DESIDOC Journal of Library & Information Technology, Vol. 31, No. 2, March 2011, pp. 103-115

Developing an Ontology for Encoding Disease Treatment Information in Medical Abstracts

Christopher S.G. Khoo, Jin-Cheon Na, Vivian Wei Wang*, and Syin Chan**

Wee Kim Wee School of Communication & Information, Nanyang Technological University
E-mail: assgkhoo, tjcna}@ntu.edu.sg

**Centre for Health Services Research, Singapore Health Services*
E-mail: vivian.wang.wei@singhealth.com.sg

***School of Computer Engineering, Nanyang Technological University*
E-mail: asschan@ntu.edu.sg

ABSTRACT

A disease-treatment ontology is being developed to model and represent treatment information found in medical abstracts. Treatment information extracted from medical abstracts and medical articles can then be encoded in this ontology and used for information retrieval, question-answering, summarisation and knowledge discovery. This paper explains the initial version of the ontology developed based on an analysis of 50 medical abstracts on colon cancer therapy retrieved from the Medline database. The ontology divides disease-treatment information into five classes—disease, treatment, condition, effect, and evidence. The sub-classes, properties and instances of these main classes are discussed with examples.

Keywords: Medical ontology, disease-treatment ontology, ontology class

1. INTRODUCTION

A disease-treatment ontology is being developed to model and represent treatment information found in the abstracts of medical articles. This is part of a bigger project to develop an automatic extraction system to extract treatment information from medical abstracts retrieved from the Medline database, to support information retrieval, question-answering, summarisation, and knowledge discovery. The purpose of the ontology is to serve as a knowledge base to store the extracted information and support these functions.

The ontology is also expected to be useful in supporting synthesis of information extracted from different publications, and inferencing of potentially new relations between chemical substances and effects on diseases, such as envisaged by Swanson[1,2] and others. Information stored in an ontology can also support evidence-based medicine[3,4]—to alert doctors to the range and quality of clinical data available to make informed treatment decisions. A disease-treatment ontology is potentially important for use in medical digital libraries/portals and medical information systems.

The disease-treatment ontology is being constructed as an enhancement to existing medical taxonomies and ontologies. We adopted the Unified Medical Language System (UMLS) semantic network[5,6], the Medical Subject Headings (MeSH)[7], and the NCI (National Cancer Institute) thesaurus (http://bioportal.bioontology.org/ontologies/39478 & http://ncit.nci.nih.gov/ncitbrowser) as the base medical ontology and enriched with relations to link potential medical treatments with diseases. The following URI was assigned to the disease-treatment ontology: http://dis.sci.ntu.edu.sg/medical/disease-treatment.v1#. The initial disease-treatment ontology was developed based on analyses of 40 medical abstracts on colon cancer therapy, and evaluation of additional 10 abstracts. Many definitions of ontologies have been proposed in the literature from different perspectives. Perhaps the most often quoted definition is by Gruber[8]: "A body of formally represented knowledge is based on a conceptualisation: the objects, concepts, and other

entities that are assumed to exist in some area of interest and the relationships that hold among them[9]. A conceptualisation is an abstract, simplified view of the world that we wish to represent for some purpose. An ontology is an explicit specification of a conceptualisation." In practice, an ontology is a set of concepts or classes, represented by terms, linked with semantic relations or properties. The most common semantic relation is the *isa* or *subclass-of* relation that organises the concepts into taxonomies with a hierarchical structure. Additional semantic relations enhance the ontology by specifying relations between concepts in different taxonomies. More complex rules, called axioms, specify constraints and relationships among two or more concepts. The XML-based representation formalism for ontologies adopted by the World Wide Web Consortium is the Web Ontology Language[10]. OWL offers three sub-languages of increasing expressiveness: OWL Lite, OWL DL (Description Logics) and OWL Full. In this study, we adopted OWL DL as the knowledge representation formalism and encoding scheme for the ontology. OWL DL is said to provide "maximum expressiveness while retainig computational completeness and decidability"[10]. An ontology represented in OWL can be divided into two parts: (i) the set of classes and properties (i.e. semantic relations), and (ii) the set of instances (objects) of the classes and instances of properties.

This paper describes the set of classes and properties in the ontology. The instances would be constructed from the information extracted from the medical abstracts. However, some of the classes are expected to be "closed classes" with a limited number of instances. We are attempting to construct a comprehensive list of instances for these closed classes.

2. CURRENT MEDICAL ONTOLOGIES

The majority of the literature on biomedical ontology development and use is focused on gene ontologies, particularly the Gene Ontology Project[11,12], to annotate genes, sequences, and gene expressions and functions. Ontologies on other biomedical domains have been developed for the following purposes:

- *For modelling biological systems*. For example, Antezana, *et al*.[13] developed a cell-cycle ontology to represent the cell cycle process and to support analysis of cell cycle-related molecular network components. Similarly, Matos[14], *et al*. developed a cell component ontology to support modelling of the cell structure and properties of functional components.

- *For supporting data or information management*. Min[15], *et al*. developed a prostate cancer ontology to support the management and integration of prostate cancer clinical data. Stenzhorn[16], *et al*. developed an ontology of concepts related to cancer care and research, to manage and integrate clinical trials data. The ontology was used to support a clinical trial management system.

- *For supporting a clinical decision support system*. Khoo and Na[17] developed an ontology to represent the knowledge-base for a clinical decision support system for wound management. Kawazoe and Ohe[18] developed an ontology for a clinical decision support system that alerted doctors to adverse drug events. The ontology was used to map clinical data to domain concepts.

- *For supporting database searching*. Mabotuwana and Warren[19] developed an ontology to support identification of hypertension patients from an electronic medical records system that matched certain profiles indicating that their clinical management could be improved. Bratsas[20], *et al*. developed an ontology to support fuzzy query expansion in an information retrieval system for medical computational problem solutions.

- *For specifying healthcare processes*. Ye[21], *et al*. developed a clinical pathway ontology to support the modelling and computerisation of clinical pathways for patient care. Din, Abidi and Jafarpour[22] developed a nursing care plan ontology to support modelling of clinical practice guidelines and patient-specific care plans that can be computerised to provide recommendations to nurses.

