

# **KR-MED 2006 Proceedings**

Editor: Olivier Bodenreider

Second International Workshop  
on Formal Biomedical Knowledge Representation

*Baltimore, Maryland  
November 8, 2006*

Organized by the National Center for Ontology Research  
(NCOR) and the Working Group on Formal (Bio-)Medical  
Knowledge Representation of the American Medical  
Informatics Association (AMIA)

**KR-MED 2006**

International Workshop - November 8, 2006 in Baltimore, MD, USA

**Biomedical Ontology in Action**







## Forewords

These are the proceedings of the KR-MED 2006, the **Second International Workshop on Formal Biomedical Knowledge Representation**, held in Baltimore, Maryland on November 8, 2006, two years after the first edition. The workshop is organized by the *National Center for Ontology Research* (NCOR) and the Working Group on Formal (Bio-)Medical Knowledge Representation of the *American Medical Informatics Association* (AMIA), and collocated with the *International Conference on Formal Ontology in Information Systems* (FOIS 2006).

Standing on the foundations of biomedical ontologies are the many applications supported by these ontologies. For example, in health care, ontologies are an important component of an interoperable health information technology infrastructure. In biology, ontologies have become essential for annotating the literature and integrating the multiple, heterogeneous knowledge bases resulting from the analysis of high-throughput experiments. The papers and posters featured at the workshop illustrate and demonstrate how current research can be brought to bear on the practical problems associated with the development of applications supported by these ontologies. In other words, they show “biomedical ontology *in action*”.

The scientific content of this workshop was elaborated by an international program committee composed of the following individuals:

- Olivier Bodenreider (National Library of Medicine, USA), Chair
- Anita Burgun (University of Rennes, France)
- Ronald Cornet (Amsterdam Academic Medical Center, The Netherlands)
- Maureen Donnelly (State University of New York at Buffalo, USA)
- Aldo Gangemi (ISTC Laboratory of Applied Ontology, Italy)
- Michael Gruninger (University of Toronto, Canada)
- Anand Kumar (IFOMIS, Germany)
- Philip Lord (University of Newcastle, UK)
- Yves Lussier (University of Chicago, USA)
- Onard Mejino (University of Washington, USA)
- Peter Mork (Mitre Corporation, USA)
- Fabian Neuhaus (State University of New York at Buffalo, USA)
- Daniel Rubin (Stanford University, USA)
- Stefan Schulz (Freiburg University Hospital, Germany)
- Lowell Vizenor (National Library of Medicine, USA)
- Chris Welty (IBM T.J. Watson Research Center, USA)
- Jennifer Williams (OntologyWorks, USA)

Lan Aronson, Tracy Craddock, Olivier Dameron, Keith Flanagan, Kin Wah Fung, John Kilbourne, Erik van Mulligen, Serguei Pakhomov, Valentina Presutti, Tom Rindflesch, Daniel Swan and Chris Wroe also helped referee the submissions.

27 papers were submitted to KR-MED 2006, of which the program committee selected ten for oral presentation. Four submissions will be presented as posters. An electronic version of the workshop proceedings will soon be available on CEUR-WS (<http://CEUR-WS.org>). In addition, an extended version of the best papers will be selected for publication in the journal *Applied Ontology*.

The Organizing Committee of KR-MED 2006 includes Stefan Schulz (Freiburg University Hospital, Germany), Barry Smith (State University of New York at Buffalo, USA) and Fabian Neuhaus (State University of New York at Buffalo, USA). Sandra Smith (NCOR) has orchestrated the logistics of the meeting and Stefan Schlachter (Freiburg University Hospital, Germany) created the KR-MED 2006 web site (<http://www.imbi.uni-freiburg.de/medinf/kr-med-2006/>).

My thanks go to all of you who helped make this workshop possible and successful.

Olivier Bodenreider

National Library of Medicine, USA  
Chair, KR-MED 2006 Scientific Program Committee



## Table of Contents

Registration in Practice: Comparing Free-Text and Compositional Terminological System Based Registration of ICU Reasons for Admission . . . . .	3
<i>Nicolette de Keizer, Ronald Cornet, Ferishta Bakhshi-Raiez, Evert de Jonge</i>	
Binding Ontologies & Coding Systems to Electronic Health Records and Messages . . . . .	11
<i>AL Rector, R Qamar, T Marley</i>	
Using Ontology Graphs to Understand Annotations and Reason about Them . . . . .	21
<i>Mary E. Dolan, Judith A. Blake</i>	
An Online Ontology: WiktionaryZ . . . . .	31
<i>Erik M. van Mulligen, Erik Möller, Peter-Jan Roes, Marc Weeber, Gerard Meijssen, Barend Mons</i>	
“Lmo-2 interacts with <i>elf-2</i> ” On the Meaning of Common Statements in Biomedical Literature . . . . .	37
<i>Stefan Schulz, Ludger Jansen</i>	
The Qualitative and Time-Dependent Character of Spatial Relations in Biomedical Ontologies . . . . .	47
<i>Thomas Bittner, Louis J. Goldberg</i>	
Towards a Reference Terminology for Ontology Research and Development in the Biomedical Domain . . . . .	57
<i>Barry Smith, Waclaw Kusnierczyk, Daniel Schober, Werner Ceusters</i>	
LinKBase®, a Philosophically-Inspired Ontology for NLP/NLU Applications . . . . .	67
<i>Maria van Gorp, Manuel Decoene, Marnix Holvoet, Mariana Casella dos Santos</i>	
The Development of a Schema for the Annotation of Terms in the Biocaster Disease Detecting/Tracking System . . . . .	77
<i>Ai Kawazoe, Lihua Jin, Mika Shigematsu, Roberto Barerro, Kiyosu Taniguchi, Nigel Collier</i>	
BioFrameNet: A Domain-Specific FrameNet Extension with Links to Biomedical Ontologies . . . . .	87
<i>Andrew Dolbey, Michael Ellsworth, Jan Scheffczyk</i>	
Poster Presentations:	
A CPG-Based Ontology Driven Clinical Decision Support System for Breast Cancer Follow-Up . . . .	97
<i>Samina Raza Abidi</i>	
Inferring Gene Ontology Category Membership via Gene Expression and Sequence Similarity Data Analysis . . . . .	98
<i>Murilo Saraiva Queiroz, Francisco Prosdocimi, Izabela Freire Goertzel, Francisco Pereira Lobo, Cassio Pennachin, Ben Goertzel</i>	
Experience with an Ontology of Pediatric Electrolyte Disorders in a Developing Country . . . . .	99
<i>Vorapong Chaichanamongkol, Wanwipa Titthasiri</i>	
Issues in Representing Biological and Clinical Phenotypes Using the Formal Models . . . . .	100
<i>Ying Tao, Chintan Patel, Carol Friedman, Yves A. Lussier</i>	



# Papers



# Registration in practice: Comparing free-text and compositional terminological-system-based registration of ICU reasons for admission

Nicolette de Keizer<sup>a</sup>, Ronald Cornet<sup>a</sup>, Ferishta Bakhshi-Raiez<sup>a</sup>, Evert de Jonge<sup>b</sup>

<sup>a</sup> Dept of Medical Informatics, <sup>b</sup> Dept of intensive Care, Academic Medical Center, Universiteit van Amsterdam, The Netherlands

## Abstract

*Reusability of patient data for clinical research or quality assessment relies on structured, coded data. Terminological systems (TS) are meant to support this. It is hardly known how compositional TS-based registration affects the correctness and specificity of information, as compared to free-text registration. In this observational study free-text reasons for admission (RfA) in intensive care were compared to RfAs that were composed using a compositional TS. Both RfAs were registered in the Patient Data Management System by clinicians during care practice. Analysis showed that only 11% of the concepts matched exactly, 79% of the concepts matched partially and 10% of the concepts did not match. TS-based registration results in more details for almost half of the partial matches and in less details for the other half. This study demonstrates that the quality of TS-based registration is influenced by the terminological system's content, its interface, and the registration practice of the users.*

**Keywords:** Terminological system, information storage and retrieval, medical records, evaluation

## 1. Introduction

Most potential advantages of electronic patient records, such as availability of patient data for decision support and the re-use of patient data for clinical research or quality assessment [1], rely on structured, coded data, not free text [2]. Structured data entry (SDE) [3] and terminological systems (TS) [4] are means to support this process of capturing patient data in a structured and standardized way. SDE is a method by which clinicians record patient data directly in a structured format based on predefined fields for data entry. Terminological systems provide terms denoting concepts and their relations from a specific domain [5] and can be used within predefined fields for data entry.

Nowadays most terminological systems do have a computer-based implementation. Terminological

systems can either enumerate all concepts (pre-coordination), or allow post-coordination, i.e. enabling to compose new concepts by qualifying pre-coordinated concepts with more detail. Generally it takes longer to select and post-coordinate concepts corresponding to a patient's findings, diagnoses, or tests from long lists of standard terms drawn from terminological system than to enter a summary in free text. Worse, the standard codes and terms provided by a terminological system may constrain clinical language [6]. Although the disadvantages of capturing structured, coded data might be outweighed by more informative data and automatic processing of data, evidence on the effect of structured and TS-based registration of patient data on the correctness and specificity of these data compared to free-text is hardly available. Many studies compared the content coverage (correctness and specificity) of a TS by retrospectively coding a set of diagnoses [7]. Studies in which the feasibility of automated coding has been investigated also usually use an experimental design in which free text from a medical record is coded retrospectively by some natural language processing algorithm (e.g. [8,9]). Cimino et al [10] use an observational, cognitive-based approach for differentiating between successful, suboptimal, and failed entry of coded data by clinicians. They used the Medical Entities Dictionary (MED) which only included pre-coordinated concepts. To our knowledge no observational field studies exist in which free-text recording in a medical record is compared with prospectively recorded compositional TS-based diagnoses.

The aim of this observational study is to evaluate how clinicians in every day care practice register reasons for admission (RfA) by using compositional TS-based systems. TS-based registration was compared to free-text registration with regard to correctness and specificity of recorded RfA.

The outcome of this study depends on three factors: the terminological system's content, its interface, and the registration practice of the users. In this study, we aim at distinguishing the effect of content from the effect of the user interface and the user. If structured

TS-based registration of diagnoses results in (at least) the same information as free-text diagnoses, TS-based registration is preferred, as retrieval will be much easier and thereby re-use of the data will be much more feasible. If TS-based registration results in information loss we need to investigate the reasons for this to search for possibilities to improve the terminological system and its use.

## 2. Materials & Methods

### 2.1 PDMS and Terminological system DICE

This study took place in an adult Intensive Care Unit with 24 beds in 3 units, with more than 1500 yearly admissions. Since 2002, this ward uses a commercial Patient Data Management System (PDMS), Metavision. This PDMS is a point-of-care Clinical Information System, which runs on a Microsoft Windows platform, uses a SQL server database and includes computerized order entry; automatic data collection from bedside devices such as a mechanical ventilator; some simple clinical decision support; and (free-text) clinical documentation of e.g. reasons for admission and complications during ICU stay. As part of the National Intensive Care Evaluation (NICE) project [11], a national registry on quality assurance of Dutch ICUs, for each patient a minimal dataset among which the reason for admission is extracted from the PDMS. Since April 1<sup>st</sup> 2005 a pilot study is running in which the compositional terminological system DICE [12] is integrated with the PDMS (see Figure 1) to evaluate its usability for structured registration of reasons for ICU admission. The main reasons for the development of DICE were the need for a terminological system that supports a) registration of intensive-care-specific reasons for admission, commonly either a severe acute medical condition or observation after a large surgical condition b) semantic definitions of concepts, enabling selection of patients by aggregating diagnoses on different features, and c) assignment of multiple synonymous Dutch and English terms to these concepts.

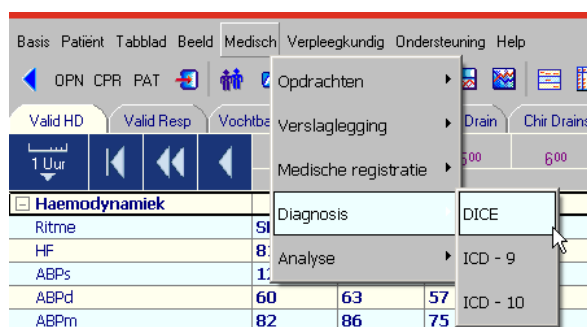


Figure 1: Activation of TS-based registration within the Patient Data Management System

DICE implements frame-based definitions of diagnostic information for the unambiguous and unified classification of patients in Intensive Care medicine. DICE defines more than 2400 concepts including about 1500 reasons for admission and uses 45 relations. DICE is implemented as a SOAP-based Java terminology service together with clients for knowledge modeling and browsing [13]. DICE is used to add controlled compositional terms to clinical records. The implementation of DICE offered the physicians two ways to search for the appropriate diagnosis concept: (a) a short list containing the most frequently occurring diagnoses, (b) entry of (a part of) its preferred or synonymous term. Once a concept is selected, DICE uses post-coordination to provide concepts with more detailed information, as shown in Figure 2. The user interface of the client by which concepts are browsed stimulates but does not enforce users to specify additional qualifiers of a concept, e.g. a Coronary Artery Bypass Graft (CABG) can be further qualified by the number of bypasses; the types of bypasses and whether it was a re-operation or not. At the start of the pilot physicians got a 15-minutes training on the use of DICE. During the pilot, registration of DICE-based reasons for admission as part of the NICE minimal dataset was voluntary. This means that after the first 24 hours of ICU admission a physician could add a controlled term from DICE into the PDMS to describe the reason for ICU admission. As the reason for admission is an essential part of the clinical documentation the regular registration of free-text-based reasons for admission into the PDMS was continued during the pilot for each patient at the time of admission.

### 2.2 Data collection and analysis

For all patients admitted between April 1<sup>st</sup> 2005 and December 1<sup>st</sup> 2005 the free-text reasons for admission



Figure 2: User interface presenting options for post-coordination

and (if available) the structured DICE-based reasons for admission were extracted from the PDMS. As free-text recording of reasons for admission is mandatory, for all patients admitted to the IC a free-text description was available. Since DICE-based registration of reason for admission was voluntary it could be possible that “difficult or complex” reasons for admissions were not registered with DICE. To investigate this possible selection bias the free-text reasons for admission were compared between the groups with and without structured DICE-based reasons for admission.

For each admission having both a free-text reason for admission and one or more DICE-based reasons for admission, these reasons for admission were compared by two independent researchers, both experienced in DICE and intensive care medicine. Each pair consisting of one free-text and one or more DICE-based reason for admission was scored as either being an exact match, partial match or mismatch. A match was considered exact when the DICE-based reason for admission was semantically equivalent to the free-text registration. For example the abbreviated free-text “AVR” was considered an exact match with the DICE concept “aortic valve replacement”. A concept pair was considered as partially matching when one concept subsumed the other (e.g. “3-fold

CABG” and “CABG”) or when the concepts were siblings with equal anatomical and pathological properties (e.g. “hepatitis A” and “hepatitis B”). A concept pair is considered a mismatch in all other cases. For each partial match the two researchers independently assessed which concepts, attributes or relations were missing or were additionally represented in the DICE-based reason for admission. Comments on missing details in the DICE-based registration were classified either as a) “not registered but available in DICE”, b) “value of relation is missing in DICE”, e.g. although a CABG can be qualified by type of graft (LIMA, RIMA, PIMA and venous) the value “LIMA-lad” is missing or c) “relation is missing in DICE”, e.g. “bleeding of the cerebellum, right side” can not completely be registered by DICE since the relation “laterality” is missing.

Different scores of the researchers were solved based on consensus and if necessary by asking an intensivist as an independent third party.

Figure 3 presents an example of a partial match. The free text “AVR-bio + CABG” coming from the clinical documentation part of the PDMS is displayed at the top of the screen dump. In the middle of Figure 3 the DICE-based reasons for admission are presented and at the bottom the scoring of agreement, in this case a partial match, is presented. A “+” indicates that the

PDMS: AVR-bio + CABG

3/19/2005 12:33:00 PM

DICE:

Terms
▶ CABG, Type:LIMA, Number:1
Valve replacement, Operation on dysfunction:Stenosis, Operation localized in:Aortic valve
Angina pectoris, Type:Stable

Record: 1 of 3

Scores:

	Matchtype	Difference	Type of difference	Reason for missing	Agreement
	Partial Match	+	concept AP stable		direct
		+	dysfunction		direct
		+	number		direct
		+	type of bypass		direct
		-	types of prosthesis	Available in Dice	direct

Figure 3: Scorings example of the agreement between free-text “AVR-bio + CABG” and the accompanying set of DICE-based reasons for admission. The bottom part represents the match type, the difference (“+” means DICE has additional detail,” -“ means DICE misses detail), the type of difference, the reason for missing (type of prosthesis is available in DICE) and if the two researchers directly agreed on the differences or after discussion.

DICE based registration includes more detail than the free-text registration on type of CABG, number of bypasses, dysfunction of the aortic valve and the Angina Pectoris diagnosis. The “-” indicates that the free-text registration includes details on the type of valve prosthesis which is not registered in the DICE-based registration, although this qualifier is available in DICE. In this example all differences between the free-text and DICE-based reasons for admission were scored by both researchers which is indicated by “direct” agreement.

In this paper a TS-based diagnosis is regarded as correct when it exactly or partially matches the free-text diagnosis. Specificity of (correct) diagnoses is expressed by as "equal" (exact match), "more specific", "less specific" or "more and less specific" depending on differences in detail of the TS-based diagnoses compared to the free-text diagnoses.

### 3. Results

During the study period 799 admissions to the ICU took place. For all these admissions a free-text reason for admission was available and for 359 (45%) of these admissions a DICE-based reason for admission was available. Those admissions for which a DICE-based registration was missing do not represent other reasons for admissions than those for which a

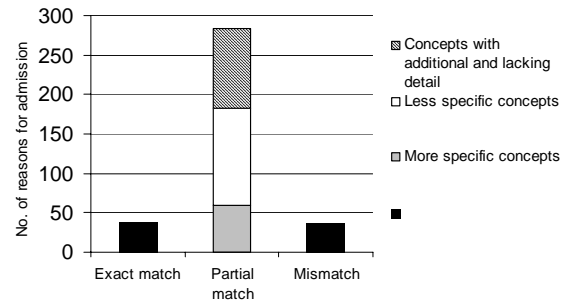


Figure 4: Distribution of exact match, mismatch and partial match (including whether the DICE based reason for admission included more and/or less specific detail).

DICE-based registration was available. One free-text reason for admission could be described by more than one DICE-based reason for admission, e.g. “CABG + AVR” is one free-text reason for admission encoded by two DICE concepts “CABG” and “Aortic valve replacement”. The 359 free-text reasons for admission were described by 457 DICE-based reasons for admission. Half of them were registered as pre-coordinated concepts such as “Pneumonia”, half of them were registered using post-coordination, e.g. “Pneumonia; has aetiology Staphylococcus aureus”.

Figure 4 shows that we found 38 (11%) exact matches, 284 (79%) partial matches and 37(10%) mismatches.

Table1. Example of 5 exact matches, 5 partial matches and 5 mismatches

	Free-text diagnoses	DICE-based diagnoses
Exact matches	THOCR	Oesophageal cardiac resection, entrance: transhiatal
	SAB	Subarchnoid bleeding
	re-CABG x2 venous	CABG, Re-operation: true, Type:Venous graft, Number:2
	Staphylococcal sepsis	Sepsis, has etiology: Staphylococcus aureus
	Stomach bleeding	GI bleeding; localized in stomach
Partial matches	SAB	Subarchnoid bleeding; closing: coil
	Respiratory insufficiency	Respiratory insufficiency; due to: pneumonia
	CABGx3 and Ao-biovalve	CABG & valve replacement
	Respiratoire insufficiency bij benzodiazepine intoxicatie	Accidental intoxication with sedatives and hypnotics
	Large posterior infarction	Acute pulmonary oedema ; due to acute myocardial infarction
Mismatches	Abdominal bleeding	Renal insufficiency
	Hypercapnia with reduced consciousness	COPD
	Hyponatremia with cerebral oedema	Self intoxication
	Resp insufficiency after cardiogenic shock	Myocardial infarction
	Respiratory insufficiency due to pneumonia	Perforated gallbladder

According to our definition 90% ((38+284)/359) of all concepts were correct but for 79% of all concepts (all partial matches), there were some discrepancies in specificity. One-third of the partial matches add some details as well as miss some details compared to the free-text reason for admission. Twenty-two percent of the partial matches was more specific and forty-four percent of the partially matches was less specific compared to the free-text reason for admission. Table 1 shows some examples of exact matches, partial matches and mismatches.

In total 582 comments were given on the 284 partially matched reasons for admission. Two hundred sixty (45%) comments were given on *additional* concepts, attributes or relations registered in the DICE-based registration of reasons for admission that were not described in the free-text reason for admission. On the other hand 325 (55%) comments were given on *missing* concepts, attributes or relations in the DICE-based registration of reasons for admission compared to the free-text reasons for admission.

Figure 5 shows the distribution of the 325 reasons why the DICE-based reasons for admission were missing detail. The majority (65%) of the details presented in free text but missing in the DICE-based registration was available in the DICE terminological system, but was not used by the clinicians.

The largest group of reasons for admission consisted of patients who were admitted to the ICU after cardiac surgery such as CABG and heart valve operations (n=112). In this patient group we found 95% correct concepts: 6(5%) exact matches, 100(90%) partial matches and 6(5%) mismatches. Among the partial matches the DICE-based registration of cardiosurgical reasons for admission contains more detail in 48% of the cases compared to the free-text registered ones. The main reason for missing detail in the remaining 52% cases is caused by the lack of a relation to describe the area of the heart to which the new graft is

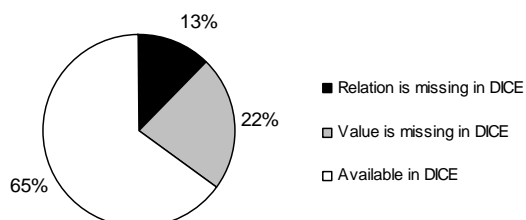


Figure 5: Reasons for missing detail in DICE-based registration of reasons for admission.

Table 2. Match scores for reasons for admission (RfA) on or not on the list of most frequently occurring reasons for admission.

	RfA on short list	RfA not on short list	All registered RfA
Mismatch	21 (7%)	17 (23%)	38 (11%)
Partial match	233 (82%)	51 (69%)	284 (79%)
Exact match	31 (11%)	6 (8%)	37 (10%)
Total	285 (100%)	74 (100%)	359 (100%)

located, e.g. “CABG, LIMA-LAD” can be coded in DICE as “CABG, Type: LIMA” but without “LAD”.

As described above the DICE user interface supports two ways to search for the appropriate diagnostic concept: using a short list or entering (a part of) a term. Table 2 shows the scores for reasons for admission split up for those that could be selected from the short list of frequently occurring reasons for admissions and those that were not on this list. Twenty percent (n=74) of all reasons for admission was not on the short list of frequently occurring reasons for admission.

Reasons for admission that could be selected from the short list were scored differently from those reasons for admission that were not represented on this list (Chi-Square  $p < 0.001$ ). Significantly more mismatches were scored among the reasons for admission that were not on the short list.

In 82% of the cases the two researchers directly agreed on the assigned scores, disagreement on the other 18% was easily resolved after short discussion.

#### 4. Discussion

Terminological systems offer the possibility to structure and standardize medical data, which improves the re-usability of these data for clinical research and quality assessment. In this study we compared the correctness and specificity between prospectively collected TS-based reasons for admission and free-text-based reasons for admission. We focused on the recorded data as such without taking into account the clinical consequences of the correctness and specificity of these data. We analyzed 359 reasons for admission to a Dutch Intensive Care registered in the PDMS by clinicians during actual care practice by using free text as well as by using the DICE terminological system. According to our definition 90% of the concepts were correctly registered based on the terminological system DICE. Only 11% of the cases had a perfect match. However, a partial match could be measured in 79% and there

were only 10% mismatches. One should be aware that if we change our definition of correctness to only “concepts with a perfect match” a completely different conclusion appears.

Among the partial matches about half of the TS-based reasons for admissions had additional detail compared to the free-text reason for admission. A possible explanation of this result could be the functionality of the terminology service in which users are encouraged to further specify a medical concept by additional qualifiers. Sixty-five percent of the information that is lacking in the other half of the partial matches was available in DICE but was not specified by the users. Further training and an improved user interface can contribute to improving these recorded reasons for admissions. Medical concepts on the short list of frequently occurring reasons for admission, counting for 80% of all reasons for admission, do have a better score than those not on this list. This is not a surprising result as the frequently occurring reasons for admission have got more attention during the modeling process of the terminological system than those not on the list. The reasons for missing concepts, attributes or relations gave us good insight into possibilities for (simple) improvements in DICE. For example the concept CABG could be extended with an attribute to describe which area of the heart is supported by the new graft. However, although we used free-text reasons for admission as they were recorded in daily care practice as a kind of golden standard, we observed many cases in which the TS-based registration included more detail than the free-text reasons for admission. Further research is necessary to determine the relevance of the details present in free-text as well as in the TS-based registration.

One weakness of our study is that the moment on which the free-text reason for admission is registered is not exactly the same as the moment on which the DICE based reason for admission has been registered. Although both reasons for admission were registered in the first 24 hours of admission, changing insight into the patient’s condition could be an explanation for the discrepancy (partial match or mismatch) between the free-text reasons for admission and the DICE-based reason for admission. We will investigate this in further research. Another weakness is the fact that TS-based registration and free-text registration have not necessarily been done by the same physician. However, when two different physicians recorded the reason for admission of a particular patient both physicians were directly involved in treating the

patient and hence both knew the patient’s condition very well. Finally, there are no clear registration rules regarding what constitutes a reason of admission of a patient. As mismatches seemed to be mainly caused by above mentioned limitations of the registration process rather than the terminological system, they have not been further investigated.

According to other studies in which the quality of structured and standardized registration of medical data was audited our study has a strong surplus value because this data comes from a real-practice situation and is not collected retrospectively in an experimental setting. Physicians in our observational study who recorded the reasons for admission treat the patients and were not informed that DICE-based reasons for admission would be compared to free-text reasons for admission. In studies such as [14-16] patient cases were selected, and structured, coded data were obtained by independent physicians or coders without a direct clinical relation with the patient.

The aim of our study corresponds most with [10] as both studies observe coding behavior of clinicians in actual practice. Although different methods are used (cognitive approach vs. document analysis) both studies compare TS-based registration with some kind of free text. We used written text while Cimino et al used video-taped spoken text. Cimino et al found a larger amount of exact matches than we did. Differences in definitions of match types partly explain this. Furthermore, the differences in results might be partly explained by the fact that in [10] TS-based registration took place at the same time as free-text registration and because of other methods used. Furthermore, in [10] not only diagnoses but also drug information is included. The main difference between the two studies, however, is that our study used a compositional TS instead of MED which only contains pre-coordinated concepts. The availability of post-coordination might have a large influence on the specificity of recorded diagnoses. Our study confirms the findings of Cimino et al. that correctness and specificity of TS-based registration depends on three factors: the terminological system’s content, its interface and the registration practice of the users.

## 5. Conclusions

This study shows that comparing free-text registration of reasons for admission with TS-based registration of reasons for admission only 11% of the concepts exactly matched and 79% of the concepts partially matched. TS-based registration added details in almost half of these partially matches and missed details in the other half. The methods used in this

study provide insight into possibilities for further improvement of the content coverage of DICE. However, 65% of the information not captured by the TS-based reasons for admission was available in DICE, indicating that user interaction with the system is more of an impediment than the contents of the TS. This study shows that availability of concepts and qualifiers in a TS does not guarantee that physicians will use them all. We expect that this result is generalizable to other terminological systems using post-coordination such as SNOMED CT. Further research is needed to investigate how physicians will be optimally supported in compositional TS-based registration.

## References

1. van Ginneken AM. The computerized patient record: Balancing efforts and benefit. *Int J Med Inf* 2002;65(2):97-119.
2. Wyatt JC. Clinical data systems, part II: components and techniques. *Lancet* 1994; 344: 1609-14.
3. Moorman PW, van Ginneken AM, van der Lei J, van Bommel JH. A model for structured data entry based on explicit descriptive knowledge. *Methods Inf Med* 1994; 33(5):454-63.
4. Rossi Mori A, Consorti F, Galeazzi E. Standards to support development of terminological systems for healthcare telematics. *Methods Inf Med* 1998; 37(4-5): 551-63.
5. de Keizer NF, Abu-Hanna A, Zwetsloot-Schonk JH. Understanding terminological systems I: Terminology and typology. *Methods Inf Med* 2000;39(1):16-21.
6. Prowsner SM, Wyatt JC, Wright P. Opportunities for and challenges of computerization. *Lancet* 1998; 352(9140): 1617-22.
7. Arts DG, Cornet R, de Jonge E, de Keizer NF. Methods for evaluations of medical terminological systems. A literature review and a case study. *Methods Inf Med* 2005;44(5):616-25.
8. Elkin PL, Brown SH, Bauer BA, Husser CS, Carruth W, Bergstrom LR, Wahner-Roedler DL. A controlled trial of automated classification of negation from clinical notes. *BMC Med Inform Decis Mak*. 2005;5;5(1):13
9. Lussier YA, Shagina L, Freidman C. Automating SNOMED coding using medical language understanding: a feasibility study. In proceedings AMIA Annual Symposium, Washington DC (2001):418-22.
10. Cimino JJ, Patel VL, Kushniruk AW. Studying the human-computer-terminology interface. *JAMIA* 2001;8(2):163-73.
11. de Keizer N.F, de Jonge E. National IC Evaluation (NICE) : a Dutch quality control system. *J ICU management* 2005; 3:62-64.
12. de Keizer NF, Abu-Hanna A, Cornet R, Zwetsloot-Schonk JHM, Stoutenbeek CP. Analysis and

design of an intensive care diagnostic classification. *Methods Inf Med* 1999; 38: 102-112.

