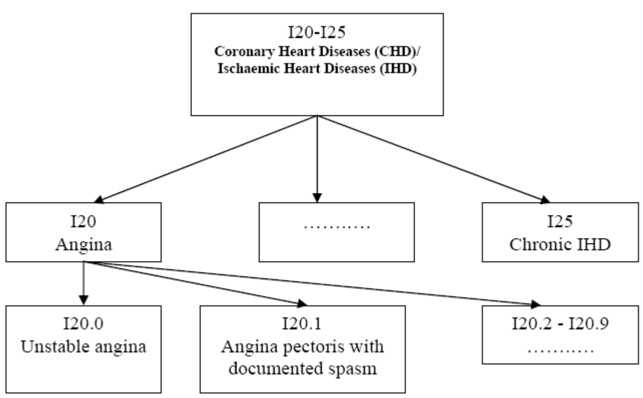

Figure 5 ICD 10 for Coronary Heart Disease in hierarchical form [24]

3.2 Proposed Framework

The next section explains the proposed framework as shown in Figure 6 below. Our emphasis is on the section surrounded by the red block.

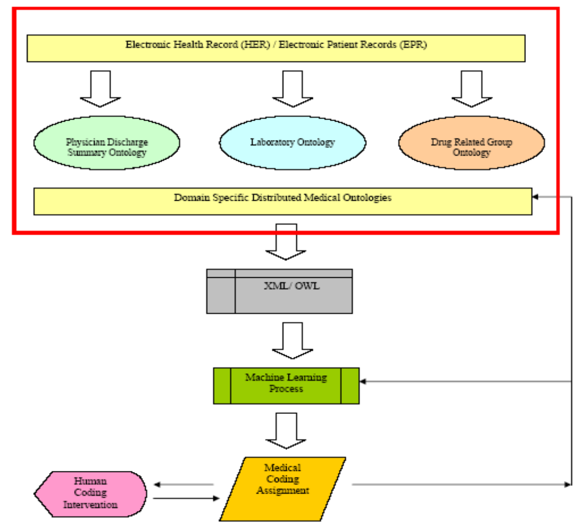

Figure 6 Task-Specific and Domain-Specific Ontology Development in a distributed environment [24]

Patient records extracted are mapped against pre-defined medical ontologies, namely, Domain-Specific Distributed Medical Ontologies derived from task-specific physician discharge summary ontology, Laboratory ontology and drug related group ontology. The mapped records are then exported in an XML/OWL format for the second phase processing task, involving machine learning technologies for text categorization and classification.

4. Conclusion and Future Works

The medical ontologies, repositories, knowledge-bases and medical resources available during this study cannot be directly applied to the automation of coding works. Each system is partly assisted by the coders during the process of obtaining the required codes. This is simply because the conceptual level or the domain characteristics under implementation and development are quite different and limited. The lack of dynamicity therefore prevents the direct reuse of the accessible growth in medical based ontological undertakings. Future work will include development and design functions for managing these ontological components based on their dependencies while maintaining the integrity and the consistency before implementing the state of art machine learning tools to represent the medical knowledge management in a form of automated medical diagnosis coding.

Acknowledgments

We would like to extend our sincere thanks to staff at the Mahidol University's Faculty of Medicine Siriraj Hospital

Division of Molecular Genetics, staff at the National Center for Genetic Engineering and Biotechnology (BIOTEC) and staff at the Human Language Technology Laboratory of the National Electronics and Computer Technology Center (NECTEC) for all resources and advices discussed in this experimental study. In addition, we acknowledge the team at the Mizoguchi Laboratory of the Institute of Scientific and Industrial Research, Osaka University for allowing us to access the software, HOZO, an ontology editor.

References

[1] Pattapong K., et al, *"Detecting adverse events in Thai hospitals using medical record reviews: agreement among reviewers"*, J. Med. Assoc. Thai, Vol.88, No.10, 2005. Also www.medassocthai.org/journal/files/Vol88_N010_1412.pdf

[2] Waraporn, P., et al. Distributed Ontological Engineering and Integrated Development of Medical Diagnosis Coding Ontology for State Hospitals in Thailand, *Proceedings of National Conference on Computer and Information Technology (NCIT 2008)* (2008)

[3] Jaiyos P., et al, *"Medical coding practices in Thailand"*, Research paper in Thai, November 2001. Also library.hsri.or.th/abs/res/ hs0913e.doc