- *For mapping the concepts and knowledge* in a particular medical area. Haghighi[23], *et al*. developed a mood disorder ontology to survey the medical and psychological approaches to mood disorders based on an analysis of 5000 articles. Rosse and Mejino[24] developed. The foundational model of anatomy ontology—a spatial-structural ontology of the entities and relations, which represent the structure of the human body at different levels of granularity.

A search of the literature did not locate any general-purpose disease-treatment ontology, especially one that can represent information found in published medical papers. However, some treatment-related semantic relations are defined in the UMLS semantic network[25] and the NCI thesaurus (http://ncit.nci.nih.gov/ncitbrowser). The UMLS semantic network includes the following semantic relations that can be used to indicate various types of treatment effects: *Affects, manages, treats, disrupts, complicates, interacts_with, prevents, brings_about, produces, causes* and *result_of*. The NCI thesaurus also has several relations that indicate the effect of drugs/treatment on diseases, including:

- *Chemical_or_Drug_FDA_Approved_for_Disease*
- *Chemical_or_Drug_Has_Accepted_Therapeutic_Use_For*
- *Chemical_or_Drug_Has_Physiologic_Effect*
- *Chemical_or_Drug_Affects_Abnormal_Cell*
- *Chemical_or_Drug_Affects_Cell_Type_or_Tissue*
- *Regimen_Has_Accepted_Use_For_Disease*
- *Technique_Has_Target_Disease*
- *Technique_Has_Target_Organism.*

There are also relations linking type of surgery or procedure with types of cancer growths:

- *Procedure_Has_Excised_Anatomy*
- *Procedure_Has_Completely_Excised_Anatomy*
- *Procedure_Has_Partially_Excised_Anatomy*
- *Procedure_May_Have_Excised_Anatomy*

Some of the relations incorporate modality indicators (e.g. "may have" and "partially"). However, the NCI and UMLS relations are binary valued and relate a treatment to a disease, but cannot model related details of the treatment, disease, effect and research evidence. Our focus is to develop an ontological structure that will link related aspects of treatment and disease together. Since the completion of this study, we have come across the PhD thesis of Williams[26] which reported the development of a breast cancer ontology based on a set of medical papers dealing with the treatment of invasive breast cancer with Tamoxifen. Willims' ontology focused on the chracteristics of the disease and patient, and the changes in various types of risks from the treatment. The treatment regimes themselves were not modelled in detail.

3. ONTOLOGY DEVELOPMENT METHODOLOGIES

Our disease-treatment ontology builds on existing medical taxonomies and ontologies—the UMLS semantic network, the Medical Subject Headings (MeSH) and the NCI Thesaurus. The additional concepts and semantic relations specified in the ontology will enrich the existing medical ontologies by linking diseases with potential medical treatments and representing additional details related to the treatment administration and dosage, the types of effects and the supporting research evidence. Several methodologies have been proposed in the literature for developing ontologies[27], such as the Cyc method[28] Methontology[29] and the On-To-Knowledge method[30] . However, most of these were designed for solving particular difficulties in developing large-scale ontologies, such as managing geographically distant cooperative teams, easing enormous knowledge acquisition, and merging/alignment of heterogeneous ontologies[27]. These methods were not adopted for developing our disease-treatment ontology, which is a small-scale ontology. Ontology design remains partly an art, regardless of the method used[31].

The ontology was developed based on the saturation principle[32] . The basic idea is that a training set of documents is used to build the ontology. Then a test set is used to evaluate whether the ontology covers most of the relevant concepts and relations in the domain. If new concepts are found in the test set, the test set documents is added to the training set and the new classes and properties are added to the ontology. A second test set is used to assess the ontology again, and the process iterated until no new concepts are found.

For the current disease-treatment ontology, 40 abstracts were used as the training set, where important concepts related to treatment were identified by manual analysis. The specific concepts were then generalised to derive broad classes and relations, which were likely to be applicable across different medical domains. An additional 10 abstracts as the test set were analysed in a similar way to identify new treatment-related concepts and relations not covered by the ontology. No new concepts were found in the test set. However, additional "instances" of concepts (e.g. particular drugs or procedures) were found. Therefore, the 40 abstracts were considered sufficient for developing a high-level disease-treatment ontology for colon cancer. As an extension of the current study, sample abstracts from the breast cancer and heart disease domains are being analysed to find out if the current ontology is sufficient.

The ontology development process was carried out by the authors. Twenty-five abstracts were first analysed jointly by all the authors, with the last 15 abstracts analysed by one of the authors. All the classes and properties were reviewed by the authors to ensure logical consistency. Evaluation using the test set of 10 abstracts was carried out by two of the authors. Admittedly, there is some amount of subjectivity involved in developing an ontology, and some design decisions were made. Identification of challenges in developing such an ontology and the design decisions that had to be made are reported later.

The focus of the effort is to ensure that the ontology covers most of the treatment concepts and relations found in the abstracts, and is logical and principled (i.e. follows certain design principles decided by the team).

4. TOP-LEVEL OF DISEASE-TREATMENT ONTOLOGY

The top-level of the disease-treatment ontology is shown in Fig. 1. Each square or node represents a class or concept. Each directed arc or arrow represents a property or relation. The top-class in the ontology is *Disease-Treatment* which represents specific treatments that are considered for a particular disease, as described in medical articles. An instance of a *Disease-Treatment* is not a specific treatment customised for a particular patient's condition, but a generic treatment that is applicable to a group of patients with a similar disease instance and condition.

Information related to a *Disease-Treatment* is divided into five aspects, represented by the following properties and top-level classes:

- *hasDisease-> Disease*
- *hasTreatment-> Treatment*
- *hasCondition-> Condition*
- *hasEffect-> Effect*
- *hasEvidence-> Evidence*
- *name->* <character string>
- *rdfs:comment-> <character string>*

Condition refers to additional conditions or attributes of the patient that may affect the efficacy of the treatment. This includes the age and gender of the patient, medical history, and other diseases the patient has. *Evidence* refers to the research design and data reported in the medical article that supports the *Effect*.