13. Cornet R, Prins AK. An architecture for standardized terminology services by wrapping and integration of existing applications. In proceedings AMIA Annual Symposium, Washington DC (2003):180-4.

14. Los RK, Roukema J, van Ginneken AM et al. Are structured data structured identically. Investigating the uniformity of pediatric patient data recorded using OpenSDE. *Method Inf Med* 2005;44:631-638.

15. Brown PJ, Warmington V, Laurence M, Prevost AT. Randomised crossover trial comparing the performance of Clinical Terms Version 3 and Read Codes 5 byte set coding schemes in general practice. *BMJ*. 2003 May 24;326(7399):1127.

16. Campbell JR, Carpenter P, Sneiderman C, Cohn S, Chute CG, Warren J. Phase II Evaluation of Clinical Coding Schemes: Completeness, Taxonomy, Mapping, Definitions and Clarity. *Journal of the American Medical Informatics Association* 1997;4:238-251

## Acknowledgement

We would like to thank Antoon Prins for implementing the DICE application.

## Address for correspondence

N.F. de Keizer, Academic Medical Center, dept Medical Informatics, POBox 22700, 1100DE Amsterdam, The Netherlands. Email: n.f.keizer@amc.uva.nl



# Binding Ontologies & Coding systems to Electronic Health Records and Messages

AL Rector MD PhD<sup>1</sup>, R Qamar MSc<sup>1</sup> and T Marley MSc<sup>2</sup>

<sup>1</sup>School of Computer Science, University of Manchester, Manchester M13 9PL, UK

<sup>2</sup>Salford Health Informatics Research, University of Salford, Salford, UK

**ABSTRACT:** *A major use of medical ontologies is to support coding systems for use in electronic healthcare records and messages. A key task is to define which codes are to be used where – to bind the terminology to the model of the medical record or message. To achieve this formally, it is necessary to recognise that the model of codes and information models are at a meta-level with respect to the underlying ontology. A methodology for defining a Code Binding Interface in OWL is presented which illustrates this point. It generalises methodologies that have been used in a successful test of the binding of HL7 messages to SNOMED-CT codes.*

## Introduction

A major use of medical ontologies is to support medical terminologies and coding systems. A major use of medical terminology and coding systems is for electronic healthcare records and messages. Specifying the validation rules for how terminology and coding systems are to be used in electronic healthcare records and messages is, therefore, a key problem for medical ontologies.

We contend that electronic healthcare records messages are data structures and refer to their models as “information models”. By contrast we contend that the model of meaning or “ontology” is a model of our conceptualisation of the world – of patients, their illnesses. The function of the information models is to make it possible to specify and test the validity of data structures so that they can be exchanged and re-used in different information systems. The function of the model of meaning is accuracy in representing our understanding of the world so that we can reason about the world in general or individual patients and their diseases in particular. Validity neither requires or guarantees accuracy, nor *vice versa*.

We contend that codes are also data structures and the model of codes is also at the level of data structures. Ideally the “model of codes” or “coding system” should be a meta model of the model of meaning. Hence, in the ideal case, we take the individuals in the model of meaning to represent patients and their illnesses. We take the individuals in the model of codes to correspond to representations of classes of illnesses or “conditions”.

Pragmatically, it is useful to decouple the coding system from the model of meaning so that reasoning about the model of meaning and model of coding system is always separated.

## Using codes in messages and EHRs

Our goal is to assist software developers in specifying information systems and the use of codes from coding systems within them. We seek to have specifications that are sufficiently precise that separately implemented systems will work together. To achieve this we need to be able to validate that the models themselves are self-consistent and that individual messages conform to the models.

Typically, we want to start with a generic information model such as the HL7 RIM<sup>1</sup> or the OpenEHR reference model<sup>2</sup>. We then want to define progressively more specialised models in which each more specialised model is consistent with the next more generic model and ultimately the reference model. We want to use the models with separately developed coding systems – e.g. SNOMED, ICD, CPT, MEDRA, etc. Since we often want to use the same information model with more than one coding system, we want the “binding” between the information and coding system to be separate from both, analogous to an “Application Programming Interface” or “API” between software modules. We call this a “Code Binding Interface”

This problem is often expressed as defining “value sets” or “code sets” or just “subsets”. For example, we might wish to specify which codes can be entered in the family history section of the record or the list of valid codes for “position” for a blood pressure measurement. For a coding system such as SNOMED-CT or GALEN that allow formal definitions by means of expressions, this includes the constraints on such expressions. The Archetype Definition Language [1] used by the CEN standard EN13606 and OpenEHR specifies an “ontology section” similar in principle to what we here call a Code Binding Interface, but provides as yet only

---

<sup>1</sup> <http://www.hl7.org>

<sup>2</sup> <http://www.openehr.org>

OWL abstract syntax	Simplified Syntax	German Syntax	DL
someValuesFrom(C)	<b>SOME C</b>	$\exists.C$	
allValuesFrom(C)	<b>ONLY C</b>	$\forall.C$	
minCardinality(n C)	<b>MIN n C</b>	$\leq n.C$	
maxCardinality(n C)	<b>MAX n C</b>	$\geq n.C$	
cardinality(n C)	<b>EXACTLY n C</b>	derived	
value(c)	<b>VALUE c</b> or <b>IS c</b>	c	
intersectionOf(C D)	<b>C AND D</b> or <b>C &amp; D</b> or <b>C, D</b>	$C \sqcap D$	
unionOf(C D)	<b>C OR D</b> or <b>C   D</b>	$C \sqcup D$	
oneOf(...)	<b>{...}</b>	$\{...\}$	
equivalentClasses	<b>↔</b>	$C \triangle D$	
subclassOf	<b>→</b>	$C \sqsubseteq D$	
Type	<b>∈</b>	<b>∈</b>	
allDifferent	<b>DIFFERENT</b>		
allDisjoint	<b>DISJOINT</b>		

**Figure 1: Manchester simplified syntax for OWL** limited mechanisms for expressing semantic constraints.

### Basic requirements and tools

This work has been performed as part of a collaboration with practical users in the UK National Health Service. Our goal is to satisfy their requirements that:

1. There be a clear interface between the model of meaning and the information model, a “Code Binding Interface” (“CBI”);
2. The binding be expressive enough to capture a) enumerated lists of codes; b) all subcodes of a given code (with or without the root); c) all boolean combinations of a) and b).
3. That it deal with expressions in SNOMED-CT, whether pre- or post-coordinated.
4. The mutual constraints between the information and coding models be explicit and testable.
5. The constraints between information and coding models be part of a coherent methodology for expressing the constraints on the information model as a whole.
6. The models and interfaces be expressed in standard languages with well defined semantics and tools.

For the standard language we have chosen OWL-DL, the description logic variant of the new W3C standard Web Ontology Language. In practice we have used some features from the new OWL 1.1 specification<sup>3</sup> which have already been widely implemented by tool builders. We use OWL here primarily as a standard syntax and toolset for description logics, a subset of first order logic. The

use of OWL does not imply that the information models are ‘ontologies’ in any strong sense of that word.

### Vocabulary and Notation

For consistency with OWL’s usage, we use the term “class” for what some others would prefer to call “types” or “universals”. We refer to “individuals” where some might use the word “instances” and reserve the word “instance” for the relation between a class and an individual belonging to that class. We use the word “illness” to refer to an individual illness – e.g. “John Smith’s diabetes” and the term “condition” to refer to a class of illnesses – e.g. “Diabetes”. We use the term “property” to refer to relations between individuals. As a typographical convention, labels for classes begin with upper case; individuals and properties with lower case, and OWL keywords are in all upper case.

All work reported was performed using the Protégé-OWL tools<sup>4</sup>. Throughout we adopt the simplified Manchester syntax for OWL, a summary of which is presented in Figure 1.<sup>5</sup>

Field	Constraint
Topic	Exact code for diabetes mellitus
Diagnosis	The code for diabetes or any kind of diabetes
Brittleness	One of the subcodes of the code for “Diabetic Brittleness”

**Figure 2: Some of the fields and constraints for a example simplified information structure for Diabetes**

### Binding the models of Meaning, Codes, and Information: Principles

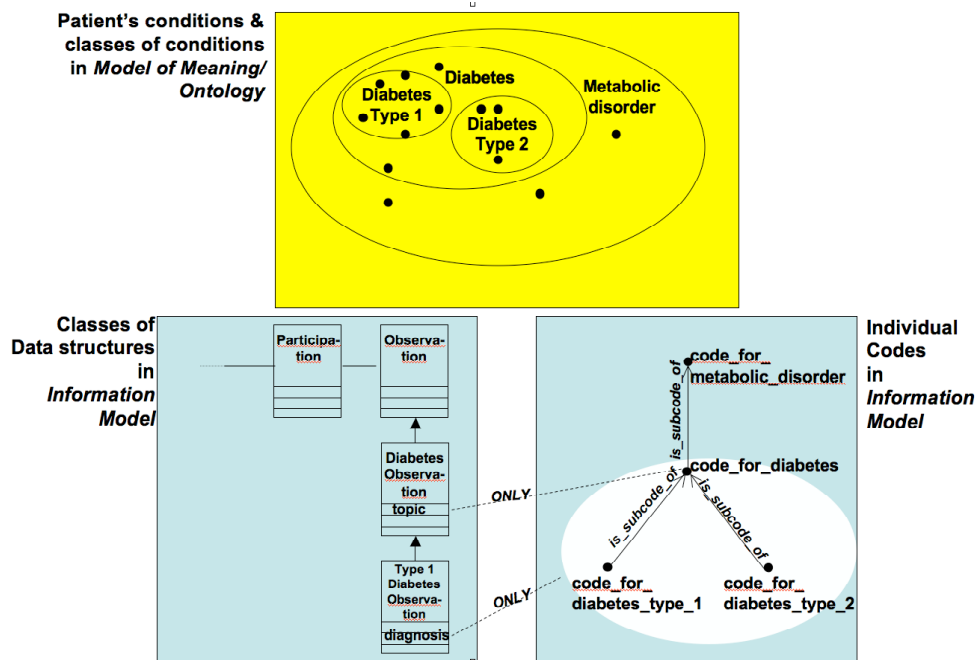
As a simplified example, we wish to specify the binding between a fragment EHR model conforming to the constraints expressed informally in Figure 2.

We show the relation of the models diagrammatically in Figure 3. The upper (yellow) square represents the model of meaning or the ontology. Dots represent individual illnesses such as “john\_smiths\_diabetes”.

<sup>3</sup> <http://www-db.research.bell-labs.com/user/pfps/owl/overview.html>

<sup>4</sup> <http://protege.stanford.edu>; <http://www.co-ode.org>

<sup>5</sup> Extensive experience in tutorials and presentations indicate that this notation is more easily understood by those less familiar with OWL as well as being more compact than either of the official OWL syntaxes. Note that the syntax includes OWL 1.1 constructs for qualified cardinality (p MIN|MAX|EXACTLY n C) and allDisjoint.



**Figure 3: Relation of Model of Meaning to classes of data structures and model of individual codes in the Information Model**

Ovals represent classes of illnesses or “conditions” such as “Diabetes\_type\_1”.

The lower two (blue) squares represent the information model on the left and the models of codes or “coding system” derived from the model of meaning on the right. The class hierarchy on the left represents classes of data structures expressed in

```

CLASS Diabetes →
  Metabolic_disorder,
  has_quality EXACTLY 1 Brittleness.
CLASS Diabetes_type_1 →
  Diabetes,
  is_caused_by SOME (Damage AND
    has_locus SOME Pancreatic_islet_cells).
CLASS Diabetes_type_2 →
  Diabetes,
  is_caused_by SOME
    (Resistance &
      has_locus SOME Insulin_metabolism) OR
    (Reduced_effectiveness &
      has_locus SOME Insulin).
CLASS Diabetic_brittleness ↔
  Brittleness,
  is_quality_of SOME Diabetes.
CLASS Diabetic_brittleness →
  has_state EXACTLY 1 Brittleness_state.
CLASS Diabetic_brittleness_state ↔
  Brittleness_state,
  is_value_of SOME Diabetic_brittleness.
CLASS Diabetic_brittleness_state →
  Brittle OR Well_controlled.

```

**Figure 4: Fragment of simplified condition model of meaning (‘ontology’) for Diabetes.**

UML diagrams. The hierarchy on the lower right represents hierarchies of codes linked by the `is_subcode_of` property, which is transitive. The oval superimposed on the hierarchy in this model represents a class of codes, in this case the class of “the code for diabetes and all its subcodes”.

In this example, the `is_subcode_of` property precisely mirrors the inferred subclass relation in the model of meaning and was derived from it by a systematic transformation. However, from the point of view of formal reasoning, each model is treated separately. The inference that, in the model of codes, `code_for_diabetes_type_1` is a member of the class `Diabetes_and_its_subcodes` is independent of the inference that, in the model of meaning, `Diabetes_type_1` is a subclass of `Diabetes`.

Why the apparent duplication? There are both theoretical and practical reasons.

- *Theoretically* – codes are not conditions and data structures are not patients. There are things that can be said of codes and data structures that are nonsense if said of conditions and patients, and *vice versa*. For example, both HL7 and OpenEHR have attributes in their data structures for “negation indicators”. Clearly, data structures have negation indicators; patients do not. It makes sense to talk about whether a patient has, or does not have, diabetes. It makes sense to talk about whether a

```

INDIVIDUAL code_for_diabetes ∈
  Code_entity,
  is_subcode_of VALUE code_for_metabolic_disorder.
INDIVIDUAL code_for_diabetes_type_1 ∈
  Code_entity,
  has_code VALUE code_for_diabetes.
INDIVIDUAL code_for_diabetes_type_2 ∈
  Code_entity,
  is_subcode_of VALUE code_for_diabetes.
INDIVIDUAL code_for_diabetic_brittleness ∈
  Code_entity,
  is_subcode_of VALUE code_for_qualifier.
INDIVIDUAL code_for_diabetic_brittle ∈
  Code_entity,
  is_subcode_of VALUE Code_for_diabetic_brittleness.
INDIVIDUAL code_for_diabetic_well_controlled ∈
  Code_entity,
  is_subcode_of VALUE Code_for_diabetic_brittleness.

```

**Figure 5a: The code individuals corresponding to Figure 4.**

```

CLASS Code_for_diabetes_and_subcodes ↔
  {code_for_diabetes} OR
  is_subcode_of VALUE code_for_diabetes.
CLASS Subcode_of_code_for_diabetic_brittleness ↔
  is_subcode_of VALUE code_for_diabetic_brittleness.

```

**Figure 5b: Classes of codes defined from code individuals. The first class corresponds to the shaded oval on the bottom right of Figure 2.**

data structure has its negation indicator set to true, false or null.

- *Pragmatically* – existing coding systems and information models contain many idiosyncrasies and errors. Many coding systems are based on no, or a flawed, model of meaning. Separating the information model and coding system from the model of meaning provides a degree of indirection that allows developers to compensate for these failings without compromising the underlying model of meaning.

## Representing the Binding in OWL

### The Model of Meaning – the “Ontology”

Figure 4 shows a fragment of a simplified ontology of conditions. The first line says that Diabetes is a kind of Metabolic disorder and that it has a quality of Brittleness. The “EXACTLY”<sup>6</sup> keyword indicate that each illness of class Diabetes has one, and only one, Brittleness quality. The definition is not closed, so there is nothing in this limited representation to say that Diabetes cannot have other qualities.

The next two clauses give simplified definitions of type 1 and type 2 diabetes.

The following clause defines Diabetic\_brittleness using the inverse of the quality relationship to say that any

```

CLASS Coded_Attribute →
  has_code MAX 1 Code.
CLASS Topic → Coded_Attribute.
CLASS Diagnosis → Coded_Attribute.
CLASS Brittleness → Coded_Attribute.
CLASS Condition_data_structure →
  has_attr EXACTLY 1 Topic,
  has_attr EXACTLY 1 Diagnosis.
CLASS Diabetes_data_structure →
  Condition_data_structure,
  has_attr EXACTLY 1 Brittleness.

```

**Figure 6a: Basic mapping of data structure model to OWL**

```

CLASS Placeholder_cls_diabetes_only_code → Code.
CLASS Placeholder_cls_diabetes_or_subcode → Code.
CLASS Placeholder_cls_for_diabetic_brittleness_subcode
  → Code.

```

**Figure 6b: Placeholder code classes for use in Code Binding Interface (CBI)**

```

CLASS Diabetes_data_structure →
  has_attr ONLY (Topic & has_code SOME
    Placeholder_cls_diabetes_only_code),
  has_attr ONLY (Diagnosis & has_code SOME
    Placeholder_cls_diabetes_or_subcode),
  has_attr ONLY (Brittleness & has_code ONLY
    Placeholder_cls_diabetic_brittleness_subcode).

```

**Figure 6c: Use of placeholder code classes and indication of whether codes are mandatory (*SOME*) or optional (*ONLY*).**

Brittleness that occurs in the context of being a quality of Diabetes is a Diabetic\_brittleness. The next clause states that each Diabetic\_brittleness quality has one, and only one Brittleness\_state. The final two clauses define Diabetic\_brittleness\_state as any Brittleness\_state in the context of Diabetic\_brittleness, and then state that it includes only the two values: Brittle and Well\_controlled.

### The model of codes – the coding system

Of the information in the ontology, only some is likely to be relevant to the coding system. For purposes of illustration we shall concentrate only on qualities and omit causation. The information as to which properties are of interest is ‘meta knowledge’ that must be held in a “profile” specifying the transformation of the of the ontology to the coding system.

From the ontology fragment in Figure 4, a mirroring profile might specify a definitions of individual codes as shown in Figure 5a. Based on these definitions of individual codes, we can define classes of codes as shown in Figure 5b. Since this model correctly mirrors a fragment of the ontology, the hierarchy the code classes will mirror the condition classes in the ontology. However, note that the additional constraints in the definitions are different in the ontology and coding system. For example, there is

<sup>6</sup> “EXACTLY” is an OWL 1.1 construct

```

CLASS Placeholder_cls_diabetes_only_code ←→
{code_for_diabetes}.
CLASS Placeholder_cls_diabetes_or_subcode ←→
{code_for_diabetes} OR
is_subcode_of VALUE code_for_diabetes.
CLASS Placeholder_cls_diabetic_brittleness_subcode ←→
is_subcode_of VALUE code_for_diabetic_brittleness.

```

**Figure 7: Code Binding Interface for Code System in Fig 5 and Information Model in Fig 6.**

no axiom in the coding system that all diabetic codes must have a brittleness qualifier, although there is an axiom in the ontology that all Diabetes have a quality Brittleness.

### The basic information model

A basic OWL model capturing the structure implied in Figure 3 is shown in Figure 6. We assume that we are modelling a class of diabetic data structures which have attributes for each item in Figure 2: topic, diagnosis, and brittleness.

The basic OWL mapping is then shown in Figure 6. We map each attribute by a class linked to the data structure by the property `has_attr`. We define a special subclass of attributes that take codes as their values, `Coded_Attribute`. Each `Coded_Attribute` is linked to a maximum of one Code as the value by the `has_code` property.

We assume that there is a generic class of `Condition_data_structures` that all have Topic and Dagnosis attributes, but that the Brittleness attribute is specific to the class of `Diabetes_data_structure`. Because the class `Diabetes_data_structure` is a subclass of `Condition_data_structure`, it “inherits” all of the attributes of its superclass.

Although a representation in which attributes are mapped to properties (as is done in the mapping specified by OMG) might seem simpler, mapping each attribute (and each association in the complete representation) to its own class makes it easier to specify cardinality and closure at the level of detail required for HL7 and OpenEHR models.

### Constraining the codes to placeholders

Given the basic information model defined in Figure 6a, we want to indicate that there are constraints on the codes to be used with each attribute. However, we do not wish to specify the coding system or the coding system specific constraints in the information model itself. Therefore, at this stage we state only that each attribute is constrained to a placeholder class of codes. These placeholder classes of codes are defined in Figure 6b.

Given the placeholder classes of codes, we can then use them in general constraints on the information model as shown in Figure 6c. In this example, we

```

CLASS Qualifier_name_code → Code.
INDIVIDUAL code_for_diabetic_brittleness_qualifier ∈
Qualifier_name_code.
CLASS Code_for_diabetes_and_subcodes →
has_qualifier ONLY
{code_for_diabetic_brittleness_qualifier}.
INDIVIDUAL code_for_diabetic_brittleness_qualifier ∈
has_code EXACTLY 1
Subcode_of_code_for_diabetic_brittleness.

```

**Figure 8a: Extension of Model of Codes to qualifiers**

```

CLASS Placeholder_diabetes_or_subcode_class ←→
({code_for_diabetes} OR
is_subcode_of VALUE code_for_diabetes),
NOT (has_qualifier VALUE
code_for_diabetic_brittleness_qualifier).

```

**Figure 8b: Extension of CBI in Figure 7 to exclude codes qualified by brittleness**

have stated the Topic and Diagnosis codes are mandatory, as indicated by the keyword “SOME”. However, by using the keyword `ONLY` for `Brittleness_code`, we have said that it is optional (because stating that a property can `ONLY` have particular codes does not imply that it need have any such codes).

### The Code Binding Interface

The model of the coding system in Figure 5 and the information model in Figure 6 might reside in separate modules. It now remains to define the Code Binding Interface between the two modules, which might likewise to reside in a third module.

The Code Binding Interface (CBI) consists of logical equivalences between the placeholder classes defined in Figure 6b and formal definitions of classes of codes in terms of the individuals in the model of codes in Figure 5. A CBI consistent with the constraints in Figure 3 is shown in Figure 7. The first line indicates that the placeholder class consists of just the codes enumerated between the curly brackets, in this case just the code for diabetes. The second line indicates that the given placeholder can be either the code for diabetes or any of its subcodes. (Remember that the property `is_subcode_of` is transitive.) The third line indicates that the code for brittleness can be any of the subcodes of the code for diabetic brittleness but not the parent code itself. They can be combined using the boolean operators `AND`, `OR`, and `NOT`. These were the three specific cases to be covered in Requirement 2.

### Extension to compositional coding systems

The previous example was limited to simple coding systems without ‘qualifiers’. However, the same principles can be extended to a coding system with qualifiers using suitably more complex constraints. In this case, since “brittleness” is to be explicitly

catered for in the information model, we want to avoid any possibility of a contradiction between the value in the information structure and the qualifier in the terminology. The simplest way to do this is to exclude the use of codes including the Brittleness qualifier from use with the Diagnosis attribute. The constraints depend only on whether the coding system model contains the necessary definitions. The methodology is the same whether it is for named, predefined (pre-coordinated) or (post-coordinated) code expressions (“code phrases” in HL7).

To represent compositional coding systems in OWL, we need to extend the definitions of the coding system to say that any code for diabetes or its subcodes may be linked to a qualifier view by the property `has_qualifier` by at most one brittleness qualifier code which, if present, must be linked to a subcode of `code_for_diabetic_brittleness`. To do this we need a new class of codes, the `Qualifier_name_code` with an instance `code_for_diabetic_brittleness_qualifier`.

Using this scheme we extend Figure 5 as shown in Figure 8a. This is an extension of the coding system model, not of the information system model (nor of the model of meaning).

Given the definitions in Figure 8a, we can extend the Code Binding Interface in Figure 7 by extending the definition of the placeholder for the for the `diabetes_or_subcode` to exclude codes qualified by brittleness as shown in Figure 8b.

A different group might develop a different information model that does not include brittleness as a separate attribute. It might, therefore, want to include brittleness with the diagnosis code. To do so, they need only change the Code Binding Interface.

#### **Absence of the Unique Name Assumption and differentiating axioms**

The above representations in OWL require a further addition. OWL does not make the “Unique name assumption”. In most formalisms, if two entities have different names they are different. In OWL, any two individuals might be the same unless declared different and any two classes might overlap unless declared disjoint.

Therefore, to represent the intentions fully, we need a set of “differentiating axioms” examples of which are shown in Figure 9abc. If these axioms are omitted, the validation in the next section will be incomplete because the reasoner will never infer that a code is incorrect because it cannot infer that it is different from the correct code, even though it has a different name.

#### **Validating information models**

OWL-DL was chosen because it allows efficient reasoners. In principle, the task of using OWL-DL to represent and validate a set of information models and bindings to a coding system simply requires that the reasoner be used to determine if the combined

```
DISJOINT Diabetes_type_1, Diabetes_type_2.
DISJOINT Brittle, Well_controlled.
```

**Figure 9a: Differentiating axioms for the model of meaning**

```
DIFFERENT
code_for_diabetes code_for_diabetes_type_1,
code_for_diabetes_type_2,
code_for_diabetic_brittleness,
code_for_diabetic_brittle,
code_for_diabetic_well_controlled).
```

**Figure 9b: Differentiating axioms for the model of codes – the coding system**

```
DISJOINT Data_structure, Attribute.
DISJOINT Topic, Diagnosis, Brittleness.
```

**Figure 9c: Differentiating axioms for the information model.**

models are consistent and the inferences as intended. Taking into account the previous discussion, the complete procedure consists of the following steps:

1. Transform the relevant parts of the model of meaning, *i.e.* the ontology, into a meta-level model of codes following the example in Figures 4 and 5.
2. Map the information model to an OWL model including the constraints on the terminology to be used as placeholders following the example in Figures 6.
3. Represent the bindings between the information model and the coding system model as a set of logical equivalences between the placeholders in the information model and class expressions in the coding system model to form the Code Binding Interface (CBI) module, following the example in Figure 7.
4. Import the three modules into a single OWL model.
5. Use the reasoner to classify the combined structure. Inconsistencies, inferred subclass relations, and inferred equivalencies will be flagged by the reasoner.
6. Examine the inferences and correct the errors.

Note that inferred subclass relations and equivalencies as well as inconsistencies may indicate errors. If an inferred subclass relation is not as intended, then either the superclass is under-constrained – *i.e.* too general – or the subclass is over-constrained – *i.e.* too specialised. If two classes that are intended to be different are inferred to be

logically equivalent, then the distinguishing features have been omitted or an axiom with unexpected consequences included. (There are a host of subtle errors that can occur in OWL models that are beyond the scope of this paper – see [2]).

### Validating individual data structures – the open and closed world assumptions

Before individual data structures can be validated, we must take into account a further feature of OWL’s semantics. Databases, logic programs, and most related systems are based on a “closed world assumption” with “negation as failure” – *i.e.* anything which cannot be found in the data base or proved true is treated as false. OWL is based on the “open world assumption” – *i.e.* things not proved true are treated as unknown; only things which can be proved false are treated as false. The open world assumption means that one can always add to an OWL model unless there is an explicit “closure axiom” to the contrary. Without the closure axiom, an OWL model or data structure means only “at least what is here”. By contrast, most message and EHR formalisms assume that the a given data structure contains “what is here and only what is here”. Without closure axioms OWL will accept a data structure with missing items because, since the representation is open, the missing item could always be added

Closure axioms are required in three places: a) in the information model to say that a particular class is complete, b) in the model of codes, to say that each code has only the subcodes explicitly asserted, and c) in each individual data structure to be validated, to say that it contains only what is explicitly present.

*Step a:* Before validating the model in Figure 6 we need to create a new subclass of “complete diabetes data structures” with the added the closure axiom. The new subclass definition is shown in Figure 10a. The second clause is the “closure axiom” that says that only these three attributes may occur.

```
CLASS Diabetes_data_structure_complete →
  Diabetes_data_structure,
  has_attr ONLY (Topic OR Diagnosis OR Brittleness).
```

**Figure 10a: “Complete” subclass of the Diabetes data structure class with closure axiom**

*Step b:* The model of codes must similarly be closed, downwards by adding closure axioms to state that each node *only* has the subcodes listed and the terminal codes have no (MAX 0) subcodes.

```
INDIVIDUAL code_for_metabolic_disorcer ∈
  has_subcode ONLY {...code_for_diabetes...}.
INDIVIDUAL code_for_diabetes ∈
  has_subcode ONLY {code_for_diabetes_type_1
                    code_for_diabetes_type_2}.
INDIVIDUAL code_for_diabetes_type_1 ∈
  has_subcode MAX 0.
```

```
INDIVIDUAL code_for_diabetes_type_2 ∈
  has_subcode MAX 0.
```

**Figure 10b: Closure axioms for code for diabetes**

*Step c:* An OWL mapping of a data structure that conforms to the model in Figures 6 is shown in Figure 10c. The final line is the closure axiom. (The use of SOME and ONLY rather than VALUE avoids the need to define individuals for each data structure’s Topic, Diagnosis and Brittleness attributes.)

```
INDIVIDUAL diabetic_data_structure_123 ∈
  has_attr SOME (Topic & has_code VALUE
                code_for_diabetes),
  has_attr SOME (Diagnosis & has_code VALUE
                code_for_diabetes_type_1),
  has_attr SOME (Brittleness & has_code VALUE
                code_for_diabetic_brittle),
  has_attr ONLY (Topic OR Diagnosis OR Brittleness).
```

**Figure 10c: The OWL mapping of a Diabetic data structure including closure axiom.**

Therefore, the steps to validate that a data structure conforms to the information model are:

1. Map the data structure to an OWL individual following the example in Figure 10c.
2. Add closure axiom as shown in Figure 10c.
3. Use the reasoner to check if the data structure is a valid instance of the intended class in the information model.