[4] Tom Gruber (2008), Ontology. Entry in the *Encyclopedia of Database Systems*, Ling Liu and M. Tamer Ö zsu (Eds.), Springer-Verlag, to appear in 2008. Also tomgruber.org/writing/ ontology-in-encyclopedia.htm

[5] Berners-Lee, T., Hendler, J. and Lassila, O. "*The Semantic Web*", Scientific American, May 2001. Also www.w3.org/2001/sw/

[6] Mizoguchi et al., *"Task Ontology for reuse of problem solving knowledge."* In: Mars N (ed) Towards Very Large Knowledge Bases: Knowledge Building and Knowledge Sharing (KBKS'95). University of Twente, Enschede, The Netherlands. IOS Press, Amsterdam, The Netherlands, pp 46–57, 1995

[7] van Heijst et al., *"Using explicit ontologies in KBS development."* International Journal of Human-Computer Studies 45:183–292,1997

[8] Gomez-Perez et al., "Ontological Engineering, Springer-Verlag, 2004

[9] Rector Al, et al, "Why do it the hard way? The Case for an Expressive Description Logic for SNOMED". J Am Med Inform Assoc., August 2008.

[10] Gangemi A, Pisanelli DM, Steve G., *"Some Requirements and Experiences in Engineering Terminological Ontologies over the WWW."* In: Gaines BR, Musen MA (eds) 11th International Workshop on Knowledge Acquisition, Modeling and Management (KAW'98). Banff, Canada, SHARE10:1–20, 1998

[11] WHO. World Health Organization. www.who.int, www.who.int/classifications/icd/en/

[12] P.Taylor, *"From patient data to medical knowledge : the principles and practice of health informatics"*, Blackwell Publishing Ltd., London, 2006.

[13] U.S. National Library of Medicine, National Institute of Health www.nlh.nih.gov,www.nlm.nih.gov/research/umls/ umlsmain.html

[14] www.nlm.nih.gov/research/umls/Snomed/snomed_main.htm l

[15] Lister Hill National Center for Biomedical Communications, a research division of the U.S. National Library of Medicine, National Institute of Health (http://lhncb.nlh.nih.gov/)

[16] HL7. Health Level Seven Organization. HL7 Reference Information Model. . www.hl7.org

[17] Russell S., Norvig P., "Artificial Intelligence: A Modern Approach", Prentice Hall, 2003.

[18] Rector Al, et al, "What's in a code? Towards a formal account of the relation of ontologies and coding systems." Studies in Health Technology and Informatics. August 2008.

[19] Davenport T, Prusak L Working Knowledge: How Organizations Manage What They Know. Harvard Business School Press, Boston, 1997.

[20] Holsapple CW (ed) Handbook on knowledge management, vols. 1 and 2. International Handbooks on Information Systems. Springer, Berlin, 2003.

[21] Abecker A, van Elst L. Ontology for Knowledge Management. In: Staab S. and Studer R. (eds.), Handbook on Ontologies, International Handbooks on Information Systems, Springer-Verlag Berlin Heidelberg, 2009.

[22] Weber R, Kaplan R. Knowledge-based knowledge management. In: Faucher C, Jain L, Ichalkaranje N (eds) Innovations in Knowledge Engineering. Physica, Heidelberg, 2002.

[23] Mizoguchi, R. et al. An Advanced Clinical Ontology. Available from Nature Precedings <http://dx.doi.org/10.1038/ npre.2009.3506.1> 2009

[24] Waraporn, P., et al., Ontology-supported processing of clinical text using medical knowledge integration for multi-label classification of diagnosis coding, International Journal of Computer Science and Information Security, Volume 7 No. 3, 2010 (In Press).

Phanu Waraporn, a PhD candidate in the Faculty of Information Technology, King Mongkut's University of Technology North Bangkok. Currently, he is the lecturer in Business Computing at the Faculty of Management Science, Suan Sunandha Rajabhat University, Bangkok, Thailand.

Phayung Meesad, Ph.D., assistant professor. He is an associate dean for academic affairs and research at the faculty of Information Technology, King Mongkut's University of Technology North Bangkok. He earned his MS and Ph.D. degrees in Electrical Engineering from Oklahoma State University, U.S.

Gareth Clayton, Ph.D., a senior lecturer in statistics at the Department of Applied Statistics, Faculty of Applied Science, King Mongkut's University of Technology North Bangkok. He earned his Ph.D. in Statistics from Melbourne University, Australia.