In the OWL language, properties (i.e. relations) can have sub-properties (sub-relations) to form a relation hierarchy. We define two sub-properties for the *hasTreatment* property to handle more complex treatments:

- *hasPrimaryTreatment*
- *hasSecondaryTreatment.*

Currently no constraints are imposed on the number of times each component can occur for a *Disease-Treatment*. So a treatment may comprise a combination of treatments (e.g. two drugs, or a drug and surgery) and may be applied to a combination of diseases in a patient, with multiple effects. There may also be multiple studies (evidence) for this disease-treatment.

We now take a closer look at each of the top-level classes of *Disease-Treatment*.

5. THE DISEASE CLASS

The *Disease* class is specified as *equivalentClass* to the *Diseases* class in MeSH. An instance of *Disease* will be an instance of one of the sub-classes of MeSH *Diseases*. Alternatively, we can import the *Disease, Disorder or Finding* taxonomy from the NCI Thesaurus.

Instances of a *Disease* would have several properties, depending on the type of disease. The list of properties of the *Disease* class is given in Fig. 2. We have specified some general properties that are applicable to many types of diseases:

- *hasLocation->Anatomy*– equivalent to the *Anatomy* class in MeSH and the *Anatomic Structure, System or Substance* class in the NCI Thesaurus
- *hasPhysicalFeature->Feature*
- *hasDuration->TemporalQuantity*
- *hasDiseaseModifier-> DiseaseModifier*

The property *hasPhysicalFeatures* that can be sub-divided into sub-properties to represent various types of physical features, including *hasSize-> Quantity.* A list of common physical features encountered in medical abstracts is being constructed. It is expected that each disease domain will have its own peculiar set of physical

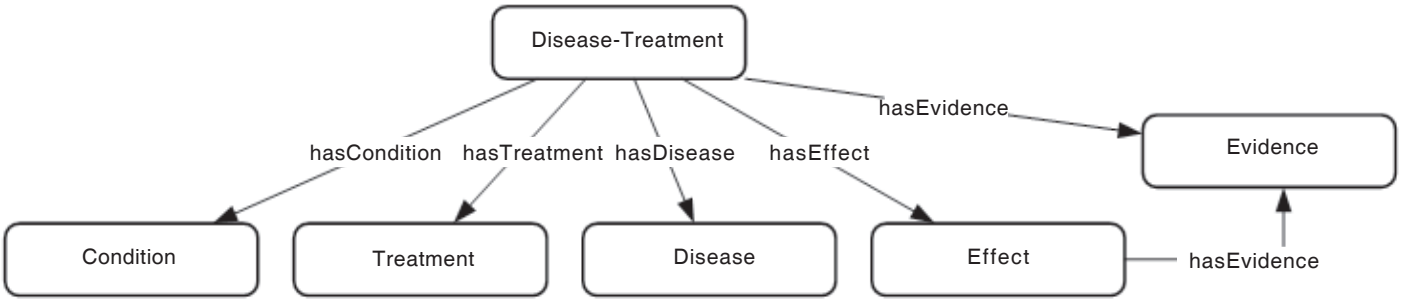

Figure 1. Top-level classes of the Disease-Treatment ontology.

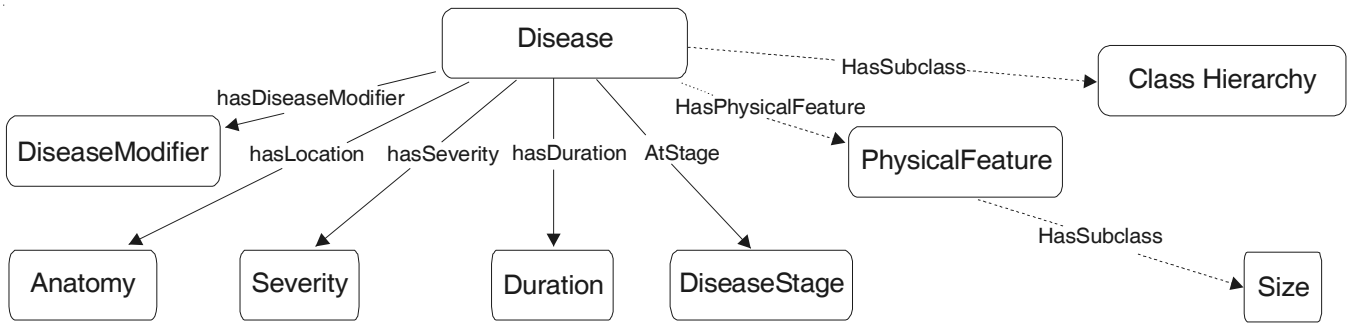

Figure 2. Properties of the disease class.

features, as exemplified by the following text describing colon polyps: “Polyps presented as sitting on a wide pedicle (25), sprawling (19) or spreading flat (7) in rectum (24), sigmoid (18) or colon (9). Polyp diameter ranged 1.0-1.9 cm (12 adenomas), 2-2.9 (18) or 3-5 cm (21).”

The *Quantity* class is used for all kinds of measures, and have the properties *quantityValue* and *quantityUnit*, e.g.

<Quantity>

<quantityValue rdf:datatype=”&xsd;float”>

59.5</quantityValue>

<quantityUnit rdf:resource=”#Kilogram”/>

</Quantity>

TemporalQuantity is a specialised type of quantity defined as a subclass of *Quantity*.

hasDiseaseModifier is adapted from the NCI Thesaurus to indicate the severity and other non-physical characteristics of the disease. We have defined the sub-properties *hasDiseaseStage* and *hasDiseaseSeverity*. The NCI Thesaurus lists various groups of disease modifiers including disease stage (e.g. Stage III), disease morphology (e.g. malignant, invasive), disease grade (e.g. low grade, poorly differentiated), disease clinical (e.g. acute, secondary, recurrent, chronic), disease gross finding (e.g. metastatic, palpable), and disease treatment (e.g. resected, unresectable), most of which have implications regarding the severity of the disease.

Specialised properties for specific types of diseases need to be developed.