### Limitations of OWL

OWL-DL is based on a subset of first order logic deliberately limited so that inference is computationally tractable. There are two main limitations relevant to the work reported here:

- *Limited support for data types.* Both HL7 and Archetypes have very elaborate structures of datatypes that go beyond the usual XML datatypes supported by OWL. This can be overcome by encapsulating datatype in “holders”. What OWL provides is a check on the constraints on which data types should be used where. Separate datatype syntax checkers will be required to check the datatype formats themselves.
- *Lack of variables.* To preserve computational tractability, OWL lacks auxiliary variables and expressions such as “same-as”. For example, one can say that the left hand must be part of the left arm, but not that hands must be part of arms on the same side. Usually, it is possible to work around this limitation by having separate axioms for each case, *e.g.* for left-sided and right-sided rather than a single axiom for “same side”. UML, and most other object oriented formalisms, share this limitation. It has not proved a serious limitation in practice in the experience reported below or in related applications.

## Experience

### Representing HL7 message fragments developed by the NHS Connecting for Health

The methods in this paper are a refinement and generalisation of methods that were developed to represent the constraints in a set of message models developed by the UK NHS Connecting for Health Programme and their binding to SNOMED CT. The set of messages related to administration of medication were represented, a total of between twenty and thirty message formats (depending on how variants are counted). The methods were successful in representing all of the constraints identified, both in the HL7 models themselves and in the accompanying documentation, including the complex constraints on compositional forms required to maintain consistency between the SNOMED Context Model and the HL7 mood and status codes.

The representation, however, was tedious. Existing OWL tools are adapted to representing ontologies and models of meaning rather than data structures. Wider use of the methods presented here would benefit from the development of alternative tools, or at least alternative front-ends. In this respect OWL is best viewed as an assembly language. A high level language adapted to the task of representing information systems and their binding to coding systems is required along with ‘compilers’ to transform it to OWL in a standard way.

## Discussion

In previous papers [3-5] we have identified the interface between models of meaning – the ontology – and models of use as critical to clinical systems. This paper clarifies the relation between the model of meaning and one sort of model of use, the information model used for validating EHRs and messages. It contends that these information models are, in fact, models of data structures, and that they are formulated at a meta level with respect to the model of use, the ontology proper. It further contends that codes are likewise data structures and that the model of codes, or coding system, is likewise at a meta-level with respect to the model of meaning – the ontology.

The paper illustrates a methodology for formulating a “Code Binding Interface” (CBI) to specify and constrain how codes are to be used in data structures. This task is essentially “syntactic” – it is concerned with whether the data structures can be processed reliably rather than with whether the information conveyed is accurate or correct. The structure of the information model is motivated by adequacy to convey meanings, but the constraints in the model are on the data structures rather than on the meaning

itself. We suggest that the controversies around coding systems and standards such as HL7 arise, in part, from lack of clarity about this distinction between validity and accuracy.

The methodology has been used in practice and proved effective in supporting a range of independently formulated constraints.

This theoretical justification and practical experience is further supported by the observation that the requirements in the introduction cannot be met by a first order model of meaning directly linked to the information model. Requirement 2 includes being able to restrict the value of an attribute to a specific code at any level of abstraction – *e.g.* to “the specific code for diabetes” – or to any of the subcodes of a parent code but not the parent code itself – *e.g.* “to any subcode of brittleness”. However, the semantics of the model of meaning are defined in terms of classes of illnesses. The class “all diabetic illnesses that are neither type 1 nor type 2” would be all those diabetic illnesses of some alternative type – a class which is quite probably empty. It would *not* be the parent class, Diabetes, as required. By contrast, if dealing with classes of codes at the meta-level, the required expressions, as shown in Figure 7, are straightforward. Implicitly, this is what most users of terminologies such as SNOMED actually do – they query the coding system in a “distribution form” which does not give access to the underlying semantics. However, without explicit recognition of the separation of the models of meaning and meta-level models of coding systems, these mechanisms remain *ad hoc* and cannot be specified formally.

This paper deals with only the first two steps in using patient information – formulating meanings and storing or transmitting meanings in data structures. The third step – using the information for clinical decisions about individual patients – requires a further model – a model of clinical action – to be discussed in a further paper.

The methodology given here meets the requirements given in the introduction for binding ontologies, coding and information models. There is great controversy over the flaws in both SNOMED and HL7. The indirection in this methodology can help provide rigorous specifications that allows systems to interoperate using valid message despite flaws in such models. However, even if the models were ideal, the ontology sound, the coding system a faithful meta model of the ontology, and the information model founded on a sound model of the information to be conveyed, a Code Binding Interface would still need to be specified to specify what constituted valid bindings of codes to the data structures. Any given message or record fragment

will provide places for only a limited view on all possible meanings and hence all possible combinations of codes. Even in a near ideal world, if the information model and ontology are developed independently, there will still be overlaps and consequent need for mutual constraints between them.

Whether the methodology presented here is the best means to do so remains open to investigation. OWL has the technical advantage of being highly expressive, of supporting inverse properties which can be used to represent context, and of having available sound and complete reasoners. Its status as a standard brings the organisational advantage of a broad community developing tools and techniques. However, potential alternatives might include F-Logic [7], broader epistemic extensions to OWL and description logics [8] or other epistemic and or higher order logics. A principled layered version of OWL similar to that in this paper has also been suggested by others [6]. We hope that the issues are presented here in sufficient detail to allow alternatives to be formulated and compared.

### Acknowledgements

This work supported in part by the UK Department of Health "Connecting for Health" programme, the UK MRC CLEF project (G0100852), the JISC and UK EPSRC projects CO-ODE and HyOntUse (GR/S44686/1) and the EU Funded Semantic Mining Network of Excellence. The

HL7 Terminfo working group stimulated and contributed to many of the ideas presented here.

### References

1. Beale T. Archetypes: Constraint-based domain models for future-proof information systems. In: *OOPSLA-2002 Workshop on behavioural semantics*; 2002;
2. Rector A, Drummond N, Horridge M, Rogers J, Knublauch H, Stevens R, et al. OWL Pizzas: Common errors & common patterns from practical experience of teaching OWL-DL. In: *EKAU-2004*; October, 2004; Northampton, England: Springer; p. 63-81.
3. Rector AL. The Interface between Information, Terminology, and Inference Models. In: Patel V, (ed) *Proc Medinfo-2001*; London, England; p. 246-250.
4. Rector AL, Johnson PD, Tu S, Wroe C, Rogers J. Interface of inference models with concept and medical record models. In: Quaglini S, Barahona P, Andreassen S, editors. *Artificial Intelligence in Medicine Europe (AIME)*; 2001; Cascais, Portugal: Springer. p. 314-323.
5. Rector A, Taweel A, Rogers J. Models and inference methods for clinical systems: A principled approach. In: *Proc Medinfo-2004*; San Francisco: North Holland; p. 79-83
6. Pan JZ, Horrocks I, Schreiber G. OWL FA: A metamodeling extensions of OWL DL. In: *Proc OWL-ED 2005*
7. Kifer, M, Lausen G. F-logic: a higher-order language for reasoning about objects, inheritance, and schemas. Portland, Oregon, United States: ACM Press; 1989.
8. Donini F, M L, Nardi D, Shaerf A, Nutt W. An epistemic operator for description logics. *Artificial Intelligence* 1998;100(1-2):225-274.



## Using ontology visualization to understand annotations and reason about them

Mary E. Dolan, Ph.D., and Judith A. Blake, Ph.D.,  
Mouse Genome Informatics [MGI], The Jackson Laboratory,  
Bar Harbor, ME 04609 USA  
mdolan@informatics.jax.org

*Biomedical ontologies not only capture a wealth of biological knowledge but also provide a representational system to support the integration and retrieval of biological information. Various biomedical ontologies are used by model organism databases to annotate biological entities to the literature and have become an essential part of high throughput experiments and bioinformatics research. We are exploring the power of ontology visualization to enhance the understanding of annotations by placing annotations in the graph context of the broader biological knowledge the ontology provides. Presenting annotations in this context provides a better understanding of the annotations because humans are adept at extracting patterns and information from graphical representations of complex data.*

### INTRODUCTION

Biological systems can be very complex but many aspects of biological system characterization have a wealth of biomedical knowledge accumulated over years of clinical and laboratory experience. Ontologies provide a shared understanding of a domain that is human intelligible and computer readable and, consequently, a representational system to support the integration and retrieval of this knowledge.

As techniques of large-scale genomic analysis and functional gene annotation have progressed and are becoming more common, it is essential to find approaches to provide a comprehensive view of annotation sets. We are exploring the power of several widely used ontologies to provide a comprehensive graphical view of annotations by presenting the annotations visualized within an ontology relationship structure. By presenting annotations in the graph context we hope to provide a better understanding of the annotations because humans are adept at extracting patterns and information from graphical representations of complex data.

### BACKGROUND

Ontologies can be used to abstract knowledge of a domain in a way that can be used by both by humans and computers by providing an explicit representation of the entities of interest and the relationships among them. In particular, biomedical ontologies representing various aspects of biology are being used for annotating entities to the literature and for integrating the diverse information resulting from the analysis of high-throughput experiments.

Open Biomedical Ontologies (OBO) is an umbrella repository for well-structured controlled vocabularies for shared use across different biological and medical domains [1]. The OBO website contains a range of ontologies that are designed for biomedical domains. Some of the OBO ontologies, such as the Gene Ontology (GO), apply across all organisms. Others are more restricted in scope; for example, the Mammalian Phenotype Ontology (MP) is a phenotype ontology designed for specific taxonomic groups.

The GO Project was established to provide structured, controlled, organism-independent vocabularies to describe gene functions [2] and, as a consequence, provides semantic standards for annotation of molecular attributes in different databases. Members of the GO Consortium supply annotations of gene products using this vocabulary. The GO and annotations made to GO provide consistent descriptions of gene products and a valuable resource for comparative functional analysis research.

Currently, the three ontologies of GO contain nearly 20,000 terms [3]. The terms are organized in structures called directed acyclic graphs (DAGs) which differ from strict hierarchies in that a more specialized (granular) child term can have more than one less specialized parent term. In the GO a child can be related to a parent by either a 'part of' or 'is a' relationship. Mouse Genome Informatics (MGI) curators use the GO to annotate mouse genes from the literature. Currently, MGI has more than 100,000 annotations to more than 17,000 genes; approximately half of the annotations are manual

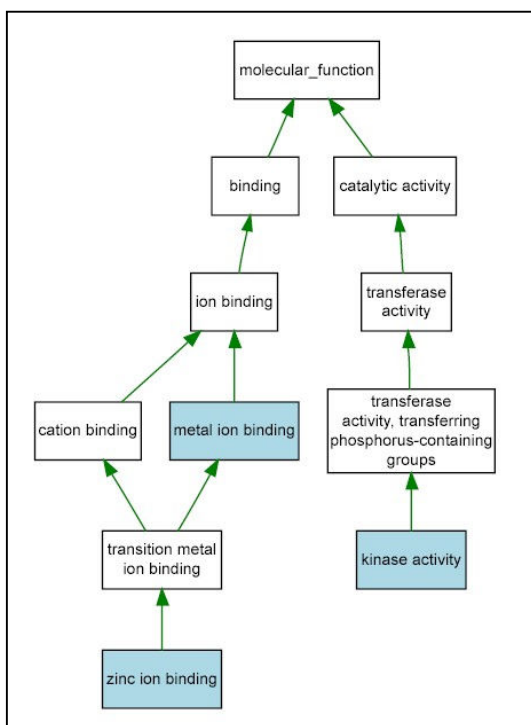


Figure 1. GO annotation graph for mouse Hgs (HGF-regulated tyrosine kinase substrate) provides an alternative to tabular or text views. Blue/shaded nodes in the GO graph indicate mouse annotations.

Full graph and annotation set available at:  
<http://www.informatics.jax.org/javawi2/servlet/WIFetch?page=GOMarkerGraph&id=MGI:104681>

annotations from the literature, the balance from automated data loads. An MGI user has the option of viewing the full set of GO annotations for a particular gene in three formats: as a table, as automatically generated text, and as a graph. The graph presents relevant parts of the GO with direct annotations indicated as colored nodes, as shown in figure 1. The graphical format allows a user to easily see, for example, whether a gene product appears to participate in a broad range of molecular functions or in only a narrow, specialized function.

Genes that share close evolutionary relationships are likely to function in similar ways. As a complement to our previous work [4] on the assessment of annotation consistency of independently developed annotation sets for curated mammalian orthologs [5], we provided comparative graphical visualizations of annotations, one graph for each mouse-human-rat ortholog triple with nodes colored according to organism annotated. Coloring nodes to distinguish among annotations extends the usefulness of the visualization for pattern recognition by users. The

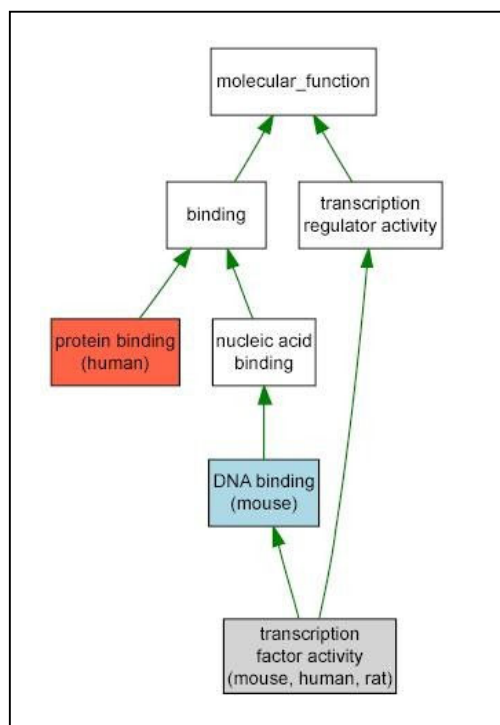


Figure 2. GO comparative graph for MGI curated orthologs to mouse Pax6 (paired box gene 6). The nodes are color-coded according to organism: mouse annotations shown in blue/lighter shading, human annotations in red/darker shading, multiple organisms in gray. Full graph available at:  
<http://www.informatics.jax.org/javawi2/servlet/WIFetch?page=GOOrthologyGraph&id=MGI:97490>

graphical format, as shown in figure 2, allows a user to assess the consistency, inconsistency and level of detail of annotations made to different model organisms.

Our examination of the comparative graphs led to the observation that annotations are often complementary, reflecting the fact that the different model organisms are used to study different aspects of biology. Since biologists are often species-blind and assemble their initial picture of a gene and its function without regard to the taxonomic origin of the gene that was studied in a particular experiment, this suggested the broader application of such graphs as ‘summary’ rather than ‘comparative’ graphs that might be used to answer the request: “Show me everything that is known about this gene.” The power of this representation is that it provides a view of the summary of information derived from species-specific experimental results.

In addition to the ability to visualize comparative annotation sets, graphs can be used to coordinate information for animal models of human

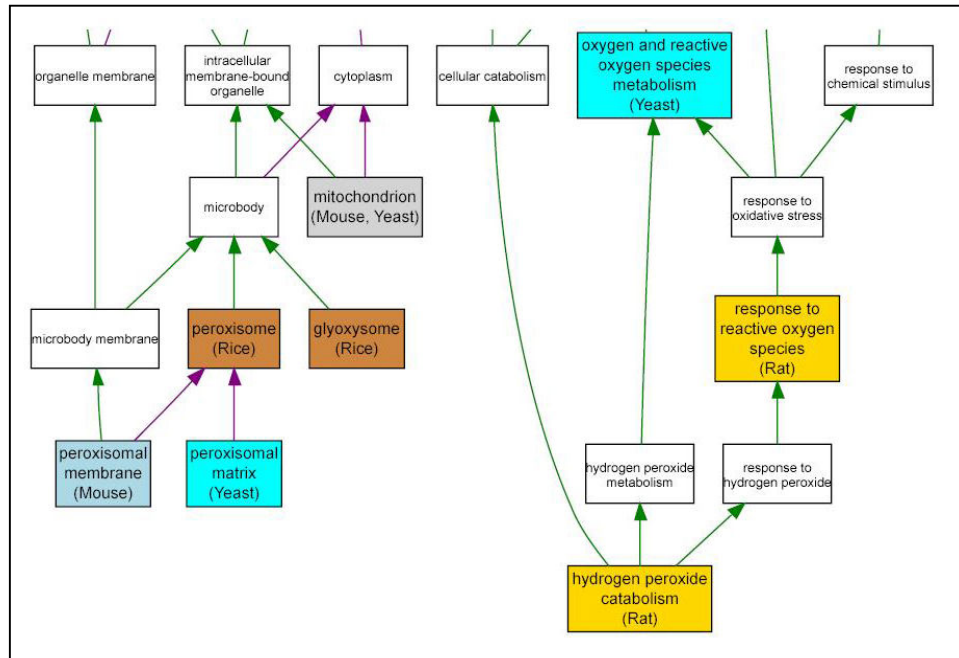


Figure 3. GO annotation graph for OMIM gene CATALASE; CAT. The graph coordinates GO annotations for thirteen model organisms with nodes colored by organism. Full graph and annotation set available at: [http://www.spatial.maine.edu/~mdolan/OrthoDisease\\_Graphs/OMIM\\_GeneGraphs/CAT.html](http://www.spatial.maine.edu/~mdolan/OrthoDisease_Graphs/OMIM_GeneGraphs/CAT.html)

diseases. The primary purpose of performing experiments that study the consequences of mutations in a particular organism is that these experiments provide valuable models for the understanding of human disease. We have extended our ontology visualization approach [6] to the orthology sets developed in the resource OrthoDisease [7], a comprehensive database of model organism genes that are orthologous to human disease genes derived from the OMIM database [8], a continuously updated catalog of human genes and inherited, or heritable, genetic diseases. We have abstracted orthology information on thirteen organisms for which curated GO annotation sets are publicly available. By combining all GO annotations for the orthologs associated with each disease gene or with each disease, we obtain a comprehensive annotation set for each disease gene and for each disease. Each annotation set is presented on the GO graph with nodes having annotation colored according to the organism that is the source of the annotation. Figure 3 shows part of the graph for OMIM gene CAT that demonstrates the degree of similarity annotations to diverse organisms can show. Of course, in some sense, it is the differences that are of more interest in this case since we are interested in collecting together as much information as possible.

## DESCRIPTION OF CURRENT WORK

While each annotation group develops curation standards to meet the needs of their community, one of the important results of various ontology projects has been an attempt to develop a common vocabulary and shared annotation standards that enhance the utility of these annotations for analysis. We have found that regardless of the ontology, presenting terms in a graphical context makes the relationships of ontology terms clear, provides context for annotations, and makes the examination of large annotation sets feasible. The long-term objective, now, is to build consensus for curation standards that will strengthen the utility of data integration capabilities of this approach.

We have generalized our GO visualization approach to other ontologies and annotation data sets. First, we construct a complete graph to represent the ontology. Second, we color nodes that have annotations and limit the graph to the sections necessary to show all annotations. By limiting the graphs to annotated sections we do not have to deal with scalability issues that might arise if we were to attempt to represent an entire ontology that includes thousands of terms. Finally, we build a web page for each gene that includes an image of the graph and a table of annotations. In addition, to facilitate the examination

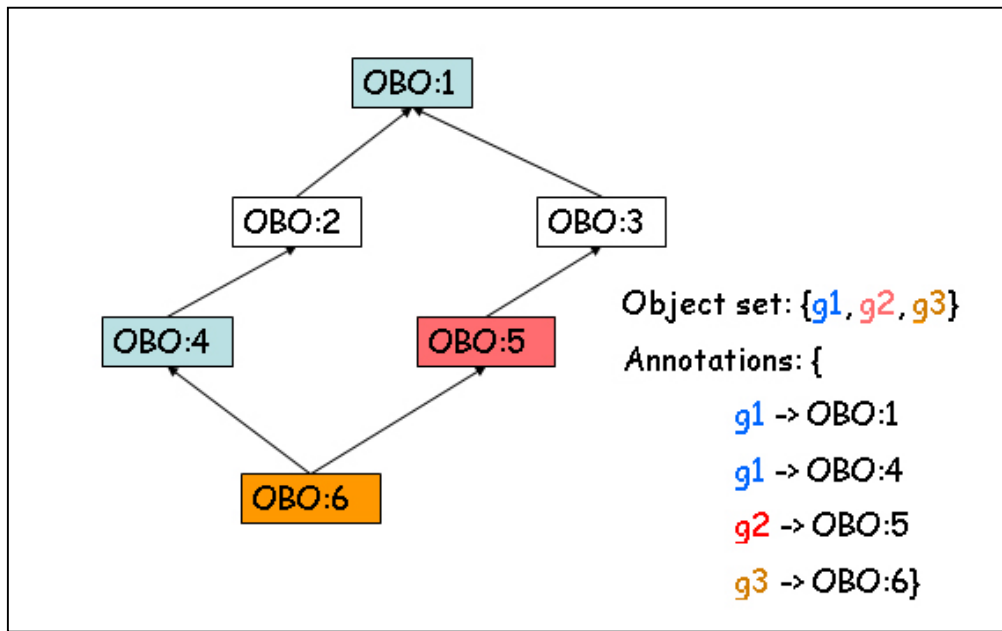


Figure 4. The comparative graph paradigm: an ontology provides the relationship structure among terms; a grouping idea defines the object set; and discriminating idea distinguishes objects whose annotations will be color-coded in the graph.

of larger graphs, we provide scalable vector graphics (SVG) images, which include pan-zoom-search functionality that allow a user to examine specific sections of the graphs. The graph images are generated using GraphViz, a freely available, open source graph layout program [9].

Gene expression data sets describe when and where particular genes are active. Providing a comprehensive picture of the level of gene expression across developmental stages and anatomical structures will facilitate investigation of regulation of gene expression.

We have applied our simple graphical display approach to gene expression data with annotations to both the Adult Mouse Anatomical Dictionary (MA) [10] and the Edinburgh atlas of mouse embryonic development (EMAP) [11]. For each gene with annotation data, the resulting graph shows the mouse anatomy ontology with anatomical structure nodes colored to indicate where that gene is expressed. In addition, in the case of the EMAP graphs, we have attempted to tease apart time dependence of gene expression patterns by separating annotations to different developmental stages by producing graphs for each Theiler stage.

The laboratory mouse is an important model organism for a broad range of human diseases and disorders, including diabetes, heart disease, and cancer. Genomic and genetic investigations of

particular mouse models (phenotypes) reveal the contribution of particular genomic variants (alleles) to the presentation of disease phenotypes. The annotation of genotype-phenotype associations is an essential part of assessing mouse models for human disease.

We have adapted our comparative GO annotation approach to phenotype annotations made to different mouse gene alleles to create Mammalian Phenotype (MP) Ontology [12] graphs. As in the case of GO comparative graphs (figure 2), the generalized approach to comparative graphs requires three things: an ontology to provide the relationship structure, a grouping idea to connect the annotated objects, and a distinguishing idea (see figure 4). First, we construct a complete graph to represent the ontology. Second, we color nodes that have annotations according to the distinguishing characteristic and limit the graph to the sections necessary to show all annotations. Finally, we build a web page for each gene that includes an image of the graph and a table of annotations.

In the case of the GO comparative graphs the grouping idea is orthology and the distinguishing idea is organism: mouse annotations in blue, human annotations in red and so forth. In the case of MP graphs the grouping idea is the gene and the distinguishing idea is the allele: each allele's annotated nodes are colored differently. In a similar way to color coding of GO nodes by organism, color-coding of MP nodes by allele allows a user to easily

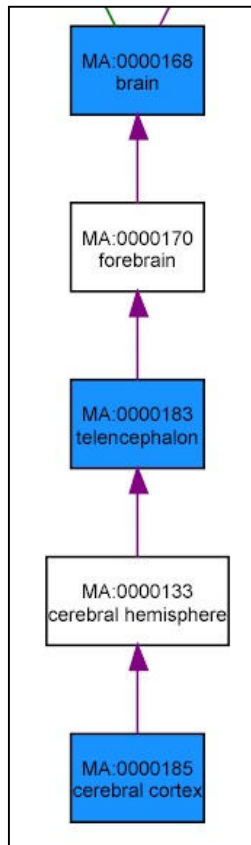


Figure 5. Part of the Adult Mouse Anatomical Dictionary (MA) annotation graph for postnatal expression data for mouse gene *Abcg2* (ATP-binding cassette, sub-family G (WHITE), member 2).

Full graph and annotation set available at:  
[http://www.spatial.maine.edu/~mdolan/GXD\\_Graphs/Abcg2.html](http://www.spatial.maine.edu/~mdolan/GXD_Graphs/Abcg2.html)

see similarities and differences in alleles annotated to different phenotypes. Our purpose in creating such graphs is to move beyond simply providing another representation of a phenotype data set to add potential value to this data set as a method of assessing mouse models for human disease.

## RESULTS

### Graphical representations of expression data sets using anatomy ontologies

The Mouse Anatomical Dictionary provides ontologies that provide a standardized nomenclature for anatomical parts to describe the complex patterns of gene expression in the developing and adult mouse and how they relate to the emerging tissue structure. Terms that describe embryonic developmental stages (Theiler Stages 1 through 26) have been developed by the Edinburgh Mouse Atlas Project (EMAP) [11]. Terms that describe mice at postnatal stages, including adult (Theiler stage 28) have been developed as the Adult Mouse Anatomical Dictionary (MA) [10].

### Adult Mouse Anatomical Dictionary graphs display relationships of annotations

The Adult Mouse Anatomical Dictionary (MA) is an anatomy ontology that can be used to provide standardized nomenclature for anatomical terms in the postnatal mouse. It was developed as part of the Gene Expression Database (GXD) resource of information from the mouse [12]. The Adult Mouse Anatomical Dictionary organizes anatomical structures for the postnatal mouse spatially and functionally. Each MGI gene detail page includes links to gene expression data; the user can select data for the postnatal mouse and obtain a tabular view of available expression data.

Our graphical representations present another view of the data, as shown in figure 5. This partial view of the graph for *Abcg2* (ATP-binding cassette, sub-family G (WHITE), member 2) clearly shows the relationship of three annotations as variations in granularity. Note that the colored nodes indicate only direct annotations made by curators from the literature, although indirect annotation can be inferred from the ontology structure.

### EMAP graphs provide information on developmental stage specific expression

The Edinburgh Mouse Atlas Project (EMAP) annotation of gene expression data can be used to capture the complex and ever-changing patterns throughout the development of the mammalian embryo and how they relate to the emerging tissue structure at each developmental stage.

We have adapted the EMAP ontology to separate annotations associated with different Theiler stages and created EMAP annotation graphs for each stage, effectively treating each stage as a separate ontology structure. With this approach we can, within the limits of incomplete annotation, see stage separated annotations as a time series of expression patterns. For example, figure 6 shows expression annotations for mouse gene *Shh* (Sonic hedgehog) for Theiler stages 11 (figure 6, upper panel) and 12 (figure 6, lower panel). A user might consult such graphs to explore changes in expression pattern between stages or determine the earliest stage at which the gene is known to be expressed in a particular anatomical structure. The way these graphs are presented at our web site, a user can move forward or back to adjacent Theiler stage.

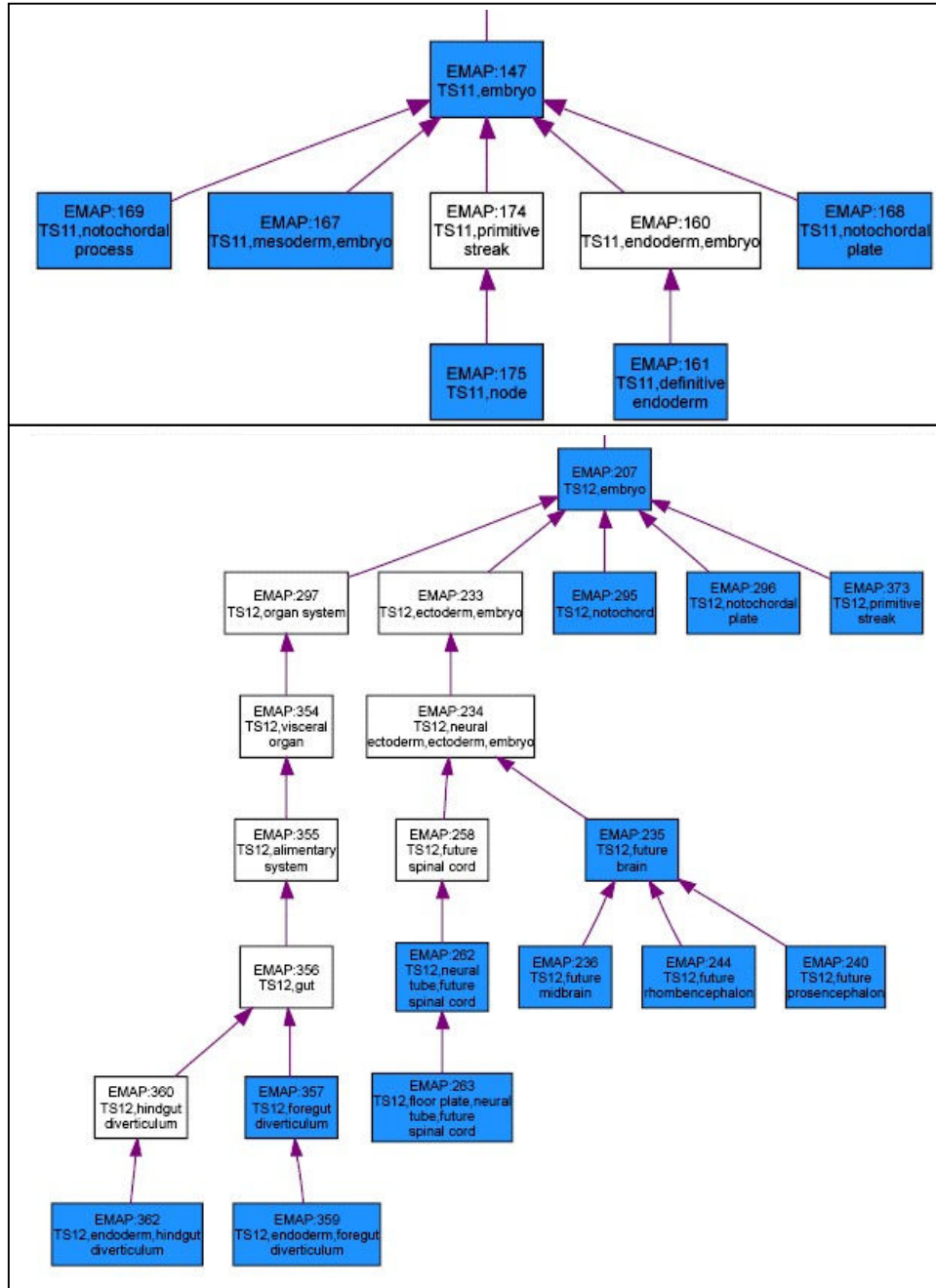


Figure 6. EMAP ontology graphs for Theiler stages 11 (upper) and Theiler stage 12 (lower) displaying expression patterns for mouse *Shh* (Sonic hedgehog). (Annotations available from GXD.) Full graph and annotation set available at: [http://www.spatial.maine.edu/~mdolan/GXD\\_Graphs/TimeSlices/TS11.html](http://www.spatial.maine.edu/~mdolan/GXD_Graphs/TimeSlices/TS11.html)

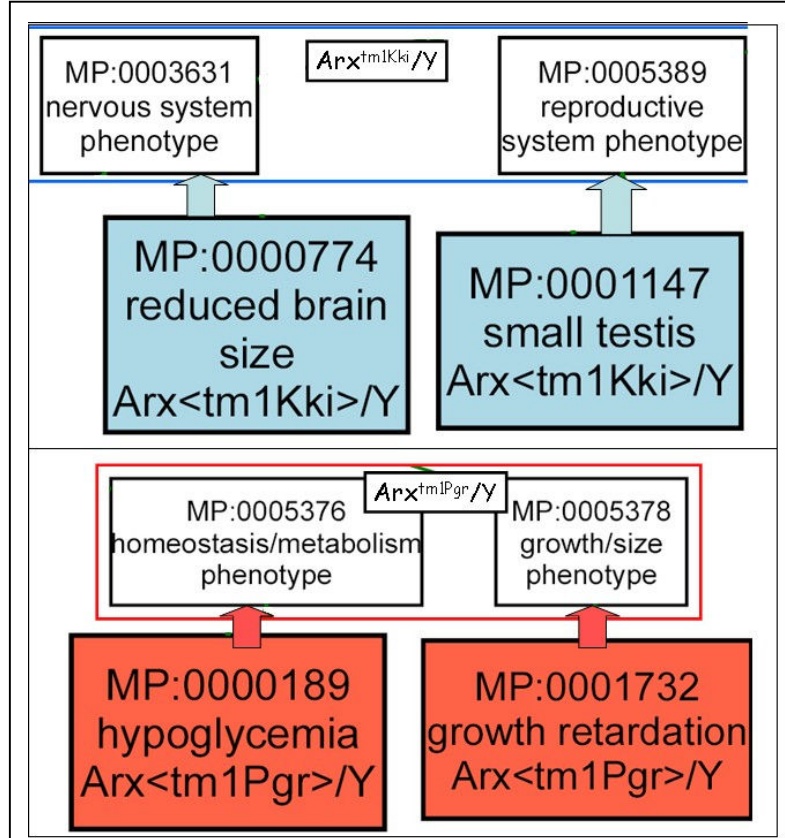


Figure 7. Detail of the Mammalian Phenotype (MP) Ontology annotation graph for two alleles of mouse gene *Arx* representing allelic compositions  $Arx^{tm1Kki}/Y$  (blue/lighter shading) and  $Arx^{tm1Pgr}/Y$  (red/darker shading). We observe that the allele annotations segregate in separate ontology branches. Only the allelic composition  $Arx^{tm1Kki}/Y$  high-level phenotypes correspond to nervous system and reproductive system phenotypes, while only the allelic composition  $Arx^{tm1Pgr}/Y$  corresponds to homeostasis/metabolism and growth/size phenotype. Full graph and annotation set available at: [http://www.spatial.maine.edu/~mdolan/GenoPheno\\_Graphs/Arx.html](http://www.spatial.maine.edu/~mdolan/GenoPheno_Graphs/Arx.html)

#### Using graphical representations to reason about annotations: assess mouse models for human disease

The Mammalian Phenotype (MP) Ontology [13] is used by MGI to represent phenotypic data. The MP Ontology enables annotation of mammalian phenotypes in the context of mutations and strains that are used as models of human disease and supports different levels of phenotypic knowledge. For example, among the highest levels of the MP Ontology are terms for: growth/size phenotype, homeostasis/metabolism phenotype, nervous system phenotype, and reproductive system phenotype.

So for example, the mouse gene *Arx* (aristaless related homeobox gene (*Drosophila*)) has 2 alleles,  $Arx^{tm1Kki}$  and  $Arx^{tm1Pgr}$ , both of which have been annotated to MP by curators at MGI. We might ask:

how do the annotations to the different alleles compare? Applying the comparative graph methodology and indicating MP annotations to terms by color-coding according to allelic composition  $Arx^{tm1Kki}/Y$  and  $Arx^{tm1Pgr}/Y$  results in the graph detail shown in figure 7. (Information on mouse strain background is not indicated in the graph but is given in a complete annotation table that accompanies the graph.) We observe that in the graph the allele annotations segregate in separate branches reflecting the fact that the phenotype annotations associated with the two alleles fall into distinct high-level phenotypes. Only the allelic composition  $Arx^{tm1Kki}/Y$  corresponds to high-level nervous system and reproductive system phenotypes, while only the allelic composition  $Arx^{tm1Pgr}/Y$  corresponds to homeostasis/metabolism and growth/size phenotypes.

?

Human Disease and Mouse Model Detail

Human Disease

Term: Lissencephaly, X-Linked, with Ambiguous Genitalia; XLAG  
OMIM ID: [300215](#)

Associated Genes

Orthologous mouse and human markers where mutations in one or both species have been associated with phenotypes characteristic of this disease.


Mouse Gene

[Arx](#)

Human Gene

[ARX](#)

Characteristics of this human disease are associated with mutations in...



...both mouse and human orthologous genes.

Mouse Models

Genotype		Ref(s)
Allelic Composition	Genetic Background	
Models with phenotypic similarity to human disease where etiologies involve orthologs. <sup>1</sup>		
<a href="#">Arx<sup>tm1Kki</sup>/Y</a>	involves: 129P2/OlaHsd * C57BL	<a href="#">1:79871</a>

<sup>1</sup>Human genes are associated with this disease. Orthologs of those genes appear in the mouse genotype(s).

Figure 8. MGI integrates data on mouse models of human disease from OMIM with existing data for mouse genes and strains. For example, as shown on this “Associated Human Diseases” information page for *Arx*, *Arx<sup>tm1Kki</sup>/Y* on the strain background 129P2/OlaHsd \* C57BL is a known mouse model for OMIM human disease, “Lissencephaly, X-Linked, with Ambiguous Genitalia; XLAG” characterized by nervous system and reproductive system phenotypes.

The visualization methodology as shown in figure 7 is consistent with the known association of this particular human disease and the *Arx<sup>tm1Kki</sup>* mouse model. (This page is available at:

<http://www.informatics.jax.org/javawi2/servlet/WIFetch?page=humanDisease&key=850912> )

This distinction is confirmed by seeing that, indeed, *Arx<sup>tm1Kki</sup>* is a known mouse model for OMIM human disease, “Lissencephaly, X-Linked, with Ambiguous Genitalia; XLAG” (see figure 8), which is characterized by nervous system and reproductive system phenotypes. The visualization methodology outlined here is consistent with the known association of this particular human disease and the *Arx<sup>tm1Kki</sup>* mouse model. Our hope is that examination of the MP graphs for specific disease associated phenotypes would help point to good mouse models. To facilitate this, we have created an index to all genes and alleles indicating high-level phenotypes. For example, a user can search the index for all genes and alleles annotated for “nervous system phenotype” and examine the linked MP graphs for segregation of allele phenotypes and a potential novel mouse model for a human disease characterized by nervous system abnormality. In this way we have extended the usefulness of the graphical representations beyond just another way of presenting the data to a method that allows a user to reason about annotations.

### Availability of graphs

All graphs presented in this work are publicly available.

- The GO graphs are available for each gene from the gene detail pages at MGI.
- The OrthoDisease graphs are available at: [http://www.spatial.maine.edu/~mdolan/OrthoDisease\\_Graphs/](http://www.spatial.maine.edu/~mdolan/OrthoDisease_Graphs/)
- The Adult Mouse Anatomical Dictionary (MA) graphs for GXD data for selected genes are available [http://www.spatial.maine.edu/~mdolan/GXD\\_Graphs/](http://www.spatial.maine.edu/~mdolan/GXD_Graphs/)
- The Theiler stage separated Edinburgh Mouse Atlas Project (EMAP) graphs displaying GXD data for *Shh* are available at: [http://www.spatial.maine.edu/~mdolan/GXD\\_Graphs/TimeSlices](http://www.spatial.maine.edu/~mdolan/GXD_Graphs/TimeSlices)
- The Mammalian Phenotype (MP) graphs for all MGI genes with phenotype annotations are available at: [http://www.spatial.maine.edu/~mdolan/GenoPheno\\_Graphs/](http://www.spatial.maine.edu/~mdolan/GenoPheno_Graphs/)

## CONCLUSIONS

Biological systems can be very complex but many aspects of biological system characterization have a wealth of biomedical knowledge accumulated over years of clinical and laboratory experience. Ontologies provide a shared understanding of a domain that is human intelligible and computer readable that can help support the integration and retrieval of this knowledge.

Here we provide a methodology to visualize sets of annotations as provided by a model organism database curation system to aid researchers in better comprehending and navigating the data. The result is a comprehensive view of available knowledge. As more annotations are made and become available, such tools will be both more necessary, to handle larger data sets, and more useful, as annotation approaches completeness. We believe that this approach to coordinating biological knowledge available in model organism resources will provide a valuable resource in medical research and contribute to understanding these systems.

## Acknowledgements

This work is funded by NIH/NHGRI (HG-002273).

## References

1. Open Biomedical Ontologies (OBO) [<http://obo.sourceforge.net/>]
2. The Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. *Nature Genetics* 2000, 5: 25-29.
3. The Gene Ontology Consortium. The Gene Ontology (GO) project in 2006. *Nucleic Acids Res* 2006, 34: D322-D326
4. Dolan ME, Ni L, Camon E, and Blake JA. A procedure for assessing GO annotation consistency. *Bioinformatics* 2005, 21(Suppl 1):i136-i143.
5. Eppig JT, Bult CJ, Kadin JA, Richardson JE, Blake JA, et al. The Mouse Genome Database (MGD): from genes to mice -- a community resource for mouse biology. *Nucleic Acids Research* 2005, 33: D471-5.
6. Dolan ME and Blake JA. Using Ontology Visualization to Coordinate Cross-species Functional Annotation for Human Disease Genes. *Proceedings Nineteenth IEEE International Symposium on Computer-based Medical Systems: Ontologies for Biomedical Systems* 2006, 583-587.
7. O'Brien KP, Westerlund I, Sonnhammer EL. OrthoDisease: a database of human disease orthologs. *Human mutation* 2004, 24(2):112-9.
8. Hamosh A, Scott AF, Amberger JS, Bocchini CA and McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*. 2005, 33: D514-D517.
9. GraphViz [<http://www.graphviz.org/>]
10. Hayamizu TF, Mangan M, Corradi JP, Kadin JA and Ringwald M. The Adult Mouse Anatomical Dictionary: a tool for annotating and integrating data. *Genome Biology* 2005, 6:R29 1-8.
11. Baldock RA, Bard JB, Burger A, Burton N, Christiansen J, Feng G, Hill B, Houghton D, Kaufman M, Rao J, et al. EMAP and EMAGE: a framework for understanding spatially organized data. *Neuroinformatics* 2003, 1:309-325.
12. Hill DP, Begley DA, Finger JH, Hayamizu TF, McCright IJ, Smith CM, Beal JS, Corbani LE, Blake JA, Eppig JT, et al. The mouse Gene Expression Database (GXD): updates and enhancements. *Nucleic Acids Res* 2004, 32: D568-D571.
13. Smith CL, Goldsmith CW and Eppig JT. The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information. *Genome Biology* 2004, 6:R7 1-9.



## An Online Ontology: WiktionaryZ

**Erik M. van Mulligen, Ph.D.<sup>1,2</sup>, Erik Möller, Peter-Jan Roes<sup>3</sup>, Marc Weeber, Ph.D.<sup>2</sup>,  
Gerard Meijssen, Christine Chichester Ph.D.<sup>2,4</sup>, Barend Mons Ph.D.,<sup>1,2,4</sup>**

**<sup>1</sup>Dept. of Medical Informatics, Erasmus Medical Center, Rotterdam, the Netherlands**

**<sup>2</sup>Knewco Inc, Rockville, United States of America**

**<sup>3</sup>Charta Software, Rotterdam, the Netherlands**

**<sup>4</sup>Human and Clinical Genetics, Leiden University Medical Center, Leiden, the Netherlands**  
e.vanmulligen@erasmusmc.nl

*There is a great demand for online maintenance and refinement of knowledge on biomedical entities<sup>1</sup>. Collaborative maintenance of large biomedical ontologies combines the intellectual capacity of millions of minds for updating and correcting the annotations of biomedical concepts with their semantic relationships according to latest scientific insights. These relationships extend the current specialization and participation relationships as currently exploited in most ontology projects. The ontology layer has been developed on top of the Wikidata<sup>2</sup> component and allows for presentation of these biomedical concepts in a similar way as Wikipedia pages. Each page contains all information on a biomedical concept with semantic relationships to other related concepts. A first version has been populated with data from the Unified Medical Language System (UMLS), SwissProt, GeneOntology, and Gemet. The various fields are online editable in a Wiki style and are maintained via a powerful versioning regiment. Next steps will include the definition of a set of formal rules for the ontology to enforce (onto)logical rigor.*

### INTRODUCTION

In order to deal with the deluge of biomedical information many projects have been initiated that aim at semantically annotating content. Many of these projects can be characterized as an attempt to exploit advanced natural language processing and text mining technology to identify the relevant semantic topics contained in a text<sup>3</sup>. By identifying these concepts in a text one can exploit available information about a concept as being formalized in an ontology for a number of tasks. One of these tasks is to improve information retrieval<sup>4</sup> (e.g., retrieval of texts on a particular concept might also include the

retrieval of documents with a more specific, narrower meaning). Another task would be semantic navigation between texts (e.g., exploring the semantic relationships between an identified concept in a text and concepts in other texts<sup>5</sup>).

Outside the biomedical domain the W3C has been working on defining exchange standards for ontologies. Their objective is to facilitate the development of technologies that enable cross-community data integration and collaborative efforts by adding semantics to the data. An example is the semantic web where webpages are semantically tagged and through these semantic tags linked to other webpages (similar to the current hyperlinked web). RDF, OWL and DAML<sup>6</sup> are examples of standards to impose semantic tags on information on the web. The meaning of these tags is captured in ontologies that contain additional information on how these semantic tags interrelate. These semantic interrelated tags can be used by applications for instance to semantically navigate between web resources.

All these tasks heavily rely on ontologies that serve as a repository of these biomedical concepts. Ontologies provide facilities to semantically relate the different biomedical topics. A first generation of ontologies (with limited scope) is available now. Good ontological principles have been a research topic and many scientific projects aim at a next generation of ontologies<sup>7</sup>. The Open Biomedical Ontologies consortium provides a platform for making available ontologies for shared use in the medical and biomedical domain that have been constructed with tools that bring in a greater degree of logical and ontological rigor<sup>8</sup>. Various tools have

been constructed that assist users with constructing these ontologies. Protégé is a freely downloadable program to construct ontologies using a strong formalism<sup>9</sup>.

OntoBuilder is another ontology editor that has been developed to automatically derive ontologies from a corpus (web pages) with support to refine and restructure them. Its focus is in particular on ontologies supporting the semantic web<sup>10</sup>. The main emphasis of all these tools is to make the development of (rigorous) ontologies easier. The whole process of collaboration, discussion and interrelating ontologies has not yet been addressed in these tools.

In this paper a mechanism is presented to harvest from existing ontologies originating from different sources and make these ontologies available for web-based refinement through a collaborative effort of the community of scientists. The hypothesis is that the online interaction, discussion and annotation of biomedical concepts will lead to wider coverage and higher quality ontologies with more semantics defined. Typically, most ontologies limit themselves to defining a hierarchy containing the specialization or participation relations. The biomedical semantic relations (a particular biomedical concept has a particular semantic relationship with another biomedical concept) require experts to interact and refine. These are important for the next generation of intelligent applications.

It is clear that an ontology has to cover a substantial part of the domain in order to be useful. In the biomedical domain, this would require that at least a substantial part of all medical concepts and of all genomic and proteomic concepts have to be in. Current vocabularies in these fields yield about 1,352K concepts for the medical domain (UMLS<sup>11</sup>) and about 200K for the genomics and proteomics domain (Swiss-Prot, EntrezGene, and Gene Ontology<sup>12</sup>).

Building a comprehensive ontology is an enormous endeavor. Bringing together all ontological knowledge from different biomedical disciplines in one environment seems to be quite impossible.

Furthermore, a biomedical ontology is not a static, one-time effort. Such an ontology should be continuously revised and updated with the latest new biomedical concepts and the latest semantic relations between the concepts<sup>1</sup>. Only imagining the rate with which genomics and proteomics data are produced yielding new information on genes and proteins it becomes clear that a comprehensive and up-to-date ontology is beyond the capabilities of any single scientific project.

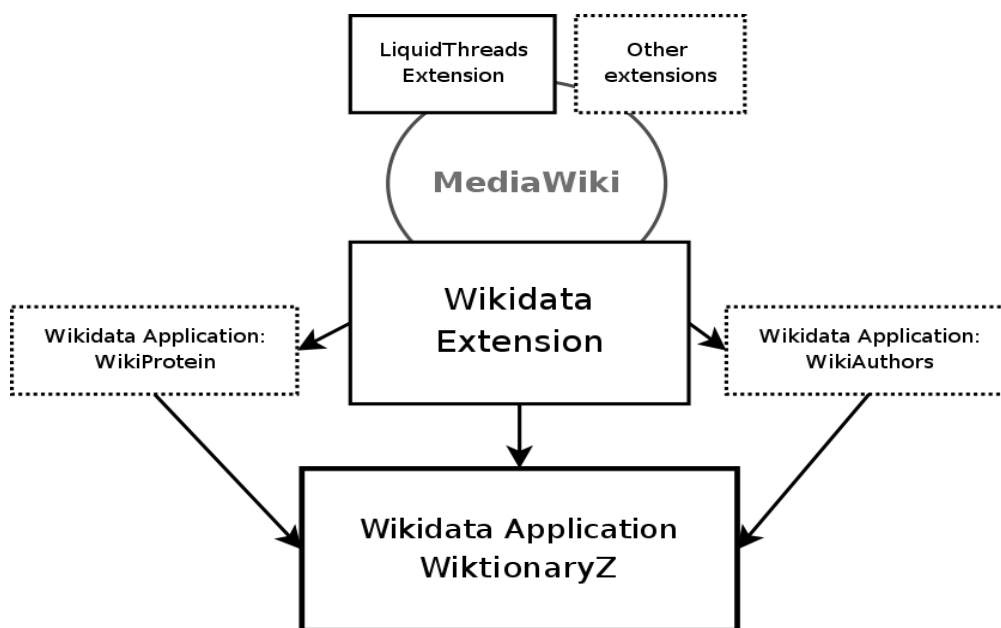
The only way to cope with such enormous amounts of data in so many different biomedical fields is to have an open environment in which all scientists can collaboratively share their knowledge on particular biomedical topics. Therefore we are currently investigating the possibilities of using a web-based approach to build and maintain biomedical ontologies. Benefiting from the pioneering work of the Wikimedia Foundation on collaborative development of web-based encyclopedias, we are exploring the possibilities to adapt a Wikimedia product in such a way that it can be used to support collaboration on ontology work: the WiktionaryZ software.

Many of the current vocabularies do not satisfy the ontological principles as current research has defined<sup>13</sup>. In addition, editing and updating ontologies should follow rules that guarantee soundness and correctness of the ontology. Description logic in combination with the specification of a separate hierarchy along the specialization and participation relation could make it possible to automatically detect errors in the concept classification. The WiktionaryZ has been developed in such a way that such an additional hierarchy can be expressed.

In addition to creating a collaborative instrument for biomedical scientists, this approach is also of interest to language engineering scientists. A systematic translation of biomedical terms is a rich source for language engineers and of great interest to them.

## METHODS

The architecture of WiktionaryZ (see Figure 1) has been based on the existing MediaWiki software. Wikidata itself is an extension of the MediaWiki



*Figure 1 - Schematic overview of the architecture of WiktionaryZ. It has been developed on top of the existing MediaWiki software.*

software that allows for structured data functionality beyond editing flat documents like Wikipedia articles. All data are stored in a MySQL relational database management system. WiktionaryZ has been built using Wikidata to store multilingual ontologies. It supports the notion of concepts, terms, synonyms, translations, definitions and alternative definitions, semantic relations, attributes, ontology class membership, and source annotations. Each of these elements is stored in the database as a separate entity. These entities can be combined in various queries supporting different applications. Specific applications (e.g., WikiProtein and WikiAuthors) can be defined as an implementation of the WiktionaryZ schema definition (with possibly some application-specific extensions).

The WiktionaryZ software provides the same functionality as the MediaWiki software with respect to online editing (talk pages) and version management. In order to distinguish between the ontology as provided by the authority - i.e. the organization that developed the thesaurus or vocabulary - and the version as maintained by the community an extended version management system is in place. The WiktionaryZ software discriminates between two version branches: the so-called authoritative version and the community version.

These two branches are more or less independent: new versions of the authoritative version can be imported without disrupting the community version. Vice versa are edits made by the community clearly (visually) distinguishable from the authoritative version avoiding any confusion with respect to accountability. The authority can monitor and selectively include community edits to refine its own authoritative version. The community can harvest from the latest release of the version maintained by the authority after its import into the authoritative branch.

Every scientist can contribute and discuss information on a concept. The version management layer treats every edit as a new version. Versions can be rolled back if such a rollback does not cause relational inconsistencies. The LiquidThreads extension supports multiple threads per Wiki page. This means that one could have a discussion thread around the definition of a concept and a separate one for the translations of terms. The WiktionaryZ software and its database are available under a free content license as defined by the Free Content Definition (<http://www.freecontentdefinition.org>).

A Wikidata application is defined by a namespace and associated functionality. Each different vocabulary can have its own namespace and attached

to its namespace can be additional tables that require specific functionality. For instance, in the WikiProtein namespace each protein can be described by its own specific features, such as amino acid sequence, the species of origin, the experimentally identified function, etc. For a gene concept, the DNA sequence could be given. Despite these specializations for each namespace, the concepts share a common set of data (and structure) for each concept.

Each biomedical concept is defined by a definition – a short and precise specification of the concept. A biomedical concept can have additional definitions: these definitions might comprise real alternatives for the definition or definitions with a slightly different perspective: aiming at a different scientific discipline or at a different community (high school students, for instance). Figure 2 shows an example of the information comprised at a WiktionaryZ page. The palette of semantic relations between the biomedical concepts has initially been defined as the set of relations defined in the Semantic Network of the Unified Medical Language System<sup>11</sup>. This set of

hierarchically organized relations can be easily extended and refined by the user.

Attached to each concept are terms (and synonyms), the language utterances used to refer to the concept. These terms are organized per language. Translations for each term can be entered and the system has been predefined with codes as defined in the ISO/FDIS 639-3 standard. Attached to each definition can be attributes. Initially these attributes will specify properties on the defined meaning: for instance the semantic type (e.g., a disease, a gene, a finding, a chemical, etc.) of the biomedical concept.

In order to benefit from the biomedical concepts as already defined in existing vocabularies and thesauri batch import facilities have been developed for the WiktionaryZ. Import facilities are now available for the UMLS files, Swiss-Prot files, Gene Ontology files, and the Gomet files. Most information contained in these vocabularies and thesauri has been successfully imported and made available in a WiktionaryZ environment.

The screenshot shows a web browser window titled "WiktionaryZ:virus - KnewCo WZ - Mozilla Firefox". The address bar shows the URL "http://wiki.mined2mind.org/ums/index.php/WiktionaryZ:virus". The page content is as follows:

**WiktionaryZ:virus**

Your user interface language: en — [Set your preferences](#)

– Language: English

– A term for a group of infectious agents which with few exceptions are capable of passing through fin...

– Definition

Language Text

English A term for a group of infectious agents which with few exceptions are capable of passing through fine filters that retain most bacteria, are usually not visible through the light microscope, lack independent metabolism, and are incapable of growth or reproduction apart from living cells. The complete particle usually contains only DNA or RNA, not both, and is usually covered by a protein shell or capsid that protects the nucleic acid. They range in size from 15 um up to several hundred um. Classification of viruses depends upon characteristics of virions as well as upon mode of transmission, host range, symptomatology, and other factors.

+ Alternative definitions

– Synonyms and translations

Expression	Language	Spelling	Identical meaning?
English	Viruses		✓
English	Viridae		✓
English	Vira		✓
English	Virus		✓
English	Viruses, General		✓
English	VIRUS		✓

– Relations

Relation type	Other defined meaning
can be qualified by	classification
can be qualified by	enzymology
can be qualified by	genetics
can be qualified by	isolation & purification
can be qualified by	metabolism
can be qualified by	pathogenicity
can be qualified by	radiation effects

## DISCUSSION

No other online editing environment has been developed that supports collaboration of scientists on annotation and semantic refinement of an ontology. The currently available tools allow for development of ontologies along some ontology design principles. However, many scientists need to be involved to refine the ontologies to a fine granular conceptual level, to annotate the concepts, and to express the semantic relationships between concepts, in short, to represent and codify the continuous advances of scientific knowledge about any biomedical subject. For effective use of ontologies in biomedical applications it is crucial to go beyond the current foundational relations of ontologies and beyond the well established and consistently described concepts.

Our first experiments with building the WiktionaryZ demonstrate that it is quite feasible to have large sets of concepts contained in a Wikidata database. The web based interface is fast enough to retrieve the concepts and combine all concept related data dispersed in different tables to the user. Pages are referenced per term. In case of a homonymous terms the page shows all the concepts for which the term is defined. The concept page can be very long. Currently WiktionaryZ does not provide any mechanism to define views on the data. A simple first approach would be to only show data for the language(s) that the user has indicated. More advanced views that are depending on the nature of the user's task can also be foreseen (i.e., differentiate between annotators, scientists, students, ontology developers, translators, high school students, etc.).

The WiktionaryZ does provide a powerful search facility: it searches for exact matches and allows for partial matches, both in the expressions associated with each concept and in their definitions. Misspellings and phonetic search are not implemented yet. It is evident that the current implementation lacks the ontological framework that allows for more sophisticated and rigorous quality control. This is essential when various users with different skill levels in ontology development are editing the ontology. Inclusion of a set of proper and well-defined relations expressed in a formal way should yield a more robust and more consistent editing of the ontology. Violation of these editing

rules should lead to alerts to the user but should not be prohibited. It is at the moment unclear how much of the potential inconsistency problems can be avoided by this framework.

The alignment of different vocabularies also requires special attention. How can identical concepts defined in different vocabularies be aligned (mapped to the same concept)? It is yet unclear how we can support automatic detection of (almost) synonymous concepts (e.g., "water" and "H<sub>2</sub>O" as being equivalent but defined in different vocabularies). This aspect has been a topic of study for already quite some years and we will explore the possibilities that have been identified.

A comprehensive biomedical ontology that can be effectively used for a number of tasks (bioinformatics, clinical medicine) will contain at least 2 million biomedical concepts. This is a rough estimate based on combining the current available thesauri, taken into account the overlap and the amount of non-medical concepts together with those parts that are still missing. Currently the National Library of Medicine, the Swiss Institute for BioInformatics, and the Gene Ontology Consortium have, apart from providing their sources, expressed their interest in this effort. An online maintained ontology will provide mechanisms to improve their authoritative sources as well.