6. THE TREATMENT CLASS

Figure 3 shows the properties defined for the *Treatment* class. The *Treatment* class has the sub-

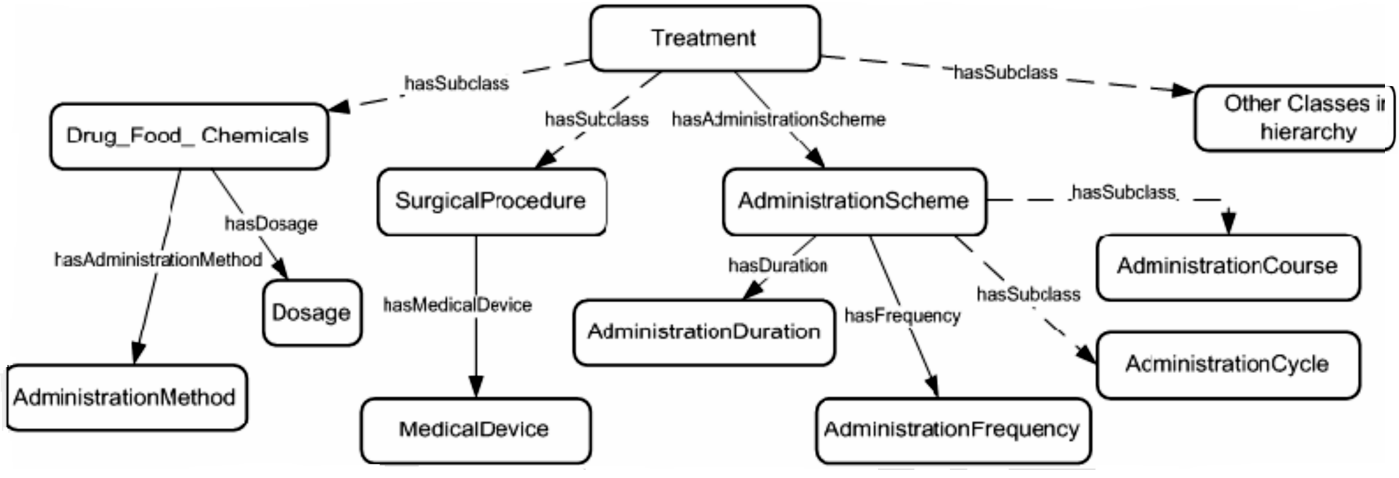

Figure 3. Properties of the treatment class.

classes *Therapeutics and SurgicalProcedures* taken from MeSH or from NCI Thesaurus—"Chemotherapy Regimen or Agent Combination", "Drug, Food, Chemical or Biomedical Material", "Gene", and "Intervention or Procedure".

The *Treatment* class is associated with a property that we call *AdministrationScheme* which is divided into two components (i.e. properties):

- *hasFrequency->AdministrationFrequency*–indicating how often the treatment is administered, e.g. 3 times a day, every 4 hours, exactly once, or whenever needed).
- *hasDuration->AdministrationDuration*–indicating how many times or for how long the treatment is administered, e.g. 10 days, or exactly once.
- *AdministrationScheme* has two subclasses:
- *AdministrationCourse* refers to one course of treatment
- *AdministrationCycle* refers to more complex treatments where a course of treatment has to be repeated for several cycles at regular intervals.

Thus *AdministrationCourse* refers to a local cycle whereas *AdministrationCycle* refers to a global or higher-level cycle. For example, in the text "one of the adjuvant treatment schemes: 5-FU 450 mg/m2 and leucovorin 100 mg/m^2 x 5 days every 4 weeks; six cycles" indicates an *AdministrationCourse* once a day lasting 5 days, with an *AdministrationCycle* of one course every 4 weeks lasting 6 cycles.

Additional properties are also specified for specific sub-classes of *Treatment*. Two sub-classes of cancer treatment are of particular interest:

- *SurgicalProcedure*–equivalent to the *Surgical Procedures, Operative* class in MeSH, or the *Surgical Procedure* class in the NCI Thesaurus.
- *Drug, Food, Chemicals*–equivalent to the *Drug, Food, Chemical or Biomedical Material* class in the NCI Thesaurus.
- Additional properties defined for *Drug, Food, Chemicals* are:
- *hasDosage->Dosage*
- *hasAdministrationMethod->AdministrationMethod.*

SurgicalProcedure has the property: *hasMedicalDevice->MedicalDevices.* The NCI Thesaurus has a similar relation, *Technique_Uses_Equipment*, to link an equipment to a technique

The example text "UFT 400 mg/day to be taken orally" indicates a *Dosage* of "400 mg", *AdministrationFrequency* of once a day, and *AdministrationMethod* of "oral". A list of 111 routes of administration can be found in the CDER DATA STANDARDS MANUAL (U.S. Food and Drug Administration, 2006).

7. THE EFFECT CLASS

The properties of the *Effect* class are shown in Fig. 4. The properties are:

- *hasMeasurementMethod->MeasurementMethod*–the method used for measuring the effect.
- *hasModality->Modality*–the degree of confidence or the truth value of the occurrence of the effect, i.e. "possible", "probable", "unlikely", "true", and "negation".

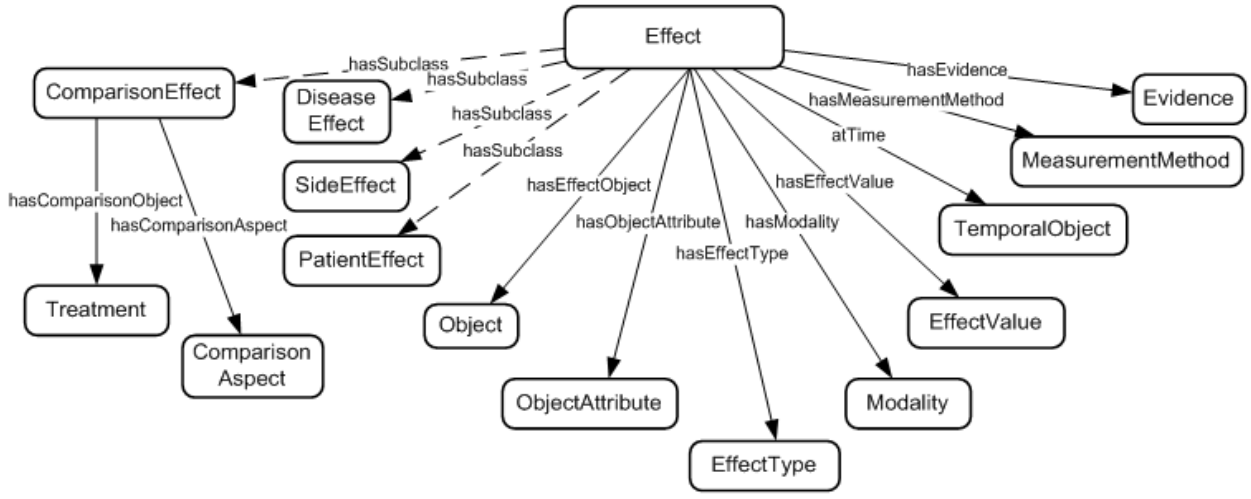

Figure 4. Properties of the effect class.