In order to be able to include other ontologies/ thesauri as well the development of a method that can both read and write ontologies expressed in a standard syntax (OBO, OWL) has to be developed. This would make it possible to easily include a wide range of ontologies that are currently available in this format. Furthermore, the export allows the source authorities to download the latest edits for inclusion in their local version of the source. The current implementation of the system shows that it is technically feasible to have all these thesauri combined in one WiktionaryZ environment. What the impact - both with respect to quality and performance - of a large scientific community will be on such an online ontology remains a topic of research and will be part of future evaluation studies.

## References

1. Wang K. Gene-function Wiki Would Let Biologists Pool Worldwide Resources. *Nature* 2006; 439-534
2. Möller E. Wikidata: Wiki-Style Databases. Available from:  
<http://mail.wikipedia.org/pipermail/wikitec-h-l/2004-September/025377.html>
3. Nagao K., Shirai Y, Squire K. Semantic Annotation And Transcoding: Making Web Content More Accessible. *IEEE Multimedia*, 2001;8(2):69-81
4. Müller H-M, Kenny EE, Sternberg PW. Textpresso: An Ontology-Based Information Retrieval and Extraction System for Biological Literature. *PloS Biology*, 2004;2(11).
5. Buitelaar P, Eigner Th, Racioppa S. Semantic Navigation With VieWs. *Proceedings of the Workshop on User Aspects of the Semantic Web at the European Semantic Web Conference*. 2005.
6. Miller E. Weaving Meaning : An Overview Of The Semantic Web. Presented at the University of Michigan, Ann Arbor, Michigan USA, 2004
7. Smith B, Rosse C: The Role Of Foundational Relations In The Alignment Of Biomedical Ontologies. *Proc. Medinf 2004*. Amsterdam: IOS Press, 2004;444-8.
8. Available from:  
<http://obo.sourceforge.net/main.html>
9. Knublauch H, Fergerson RW, Noy NF, Musen MA. The Protégé OWL Plugin: An Open Development Environment For Semantic Web Applications. *Third International Semantic Web Conference*, Hiroshima, Japan, 2004.
10. Roitman H, Gal A. OntoBuilder: Fully Automatic Extraction And Consolidation Of Ontologies From Web Sources Using Sequence Semantics. *Proceedings of the International Conference on Semantics of a Networked World (ICSNW)*, 2006
11. Lindberg DA, Humphreys BL, McCray AT. The Unified Medical Language System. *Methods Inf Med*. 1993;32(4):281-91.
12. Bada M, Stevens R, Goble C, Gil Y, Ashburner M, Blake JA, et al: A Short Study On The Success Of The GeneOntology. *J Web Semantics* 2004;1:235-40.
13. Smith B, Ceusters W, Klagges B, Köhler J, Kumar A, Lomax J, et al. Relations In Biomedical Ontologies. *Genome Biology* 2005; 6(5)

## **“*Lmo-2* interacts with *Elf-2*” On the Meaning of Common Statements in Biomedical Literature**

**Stefan Schulz<sup>1</sup>, Ludger Jansen<sup>2</sup>**

<sup>1</sup> Department of Medical Informatics, Freiburg University Hospital, Germany

<sup>2</sup> Department of Philosophy, University of Rostock, Germany

### **Abstract**

*Statements about the behavior of biological entities, e.g. about the interaction between two proteins, abound in the literature on molecular biology and are increasingly becoming the targets of information extraction and text mining techniques. We show that an accurate analysis of the semantics of such statements reveals a number of ambiguities that is necessary to take into account in the practice of biomedical ontology engineering. Several concurring formalizations are proposed. Emphasis is laid on the discussion of biological dispositions.*

### **Introduction**

The study of so-called protein-protein interactions is essential for a better understanding of biological processes, from replication and expression of genes to the morphogenesis of organisms. Statements such as “*Lmo-2* interacts with *Elf-2*” – with *Lmo-2* and *Elf-2* being proteins – occur in biomedical literature abstracts with a very high frequency and represent, in many cases, the core message of a scientific paper.

There are several kinds of biomolecular interactions, e.g. binding, inhibition, activation, and transport. They all involve (1) at least two biomolecules and (2) the spatial vicinity of these, which leads to (3) a causal influence that they exert on each other.

Text mining, i.e. the process of extracting structured knowledge from unstructured text, primarily targets statements such as these, and there is a major interest by the text mining community in obtaining ontological support for their information and knowledge extraction activities. This is one of the reasons why so-called bio-ontologies have emerged, and the use of formal ontological criteria

has been repeatedly advocated in order to facilitate the process of automatic processing of domain information.

Much work in this area has already been done in the form of ontological investigations on material continuants, such as organs, cells, molecules [1, 2, 3]. However, there has been much less emphasis on biologically relevant functions and processes. Furthermore, biomedical ontology engineering has been mainly committed to traditions of semantic networks, lexical semantics, and cognitive science. Thus rather than being construed as describing real world entities by means of logical expressions, ontology has been understood as relating concepts (i.e. representations of word meanings) by means of conceptual relations. On this assumption, “*Lmo-2* interacts with *Elf-2*” would simply signify that there is some plausible linkage between the concepts (conceived of as mental representations) “*Interaction*”, “*Lmo-2*”, and “*Elf-2*”. As much as this approach might be adequate for communicating knowledge about the world by means of natural language or some kind of abstraction (e.g. semantic networks), it fails where exact statements and reasoning about biological entities such as molecules, functions, or pathways are required.

Interestingly enough, scientists and other human agents are perfectly able to communicate by means of such sentences, although there is only a vague consensus about the referents (the entities in the world) which are denoted by these linguistic expressions. Because of the ambiguities of natural language, a natural language statement like “*Lmo-2* interacts with *Elf-2*” may have more than one possible interpretation and thus more than one formalization in, say, first order predicate logic. In

formally representing the meaning of such statements, we have thus to make explicit the ontological assumptions intended by the speakers or authors of that sentence.

In this paper we will demonstrate that even the formalization of an apparently simple but prototypical statement about protein interaction like “*Lmo-2* interacts with *Elf-2*” can yield totally different ontological assumptions.

## Basic Ontological Assumptions

It is widely recognized that the construction of biomedical ontologies should obey strict logical and ontological criteria. To this end, several top-level ontologies have been devised, such as DOLCE [4], BFO [5], and GOL [6]. These ontologies mainly coincide in their fundamental division between continuants (endurants, e.g. material objects) and occurrents (perdurants, e.g. events, processes). The distinction is that occurrents have temporal parts (they are never fully present at a given time) and they are existentially dependent on continuants. Continuants are split into independent and dependent ones. Examples of independent continuants are material objects and spaces. Dependent continuants, on the other hand, are entities which inhere in something and are thus ontologically dependent on their bearer. Examples of dependent continuants are masses, colors, and tendencies: A particular mass may inhere in a particular molecule, a particular color may inhere in a particular flower. The tendency to divide may inhere in a cell and the tendency to relieve headache may inhere in an aspirin tablet. Tendencies are related to occurrents by the relation of *realization*. They are special kinds of dependent entities, in that they need not be realized in order to exist. There are cells which never divide, and aspirin tablets that never relieve a headache.

Our ontological framework for describing molecular interaction patterns includes entities of all these kinds. For instance, a protein molecule, which is a material continuant, has a disposition to perform a certain function, e.g. binding, which is a dependent continuant, and an actual realization of this disposition, *viz.* the process of binding a protein molecule, which is an occurrent.

Aware of the need to comply with existing stan-

dards for ontologies, especially in the light of the Semantic Web and the various specifications of Description Logic (DL) [7], we keep our logic simple. So we refrain from higher-order logics, as well as temporal or modal logics. We also use a parsimonious set of relations, following the OBO (Open Biological Ontologies) recommendation [8]. As a primitive formal relation we introduce the irreflexive, non-transitive and asymmetric instantiation relation *inst* which relates particular entities to their universal properties. In addition, we need a formal relation for class subsumption between universals, expressing scientific findings about relations between the kinds of things that are in the world. As scientific laws are meant to range not only over all present instances of a given kind, but also over all past and future instances and, moreover, also over merely possible instances [9], such a relation is not easily defined. We will here follow the OBO standard and introduce, to this end, the taxonomic subsumption relation *Is-a* by means of the *inst*<sup>1</sup> relation [8]. We will neglect the time parameter, which is not important for present purposes. On this basis, we define *Is-a* as a reflexive, transitive, and antisymmetric relation between universals *A* and *B*, as follows:

$$\begin{aligned} Is-a(A, B) &=_{def} \\ \forall x : (inst(x, A) \rightarrow inst(x, B)) \end{aligned} \quad (1)$$

Furthermore, we make the following ontological subdivision: When we deal with things of a certain kind, we have to distinguish between *individuals* belonging to this kind and *collectives* of individuals that belong to the same kind [1]. This very natural ontological distinction, which is mirrored by the singular / plural division in most natural languages, must be addressed wherever collectives or pluralities of individual objects occur. However, this distinction is often obscured when referring to mass entities (e.g. water vs. water molecules). Given the atomicity of material continuants, we do not admit material mass entities in our present framework, but consider them as collectives of particles instead.

---

<sup>1</sup>We use capitalized initial letters for the names of relations between universals as well as for the names of universals.

A collective is given, e.g., by all *Lmo-2* molecules involved in an experiment, as opposed to exactly one individual *Lmo-2* molecule. In the following, we will use the subscript "COLL" to refer to collectives. Thus, for each universal *X* we principally admit the existence of a corresponding collective  $X_{COLL}$ , the class of collections of instances of *X*. For instance, "*ProteinMolecule*<sub>COLL</sub>" denotes the class of collectives of protein molecules as well as "*Lmo-2*<sub>COLL</sub>" the class of collectives of *Lmo-2* molecules. We also admit collectives of occurrents, such as *Interaction*<sub>COLL</sub>.

### Concurrent interpretations I: Event Readings

Let us come back to our example: "*Lmo-2* interacts with *Elf-2*". Such statements are generally formulated by researchers who collect scientific evidence by empirical observations. These observations are commonly made in an indirect way, since the objects under scrutiny are below the threshold of visibility. For this reason measurement procedures of varying degrees of sophistication are applied, the results of which can be used to draw conclusions about the significance of an experiment. These conclusions may vary in their degrees of certainty. This certainty is affected by measurement errors as well as by errors in the design of the experiment which then may lead to false conclusions. If a statement like "*Lmo-2* interacts with *Elf-2*" is being uttered in a laboratory or written in a scientific paper or textbook, the minimal thing that can be inferred is that there are molecules of type *Lmo-2* and *Elf-2*. By way of contrast, this inference is not possible if such a sentence appears in a science fiction novel. As we are interested here in the scientific context only, we assume in what follows that there are universals *Lmo-2* and *Elf-2* that are kinds of protein molecules. These types of protein molecules do, of course, belong to the genus of protein molecules, which in turn are molecules, which are a kind of continuants. If one has an Aristotelian theory of universals, universals only exist if they are instantiated; that is, the existence of the universals *Lmo-2* and *Elf-2* implies that there are individual molecules that instantiate these universals. (Whoever has a different theory of universals may have to add this as a further

assumption.) All the possible readings of "*Lmo-2* interacts with *Elf-2*" to be discussed in the remainder of this paper have thus the following as a common ground:

$$\begin{aligned} &Is-a(Lmo-2, ProteinMolecule) \wedge \quad (2) \\ &Is-a(Elf-2, ProteinMolecule) \wedge \\ &Is-a(ProteinMolecule, Molecule) \wedge \\ &Is-a(Molecule, Continuant) \wedge \\ &\exists l, e : inst(l, Lmo-2) \wedge inst(e, Elf-2) \end{aligned}$$

Despite this common ground, the sentence remains highly ambiguous even within a scientific context. First we will discuss interpretations of "*Lmo-2* interacts with *Elf-2*" that interpret it as a report of events. Here are some possible interpretations of the sample statement that belong to this group:

1. One individual *Lmo-2* molecule interacts with one individual *Elf-2* molecule.
2. A collection of *Lmo-2* molecules interacts with one individual *Elf-2* molecule.
3. One individual *Lmo-2* molecule interacts with a collection of *Elf-2* molecules.
4. A collection of *Lmo-2* molecules interacts with a collection of *Elf-2* molecules.

Our sample statement appears to describe the fact that exactly one such interaction happened. Alternatively, it can describe the fact that a multitude of such interactions (as described in 1-4) happens, which would be the normal thing in many biochemical contexts. This adds up to eight different interpretations. But any of these interpretations is still ambiguous in a very important respect. With each of these interpretations, the speaker may mean either that such interaction(s) did actually happen, or the speaker may mean that the molecules in question have the disposition or the tendency to interact in such a way. This gives way to even more possible interpretations. Thus, "*Lmo-2* interacts with *Elf-2*" turns out to be a highly ambiguous sentence. We will now discuss the different possible interpretations of this sentence in turn and suggest methods for representing them formally.

## Occurrences involving individual continuants

On the first interpretation, “*Lmo-2* interacts with *Elf-2*” describes the fact that an individual *Lmo-2* molecule interacts with an individual *Elf-2* molecule. A standard way to render such a situation formally would be the use of the existence quantifier of first order predicate logic:

$$\begin{aligned} \exists l, e : inst(l, Lmo-2) \wedge \\ inst(e, Elf-2) \wedge interacts(l, e) \end{aligned} \quad (3)$$

This formalization ensures that there is *at least* one individual *Lmo-2* molecule which interacts with *at least* one individual *Elf-2* molecule at *at least* one instant. This interpretation can now be modified in various ways. We could, e.g., add exclusivity postulates like in (4) that ensure that *exactly* one individual molecule of each kind are interacting with each other. Though such a solitary event might be rarely observed in experiments, there may be contexts where this is the intended meaning:

$$\begin{aligned} \exists l, e : inst(l, Lmo-2) \wedge inst(e, Elf-2) \wedge \\ interacts(l, e) \wedge \\ \forall l^*, e^* : (inst(l^*, Lmo-2) \wedge inst(e^*, Elf-2) \wedge \\ interacts(l^*, e^*)) \rightarrow (l^* = l \wedge e^* = e) \end{aligned} \quad (4)$$

Normally, however, this formalization will be much too strong an interpretation of our sample statement. For any statement of this form will be *false*, if at any other time another *Lmo-2* molecule interacts with an *Elf-2* molecule – or if at the very same time another *Lmo-2* molecule interacts with an *Elf-2* molecule at any other place. Therefore, we do not consider it as a useful interpretation of our sample sentence. We will, however, refer back to this formula and the exclusivity clauses used in it in the following discussion.

In Formulae 3 and 4 we have expressed the interaction event by means of a binary relation *interacts* between individual continuants. This relation on the level of instances is irreflexive (nothing ever interacts with itself), symmetric, and non-transitive. The OBO (Open Biological Ontologies) relation ontologies, however, recommends to restrict ourselves to a parsimonious array of basic relations.

Therefore, we will eliminate the *interacts* relation, using the technique introduced by Davidson [10] to quantify over events. This means that we represent the interaction process as an occurrent entity in its own right rather than by the relation *interacts* as in Formulae 3 or 4. This move is made possible through our admission of occurrent entities, and it corresponds to common practice in biomedical ontologies. The relation between the particular process and the participating particular continuants is then given by the relation *has-participant* [8]. The *has-participant* relation is a relation between a particular occurrent and a particular continuant, in this order. It is irreflexive, asymmetric, and non-transitive. For nothing participates in a continuant and no occurrent participates in anything. Again, we dispense with a time index for sake of simplicity. Within a fully-fledged implementation, a time index should be included, as an occurrent may have different participants at different stages.

$$\begin{aligned} \exists l, e, i : inst(l, Lmo-2) \wedge inst(e, Elf-2) \wedge \\ inst(i, Interaction) \wedge \\ has-participant(i, l) \wedge has-participant(i, e) \end{aligned} \quad (5)$$

This formalization makes it easier to represent occurrences with more than two participants, as with the representation in Formula 3, where we would have to deal with *n*-ary relations for *n* participants. According to this formal representation, “*Lmo-2* interacts with *Elf-2*” is to be understood as stating that there is at least one interaction process, in which at least one protein molecule of the given kinds is involved. It does not exclude that other molecules are involved in this very interaction process. If we want to secure that *Lmo-2* and *Elf-2* are the only participants of the molecular interaction, we have to employ exclusivity conditions similar to 4:

$$\begin{aligned} \exists l, e, i : inst(l, Lmo-2) \wedge inst(e, Elf-2) \wedge \\ inst(i, Interaction) \wedge \\ has-participant(i, l) \wedge has-participant(i, e) \wedge \\ \forall x : (has-participant(i, x) \rightarrow \\ inst(x, Lmo-2) \vee inst(x, Elf-2)) \end{aligned} \quad (6)$$

If we want to keep the requirement of pairwise interaction, if have to add uniqueness conditions

in the fashion of Formula 4 for this purpose:

$$\begin{aligned}
& \exists l, e, i : inst(l, Lmo-2) \wedge inst(e, Elf-2) \wedge \quad (7) \\
& \quad inst(i, Interaction) \wedge \\
& \quad has-participant(i, l) \wedge has-participant(i, e) \wedge \\
& \quad \forall x : (has-participant(i, x) \rightarrow \\
& \quad \quad inst(x, Lmo-2) \vee inst(x, Elf-2)) \wedge \\
& \quad \forall l^*, e^* : (inst(l^*, Lmo-2) \wedge inst(e^*, Elf-2) \wedge \\
& \quad \quad has-participant(i, l^*) \wedge has-participant(i, e^*)) \\
& \quad \rightarrow (e^* = e \wedge l^* = l))
\end{aligned}$$

In contrast to Formula 4, such a formalization that quantifies over events is still much more realistic, because its truth is compatible with more than one interaction process of the same kind happening at the same time or at other times.

### Occurrents involving collectives of continuants

As mentioned above, it is important to distinguish between individuals of a kind and collectives of individuals of that kind. Rector and Bittner [1] have accounted for this by introducing the formal relation *has-grain* which relates a collective *c* to each of its constituents *e*. In [3] this account has been further developed by introducing a collective universal  $X_{COLL}$  whose instances are constituted by two or more constituents which are instances of *X*:

$$\begin{aligned}
& \forall c : inst(c, X_{COLL}) \rightarrow \exists e_1, e_2, \dots, e_n, n > 1 : \quad (8) \\
& \quad \bigwedge_{\nu=1}^n inst(e_\nu, X) \wedge has-grain(c, e_\nu)
\end{aligned}$$

Note that *has-grain* is a subrelation of *has-part*. As a consequence, we identify a collection as a mereological sum of its constituents (regardless of their spatiotemporal arrangement), and not as a mathematical set. The reason for rejecting the set approach is two-fold. Firstly, because mathematical sets are extensional and therefore not robust with regard to the gain and loss of constituents. Secondly, because collectives should be of the same ontological category as their constituents: A collective of material objects should be a material object, and a collective of events should be an event. Sets, however, are abstract objects that do neither

exist in space nor in time. We do not use the *has-part* relation, because participants in interactions may have parts that do not themselves participate in the interaction. A *Lmo-2* molecule, e.g., may participate in an interaction without every of its electrons being a participant in this interaction. Whereas *has-part* is transitive, *has-grain* is not. It is a irreflexive, asymmetric, and intransitive relation that holds between particular collectives and individuals.

We therefore modify our formalism substituting individuals by collectives:

$$\begin{aligned}
& \exists l, e, i : inst(l, Lmo-2_{COLL}) \wedge \quad (9) \\
& \quad inst(e, Elf-2_{COLL}) \wedge inst(i, Interaction) \wedge \\
& \quad has-participant(i, l) \wedge has-participant(i, e)
\end{aligned}$$

### Collectives of occurrents

Formalism 7 and 9 use the same occurrent type *Interaction* for different scenarios: In the first case, a particular interaction has individual protein molecules as participants, in the second case collectives of molecules. This ambiguity may be acceptable when we talk about such a generic process as interaction. It would not be tolerable in the case of a more specific one, such as *binding*. A binding can only happen between two individual molecules, not between two collectives of molecules. Thus, if we encounter a plurality of bindings within a plurality of molecules, it would not be admissible to describe this as a binding between two collectives of molecules but rather a collective of bindings between pairs instances of the kinds of molecules in question<sup>2</sup>. Thus we have to deal with a collective of processes rather than with collectives of continuants.

In order to represent such a situation, let us first introduce the collective interaction universal  $I_{COLL}$  which is constituted by individual constituents which are instances of *I*, analogously to Formula 8. Then we have to determine how each of the grain interactions look like. If they are pairwise interactions between an *Lmo-2* molecule and an *Elf-2* molecule, each of these interactions fits

<sup>2</sup>A counterexample is the interaction between solutes and solvents in a solution which necessarily involves collectives of both solvents and solutes.

Formula 7. Combining Formulae 7 and 8, we get:

$$\begin{aligned}
& \exists p, i_1, i_2, \dots, i_n, n > 1 : \\
& \bigwedge_{\nu=1}^n (inst(i_\nu, I) \wedge has-grain(p, i_\nu) \wedge \\
& \exists l_\nu, e_\nu : inst(l_\nu, Lmo-2) \wedge inst(e_\nu, Elf-2) \wedge \\
& \quad has-participant(i_\nu, l_\nu) \wedge \\
& \quad has-participant(i_\nu, e_\nu) \wedge \\
& \forall x : (has-participant(i_\nu, x) \rightarrow \\
& \quad inst(x, Lmo-2) \vee inst(x, Elf-2)) \wedge \\
& \forall l_\nu^*, e_\nu^* : ((inst(l_\nu^*, Lmo-2) \wedge \\
& \quad inst(e_\nu^*, Elf-2) \wedge has-participant(i_\nu, l_\nu^*) \wedge \\
& \quad has-participant(i_\nu, e_\nu^*)) \rightarrow (e_\nu^* = e_\nu \wedge l_\nu^* = l_\nu))
\end{aligned} \tag{10}$$

## Concurrent Interpretations II: Dispositional Readings

The above interpretations stated the existence of one or more interaction events. However, messages of the style “*Lmo-2* interacts with *Elf-2*” very often do not focus on the accidental occurrence of an event but are rather meant to express some inherent property of the objects under investigation. On the one hand it is likely that a biologist would mean “An interaction between *Lmo-2* and *Elf-2* happened” while describing the outcome of a specific experiment. On the other hand a biology textbook would rather want to communicate something like “*Lmo-2* molecules have the disposition or tendency to interact with *Elf-2* molecules”. This ambiguity, of course, matches Aristotle’s famous distinction between act and potency, and Aristotle himself observed that “potency” is in itself an ambiguous term [11]. Thus the ambiguity of our sample statement increases even more, because the dispositional reading of our sample sentence is ambiguous in itself. Obviously, such a reading of “*Lmo-2* interacts with *Elf-2*” is intended to ascribe some causal or statistical property, a disposition or tendency. But even if this is the common ground of the dispositional reading, three questions remain open and have to be answered:

1. Which event is it exactly that the property in question is meant to cause?

2. What is thought to be the bearer of this property?
3. Which kind of property is in fact intended to be ascribed?

The first question can be answered by pointing to one of the many event readings we discussed (and formalized) thus far. Our answer to the second question will at least in part depend on our response to question 1. Are all instances of a given universal bearers of the disposition in question? Or only some of the instances? Are the individual molecules the bearers of the disposition, or rather collectives of such molecules? The third question, however, leads us in to the middle of the lively debate going on in philosophy on the ontology of disposition [12, 13, 14]. The dispositional properties most often discussed in the literature are so-called *surefire dispositions*: dispositions to react invariably in a certain way under specific circumstances. They are one candidate for an answer to question 3. From the point of view of knowledge representation, however, there are some problems connected with surefire dispositions. First, things may react differently in different circumstances. Thus to say that *Lmo-2* molecules have the disposition to interact with *Elf-2* molecules still leaves it open under which circumstances such an interaction will occur. We could account for this by explicitly mentioning the conditions of realization for each disposition. We may, of course, not *know* all these conditions, but this is an epistemic problem only. A more significant problem is that there may be infinitely many causally relevant conditions that have to be taken into account, and such an infinite list would be impossible for principled reason. We could try to circumvent this problem by adding (implicitly or explicitly) quantification phrases like “In all circumstances” or “In some circumstances”. The *all*-phrase, however, will not do. For if a certain disposition would be realized under all circumstances, it will never be unrealized. Such cases may exist, but normally a disposition will only be realized under certain circumstances and not realized under others. When we use the *some*-phrase, on the other hand, many statements about dispositions for molecule interactions will become trivial, since nearly any mole-

cule may interact with any other molecule in some peculiar way under certain (possibly very extreme) conditions. A usual way to deal with this problem is to introduce a set of standard or normal conditions [15]. In biology, this could mean that the “disposition to interact with *Elf-2* molecules” is only ascribed to *Lmo-2* if the interaction commonly occurs under biological conditions, such as physiological pH and temperature intervals. But the problem is not solved by referring to normal conditions. For, first, the problem that infinitely many conditions cannot be described in necessarily finite lists recurs with normal conditions. And, second, biomedical knowledge may also include the behavior of molecules in non-normal or even extreme circumstances, like low or high temperatures, exposure to intensive sunlight or atomic radiations. One option at this point would be to choose a different answer to question 3. Instead of ascribing surefire dispositions we could ascribe probabilistic dispositions, i.e. dispositions to do something (under certain circumstances) with a certain probability [16]. Such causal properties are also sometimes called “tendencies” [17] or “propensities” [18]. While with surefire dispositions a certain event will happen invariably in given circumstances, the event in question will only happen with a certain probability when a tendency is ascribed. It will, of course, be crucial to know with *which* probability the event will happen. Following standard procedures in mathematical probability theory, we can represent the quantities of the probabilities in question by real numbers between 0 and 1 satisfying the Kolmogorov axioms. In biomedical experiments, the observed result is often such a probability. Tendencies are thus of vital importance for the representation of biomedical knowledge [17]. There can, however, be several ontological groundings for such a probability. Suppose that we observed a hundred instances of a given universal *U* in situations in which all conditions necessary for the realization *R* of a certain disposition were present, but that in only fifty cases *R* happened, i.e. in only 50 % of all cases the disposition realized itself. There are several ontological scenarios that would explain this result. Here are two of them:

- (A) Every instance of *U* has a tendency to *R* with a probability of 0.5.
- (B) Every second instance of *U* has a surefire disposition to *R*; the other instances of *U* do not have any disposition to *R*.

Both of these scenarios would explain the assumed observations. Which of these scenarios we choose for our account of the observation will depend on other observations and causal assumptions. If we, e.g. knew that nearly always the same instances of *U* display *R* and nearly always the same instances of *U* do not display *R*, this would *prima facie* count as a reason to embrace (B). If, on the other hand, we know that the same instances of *U* sometimes do display *R* and sometimes do not display *R*, this would *prima facie* count as a reason to embrace (A). For such reasoning, however, we need background assumptions about the stability of the causal properties in question: how they can be stable over time, how (if at all) they can be acquired and how (if at all) they can get lost. Last but not least, (B) can indicate that the instances of the universal *U* differ in certain features, which are crucial to the ability to display *R*. An important example of this is the observation of modified proteins produced by mutated genes opposed to the observation of normal (wild-type) proteins. Considering all this, there is quite a long and complex list of entities that we implicitly refer to when ascribing a disposition or tendency to a molecule:

- (independent) continuants (i.e. the bearer of the disposition),
- dependent continuants or occurrents (i.e. the realization),
- quantities (of probabilities), and
- state of affairs (of realization conditions).

## Conclusion

Our deliberations shed light on the need for a more principled account of dispositions and processes in biomedical ontologies. Machine supported information extraction and knowledge acquisition techniques from scientific texts have become a cornerstone in molecular biology and genomics due to the increasing scientific productivity in this field.

The necessity of logic based ontologies for this purpose has been controversially discussed [19]. If we subscribe to a formally principled account as a basis for the semantic representation of the content of scientific texts then we have to take into account that the most common type of statements that are of interest in texts describing biochemical regularities do not have a clear and unambiguous meaning. Assertions of the type “A interacts with B” are generally more than accounts of a single event. Rather they refer to a plurality of events of the same kind, or an event involving pluralities (collectives) of participants. A universal interpretation such as “For each instance of *A* there is an interaction with some *B*” can easily be discarded. The need for universal quantifications can be satisfied by introducing dispositions: “Every *A* has the disposition to interact with some *B*.” However, not every occurrence of the participation of some continuant in some process is proof of the existence of a related disposition.

Since interaction is a very general term, it is difficult to express a clear preference in favor of any of the proposed approaches without analyzing the nature of interaction on a molecular level, as well as the study of the “normal” behavior of biomolecules. The question when to ascribe a disposition or tendency – and which one – can not be discussed here (but cf. [16] on this).

We demonstrated that sentences like “A interacts with B” exhibit indeed a wide range of ambiguity. We offered several possible analyses to formally represent the different meanings of sentences of this type. Now, which one should we choose? One strategy would be to say: Which strategy you choose depends on the intended meaning of the particular occurrence of the sentence you deal with. For text mining purposes, however, that have to digest large amounts of texts in short periods of time and with as much automatization as possible, this strategy would be scarcely feasible. To cope with this situation, several strategies are conceivable. One strategy would be to choose the highest common factor of all interpretation – that what is included in all. Another strategy would be to set as a standard interpretation that is *most likely* the intended meaning. In order to determine which interpretation is the best candidate, empir-

ical work on relevant text corpora may be helpful. This, however, is already beyond the scope of the present paper.