- *hasEffectObject->Object*–the object that is affected.
- *hasObjectAttribute->ObjectAttribute*–the specific attribute of the object affected.
- *hasEffectType->EffectType*–for example, "reduce", "inhibit", "prevent", and "remove".
- *hasEffectValue->EffectValue*–the direction and extent of the effect. *EffectValue* has the properties *value*, *comparisonValue differenceValue* and *percentDifference*, which are linked to some *Quantity*.
- *atTime->TemporalObject*–to indicate when or how long after treatment the effect occurs.
- *hasEvidence*->Evidence to link to details of the research study and research results.

Table 1. List of effecttypes

cure	chemopreventive
diseaseFree	eliminate risk
backToNormal	reduce risk
effective (potent)	improve
anti-	worsen
destroy	increase
maintain	enhance
control	reduce
delay	induce
prevent	quicken
suppress	no difference
inhibit	no adverse effects
positive outcome	

Table 1 lists the *EffectType* we have encountered in the 50 medical abstracts analysed. Figure 5 gives an example text illustrating some of the elements of *Effect*.

The class *TemporalObject* has two sub classes:

- *RelativeTime* refers to time relative to a *ReferenceEvent*.
- *AbsoluteTime* refers to an absolute time point or a range of time.

For example, the text "the patient died 1 year after the first medical examination" indicates that death (*PatientEffect*, explained later) occurred at a *RelativeTime.value* of 1 year after the first medical examination (*ReferenceEvent*).

The growth rate [*ObjectAttribute*] of C26 tumours [*EffectObject*] treated with CEFUS was significantly reduced [*EffectType*] with respect to untreated controls at day 7 [*TimeAfterTreatment*] (96% [*EffectValue.value*] of average initial tumour volume in CEFUS group versus 615% [*EffectValue.comparisonValue*] for controls, P < 0.05).

Figure 5. Example text illustrating some properties of effect.

The *Effect* class is also divided into a taxonomy of sub-classes:

- *Effect*
  - *ComparisonEffect* compares the effect of one treatment with that of another
  - *PatientEffect*–the effect is expressed holistically from the perspective of the patient, e.g. mortality, 5-year survival rate, quality of life in the first postoperative month, mental health, and minimal follow-up period.
  - *DiseaseEffect*–specifies the effect on the disease, e.g. "the colon polyps were removed completely through polypectomies."
  - *SideEffect*–a secondary and usually adverse effect/ reaction/complication, e.g. "The majority of our complications were bacterial infections, followed by rejections and relaparotomies."

ComparisonEffect has two properties:

- *hasComparisonObject->Treatment*–the treatment for the same disease being compared against)
- *hasComparisonAspect->ComparisonAspect*–specifies the criterion on which the treatments are compared. Examples (i.e. instances) of *ComparisonAspect* includes s*afety, operating time*, *hospital length of stay*, and economic and technological *feasibility*.

For example, in an article that compares laparoscopic surgery with open surgery for rectal cancer, the text "the mean operative time was significantly increased in the laparoscopic group; however, this group showed faster recovery of bowel function. Although anastomotic leakage was comparable between 2 groups, surgical wound infection was significantly higher in open surgery group" indicates evaluation criteria of "operative time", "recovery" rate, "anastomotic leakage" and "infection" during the comparison.

8. THE CONDITION CLASS

The *Condition* class has three sub-classes and one property (Fig. 6):

- *PatientCondition* refers to information on age, gender and other personal attributes, and the patient's medical history. Personal attributes is equivalent to the *Personal Attributes* class in the NCI thesaurus.
- *DiseaseCondition* specifies other co-existing diseases, e.g. "accompanied with colon ileus" and "the neighbouring organs are affected."

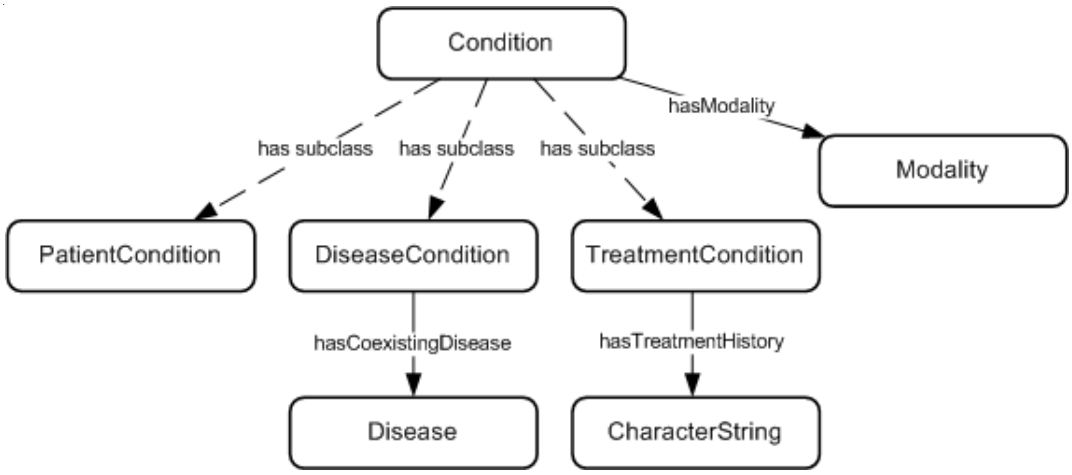

Figure 6. Properties of the condition class.

- *TreatmentCondition* indicates the supportive/optional/recommended medical environment prior to or during the treatment, including treatment history, e.g. "treatment applied after resection surgery" and "in combination with surgery and chemotherapy."

- *hasModality->Modality* refers to modal modifiers such as negation and possibility.

For instance, the text "laparoscopic approach is only contraindicated if the neighbouring organs are affected" could be represented as an instance of *DiseaseCondition*, i.e. "the neighbouring organs are affected" with *Modality* "negation". The following text serves as an example for three kinds of conditions: "A 50-year-old man was diagnosed with non-resectable scirrhous gastric cancer of antrum accompanied with colon ileus due to direct invasion of the transverse colon. As the ileus improved after cecostomy, chemotherapy with TS-1/cisplatin(CDDP) was first performed. Because of no response, 4 cycles of paclitaxel (PTX)/doxifluridine (5'-DFUR) therapy was performed as second-line chemotherapy." The following conditions are specified:

PatientCondition: Age "50 years", gender "male"

DiseaseCondition: Co-existing disease "colon ileus"

TreatmentCondition: TreatmentHistory "cecostomy" and "chemotherapy with TS-1/cisplatin(CDDP)".

9. THE EVIDENCE CLASS

The *Evidence* class has the following properties (Fig. 7):

- *providedBy->Organisation*

- *hasResearchLocation->Location*

- *hasResearchType->ResearchType* e.g. systematic review, randomised controlled trial, cohort studies, case control studies, case reports/series, animal research/laboratory studies.

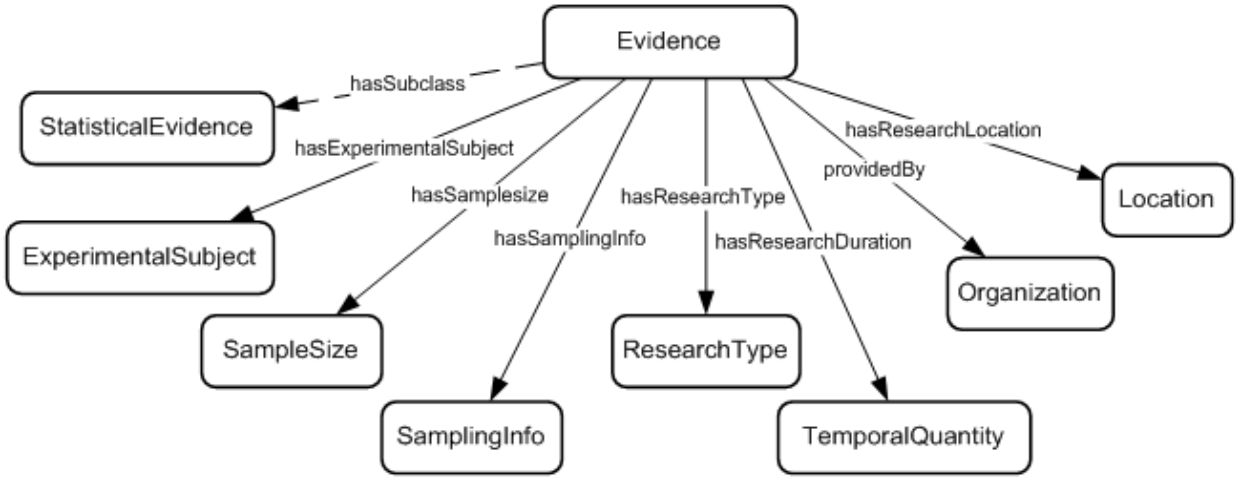

Figure 7. Properties of the evidence class.

- *hasExperimentalSubject->ExperimentalSubject*-for example, person, nude mouse, and human cells
- *hasSampleSize->SampleSize*
- *hasSamplingInfo->SamplingInformation*
- *hasDuration-> TemporalQuantity*

The *Evidence* class has subclass *StatisticalEvidence* to represent detailed statistical information including significance, p-value and 95% confidence interval.

10. CONCLUSION

We have described the initial version of the disease-treatment ontology that we are developing to encode treatment information reported in medical abstracts in the Medline database. The ontology was developed from an analysis of 40 abstract in the domain of colon cancer therapy. A further 10 abstracts were used to verify that the ontology is reasonably complete and can represent most of the information related to colon cancer therapy. We found that the ontology can cover all the types of information (classes) associated with colon cancer therapy in the 10 test abstracts. An example of a disease-treatment instance extracted from a medical abstract and represented using the ontology is given in the *Appendix 1.* The ontology instance is presented in a template format, rather than in OWL, for readability. Current work is focused on constructing a comprehensive list of the subclasses of the main classes and instances of the closed classes. Closed classes are classes that have a limited number of instances that can be enumerated completely, e.g. *Modality* and *EffectType*.

The ontology is being extended to cover treatment of two other domains—breast cancer and heart disease. It was found that these two domains have different kinds of treatments and the following additional treatment concepts need to be added to the ontology: *Radiation Therapy* (a subclass of cancer treatment and equivalent to the *Radiation Therapy* class in the NCI Thesaurus), *hasProtein & gene expression status-> Protein & gene expression status* (a sub-property of *hasDiseaseModifier*).