#### Acknowledgments:

This work was supported by the EU Network of Excellence *Semantic Interoperability and Data Mining in Biomedicine* (NoE 507505), the project *Forms of Life* sponsored by the Volkswagen Foundation, and the *Wolfgang Paul Award* of the Alexander-von-Humboldt-Foundation. We are indebted to Andrew D. Spear and to the anonymous referees of KR-MED for valuable comments.

#### Address for Correspondence:

Stefan Schulz, Department of Medical Informatics, Freiburg University Hospital, Stefan-Meier-Str. 26, 79104 Freiburg (Germany), phone: +49 761 203 6702, e-mail: stschulz@uni-freiburg.de

## References

- [1] Alan Rector, Jeremy Rogers, and Thomas Bittner. Granularity scale and collectivity: When size does and doesn’t matter. *Journal of Biomedical Informatics*, 38, 2005.
- [2] Stefan Schulz and Anand Kumar. Biomedical ontologies: What part-of is and isn’t. *Journal of Biomedical Informatics*, 38, 2005.
- [3] Stefan Schulz, Elena Beisswanger, Udo Hahn, Joachim Wermter, Anand Kumar, and Holger Stenzhorn. From GENIA to BioTop towards a top-level ontology for biology. *FOIS 2006 – International Conference on Formal Ontology in Information Systems*, 2006. Accepted for publication.
- [4] Aldo Gangemi, Nicola Guarino, Claudio Masolo, and Alessandro Oltramari. Sweetening ontologies with DOLCE. In Asunción Gómez-Pérez and V. Richard Benjamins, editors, *Proceedings of the 13th International Conference – EKAW 2002*, volume 2473 of *Lecture Notes in Artificial Intelligence*, pages 166–181. Berlin: Springer, 2002.

- [5] Barry Smith and Pierre Grenon. The cornucopia of formal-ontological relations. *Dialectica*, 58(3):279–296, 2004.
- [6] Barbara Heller and Heinrich Herre. Ontological categories in GOL. *Axiomathes*, 14(1):57–76, 2004.
- [7] Franz Baader, Diego Calvanese, Deborah L. McGuinness, Daniele Nardi, and Peter F. Patel-Schneider, editors. *The Description Logic Handbook. Theory, Implementation, and Applications*. Cambridge, U.K.: Cambridge University Press, 2003.
- [8] Barry Smith, Werner Ceusters, Bert Klagges, Jacob Köhler, Anand Kumar, Jane Lomax, Chris Mungall, Fabian Neuhaus, Alan L. Rector, and Cornelius Rosse. Relations in biomedical ontologies. *Genome Biology*, 6(5), 2005.
- [9] Karl. R. Popper. *Logic of Scientific Discovery*. Hutchinson, London, 1959.
- [10] Donald Davidson. The logical form of action sentences. In N. Rescher, editor, *The Logic of Decision and Action*, pages 81–95. Pittsburgh, PA: University of Pittsburgh Press, 1967.
- [11] Ludger Jansen. *Tun und Können. Ein systematische Kommentar zu Aristoteles’ Theorie dispositionaler Eigenschaften im neunten Buch der Metaphysik*. Hänsel-Hohenhausen, Frankfurt am Main, 2002.
- [12] Raimo Tuomela. *Dispositions*. Reidel, Dordrecht, 1978.
- [13] Stephen Mumford. *Dispositions*. Oxford University Press, Oxford, 1998.
- [14] Bruno Gnassounou and Max Kistler. *Les dispositions en philosophie et en sciences*. CNRS Editions, Paris, 2006. An English translation is to be published with Ashgate Publishers.
- [15] Wolfgang Spohn. Begründungen a priori – oder: ein frischer Blick auf Dispositionsprädikate. In Wolfgang Lenzen, editor, *Das weite Spektrum der analytischen Philosophie. Festschrift Franz von Kutschera*, pages 89–106. de Gruyter, Berlin/New York, 1997.
- [16] Ludger Jansen. Attribuer des dispositions. In Bruno Gnassounou and Max Kistler, editors, *Les dispositions en philosophie et en sciences*, pages 89–106. CNRS Editions, Paris, 2006.
- [17] Ludger Jansen. The ontology of tendencies and medical information sciences. In Ingvar Johansson, Bertin Klein, and Thomas Roth-Berghofer, editors, *WSPI 2006: Contributions to the Third International Workshop on Philosophy and Informatics*, volume 14 of *IFOMIS Reports*, pages 89–106. IFOMIS, Saarbrücken, 2006.
- [18] Karl. R. Popper. *A World of Propensities*. Thoemmes, Bristol, 1990.
- [19] Sophia Ananiadou and Jun’ichi Tsujii. The saurus or logical ontology, which one do we need for text mining? *Language Resources and Evaluation, Springer Science and Business Media B.V.*, 39(1):77–90, 2005.



# The qualitative and time-dependent character of spatial relations in biomedical ontologies

Thomas Bittner<sup>1,3,4</sup> and Louis J. Goldberg<sup>2,3</sup>

<sup>1</sup>Departments of Philosophy and Department of Geography,

<sup>2</sup>Departments of Oral Biology and Oral Diagnostic Sciences, School of Dental Medicine,

<sup>3</sup>New York State Center of Excellence in Bioinformatics and Life Sciences

<sup>4</sup>National Center of Geographic Information and Analysis (NCGIA)

State University of New York at Buffalo

{bittner3, goldberg}@buffalo.edu

## Abstract

*The formal representation of mereological aspects of canonical anatomy (parthood relations) is relatively well understood. The formal representation of other aspects of canonical anatomy like connectedness relations between anatomical parts, shape and size of anatomical parts, the spatial arrangement of anatomical parts within larger anatomical structures are, however, much less well understood and only partial represented in computational anatomical ontologies. In this paper we propose a methodology of how to incorporate this kind of information into anatomical ontologies by applying techniques of qualitative spatial representation and reasoning from Artificial Intelligence. As a running example we use the human temporomandibular joint (TMJ).*

## INTRODUCTION

Anatomical ontologies are formal representations of facts about the major parts of anatomical structures, the qualitative shapes of those parts, and qualitative relations between them [19, 13, 30].

The formal representation of mereological aspects of canonical anatomy (parthood relations) is relatively well understood [16, 31, 13], and has been implemented in computational medical ontologies like the FMA [23], GALEN [22], and SNOMED [32]. On the other hand, the formal representation of other aspects of canonical anatomy like connectedness relations between anatomical parts, shape and size of anatomical parts, the spatial arrangement of anatomical parts within larger anatomical structures are less well understood and only partially represented in computational anatomical ontologies. In this paper we propose a methodology of how to incorporate this kind of information into anatomical ontologies.

We stress here the importance of recognizing the qualitative nature of *all* facts represented

in anatomical ontologies such as the FMA. It is impossible to quantitatively describe aspects of shape and spatial arrangement of canonical anatomy. There is too much variation between the actual shapes and metric arrangements of particular structures among particular human beings. Moreover it is the very nature of many anatomical structures to change in shape and spatial arrangement over time: the heart beats, the jaw opens and closes, etc.

Qualitative representations of canonical anatomy take advantage of the fact that despite the variations and changes in size, shape, distance, and spatial arrangement, at the gross anatomical level, all normal instances of the same biological species are qualitative copies of each other. In all canonical anatomical structures certain parts need to be present. These parts need to have certain qualitative shape features (convex parts, concave parts, other landmark features, etc.), their size must be within certain limits, and certain qualitative relations need to hold between those parts: some parts are connected to others, some part are disconnected from others, some parts (like articular discs) need to be between other parts (like the bones in synovial joints) etc.

In this paper we give an overview of the most important of those relations. We also demonstrate how the changes in shape and arrangement can be specified using qualitative spatial relations. In addition, we claim that most pathological cases can also be characterized and distinguished from non-pathological cases in terms of qualitative relations: there may be too many or too few parts, parts that are supposed to be connected are disconnected, parts that are supposed to be between other parts fail to be so, etc.

Qualitative representation of, and reasoning about complex systems has a long tradition in Artificial Intelligence [34, 5, 10]. Cohn and Hazarika [8]

stress that the essence of qualitative representations is to find ways to represent continuous properties of the world by discrete systems of symbols. As Forbus [14] points out, one can always quantize something continuous, but not all quantizations are equally useful because the distinctions made by a quantization must be relevant for the kind of reasoning performed. This is where *formal ontology* comes into play [29]. It will be an important aspect of this paper to show how to discretize continuous domains in such a way that ontologically significant properties are preserved.

For example, to qualitatively model the behavior of water at different temperatures the continuous domain of temperature is discretized by introducing landmark values: temperature landmark 1 (TLM1) the temperature at which water changes from its solid state to its liquid state and (TLM2) the temperature where water changes from its liquid state to being a gas. These landmark values bound intervals: for example, (TI1) the interval of temperatures at which water is solid, (TI2) the interval of temperatures at which water is liquid, and (TI3) the (half open) interval at which water is a gas. In a qualitative model the behavior of water at different temperatures is described only by referring to the landmark values and the intervals bounded by those values.

An important point is that the landmarks are not chosen arbitrarily. The landmarks represent *significant changes* in the domain at hand, while within the intervals between landmarks no significant changes occur. Thus qualitative representations focus on *ontologically salient features*. For many purposes this qualitative representation of water at different temperatures will be sufficient. For example, in order to transport bottled water from one place to another the exact temperature of the water is irrelevant as long as it does not freeze or change to its gas state since in both cases the bottled water will destroy their containers.

We propose the following methodology for building qualitative representations of canonical anatomical structures that preserve ontologically significant distinctions:

1. Specify and classify the major canonical parts of the structure at hand and establish canonical mereotopological (parthood and connectedness) relations between them;
2. Identify ordering relations between the major parts anatomical structures to qualitatively

characterize the spatial arrangement of the parts within the structures;

3. Refine ordering relations between parts by identifying anatomical landmarks and by using landmarks as a frame of reference;
4. Specify qualitative distance relations between landmarks to qualitatively characterize shape and arrangement of the parts.

We will discuss each step below in sequence and use the human temporomandibular joint (TMJ) as a running example. We go into a detailed discussion of how existing techniques of qualitative spatial representation and reasoning from Artificial intelligence can be used and extended to formally and qualitatively represent the mereotopology of anatomical structures, the shape and size of anatomical parts, and the spatial arrangement of anatomical parts within larger anatomical structures. The methods we present here we believe will provide the foundations for the next generation of anatomical ontologies.

## ANATOMICAL PARTS AND MEREOTOPOLOGICAL RELATIONS

### Parthood relations

At the most basic level of the study of the canonical structure of the TMJ we consider its anatomical parts. Anatomical parts here means, maximally connected parts of non-negligible size (thus cells and molecules are parts of anatomical structures but not anatomical parts). At this gross anatomical granularity we will distinguish two kinds of anatomical parts: material parts and cavities. The material anatomical parts of the TMJ at the gross anatomical level of granularity according to [18] are depicted in Figure 1, which shows, in a sagittal section through the middle of the condyle, a TMJ in closed (a) and open (b) jaw position: temporal bone (1), head of condyle (2), articular disc (3), posterior attachment (4), lateral pterygoid muscle (5). Immaterial anatomical parts (cavities) are the superior and inferior synovial cavities, which are depicted as white spaces above and below the articular disc and the posterior attachment. Here we will focus on material parts. For a discussion of immaterial anatomical parts see [12, 26, 19].

A clear understanding of the number and kinds of canonical parts of an anatomical structure is

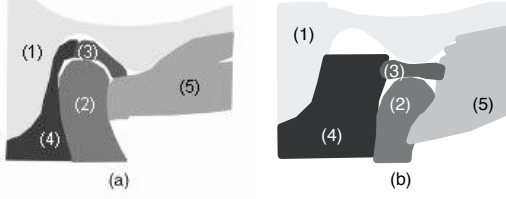


Figure 1: Drawings of (a) the major parts of a TMJ in the jaw closed position and (b) the major parts of the same TMJ in the jaw open position.

critical for identifying non-canonical (and potentially pathological) parts such as tumors. Moreover, without a clear understanding of the number of canonical parts it is not possible to recognize the absence of certain parts. In the remainder of this paper we refer to individual anatomical structures and their material anatomical parts as objects.

Parthood is a ternary relation (a relation with three arguments) that holds between two objects  $x$  and  $y$  and a time instant  $t$ . Parthood is a time-dependent relation since anatomic structures can have different parts at different times. For example, in the course of their transition from children to adults, it is normal for people to have different teeth at different times. See, for example, [27] for axiomatic formalizations time-dependent parthood.

In terms of parthood we define the relations of proper parthood and overlap. Object  $x$  is a *proper part* of object  $y$  at  $t$  if and only if  $x$  is a part of  $y$  at  $t$  and  $y$  is not part of  $x$  at  $t$ . For example, at time  $t$  the head of Joe's condyle is a proper part of his condyle. Object  $x$  *overlaps* object  $y$  at time  $t$  if and only if there is an object  $z$  such that  $z$  is part of  $x$  at  $t$  and  $z$  is part of  $y$  at  $t$ . If  $x$  is a (proper) part of  $y$  at  $t$  then  $x$  and  $y$  overlap at  $t$ . Thus, at time  $t$  Joe's condyle and the head of his condyle overlap.

### Connectedness relations

The ternary relation of connectedness holds between two objects  $x$  and  $y$  at a time instant  $t$ . Intuitively,  $x$  is connected to  $y$  at  $t$  if and only if  $x$  and  $y$  overlap at  $t$  or  $x$  and  $y$  are in direct external contact at  $t$ . Two regions are connected at  $t$  if and only if they share at least a boundary point at  $t$  (they may share interior points at  $t$ ). For a discussion of the wide range of possible formalizations see [33].

Objects  $x$  and  $y$  are *externally connected* at time  $t$  if and only if  $x$  and  $y$  are in direct external contact

at  $t$  but  $x$  and  $y$  do not overlap at  $t$ . Externally connected regions share boundary points but no interior points. Objects  $x$  and  $y$  are *disconnected* at time  $t$  if and only if  $x$  and  $y$  are not connected at  $t$ .

We introduce connectedness as a time-dependent relation since anatomic structures can be connected to different (parts of) structures at different times. As depicted in Figure 1(a), at time  $t_1$  the articular disc is (externally) connected to the fossa (a fiat part<sup>1</sup> of the temporal bone). At time  $t_2$ , as depicted in Figure 1(b) the articular disc is connected to the articular eminence (another fiat part of the temporal bone).

The following topological relations hold between the five major parts of the TMJ depicted in Figures 1(a) and (b): the temporal bone (1) is externally connected to the posterior attachment (4) and to the lateral pterygoid muscle (5). The condyle (2) is externally connected to the posterior attachment (4) and to the lateral pterygoid muscle (5). The articular disc (3) is externally connected to the posterior attachment (4) and the lateral pterygoid muscle (5).

### Permanent parthood and connectedness

Consider the relation of external connectedness between the articular disc and the temporal bone. Clearly, at every time  $t$  the articular disc is externally connected (in external contact) to *some* part of the temporal bone. However at different times the articular disc is externally connected (in external contact) to *different* parts of the temporal bone. In Figure 1 (a) the articular disc is externally connected (in external contact) to the fossa, while in Figure 1 (b) the articular disc is externally connected (in external contact) to the articular eminence (another fiat part of the temporal bone).

It is important to make explicit that the connectedness relation between the articular disc and the temporal bone is different from the connectedness relation between the articular disc the posterior attachment and the lateral pterygoid muscle: at all times at which the articular disc is connected to the posterior attachment it is connected to the *same* part of the posterior attachment and similarly for the lateral pterygoid muscle. The relation between articular disc and posterior attachment is a relation of *constant or permanent* con-

<sup>1</sup>A fiat part is a part which boundaries are (partly) the result of human demarcation and do not correspond to discontinuities in reality [28].

nection (articular disc and posterior attachment are ‘glued’ together by direct connective tissue attachments). On the other hand the relationship between articular disc and temporal bone is such that both are externally connected (in external contact) but the articular disc has the freedom to slide along the surface of the bone.<sup>2</sup>

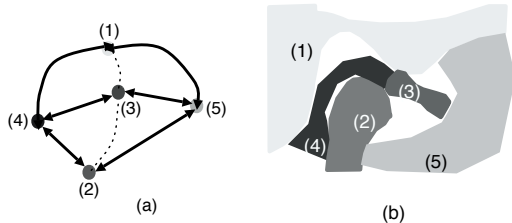


Figure 2: (a) Graph structure which represents the relations of external connectedness between the major parts of the TMJ, (b) TMJ with articular disc not positioned between condyle and temporal bone.

We define the following constant mereotopological relations: Object  $x$  is a *constant* part of object  $y$  if and only if whenever  $y$  exists,  $x$  is a part of  $y$ . Object  $x$  is a *constant proper part* of object  $y$  if and only if whenever  $y$  exists,  $x$  is a proper part of  $y$ . Object  $x$  is a *constantly* connected to object  $y$  if and only if whenever  $y$  exists,  $x$  is connected to  $y$ . Object  $x$  is a *constantly* externally connected to object  $y$  if and only if whenever  $y$  exists,  $x$  is externally connected to  $y$ . Object  $x$  is a *constantly* disconnected from object  $y$  if and only if whenever  $y$  exists,  $x$  is disconnected to  $y$ .

Consider Figure 2 (a). Every part of the TMJs in Figure 1 (a) and (b) is topologically equivalent to a filled circle which is indicated by the corresponding labels of the dots in Figure 2. Moreover, the nodes (the labeled circles) in the graph represent constant proper parts of the TMJ: at all times at which the TMJ as a whole exists, the condyle (2) is a proper part of it. Similarly the temporal

<sup>2</sup>Strictly speaking, this ability to slide is due to the fact that the articular disc is separated from the temporal bone by a film of fluid which fills the superior synovial cavity. As stated previously, for the purpose of this paper we will not consider cavities or holes, and so will consider that the articular disc is effectively free to slide to various positions along the surface of the temporal bone. Notice, however, that we could introduce a relation of adjacency. We would then have to distinguish between constant adjacency and temporary adjacency in the same way we distinguish constant external connectedness and temporary external connectedness.

bone (1), the articular disc (3), the posterior attachment (4), and the lateral pterygoid muscle (5) are constant proper parts of the TMJ.

The solid edges in the graph in Figure 2(a) represent constant connectedness relations between parts of the TMJs depicted in Figure 1 (a) and (b): at all times at which the TMJ as a whole exists the condyle (2) is (externally) connected to the posterior attachment (4) and to the lateral pterygoid muscle (5). By contrast, a (with respect to time) different connectedness relation holds between articular disc (3) and the temporal bone (1) and the articular disc and the head of the condyle (2): the disc is externally connected to different parts of the temporal bone and the head of the condyle at different times. In the graph in Figure 2(a) this is represented by dotted edges between the respective nodes.

## ORDERING RELATIONS BETWEEN EXTENDED OBJECTS

Mereotopology alone is not powerful enough to sufficiently characterize the important properties of TMJs. Consider the graph in Figure 2(a), which is a graph-theoretical representation of the mereotopological properties of the TMJs depicted in Figures 1(a), 1(b), and 2(b). The fact that the TMJs depicted in the three figures have the same graph-theoretic representation shows that in terms of mereotopological properties we cannot distinguish the TMJs in Figures 1(a), 1(b), and 2(b).

Obviously it is critical to distinguish the TMJ in Figure 2(b) from the TMJs in Figures 1(a) and 1(b). It is the purpose of the articular disc in a TMJ to be *between* the condyle and temporal bone at all times. If we take the ordering relation of betweenness into account then the TMJs in Figures 1(a) and 1(b) can be distinguished from the clearly pathological TMJ in Figure 2(b) where the posterior attachment is between the condyle and the temporal bone and not the articular disc.

Ordering relations like betweenness describe the location of disjoint objects relatively to one other. Besides betweenness, ordering relations include: left-of, right-of, in-front-of, above, below, behind, etc. The science of anatomy has developed a whole set of ordering relation terms to describe the arrangement of anatomical parts in the human body: superior, inferior, anterior, posterior, lateral, medial, dorsal, ventral, rostral, proximal, distal, etc. The FMA, for example, has an ‘orientation network’ in which these kinds of relations are represented [23].

Unfortunately, ordering relations between spatially extended objects are difficult to formalize. As [11] points out in her treatment of relation of betweenness: ‘The problem with trying to characterize the betweenness relation on extended objects is that we typically use the betweenness relation only on objects that have fairly uniform shapes and are nearly the same size. It is unclear whether or not the betweenness relation should hold in certain cases involving irregularly shaped objects and differently sized objects.’ Similar problems face attempts to formalize qualitative direction relations between spatially extended objects, e.g., [20]. Similarly it is very difficult to qualitatively describe distances between extended objects particularly if they are of different size and shape, e.g., [36, 35].

## LANDMARKS

To avoid problems that occur when describing ordering relations between extended objects we will choose a different approach: we will characterize shape, extent, and spatial arrangement of anatomical structures and their anatomical parts using (point-like) *anatomical landmarks* [6] and qualitative ordering relations between the landmarks.

### Landmarks of anatomical structures

Intuitively, anatomical landmarks are *special salient points* on the surface of anatomical structures or their anatomical parts [6]. Consider the temporal bone in Figure 3. Salient points on the inferior surface of the temporal bone are local minima (LM3, LM7), local maxima (LM1, LM5) as well as points at which changes from convexity to concavity occur (LM2, LM4, LM6).

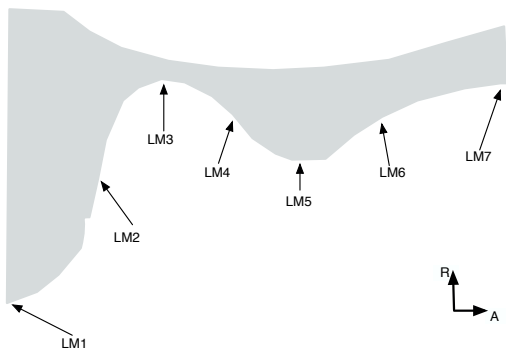


Figure 3: Landmarks on Joe’s temporal bone.

However not all salient points on the surface of a given anatomical structure are landmarks. Salient

points are landmarks of anatomical structures *of a given kind* if and only if:

1. They exist as parts of every anatomical structure of that kind;
2. They are critical for the normal function of all anatomical structures of that kind.

Thus the salient points LM1-LM6 in Figure 3 are anatomical landmarks of temporal bones of normal human TMJs, since (a) they exist as parts of every temporal bones of a normal human TMJ and (b) they are important for the function of a human TMJ as a whole. Consequently, independently of the normal variations between the actual shape of temporal bones in different human beings, all normal temporal bones will have the landmarks LM1-LM7 as depicted in Figure 3.

### Qualitative distances between landmarks

Although normal temporal bones in human TMJs will have the landmarks LM1-LM7, the particular metric properties like the actual height of the maximum, the actual depth of the minimum, as well as their actual distance, will vary from individual to individual.

Consider the landmarks of the temporal bone depicted in Figure 3. Rather than quantitatively characterizing shape differences in terms of coordinate differences among the landmarks, we can characterize the shape differences qualitatively by specifying *qualitative* distance relations between those landmarks. Consider, for example, the anatomical landmarks LM1 and LM3. In Figure 3 the coordinate difference along the anterior (horizontal) axis is smaller than the coordinate difference along the rostral (vertical) axis. Similarly the coordinate difference between LM3 and LM5 along the anterior axis is roughly twice as large as the coordinate difference along the rostral axis.

Since all TMJs will have the same landmarks on their temporal bones (assuming a certain degree of anatomical normality), we can classify TMJs according to qualitative coordinate differences between their landmarks. There are many ways of doing this. Here we only discuss some examples to demonstrate the power of the qualitative methodology. In particular we focus on the landmarks LM1, LM3, and LM5.

Given a coordinate system<sup>3</sup> existing coordinate

<sup>3</sup>We do not need the coordinate system for measurement. We only use it to distinguish coordinate differences in anterior (horizontal) direction ( $\delta h$ ) from coordinate differences in rostral (vertical) direction ( $\delta v$ ).

differences between LM1 and LM3 along the anterior axis ( $\delta a_3^1$ ) and along the rostral axis ( $\delta r_3^1$ ) can be used to distinguish the following cases:  $\delta a_3^1 = \delta r_3^1$ ,  $\delta a_3^1 < \delta r_3^1$ , and  $\delta a_3^1 > \delta r_3^1$ . Here  $\delta a_3^1 = \delta r_3^1$  means that  $\delta a_3^1$  is as large as  $\delta r_3^1$ ,  $\delta a_3^1 < \delta r_3^1$  means that  $\delta a_3^1$  is smaller than  $\delta r_3^1$ , and  $\delta a_3^1 > \delta r_3^1$  means that  $\delta a_3^1$  is larger than  $\delta r_3^1$ . Notice that this classification is *jointly exhaustive and pairwise disjoint*. That is, for any possible constellation of the anatomical landmarks LM1 and LM3 exactly one of those relations holds. In Figure 3 the rostral coordinate difference between LM1 and LM3 is larger than the anterior coordinate difference between LM1 and LM3, i.e.,  $\delta a_3^1 < \delta r_3^1$ .

Of course we can in addition classify the anterior and rostral coordinate differences between the landmarks LM3 and LM5 in the same way. If we take both classifications together then the following nine combinations are combinatorially possible:

$R \in \{=, <, >\}$	1	2	3	4	5	6	7	8	9
$\delta a_3^1 R \delta r_3^1$	=	=	=	<	<	<	>	>	>
$\delta a_5^3 R \delta r_5^3$	<	>	=	<	>	=	<	>	=

Any possible constellation of LM1, LM2, and LM3 is characterized by exactly one column in this table. In Figure 3 we have  $\delta a_3^1 < \delta r_3^1$  and  $\delta a_5^3 > \delta r_5^3$ , which corresponds to column 5 in the above table. Since this classification is exhaustive we now can analyze which of the nine possibilities are normal and which are pathological or which correlate with certain clinical symptoms. This analysis may show that distinguishing nine cases is insufficient to make the necessary distinction to distinguish normal anatomy from various kinds of pathologies. In this case we have three options: (a) take more landmarks into account; (b) distinguish more relations; (c) do both (a) and (b).

Consider option (b) instead of distinguishing three relations =, <, and > we could add two more relations:  $\ll$  and  $\gg$  interpreted as much smaller and much bigger respectively. Another way of distinguishing more relations would be to refine > by distinguishing twice as big, three times as big, etc. There are no limits to this method provided the resulting set of relations is jointly exhaustive and pairwise disjoint.

Notice that it might be more realistic to replace the identity relation = by the relation  $\sim$ , were  $\delta a \sim \delta r$  means that  $\delta a$  is *roughly* as large as  $\delta r$ . The exact definitions of the relations  $\sim$ ,  $\ll$ , and  $\gg$  are not trivial and their formalization is beyond the scope of this paper. For discussions of existing approaches see [21, 9, 7, 4].

## Qualitative directions and orientation relations between landmarks

There exist a variety of approaches to qualitatively represent angles between landmarks and to use landmarks as origins for qualitative frames of references. For example, the landmark ‘LM’ in Figure 4(a) could serve as the origin of the qualitative frame of reference in Figure 4(b). We then could specify the location of anatomical landmarks of the heart within this frame of reference.

Most of the approaches to qualitative orientation and directions also incorporate qualitative distance relations like close, near, far, etc. (where close, near, and far roughly correspond to the relations  $\sim$ , <, and  $\ll$  – see for example, [7, 4] for details). In Figure 4 we then could say that all anatomical landmarks of the heart are near and in front with respect to the frame of reference which is centered at the landmark LM. More sophisticated ways of representing qualitative order relations between landmarks were proposed in [15, 24, 25].

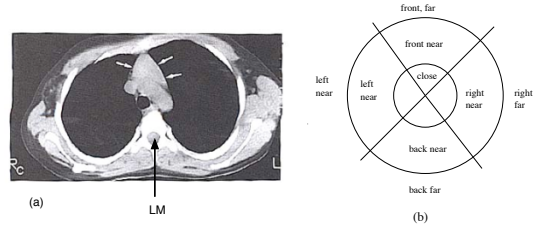


Figure 4: (a) a radiographic section taken through a human thorax. Arrows point to the heart. LM, Is a point in the center of the spinal cord. (b) qualitative ordering and qualitative distance relations according to Hernandez [17].

## APPROXIMATE LOCATION IN FRAMES OF REFERENCE

There are many ways to represent approximate location in qualitative frames of references. (See, for example [3].) Here we discuss a specific technique which is useful in the context of our TMJ example. Consider the boundary of Joe’s temporal bone as depicted in Figure 3. Topologically, the boundary is a one-dimensional curve. Since the landmarks LM1-LM7 are points on this curve, each landmark is a boundary of at least one interval (a one-piece part of the underlying curve). For example, in Figure 3 the landmarks LM2 and LM3 bound the interval which is formed by the part of the curve between them. We use the landmarks that bound

a given interval to refer to this interval. For example, we write  $\overline{L2L3}$  to refer to the interval bounded by LM2 and LM3 in Figure 3.

In our mereotopological framework we can represent the topological relations between the intervals formed by the anatomical landmarks of Joe's temporal bone as: Interval  $\overline{L1L2}$  is constantly externally connected to interval  $\overline{L2L3}$ , interval  $\overline{L2L3}$  is constantly externally connected to interval  $\overline{L3L4}$ , and so on.