REFERENCES

1. Swanson, D.R. & Smalheiser, N.R. An interactive system for finding complementary literatures: A stimulus to scientific discovery. *Artificial Intelligence,* 1997, **91**, 183-203.
2. Bekhuis, T. Conceptual biology, hypothesis discovery, and text mining: Swanson's legacy. *Biomed. Digit. Libr.,* 2006, **3**(2).
3. Sackett, D.L.; Rosenberg, W.M.C.; Gray, J.A.M.; Haynes, R.B. & Richardson, W.S, Evidence based medicine: What it is and what it isn't (Editorial). *Brit. Medi. J.,* 1996, **312**, 71-72.
4. Guyatt, G.; Cook, D.; & Haynes, B. Evidence-based medicine has come a long way. *Brit. Medi. J.,* 2004, **329**, 990-91.
5. U.S. National Library of Medicine. Unified Medical Language System (Fact Sheet). 2006, http://www.nlm.nih.gov/pubs/factsheets/umls.html (accessed on 1 October 2010)
6. U.S. National Library of Medicine. UMLS reference manual, 2009 http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=nlmumls. (accessed on 1 October 2010).
7. U.S. National Library of Medicine. Medical Subject Headings (MeSH) (Fact Sheet), 2010. http://www.nlm.nih.gov/pubs/factsheets/mesh.html.
8. Gruber, T.R. A translation approach to portable ontology specification. *Knowledge Acquisition,* **5**(2), 199-220.
9. Genesereth, M.R.; Nilsson, N.J. Logical foundations of artificial intelligence. Morgan Kaufmann, 1987.
10. World Wide Web Consortium, OWL Web Ontology Language: Overview (W3C Recommendation 10 February 2004). http://www.w3.org/TR/2004/REC-owl-features-20040210/. (accessed on 1 June 2007)
11. The Gene Ontology. Gene Ontology documentation, 2010. http://www.geneontology.org/GO.contents.doc.shtml (accessed on 1 October 2010).
12. Gene Ontology Consortium. The Gene Ontology (GO) project in 2006. *Nucleic Acids Res.,* 2006, **34**.
13. Antezana, E.; Egaña, M.; Blondé, W.; Illarramendi, A.; Bilbao, I.; De Baets, B.; Stevens, R.; Mironov, V. & Kuiper, M. The cell cycle ontology: An application ontology for the representation and integrated analysis of the cell cycle process. *Genome Biology,* 2009, **10**(5), R58.
14. Matos, E. E.; Campos, F.; Braga, R. & Palazzi, D. CelOWS: An ontology based framework for the provision of semantic web services related to biological models. *J. Biomed. Inf.,* 2010, **43**(1), 125-36.
15. Min, H.; Manion, F. J.; Goralczyk, E.; Wong, Y.N.; Ross, E. & Beck, J.R. Integration of prostate cancer clinical data using an ontology. *J. Biomed. Inf.,* 2009, **42**(6), 1035-45.

16. Stenzhorn, H.; Weiler, G.; Brochhausen, M.; Schera, F.; Kritsotakis, V.; Tsiknakis, M.; Kiefer, S. & Graf, N. The ObTiMA system: Ontology-based managing of clinical trials. *Stud. Health Technol. Infor.,* 2010, **160**, 1090-94.

17. Khoo, C.S.G. & Na, J.C. Issues in ontology design for a clinical decision support system. *In* Asia-Pacific Conference on Library & Information Education and Practice (A-LIEP 2009) University of Tsukuba, Tsukuba. http://a-liep.kc.tsukuba.ac.jp/proceedings/index.html (accessed on 1 October 2010).

18. Kawazoe, Y. & Ohe, K. An ontology-based mediator of clinical information for decision support systems: A prototype of a clinical alert system for prescription. *Meth. Inf. Med.,* 2008, **47**(6), 549-59.

19. Mabotuwana, T. & Warren J. An ontology-based approach to enhance querying capabilities of general practice medicine for better management of hypertension. *Art. Intell. Med.,* 2009, **47**(2), 87-103.

20. Bratsas, C., Koutkias, V., Kaimakamis, E., Bamidis, P., & Maglaveras, N. (2007). Ontology-based vector space model and fuzzy query expansion to retrieve knowledge on medical computational problem solutions. *In* Engineering in Medicine and Biology Society, 2007 (EMBS 2007): 29th Annual International Conference of the IEEE, 2007. pp. 3794-97.

21. Ye, Y.; Jiang, Z.; Diao, X.; Yang, D. & Du G. An ontology-based hierarchical semantic modeling approach to clinical pathway workflows. *Comp. Bio. Med.,* 2009, **39**(8), 722-32.

22. Din, M.A.; Abidi, S.S. & Jafarpour, B. Ontology-based modelling and execution of nursing care plans and practice guidelines. *Stud. Health Technol. Inform.*, 2010*,* **160**, 1104-08.

23. Haghighi, M.; Koeda, M.; Takai, T. & Tanaka, H. Development of clinical ontology for mood disorder with combination of psychomedical information. *J. Med. Dental Sci.,* 2009, **56**(1), 1-15.

24. Rosse, C. & Mejino, J.L.V. The foundational model of anatomy ontology. *Anat. Onto. Bioinfor.: Princip. and Prac.*, 2007, **6**, 59-117.

25. U.S. National Library of Medicine. Unified Medical Language System: Current relations in the semantic network, 2010, http://www.nlm.nih.gov/research/umls/META3_current_relations.html (accessed on 1 October 2010)

26. Williams, M.H. Integrating ontologies and argumentation for decision-making in breast cancer. Unpublished PhD thesis, Department of Computer Science, University College London. http://www.acl.icnet.uk/~mw/

27. Fernandez-Lopez, M., & Corcho, O. Ontological engineering: With examples from the areas of knowledge management, e-Commerce and the semantic Web. Springer-Verlag, Berlin, 2004.

28. Lenat, D.B. & Guha, R.V. Building large knowledge-based systems: Representation and inference in the Cyc project. Addison-Wesley. Boston, 1990.

29. Fernandez-Lopez, M. & Gomez-Perez, A. Overview and analysis of methodologies for building ontologies. *Know. Engi. Rev.,* 2002, **17**, 129-56.

30. Staab, S.; Studer, R.; Schnurr, H.P. & Sure, Y. Knowledge processes and ontologies. *IEEE Intell. Syst.* 2001*,* **16**(1): 26-34.

31. Ferreira, C.R., *et al*. Ontology design risk analysis. *In* Proceedings of the 2007 OTM Confederated International Conference on On the move to Meaningful Internet Systems (Part 1), Springer, Berlin, 2007. pp 522-33.

32. Holloway, I. Basic concepts for qualitative research. Wiley-Blackwell, Oxford, 1997.

About the Authors

Dr Christopher Khoo is an Associate Professor and Head of the Division of Information Studies at Nanyang Technological University, Singapore. He teaches courses in knowledge organisation, information architecture, data mining and Web-based information systems. He obtained his PhD from Syracuse University and his MSc in Library and Information Science at the University of Illinois, Urbana-Champaign. He has also worked for several years as a science reference librarian, cataloger and online information searcher at the National University of Singapore Libraries. His main research interests are in text mining (information extraction, automated sentiment analysis and text summarisation), medical decision support system, knowledge organisation, and human categorisation behaviour.