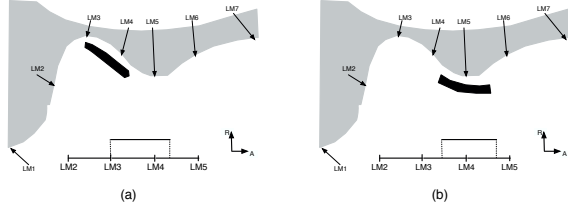


Figure 5: Relations between articular disc and landmark intervals of the temporal bone at times  $t_1$  (a) and  $t_2$  (b).

Consider Figures 5(a) and (b) which depict the relative location of Joe's articular disc with respect to his temporal bone at times  $t_1$  and  $t_2$  respectively. Figure 5(a) corresponds to Figure 1(a) and both show Joe's TMJ in the jaw closed position. Similarly, Figure 5(b) corresponds to Figure 1(b) and both show Joe's TMJ in the jaw open position. On the bottom of both images in Figure 5 the projection of Joe's articular disc onto the boundary of his temporal bone is depicted. From this point on, we will write  $Prj(D, t)$  to refer the interval that is the projection of Joe's articular disc on the boundary of his temporal bone in a sagittal section through the middle of his condyle at time  $t$ .

The interval  $Prj(D, t)$  stands in mereotopological relationships to the intervals bounded by the landmarks LM1-LM7. For example, at time  $t_1$  the projection of Joe's articular disc completely covers the interval  $\overline{L3L4}$ , i.e.,  $COV(Prj(D, t_1), \overline{L3L4}, t_1)$ . In other words the interval  $\overline{L3L4}$  is a part of the projection of Joe's articular disc, i.e.,  $PartOf(\overline{L3L4}, Prj(D, t_1), t_1)$ . Notice that at time  $t_2$  the projection of Joe's articular disc and the interval  $\overline{L3L4}$  are disconnected, i.e.,  $DC(\overline{L3L4}, Prj(D, t_2), t_2)$ .<sup>4</sup>

Thus at every time  $t$  we can specify the location of Joe's articular disc with respect to the landmarks of his temporal bone in terms of the rela-

<sup>4</sup>For details of the exact definitions of the relations between the intervals see [1, 2].

tions which hold at time  $t$  between the projection of the articular disc at  $t$  and the intervals bounded by the landmarks. These mereotopological relations at time  $t_1$  and  $t_2$  can be summarized as:

Joe's disc	$\overline{L1L2}$	$\overline{L2L3}$	$\overline{L3L4}$	$\overline{L4L5}$	$\overline{L5L6}$	$\overline{L6L7}$
$t_1$	$DC$	$EC$	$COV$	$PO$	$DC$	$DC$
$t_2$	$DC$	$DC$	$DC$	$PO$	$PO$	$DC$

The first row reads as  $DC(Prj(D, t_1), \overline{L1L2}, t_1)$ ,  $EC(Prj(D, t_1), \overline{L2L3}, t_1)$ , ... and similarly for the second row.

Consider the images shown in Figures 6(a) and (b) which depict the relative location of Joe's condyle with respect to his temporal bone at times  $t_1$  and  $t_2$  respectively. Figure 6(a) corresponds to Figure 1(a) and Figure 6(b) corresponds to Figure 1(b). In the same way we projected Joe's disc onto the boundary of his temporal bone to identify an interval that can be related to the intervals bounded by the landmarks LM1-LM7, we can project the head of his condyle onto the boundary of his temporal bone as indicated by the dotted lines in Figures 6(a) and (b).

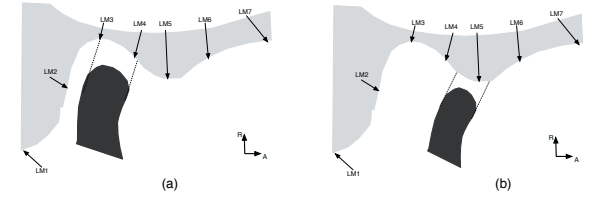


Figure 6: Mereotopological relations between the head of the condyle and landmark intervals of the temporal bone at times  $t_1$  (a) and  $t_2$  (b).

As in the case of Joe's disc, at every time  $t$  we can specify the location of the head of Joe's condyle with respect to the landmarks of his temporal bone in terms of the relations which hold at time  $t$  between the projection the head of the condyle at  $t$  and the intervals bounded by the landmarks. The spatial relations at time  $t_1$  and  $t_2$  can be summarized as:

Joe's condyle	$\overline{L1L2}$	$\overline{L2L3}$	$\overline{L3L4}$	$\overline{L4L5}$	$\overline{L5L6}$	$\overline{L6L7}$
$t_1$	$DC$	$EC$	$PO$	$DC$	$DC$	$DC$
$t_2$	$DC$	$DC$	$DC$	$PO$	$PO$	$DC$

If we use  $C$  to denote the head of Joe's condyle then the first row reads as  $DC(Prj(C, t_1), \overline{L1L2}, t_1)$ ,  $EC(Prj(C, t_1), \overline{L2L3}, t_1)$ , ..., and similarly for the second row. Notice that the table with the relations of Joe's articular disc corresponds nicely to the table with the relations of the head of Joe's

condyle, i.e., the articular disc is at both times *between* the head of the condyle and the temporal bone.

Clearly, for every possible location of an articular disc in a TMJ with respect to the temporal bone of this TMJ there is a unique sequence of relations similar to those in the table of Joe's disc. Similarly, for every possible location of the head of a condyle in a TMJ with respect to the temporal bone of this TMJ there is a unique sequence of relations similar to those in the table of Joe's condyle. Moreover, since we have, (i) the same anatomical landmarks on the temporal bones of every normal TMJ and, (ii) there are only a finite number of mereotopological relations that can hold between two intervals, we can therefore, compose two finite tables: one table in which each row corresponds to one anatomically possible location of some articular disc with respect to the corresponding temporal bone; a second table in which each row corresponds to one anatomically possible location of the head of some condyle with respect to the corresponding temporal bone.<sup>5</sup> Both tables together contain all possible combinations of locations of the head of a condyle and an articular disc with respect to the landmarks of a temporal bone in any possible TMJ. Some of these combinations we can classify as normal (among these are the two tables above) others are pathological and again others will be anatomically impossible and thus can be ruled out.

## CONCLUSIONS

The purpose of this paper is to show that there can be obtained, by following the methodology we have presented here, a series of well understood qualitative formalisms which can be used to create a formal representation of canonical anatomy. This is accomplished by incorporating into the representation, using the qualitative methods of analysis we describe in this paper, information about, a) the mereological (parthood) relationships of anatomical structures, b) the topology (e.g., connectedness) of anatomical structures, and c) the shape of anatomical parts and the spatial arrangement of anatomical structures.

The five cornerstones of the proposed methodology are:

1. The grounding of the formalization of canonical anatomy in mereotopology (rather than mereology alone);

---

<sup>5</sup>For formal details of how to construct the tables see [2].

2. The strict distinction of time-dependent and time-independent relations;
3. The identification of anatomical landmarks for the representation of the shape of anatomical parts and the spatial arrangement of anatomical structures;
4. The identification of sets of jointly exhaustive and pairwise disjoint relations to describe relations between anatomical parts and anatomical landmarks;
5. The establishment of landmarks and qualitative distinctions that reflect the ontologically significant aspects of the canonical anatomy of biomedical structures as well as relevant pathological cases.

This methodology permits, in principle, the exhaustive qualitative characterization of all anatomically possible instantiations of anatomical structures. These then can be classified as normal or pathological and correlated with other clinical findings.

The discussion in this paper exclusively focused on relations between particulars (Joe Doe's TMJ). It is well known that anatomical ontologies are mostly about relations between universals or classes [31, 30]. However it is also well known that relations between universals or classes are defined in terms of relations between particulars [13].

## Address for Correspondence

Thomas Bittner, State University of New York, Department of Philosophy, 135 Park Hall, Buffalo (NY), 14260, USA

## References

- [1] J.F. Allen. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11):832–843, 1983.
- [2] T. Bittner. Approximate qualitative temporal reasoning. *Annals of Mathematics and Artificial Intelligence*, 35(1–2):39–80, 2002.
- [3] T. Bittner. A mereological theory of frames of reference. *International Journal on Artificial Intelligence Tools*, 13(1):171–198, 2004.
- [4] T. Bittner and M. Donnelly. A theory of granular parthood based on qualitative cardinality and size measures. In B. Bennett and C. Fellbaum, editors, *Proceedings of the fourth International Conference on Formal Ontology in Information Systems, FOIS06*, 2006.
- [5] R.J. Brachman and H.J. Levesque, editors. *Readings in Knowledge Representation*. Morgan Kaufmann, Los Altos, Calif., 1985.

- [6] L.G. Brown. A survey of image registration techniques. *ACM Comput. Surv.*, 24(4):325–376, 1992.
- [7] E. Clementini, P. Di Felice, and D. Hernández. Qualitative representation of positional information. *Artificial Intelligence*, 95(2):317–356, 1997.
- [8] A G Cohn and S M Hazarika. Qualitative spatial representation and reasoning: An overview. *Fundamenta Informaticae*, 46(1-2):1–29, 2001.
- [9] P. Dague. Numeric reasoning with relative orders of magnitude. In *Proceedings of the National Conference on Artificial Intelligence*, pages 541–547, 1993.
- [10] E. Davis. *Representations of Commonsense Knowledge*. Morgan Kaufmann Publishers, Inc., 1990.
- [11] M. Donnelly. *An Axiomatization of Common-Sense Geometry*. PhD thesis, University of Texas at Austin, 2001.
- [12] M. Donnelly. On parts and holes: The spatial structure of the human body. In M. Fieschi, E. Coiera, and Y. J. Li, editors, *Proceedings of the 11th World Congress on Medical Informatics (MedInfo-04)*, pages 351–356, 2004.
- [13] M. Donnelly, T. Bittner, and C. Rosse. A formal theory for spatial representation and reasoning in bio-medical ontologies. *Artificial Intelligence in Medicine*, 36(1):1–27, 2006.
- [14] K. Forbus. Qualitative process theory. *Artificial Intelligence*, 24:85–168, 1984.
- [15] J.E. Goodman and R. Pollack. Allowable sequences and order types in discrete and computational geometry. In J. Pach, editor, *New Trends in Discrete and Computational Geometry*, volume 10 of *Algorithms and Combinatorics*, pages 103–134. Springer-Verlag, 1993.
- [16] U. Hahn, S. Schulz, and M. Romacker. Partonomic reasoning as taxonomic reasoning in medicine. In *Proceedings of the 16th National Conference on Artificial Intelligence and 11th Innovative Applications of Artificial Intelligence Conference*, pages 271–276, 1998.
- [17] D. Hernandez. *Qualitative Spatial Reasoning*. Springer-Verlag, 1994.
- [18] D. M. Laskin, C. S. Greene, and W. L. Hylander, editors. *TMJ's - An Evidence Based-Approach to Diagnosis and Treatment*. Quintessence Books, Chicago, 2006.
- [19] José L. V. Mejino and Cornelius Rosse. Symbolic modeling of structural relationships in the Foundational Model of Anatomy. In *Proceedings of the KR 2004 Workshop on Formal Biomedical Knowledge Representation*, Whistler, BC, Canada, 1 June 2004, pages 48–62, 2004.
- [20] D. Papadias and T. Sellis. On the qualitative representation of spatial knowledge in 2d space. *VLDB Journal, Special Issue on Spatial Databases*, pages pp. 479–516, 1994.
- [21] O. Raiman. Order of magnitude reasoning. *Artificial Intelligence*, 51:11–38, 1991.
- [22] J. Rogers and A. Rector. GALEN’s model of parts and wholes: experience and comparisons. In *Proceedings of the AMIA Symp 2000*, pages 714–8, 2000.
- [23] C. Rosse and J. L. V. Mejino. A reference ontology for bioinformatics: The Foundational Model of Anatomy. *Journal of Biomedical Informatics*, 36:478–500, 2003.
- [24] C. Schlieder. Reasoning about ordering. In A.U. Frank and W. Kuhn, editors, *Spatial Information Theory - A Theoretical basis for GIS*, volume 988 of *LNCS*, pages 341–349, Semmering, Austria, 1995. Springer-Verlag.
- [25] C. Schlieder. Ordering information and symbolic projection. In *Intelligent image database systems*, pages 115–140. World Scientific, Singapore, 1996.
- [26] S. Schulz and U. Hahn. Mereotopological reasoning about parts and (w)holes in bio-ontologies. In C. Welty and B. Smith, editors, *Formal Ontology in Information Systems. Collected Papers from the 2nd International Conference*, pages 210 – 221, 2001.
- [27] P. Simons. *Parts, A Study in Ontology*. Clarendon Press, Oxford, 1987.
- [28] B. Smith. On drawing lines on a map. In A.U. Frank and W. Kuhn, editors, *Conference on Spatial Information Theory, COSIT*, volume 988, pages 475–484. Springer-Verlag, Semmering, Austria, 1995.
- [29] B. Smith and B. Brogaard. Quantum mereotopology. *Annals of Mathematics and Artificial Intelligence*, 35(1–2), 2002.
- [30] B. Smith, W. Ceusters, B. Klagges, J. Köhler, A. Kumar, J. Lomax, C. Mungall, F. Neuhaus, A. Rector, and C. Rosse. Relations in biomedical ontologies. *Gnome Biology*, 6(5):r46, 2005.
- [31] B. Smith and C. Rosse. The role of foundational relations in the alignment of biomedical ontologies. In M. Fieschi, E. Coiera, and Y. J. Li, editors, *Proceedings of the 11th World Congress on Medical Informatics*, pages 444–448, 2004.
- [32] K.A. Spackman, K.E. Campbell, and R.A. Cote. SNOMED RT: A reference terminology for health care. In *Proceedings of the AMIA Annual Fall Symposium*, pages 640–4, 1997.
- [33] A. Varzi. Parts, wholes, and part-whole relations: The prospects of mereotopology. *Data and Knowledge Engineering*, 20(3):259–86, 1996.
- [34] D.S. Weld and J. de Kleer, editors. *Readings in Qualitative Reasoning about Physical Systems*. The Morgan Kaufmann Series in Representation and Reasoning. Morgan Kaufmann Publishers, INC, San Mateo, California, 1990.
- [35] M. Worboys. Metrics and topologies for geographic space. In M.J. Kraak and M. Moleenaar, editors, *Advances in Geographic Information Systems Research II: Proceedings of the International Symposium on Spatial Data Handling, Delft*, pages 7A.1–7A.11. International Geographical Union, 1996.
- [36] M. F. Worboys. Nearness relations in environmental space. *International Journal of Geographical Information Science*, 15(7):633–651, 2001.



## Towards a Reference Terminology for Ontology Research and Development in the Biomedical Domain

<sup>1</sup>Barry Smith, Ph.D., <sup>2</sup>Waclaw Kusnierczyk, M.D., <sup>3</sup>Daniel Schober, Ph.D.,  
<sup>1</sup>Werner Ceusters, M.D.

<sup>1</sup>Center of Excellence in Bioinformatics and Life Sciences, Buffalo NY/USA

<sup>2</sup>Department of Computer Computer and Information Science, NTNU, Trondheim, Norway

<sup>3</sup>European Bioinformatics Institute (EMBL-EBI), Hinxton, Cambridge, UK

phismith@buffalo.edu, waku@idi.ntnu.no, schober@ebi.ac.uk, ceusters@buffalo.edu

*Ontology is a burgeoning field, involving researchers from the computer science, philosophy, data and software engineering, logic, linguistics, and terminology domains. Many ontology-related terms with precise meanings in one of these domains have different meanings in others. Our purpose here is to initiate a path towards disambiguation of such terms. We draw primarily on the literature of biomedical informatics, not least because the problems caused by unclear or ambiguous use of terms have been there most thoroughly addressed. We advance a proposal resting on a distinction of three levels too often run together in biomedical ontology research: 1. the level of reality; 2. the level of cognitive representations of this reality; 3. the level of textual and graphical artifacts. We propose a reference terminology for ontology research and development that is designed to serve as common hub into which the several competing disciplinary terminologies can be mapped. We then justify our terminological choices through a critical treatment of the 'concept orientation' in biomedical terminology research.*

### PREAMBLE

Ever since the invention of the computer, scientists and engineers have been exploring ways of 'modeling' or 'representing' the entities about which machines are expected to reason. But what do 'modeling' and 'representing' mean? What is a 'conceptual model' or an 'information model' and how can they and their components be unambiguously described?

Two questions here arise: To what do expressions such as 'concept', 'information', 'knowledge', etc. precisely refer? And what is it to 'model' or 'represent' such things? If information and knowledge themselves consist in representations, then what could an *information representation* or a *knowledge representation* be? There is, to say the least, some suspicion of redundancy here.

As we have argued elsewhere, the term 'concept' is marked in a peculiarly conspicuous manner by problems in this regard.<sup>1</sup> But the problem of multiple conflicting meanings arises also in regard to other

terms, such as 'class', 'object', 'instance', 'individual', 'property', 'relation', etc., all of which have established, but unfortunately non-uniform, meanings in a range of different disciplines.

Among philosophical ontologists, the term 'instance' means an individual (for example this particular dog Fido), which is an instance of a corresponding universal or kind (*dog*, *mammal*, etc.). In OWL, 'instance' means 'element' or 'member' of a class (where 'class' means 'general concept, category or classification ... that belongs to the class extension of owl:Class'<sup>2</sup>).

Standardization agencies such as ISO, CEN and W3C have been of little help in engendering cross-disciplinary uniformity in the use of such terms, since their standards are themselves directed towards specific communities. Standardization efforts under the auspices of W3C or UML or Dublin Core, too, have not addressed these problems. For while OWL-DL, for example, has a rigorously defined semantics,<sup>3</sup> this does not by any means guarantee that an ontology formulated using OWL-DL is an error-free representation of its intended domain, and nor – until the day when the use of OWL or of some successor becomes uniform common practice – will it do anything to resolve the problems of semantic ambiguity adverted to in the above.

In the domain of biomedical informatics a number of attempts have been made to resolve these problems<sup>4,5,6</sup> in light of an increasing recognition that many ambitious terminological systems developed in this field are marked by unclarity over what, precisely, they have been designed to achieve. Are biomedical controlled vocabularies 'concept representations' or 'knowledge models'? And if they are either of these things, how, if at all, do they relate to the reality – the tumors, diseases, treatments, chemical interactions – on the side of the patient?

### OBJECTIVES AND METHODS

The purpose of this communication is to initiate a process for resolving such problems by drawing on the best practices in ontology which are now beginning to take root through the efforts of

organizations such as the National Center for Biomedical Ontology,<sup>7</sup> the Open Biomedical Ontologies (OBO) Consortium,<sup>8</sup> the OBO Foundry,<sup>9</sup> and others.<sup>10</sup>

What is needed is a set of terms referring in unambiguous fashion to the different kinds of entities surveyed above, which can serve as common target for mappings from other discipline- and computational idiom-centric terminologies, thereby mediating efficient pairwise translations between these terminologies themselves.

Our strategy is to advance precision via clear informal definitions rooted in what we assume are commonly accepted intuitions, providing references to associated formal treatments where possible. In selecting terms we have sometimes chosen expressions precisely because they have not been used by others and hence do not have established (and potentially conflicting) meanings. In other cases we have adapted existing terms to our purposes by providing them with more precise definitions or (in case of primitive terms) elucidations.

These proposals are focused primarily on the ontology-related needs of natural science, including the clinical basic sciences, though we believe them to be of quite general applicability.

We start out from a distinction of three levels of entities which have a role to play wherever ontologies are used:

- Level 1: the objects, processes, qualities, states, etc. in reality (for example on the side of the patient);
- Level 2: cognitive representations of this reality on the part of researchers and others;
- Level 3: concretizations of these cognitive representations in (for example textual or graphical) representational artifacts.

This tripartite distinction will awaken echoes of the Semantic Triangle of Ogden and Richards, to which we return in the sequel. For present purposes we note that the indispensability of Level 1 reflects the fact that even those who see themselves as building for example ‘data models’ in the domain of the life sciences are attempting to create thereby artifacts which stand in some representational relation to entities in the real world. Level 2 reflects the fact that a crucial role is played in ontology and terminology development by the cognitive representations of human subjects. Level 3 reflects the fact that cognitive representations can be shared, and serve scientific ends, only when they are made communicable in a form whereby they can also be subjected to criticism and correction, and also to implementation in software.

Note that the three levels overlap; thus the textual and graphical artifacts distinguished in Level 3 are themselves objects on Level 1. Our talk of ‘levels’

should thus be interpreted by analogy with talk of ‘levels of granularity’: if we have apprehended all the liquid in a vessel, then in a sense we have thereby apprehended also all the molecules. Yet for scientific purposes molecules and liquids must be distinguished nonetheless, and the same applies, for the purposes of clarity in our thinking about ontologies, to the three levels delineated in the above.

## FOUNDATIONS

Here we give precise definitions to a number of central terms, which will then be used in conformity thereto in the remainder of the paper. Really existing ontologies and related artifacts are typically constructed to realize a mixture of different sorts of ends (terminologies, for example, to support clinical record keeping and large-scale epidemiological studies, and to serve as controlled vocabularies for the expression of research results). Hence they typically combine the features of artifacts of different basic types. Our reference terminology is designed to reflect these basic types. Hence the definitions we propose for terms such as ‘ontology’ or ‘class’ do not imply any claim to the effect that everything called an ‘ontology’ or ‘class’ in the literature exhibits just the characteristics referred to in the definition..

An **ENTITY** is anything which exists, including objects, processes, qualities and states on all three levels (thus also including representations, models, beliefs, utterances, documents, observations, etc.)

A **REPRESENTATION** is for example an idea, image, record, or description which refers to (is *of* or *about*), or is intended to refer to, some entity or entities external to the representation. Note that a representation (e.g. a description such as ‘the cat over there on the mat’) can be of or about a given entity even though it leaves out many aspects of its target. A **COMPOSITE REPRESENTATION** is a representation built out of constituent sub-representations as their parts, in the way in which paragraphs are built out of sentences and sentences out of words. The smallest constituent sub-representations are called **REPRESENTATIONAL UNITS**; examples are: icons, names, simple word forms, or the sorts of alphanumeric identifiers we might find in patient records. Note that many images are not composite representations since they are not built out of smallest representational units in the way in which molecules are built out of atoms. (Pixels are not representational units in the sense defined.)

If we take the graph-theoretic concretization of the Gene Ontology<sup>11</sup> as our example, then the representational units here are the nodes of the graph (taken to comprehend terms and unique IDs), which are intended to refer to corresponding entities in reality. But the composite representation refers,

through its graph structure, also to the relations between these entities, so that there is reference to entities in reality both at the level of single units and at the structural level.<sup>12</sup>

A **COGNITIVE REPRESENTATION** (Level 2) is a representation whose representational units are ideas, thoughts, or beliefs in the mind of some cognitive subject – for example a clinician engaged in applying theoretical (and practical) knowledge to the task of establishing a diagnosis.

A **REPRESENTATIONAL ARTIFACT** (Level 3) is a representation that is fixed in some medium in such a way that it can serve to make the cognitive representations existing in the minds of separate subjects publicly accessible in some enduring fashion. Examples are: a text, a diagram, a map legend, a list, a clinical record, or a controlled vocabulary. Clearly such artifacts can serve to convey more or less adequately the underlying cognitive representations – and can be correspondingly more or less intuitive or understandable.

Because representational artifacts such as SNOMED CT give textual form to cognitive representations which pre-exist them, some have taken this to mean that these artifacts are in fact made up of representations which *refer to* (are *of* or *about*) these cognitive representations (the ‘concepts’) from out of which the latter are held to be composed.

We shall argue below that this reflects a deep confusion, and that the constituent units of representational artifacts developed for scientific purposes should more properly (and more straightforwardly) be seen as referring to the very same entities in reality – the diseases, patients, body parts, and so forth – to which the underlying cognitive representations of clinicians and others refer. Such artifacts are in this respect no different from scientific textbooks. They are windows on reality, designed to serve as a means by which representations of reality on the part of cognitive agents can be made available to other agents, both human and machine. A simple phrase, such as ‘the cat over there on the mat’, can be used to refer more or less successfully to what is, in reality, a portion of reality of a highly complex sort – and the same applies to all of the types of artifacts referred to above. The window on reality which each provides is, to be sure, in every case from a certain perspective and in such a way as to embody a certain granularity of focus. Yet the entities to which it refers are full-fledged entities in reality nonetheless – the very same, full-fledged entities in reality with which we are familiar also in other ways, for example because they provide us with food or companionship.

## REALITY

The clinician is concerned first and foremost with

**PARTICULARS** in reality (Level 1), (in the vernacular also called ‘tokens’ or ‘individuals’), that is to say with individual patients, their lesions, diseases, and bodily reactions, divided into **CONTINUANTS** and **OCCURRENTS**.<sup>13</sup> Some particulars, such as human beings, planets, ships, hurricanes, receive **PROPER NAMES** (they may also receive unique identifiers, such as social security numbers) which are used in representational artifacts of various sorts. But we can refer to particulars also by means of complex expressions – *that man on the bench*, *this oophorectomy*, *this blood sample* – involving **GENERAL TERMS** of different sorts, including:

i. General terms such as ‘*apoptosis*’, ‘*fracture*’, ‘*cat*’, which represent structures or characteristics in reality which are exemplified – the very same structures or characteristics; over and over again – in an open-ended collection of particulars in arbitrarily disconnected regions of space and time. Consider for example the way in which a certain DNA structure is instantiated as a transcript (RNA-structure) over and over again in cells of our body.

ii. General terms such as ‘danger’, ‘gift’, ‘surprise’, which draw together entities in reality which share common characteristics which are not intrinsic to the entities in question.

iii. General terms such as ‘Berliner’, ‘Paleolithic’, which relate to specific collections of particulars tied to specific regions of space and time.

General terms of the first sort refer to **UNIVERSALS** (in the vernacular also called ‘types’ or ‘kinds’). A universal is something that is shared in common by all those particulars which are its **INSTANCES**. The universal itself then exists in Level 1 reality as a result of existing in its particular instances. When a clinician says ‘*A* and *B* have the same disease’, she is referring to the universal; when she says ‘*A*’s diabetes is more advanced than *B*’s,’ then she is referring to the respective instances.

It is overwhelmingly universals which are the entities represented in scientific texts, and a good *prima facie* indication that a general term ‘*A*’ refers to a universal is that ‘*A*’ is used by scientists for purposes of classification and to make different sorts of law-like assertions about the individual instances of *A* with which they work in the lab or clinic.

<universal, universal>	<i>nose part_of body</i>
<particular, particular>	Mary’s nose <b>part_of</b> Mary
<particular, universal>	Mary’s nose <b>instance_of</b> <i>nose</i>

Table 1 – Three Basic Sorts of Binary Relation

Both particulars and universals stand to each other in various **RELATIONS**. Thus particulars stand to the corresponding universals in the relation of

**INSTANTIATION.** This and other binary relations (of parthood, adjacency, derivation) used in biomedical ontologies<sup>13</sup> can be divided into groups as in Table 1, which uses Roman for particulars, **bold type** for relations involving particulars, and *italics* for universals and for relations between universals.

A **COLLECTION OF PARTICULARS** (of molecules in John's body, of pieces of equipment in a certain operating theater, of operations performed in this theater over a given period of months) is a Level 1 particular comprehending other particulars as its **MEMBERS**.<sup>14</sup> We note that confusion is spawned by the fact that we can use the very same general terms to refer both to universals and to collections of particulars. Consider:

- *HIV* is an infectious retrovirus
- *HIV* is spreading very rapidly through Asia

A **CLASS** is a collection of all and only the particulars to which a given general term applies. Where the general term in question refers to a universal, then the corresponding class, called the **EXTENSION** of the universal (at a given time), comprehends all and only those particulars which as a matter of fact instantiate the corresponding universal (at that time).

The totality of classes is wider than the totality of extensions of universals since it includes also **DEFINED CLASSES**, designated by terms like 'employee of Swedish bank', 'daughter of Finnish spy'. Languages like OWL are ideally suited to the formal treatment of such classes, and the popularity of OWL has encouraged the view that it is classes which are designated by the general terms in terminologies. (OWL classes are not, however, identical with classes in the usual set-theoretic sense on which we draw also here.)

Some OWL classes (above all *Thing* and *Nothing*) are 'primitive' (which means: not defined), and these classes are sometimes asserted to constitute an OWL counterpart of universals ('natural kinds') in the sense here defined.<sup>15</sup> Because OWL identifies the relation of instantiation with that of membership, however, it in effect identifies universals with their extensions.

Through relations of greater and lesser generality both classes and universals are organized into trees, the former on the basis of the subclass relation, the latter on the basis of the *is\_a* relation (whereby, again, in the OWL framework the two relations are identified). Because the instances of more specific universals are *ipso facto* also instances of the corresponding more general universals, the latter hierarchy is, when viewed extensionally, a proper part of the former. As we shall discuss further in our treatment of *the argument from borderline cases* below, it is difficult to draw a sharp line between terms designating universals and those designating defined classes. This does not mean, however, that

the distinction is of no import. Indeed we believe that taking account of this distinction is indispensable to creating an path to improvement of ontologies.<sup>16</sup>

We use the term **PORION OF REALITY** to comprehend both single universals and particulars and their more or less complex combinations. Some portions of reality – for example single organisms, planets – reflect autonomous joints of reality (that is, they would exist as separate entities even in a world denuded of cognitive subjects). Other portions of reality are products of fiat demarcations of one or other sort,<sup>17</sup> as when we delineate a portion of reality by focusing on some specific granular level (of molecules, or molecular processes), or on some specific family of universals (for example when we view the human beings living in a given county in light of their patterns of alcohol consumption).