Dr Jin-Cheon Na is currently Associate Professor at the Wee Kim Wee School of Communication and Information, Nanyang Technological University (NTU), Singapore. Before he joined NTU, he was a senior researcher at the Agency for Defence Development, Korea. He obtained his PhD from the Department of Computer Science at Texas A&M University. He has published more than 50 papers mainly in the areas of digital libraries, hypertext and hypermedia, document engineering, and knowledge organisation. He is a member of the Editorial Review Board of the *International Journal of Organisational and Collective Intelligence* (*IJOCI*). He was also a non-executive director of DideoNet Co. Ltd, Korea.

Ms Vivian Wang obtained her Bachelor's degree in Computer Science from Wuhan University, China, and her Master of Applied Science degree by research from Nanyang Technological University (NTU). She was a recipient of an NTU Research Scholarship. Vivian contributed to the design and development of the Shared Services Platform. Prior to HP, she was the Head of Product Development at DecisionWare, a computer simulation company based at the National University of Singapore. Vivian is currently a research analyst at the Centre for Health Services Research, Singapore Health Services, where she researches in the area of health economics. Her current projects include economic evaluation related to HPV vaccination, diabetic retinopathy screening, genetic testing for colorectal cancer, gefitinib second-line treatment for lung cancer and transradial percutaneous coronary intervention.

Dr Syin Chan received her Bachelor's degree (First Class Honours) in Electrical Engineering from the National University of Singapore in 1987, and PhD in Computer Science from the University of Kent, United Kingdom in 1993. She is an Associate Professor in the School of Computer Engineering, Nanyang Technological University. Her research interests include mobile applications and multimedia information systems.

Example of Disease-Treatment Information Extracted from a Medical Abstract

PubMed Record No.: PMID-16367907

Citation: Inamine, M., et al. (2005). Inhibitory effect of dietary monoglucosylceramide 1-O-beta-glucosyl-N-2'-hydroxyarachidoyl-4,8-sphingadienine on two different categories of colon preneoplastic lesions induced by 1,2-dimethylhydrazine in F344 rats. *Cancer Science, 96*(12), 876-881.

Abstract: we evaluated the preventive efficacy of enriched dietary monoglucosylceramide 1-O-beta-glucosyl-N-2'-hydroxyarachidoyl-4,8-sphingadienine (G(1)CM) on 1,2-dimethylhydrazine (DMH)-induced aberrant crypt foci (ACF) and beta-catenin-accumulated crypt (BCAC) formation in F344 rats during initiation stage. ... Pure G(1)CM was isolated from rice bran. … Forty-two rats were divided randomly into five experimental groups. ... One week before the first injection of DMH, rats in groups 2 and 3 were fed a diet containing 200 and 1,000 p.p.m. G(1)CM, respectively, for 5 weeks. … The experiment was terminated 5 weeks after the start. Dietary G(1)CM at both doses (groups 2 and 3) significantly inhibited the induction of ACF and BCAC ($P<0.001$) when compared to group 1 treated with DMH alone. In groups 2 and 3, the proliferating cell nuclear antigen labeling indices of epithelial cells in ACF and BCAC were also lower than in group 1 ($P<0.0001$ for ACF, $P<0.05$ for BCAC). These results, that dietary G(1)CM has possible chemopreventive effects in the present short-term colon carcinogenesis bioassays, suggest that longer exposure may cause suppression of tumor development.

Notes:

1. Two types of colon lesions are mentioned in the abstract. Each lesion would be represented as a separate disease-treatment instance. However, they are placed together in this example for easy reading.

2. # indicates an anonymous instance of the class

Table 1. Disease-Treatment: PMID-16367907

Disease	
hasDisease-> aberrant crypt foci (ACF) formation: #	Comment-> "Subclass of colon preneoplastic lesion" hasDiseaseStage - > DiseaseStage: initiation stage hasLocation - > Colon: #
hasDisease-> beta-catenin-accumulated crypt (BCAC) formation: #	atStage - > DiseaseStage: initiation stage hasLocation - > Colon: #
Treatment	
hasTreatment-> monoglucosylceramide 1-O-beta-glucosyl-N-2' hydroxyarachidoyl-4,8-sphingadienine G(1)CM: #	hasAdministrationMethod - > AdministrationMethod: oral hasDosage - > Dosage: # quantityValue -> 200 quantityUnit -> p.p.m hasDosage - > Dosage: # quantityValue -> 1000 quantityUnit -> p.p.m hasAdministrationScheme - > AdministrationCourse: # hasFrequency -> AdministrationFrequency: # quantityValue -> 1 qyantityUnit -> per day hasDuration -> AdministrationDur ation: # quantityUnit -> week
Condition	
hasCondition-> Treatment Condition: #	preparationMethod -> "Pure G(1)CM was isolated from rice bran"
hasCondition-> DiseaseCondition: #	inducedBy -> "1,2-dimethylhydrazine (DMH)"
Effect	
hasEffect-> Disease Effect: #	hasEffectType - > EffectType: chemopreventive hasModality - > Modality: possible hasEvidence -> StatisticalEvidence: #1
hasEffect-> Disease Effect: #	hasEffectType - > EffectType: suppress hasEffectObject - > Tumor: # hasObjectAttribute - > Development: # hasModality - > Modality: possible
hasEffect-> SideEffect: #	hasEffectType - > EffectType: reduce hasEffectObject - > Epithelial cells: # hasObjectAttribute - > Proliferating cell nuclear antigen labeling indices: # hasEvidence -> StatisticalEvidence: #2
Evidence	
StatisticalEvidence: #1	hasExperimentalSubject - > ExperimentalSubject: F344 rats hasSampleSize - > 42 hasResearchType - > ResearchType: animal trials hasRearchDuration - > TemporalQuantity quantityValue -> 5 quantityUnit -> week hasSignificanceLevel -> 0.001
StatisticalEvidence: #2	hasSignificanceLevel -> 0.0001