A **DOMAIN** is a portion of reality that forms the subject-matter of a single science or technology or mode of study; for example the domain of proteomics, of radiology, of viral infections in mouse. Representational artifacts will standardly represent entities in domains delineated by level of granularity. Thus entities smaller than a given threshold value may be excluded from a domain because they are not salient to the associated scientific or clinical purposes.<sup>18</sup>

## REPRESENTATIONAL ARTIFACTS

In developing theories, biomedical researchers seek representations of the universals existing in their respective domain of reality. They first develop cognitive representations, which they then transform incrementally into representational artifacts of various sorts.

In developing *diagnoses*, and in compiling such diagnoses into clinical records, clinicians seek a representation of salient particulars (diseases, disease processes, drug effects) on the side of their patients. Drawing on their theoretical understanding of the universals which these particulars instantiate (which in turn draws on prior representations formed in relation to earlier particulars<sup>19</sup>), they first develop a cognitive representation of what is taking place within a given collection of particulars in reality, which they then transform into representational artifacts such as clinical documents, entries in databases, and so forth, which may then foster more refined cognitive representations in the future.

The mentioned representations are typically built up out of sub-representations each of which, in the best case, mirrors a corresponding salient portion of reality. The most simple representations ('*blood!*') mirror universals or particulars taken singly; more complex representations – such as therapeutic schemas, diagnostic protocols, scientific texts, pathway diagrams – mirror more complex portions of

reality, their constituent sub-representations being joined together in ways designed to mirror salient relations on the side of reality.

In the ideal case a representation would be such that all portions of reality salient to the purposes for which it was constructed would have exactly one corresponding unit in the representation, and every unit in the representation would correspond to exactly one salient portion of reality.<sup>19</sup> Unfortunately, in a domain like biomedicine, ideal case will likely remain forever beyond our grasp. Researchers working on the level of universals may fall short by creating representations which either (i) fail to include general terms for universals which are salient to their domain, or (ii) include general terms which do not in fact denote any universals at all. Similarly, clinicians working on the level of particulars may fall short of the best case by creating misdiagnoses, either (i) by failing to acknowledge particulars which do exist and which are salient to the health of a given patient, or (ii) by using representational units assumed to refer to particulars where no such particulars exist.

A **TAXONOMY** is a tree-form graph-theoretic representational artifact with nodes representing universals or classes and edges representing *is\_a* or subset relations.

An **ONTOLOGY** is a representational artifact, comprising a taxonomy as proper part, whose representational units are intended to designate some combination of universals, defined classes, and certain relations between them.<sup>13</sup>

A **REALISM-BASED ONTOLOGY** is built out of terms which are intended to refer exclusively to universals, and corresponds to that part of the content of a scientific theory that is captured by its constituent general terms and their interrelations.

A **TERMINOLOGY** is a representational artifact consisting of representational units which are the general terms of some natural language used to refer to entities in some specific domain.

An **INVENTORY** is a representational artifact built out of singular referring terms such as proper names or alphanumeric identifiers. Electronic Health Records (EHRs) incorporate inventories in this sense, including both terms denoting particulars ('patient #347', 'lung #420') and more complex expressions involving terms designating universals and defined classes ('the history of cancer in patient #347's family').<sup>20</sup>

In the best case, again, each of the representational artifacts listed above (ontologies, taxonomies, inventories) will be such that its representational units stand in a one-to-one correspondence with the salient entities in its domain. In practice, however, such artifacts can be classified on the basis of the various ways in which they fall short of this best case, in terms of properties such as correctness, degree of

structural fit, degree of completeness and degree of redundancy.<sup>16,18</sup> By exploiting such classifications we can measure the quality improvements made in successive versions, and also use such measures as a basis for further improvement.<sup>20</sup>

To make a representation interpretable by a computer, it must be published in a language with a formal semantics and so converted into a **FORMALIZED REPRESENTATION**. The choice of language will depend on the complexity of what one needs to express and on the sorts of reasoning one needs to perform. While OWL, for example, can cope well with defined classes, it may not have sufficient expressive power to meet the needs of ontologies in the life sciences domain. Thus it seems to be incapable, for example, of capturing the relations involved even in simple interactions among pluralities of continuants, or of capturing the changes which take place in such continuants (for example growth of a tumor) over time.<sup>21,22</sup>

Most inventories in the biomedical field (including most EHRs) have still exploited hardly at all the powers of formal reasoning. The paradigm of Referent Tracking represents an exception to this rule,<sup>20</sup> since it involves precisely the embedding of a highly structured representation of particulars in a formalized representation of the corresponding universals.

## THE CONCEPT ORIENTATION

We believe that ontologies, inventories and similar artifacts should consist exclusively of representational units which are intended to designate entities in Level 1 reality. Defenders of the concept orientation in medical terminology development have offered a series of arguments against this view, to the effect that such terminologies should include also (or exclusively) representational units referring to what are called 'concepts'.<sup>23</sup>

First, is what we can call the *argument from intellectual modesty*, which asserts that it is up to domain experts, and not to terminology developers, to answer for the truth of whatever theories the terminology is intended to mirror. Since domain experts themselves disagree, a terminology should embrace no claims as to what the world is like, but reflect, rather, the coagulate formed out of the concepts used by different experts.

Against this, it can be pointed out that communities working on common domains in the medical as in other scientific fields in fact accept a massive and ever-growing body of consensus truths about the entities in these domains. Many of these truths are, admittedly, of a trivial sort (that mammals have hearts, that organisms are made of cells), but it is precisely such truths which form the core of science-

based ontologies. Where conflicts do arise in the course of scientific development, these are highly localized, and pertain to specific mechanisms, for example of drug action or disease development, which can serve as the targets of conflicting beliefs only because researchers share a huge body of presuppositions.

We can think of no scenario under which it would make sense to postulate special entities called ‘concepts’ as the entities to which terms subject to scientific dispute would refer. For either, for any such term, the dispute is resolved in its favor, and then it is the corresponding level 1 entity that has served as its referent all along; or it is established that the term in question is non-designating, and then this term is no longer a candidate for inclusion in a terminology. We cannot solve the problem that we do not know, at some given stage of scientific inquiry, to which of these groups a given term belongs, by providing such terms instead with guaranteed referents called ‘concepts’. It may, finally, be the case that it is not the disputed term itself which is at issue, but rather some more complex expression, as when we talk about ‘G. E. Stahl’s concept of phlogiston’, but that the latter refers to some entity – a concept – in (psychological) reality is precisely *not* subject to scientific dispute.

Sometimes the argument from intellectual modesty takes an extreme form, for example on the part of those for whom reality itself is seen as being somehow unknowable (‘we can only ever know our own concepts’). Arguments along these lines are of course familiar from the history of philosophy. Stove provides the definitive refutation.<sup>24</sup> Here we need note only that they run counter not just to the successes, but to the very existence, of science and technology as collaborative endeavors.

Second, is the *argument from creativity*. Designer drugs are conceived, modeled, and described long before they are successfully synthesized, and the plans of pharmaceutical companies may contain putative references to the corresponding chemical universals long before there are instances in reality. But again: such descriptions and plans can be perfectly well apprehended even within terminologies and ontologies conceived as relating exclusively to what is real. Descriptions and plans do, after all, exist. On the other hand it would be an error to include in a scientific ontology of drugs terms referring to pharmaceutical products which do not yet (and may never) exist, solely on the basis of plans and descriptions. Rather, such terms should be included precisely at the point where the corresponding instances do indeed exist in reality, exactly in accordance with our proposals above.

Third, is what we might call the *argument from unicorns*. Some of the terms needed in medical terminologies refer, it is held, to what does not exist.

Some patients do, after all, believe that they are James Bond, or that they see unicorns. The realist approach is however perfectly well able to comprehend also phenomena such as these, even though it is restricted to the representation of what is real. For the beliefs and hallucinatory episodes in question are of course as real as are the persons who suffer (or enjoy) them. And certainly such beliefs and episodes may involve concepts (in the properly psychological sense of this term). But they are not *about* concepts, they do not have concepts *as their targets* – for they are intended by their subjects to be about entities in flesh-and-blood external reality.

Fourth, is the *argument from medical history*. The history of medicine is a scientific pursuit; yet it involves use of terms such as ‘diabolic possession’ which, according to the best current science, do not refer to universals in reality. But again: the history of medicine has as its subject-domain precisely the beliefs, both true and false, of former generations (together with the practices, institutions, etc. associated therewith). Thus a term like ‘diabolic possession’ should be included in the ontology of this discipline in the first place as component part of terms designating corresponding classes of beliefs. In addition it may appear also as part of a term designating some fiat collection of those diseases from which the patients diagnosed as being possessed were in fact suffering. The evolution of our thinking about disease can then be understood in the same way that we deal with theory change in other parts of science, as a reordering of our beliefs about the ontological validity and salience of specific families of terms – and once again: concepts themselves play no role as referents.<sup>20,26</sup>

Fifth, is the *argument from syndromes*. The subject-matters of biology and medicine are, it is held, replete with entities which do not exist in reality but are rather convenient abstractions. A syndrome such as congestive heart failure, for example, is nothing more than a convenient abstraction, used for the convenience of physicians to collect together many disparate and unrelated diseases which have common final manifestations. Such abstractions are, it is held, mere concepts.

According to the considerations on fiat demarcations advanced above, however, syndromes, pathways, genetic networks and similar phenomena are indeed fully real – though their reality is that of defined (fiat) classes rather than of universals. A similar response can be given also in regard to the many human-dependent delineations used in expressions like ‘obesity’ or ‘hypertension’ or ‘abnormal curvature of spine’. These terms, too, refer to entities in reality, namely to defined classes which rest on fiat thresholds established by consensus among physicians.

Sixth is the *argument from error*. When erroneous entries are entered into a clinical record and interpreted as being about level 1 entities, then logical conflicts can arise. For Rector *et al.*, this implies that the use of a meta-language should be made compulsory for all statements in the EHR, which should be, not about entities in reality, but rather about what are called ‘findings’.<sup>25</sup> Instead of *p* and *not p*, the record would contain entries like: *McX observed p* and *O’W observed not p*, so that logical contradiction is avoided. The terms in terminologies devised to serve such EHRs would then one and all refer not to diseases themselves, but rather to mere ‘concepts’ of diseases. This, however, blurs the distinction between entities in reality and associated findings, and opens the door to the inclusion in a terminology of problematic findings-related expressions such as SNOMED’s ‘absent nipple’, ‘absent leg’, etc. Certainly clinicians need to record such findings. But then their findings are precisely that a leg is absent; not that a special kind of (‘absent’) leg is present.

In the domain of scientific research we do not embargo entirely the making of object-language assertions simply because there might be, among the totality of such assertions, some which are erroneous. Rather, we rely on the normal workings of science as a collective, empirical endeavor to weed out error over time, providing facilities to quarantine erroneous entries and resolve logical conflicts as they are identified. We have argued elsewhere that these same devices can be applied also in the medical context.<sup>26</sup>

The argument for the move to the meta-level is sometimes buttressed by appeal to medico-legal considerations seen as requiring that the EHR be a record not of what exists but of clinicians’ beliefs and actions. Yet the forensic purposes of an audit trail can equally well be served by an object-language record if we ensure that meta-data are associated with each entry identifying by whom the pertinent data were entered, at what time, and so forth.

On the other side, moreover, even the move to meta-level assertions would not in fact solve the problems of error, logical contradiction and legal liability. For the very same problems arise not only when human beings are describing, on the object-level, fractures, or pulse rates, or symptoms of coughing or swelling, but also on the meta-level when they are describing what clinicians have heard, seen, thought and done. The latter, too, are subject to error, fraud, and disagreement in interpretation.

Seventh is the *argument from borderline cases*. As we have already noted above, there is at any given stage no bright line between those general terms properly to be conceived as designating universals and those designating merely ‘concepts’ (or defined classes). Certainly there are, at any given stage in the development of science, clear cases on either side:

‘electron’ or ‘cell’, on the one hand, and ‘fall on stairs or ladders in water transport NOS, occupant of small unpowered boat injured’ (Read Codes) on the other. But there are also borderline cases such as ‘alcoholic non-smoker with diabetes’, or ‘age-dependent yeast cell size increase’, which call into question the very basis of the distinction.

In response, we note first the general point, that arguments from the existence of borderline cases in general have very little force. For otherwise they would allow us to prove from the existence of people with borderline complements of hair that there is no such thing as baldness or hairiness.

As to the specific problem of how to classify borderline expressions, this is a problem not for terminology, but rather for empirical science. For borderline terms of the sorts mentioned will, as an inevitable concomitant of scientific advance, be in any case subjected to a filtering process based on whether they are needed for purposes of (for example therapeutically) fruitful classifications, and thus for the expression of scientific laws.

Science itself is thereby subject to constant update. A term taken to refer to a universal by one generation of scientists may be demoted to the level of non-designating term (‘phlogiston’) by the next. This means also that representational artifacts of the sorts considered in the above, because they form an integral part of the practice of science, should themselves be subject to continual update in light of such advance. But again: we can think of no circumstance in which updating of the sort in question would signify that phlogiston is itself a concept, or that some expression was at one or other stage being used by scientists with the intention of referring to ‘concepts’ rather than to entities in reality.

## THE SEMIOTIC TRIANGLE

Finally is what we might call the *argument from multiple perspectives*. Different patients, clinicians and biologists have their own perspectives on one and the same reality. To do justice to these differences, it is argued, we must hold that their respective representations point, not to this common reality, but rather to their different ‘concepts’ thereof.

This argument has its roots in the work of Ogden and Richards, and specifically in their discussion of the so-called ‘semiotic triangle’, which is of importance not least because it embodies a view of meaning and reference that still plays a fateful role in the terminology standardization work of ISO.<sup>26</sup>

As Figure 1 makes clear, the triangle in fact refers not to ‘concepts’, but rather to what its authors call ‘thought or reference’,<sup>27</sup> reflecting the fact that Ogden and Richards’ account is rooted in a theory of psychological causality. When we experience a

certain object in association with a certain sign, then memory traces are laid down in our brains in virtue of which the mere appearance of the same sign in the future will, they hold, ‘evoke’ a ‘thought or reference’ directed towards this object through the reactivation of impressions stored in memory.

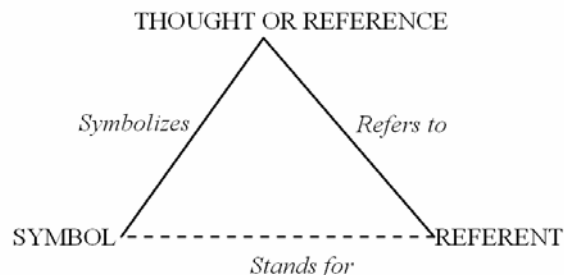


Figure 1 – Ogden and Richards' Semiotic Triangle

The two solid edges of the triangle are intended to represent what are held to be *causal* relations of ‘symbolization’ (roughly: evocation), and ‘reference’ (roughly: perception or memory) on the part of a symbol-using subject. The dashed edge, in contrast, signifies that the relation between term and referent – the relation that is most important for the discussion of terminology – is merely ‘imputed’.

The background assumption here is that multiple perspectives are both ubiquitous and (at best) only locally and transiently resolvable. The meanings words have for you or me depend on our past experiences of uses of these words in different kinds of contexts. Ambiguity must be resolved anew (and a new ‘imputed’ relation of reference spawned) on each successive occasion of use. From this, Ogden and Richards infer that a symbolic representation can never refer *directly* to an object, but rather only *indirectly*, via a ‘thought or reference’ within the mind.

It is a depsychologized version of this latter thesis which forms the basis of the concept orientation in contemporary terminology research. The terms in terminologies refer not to entities in reality, it is held, but rather to ‘concepts’ in a special ‘realm’. The latter are not *transparent mediators* of reference; rather they are its *targets*, and the job of the terminologist is to calibrate his list of terms in relation not to reality but to this special ‘realm of concepts’.<sup>26</sup>

The relation between terms in a terminology and the reality beyond becomes hereby obscured. Reality exists, if at all, only behind a conceptual veil – and hence familiar confusions according to which for example *the concept of bacteria would cause an experimental model of disease*, or *the concept of vitamin would be ‘essential in the diet of man’*.<sup>28</sup>

## ‘CONCEPTS’ AND ‘MODELS’

How, then, should ‘concept’ be properly treated in the

terminology literature henceforth? There are of course sensible uses of this term, for example in the literature of psychology. In the terminology literature, however, ‘concept’ has been used in such a bewildering variety of confused and confusing ways that we recommend that it be avoided altogether.

It is tempting to suppose that, when considered extensionally, all of the mentioned alternative readings come down to one and the same thing, namely to an identification of ‘concept’ with what we have earlier called ‘defined class’. If ‘concept’ could be used systematically in this way in terminological circles, then this would, indeed, constitute progress of sorts, though the question would then arise why ‘defined class’ itself should not be used instead. Unfortunately, however, the proposal in question stands in conflict with the fact that ‘concept’ is used by its adherents to comprehend also putative referents even for terms – such as ‘surgical procedure not carried out because of patient’s decision’ – which do not designate defined classes because they designate *nothing at all*. Here again, we believe, a proper treatment would involve appeal to appropriate fiat classes, defined in terms of utterances, interrupted plans, expectations, etc. on the part of the subjects involved.

What, now is to be said of terms such as ‘concept model’, ‘knowledge representation’, ‘information model’, and so forth referred to in our prelude above? To the extent that concept-based terminological artifacts consist in representations not of the reality on the side of the patient but rather of the entities in some putative ‘realm of concepts’, the term ‘concept model’ may be justified. This term is indeed used by SNOMED CT in its own self-descriptions, though given SNOMED’s scientific goals, we believe that, on the basis of the arguments given above, it should be abandoned. Still more problematic is the term ‘knowledge model’ or ‘knowledge representation’ (GALEN). For in the absence of a reference to reality to serve as benchmark, what could motivate a distinction between *knowledge* and mere *belief*?<sup>19</sup> And what, in the absence of a reference to reality, could motivate adding or deleting terms in successive versions of a terminology, if every term is in any case guaranteed a reference to its own specially tailored ‘concept’.

As to ‘information model’, here one standard uncertainty concerns the relation between an entity in reality and the body of information used to ‘represent’ this entity in some information system. Is it information which is being ‘modeled’ in an information model, or the reality which this information is about? The documentation of the HL7 Reference Information Model (RIM)<sup>29</sup> adds extra layers of uncertainty by conceiving its principal formulas as referring to the *acts* in which entities are observed for

example in a clinical context. Simultaneously, however, it conceives these formulas as referring also to the *documentation* of such acts for example in an information system. The apparent contradiction is to some degree resolved by the RIM on the basis of its assertion that there is in any case ‘no distinction between an activity and its documentation’.<sup>30</sup>

## CONCLUSION

Drawing on our distinction of the three levels of *reality*, *cognition* and *representational artifact* we have sought to formulate an unambiguous terminology for describing ontologies and related artifacts. The proposed terminology allows us to characterize more precisely the sorts of things which go wrong when the distinction between these levels is ignored, or when one or other level is denied, so that the approach may also help in improving such artifacts in the future.

## Acknowledgements

This work was supported by the Wolfgang Paul Program of the Humboldt Foundation, the Volkswagen Foundation, the European Union Semantic Mining Network, by BBSRC Grant BB/D524283/1, and by the NIH Roadmap Grant U54 HG004028. Thanks are due also to Jim Cimino, Chris Chute, Gunnar Klein, Alan Rector, Stefan Schulz, and Kent Spackman for fruitful discussions.

## References

(URLs last accessed July 1, 2006)

- Smith B. Beyond concepts, or: Ontology as reality representation, *Formal Ontology in Information Systems (FOIS 2004)*, p. 73-84.
- <http://www.w3.org/2003/glossary>.
- Patel-Schneider PF, Hayes P, Horrocks I. OWL Web Ontology Language. 2004. <http://www.w3.org/TR/owl-semantics>.
- Spackman KA, Reynoso G. Examining SNOMED from the perspective of formal ontological principles. *Workshop on Formal Biomedical Knowledge Representation (KR-MED 2004)*, p. 72-80.
- Johansson I. Bioinformatics and biological reality. *J Biomed Inform.* 2006;39(3):274-87.
- Klein GO, Smith B. Concept systems and ontologies. <http://ontology.buffalo.edu/concepts/ConceptsandOntologies.pdf>.
- <http://ncbo.us/>.
- <http://obo.sourceforge.net/>.
- <http://obofoundry.org/>.
- Rosse C, Mejino JL, Jr. A reference ontology for biomedical informatics: the Foundational Model of Anatomy. *J Biomed Inform* 2003;36:478-500.
- <http://geneontology.org/>.
- Wittgenstein L. 1921 *Tractatus Logico-Philosophicus*, London: Routledge, 1961.
- Smith B, Ceusters W, Klagges B et al.. Relations in biomedical ontologies. *Genome Biol*, 2005;6(5):R46.
- Bittner T, Donnelly M, Smith B. Individuals, universals, collections. *Formal Ontology in Information Systems (FOIS 2004)*, p. 37-48.
- Drummond N. Introduction to ontologies. <http://www.cs.man.ac.uk/~drummond/presentations/IntroductionToOWL50mins.ppt>.
- Ceusters W, Smith B. A realism-based approach to the versioning and evolution of biomedical ontologies. *Proc AMIA Symp 2006*, in press.
- Smith B. Fiat objects. *Topoi*, 2001;20(2):131-48.
- Bittner T, Smith B. A theory of granular partitions. *Foundations of Geographic Information Science*, London, 2003, p. 117-51
- Smith B. From concepts to clinical reality, *J Biomed Inform.* 2006 Jun;39(3):288-98.
- Ceusters W, Smith B. Strategies for referent tracking in Electronic Health Records. *J Biomed Inform.* 2006 Jun;39(3):362-78.
- Bera P, Wand Y. Analyzing OWL using a philosophy-based ontology. *Formal Ontology in Information Systems (FOIS 2004)*, p. 353-62.
- Kazic T. Putting semantics into the semantic web: How well can it capture biology? *Pac Symp Biocomputing 2006*;11:140-51.
- Cimino JJ. In defense of the desiderata. *J Biomed Inform.* 2006;39:299-306.
- Franklin J. Stove's discovery of the worst argument in the world. *Philosophy* 2002;77:615-24. [www.maths.unsw.edu.au/~jim/worst.pdf](http://www.maths.unsw.edu.au/~jim/worst.pdf).
- Rector A, Nolan W, Kay S. Foundations for an electronic medical record. *Methods Inf Med*, 1991;30:179-86.
- Smith B, Ceusters W, Temmerman R. Wüsteria, *Stud Health Technol Inform.* 2005;116:647-652.
- Ogden CK, Richards IA. *The Meaning of Meaning*. 3rd ed. New York, 1930.
- The UMLS Semantic Network. <http://semantic.network.nlm.nih.gov/>.
- HL7 V3 Reference Information Model: Version V 01-20. Normative Ballot 11/22/2005.
- Smith B, Ceusters W. HL7 RIM: An incoherent standard, *Proc MIE*, 2006, p. 133-138



## LinKBase®, a Philosophically-inspired Ontology for NLP/NLU Applications

Maria van Gurp, PhD, Manuel Decoene, MD, Marnix Holvoet and Mariana Casella dos Santos, MD.

Language and Computing NV, Sint-Denijs-Westrem, Belgium

Tel: +32-(0)9-2808400; <http://www.landc.be>; {marjan, mariana}@landc.be

### ABSTRACT

*LinKBase® is a biomedical ontology. Its hierarchical structure, coverage, use of operational, formal and linguistic relationships, combined with its underlying language technology, make it an excellent ontology to support Natural Language Processing and Understanding (NLP/NLU) and data integration applications. In this paper we will describe the structure and coverage of LinKBase®. In addition, we will discuss the editing of LinKBase® and how domain experts are guided by specific editing rules to ensure modeling quality and consistency. Finally, we compare the structure of LinKBase® to the structure of third party terminologies and ontologies and discuss the integration of these data sources into LinKBase®.*

### INTRODUCTION TO LINKBASE®

To achieve full benefit of health information technology, a health information network, enabling interoperability across different facilities and countries, is essential. However, different and diverse medical information systems hamper the process of data sharing. One solution to this problem is to use a central ontology, with a strict hierarchical structure and a consistent semantic network of relationships between its types that can support NLP/NLU and data integration applications and that can serve as the link between the different medical information sources and systems. LinKBase® is such an ontology. LinKBase® has been designed with the main goal of integrating terminologies and databases with applications designed for NLP and information management and retrieval and has been built up from the ground over the past 10 years. It covers various aspects of medicine, including procedures, anatomy, pharmaceuticals and various disorders and anomalies delivering over 9 million knowledge elements making it the largest biomedical knowledge base in the world. The core ontological elements, being its types and relationships, have no embedded grammatical information and are as such language independent, but they are cross-referenced to terms and lexemes in various languages. Several features make LinKBase® the preferred ontology to eliminate some of the barriers to creating health information organizations; 1) LinKBase® is a language and

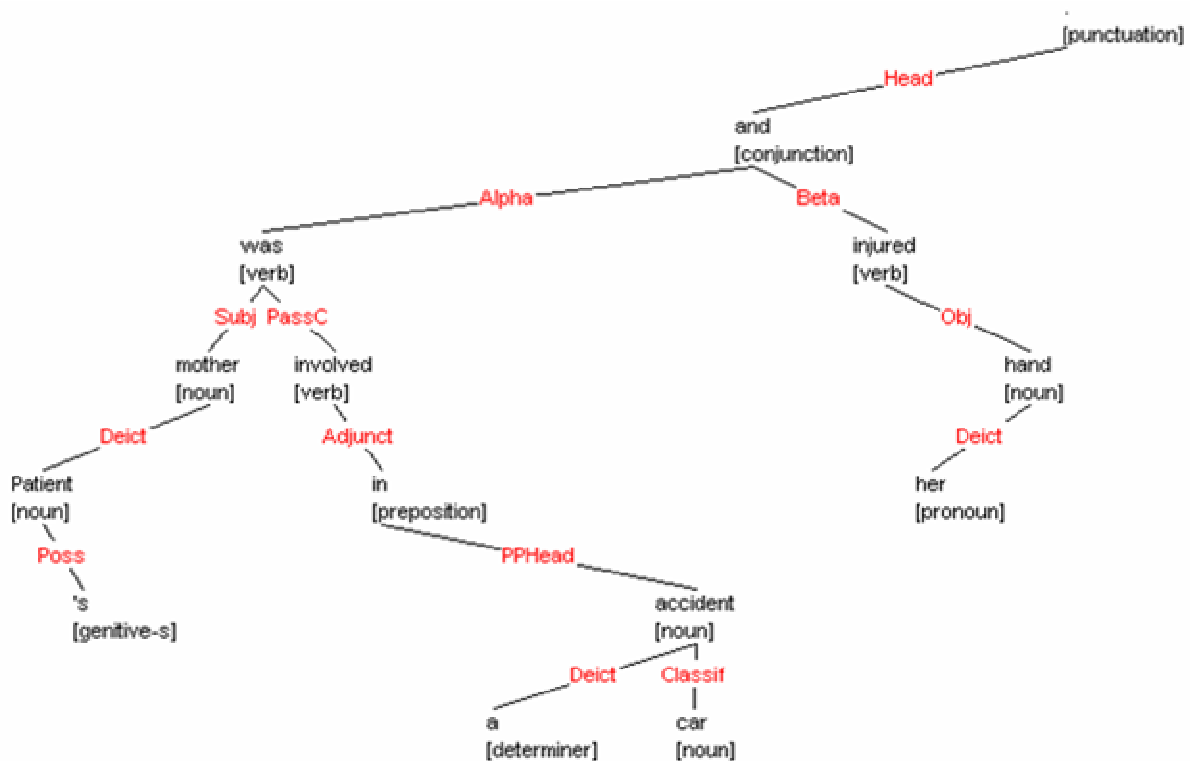
application independent ontology 2) LinKBase® is integrated to and under the guidance of a formal upper level framework Basic Formal Ontology (BFO)<sup>1</sup>, 3) LinKBase® embedded the linguistic ontology framework Generalized Upper Model (GUM)<sup>2</sup>, 4) the types within LinKBase® are interconnected by a rich set of hierarchical relationship types, 5) LinKBase® unambiguity is supported by full definitions and 6) the LinKBase® ontology is connected to a lexicon of terms in various languages.

Inherent to the interoperability of medical information systems, is the integration of the medical data within those systems. This task turns out to be a complex endeavor, not least because the different terminologies or databases that are to be integrated are often internally and mutually inconsistent. In this respect, LinKBase® can serve as a 'translation hub' between diverse third party terminologies, based on the fact that all these terminologies essentially speak about the same reality. This makes it possible to integrate them on the basis of a sound understanding of those basic categorical distinctions that are common to them all.

### STRUCTURE OF LINKBASE®

To achieve a coherent framework, able to support reasoning applications, NLP and NLU, the LinKBase® ontology is founded on philosophical and linguistic theories.

BFO<sup>1</sup>, a philosophically inspired upper-level ontology that focuses on the entities in reality at different levels of granularity and not on the human conceptualization of this reality, was used to structure the upper level of LinKBase®. Theories of endurants and perdurants<sup>3</sup>, mereology, topology, universals and particulars, biological classes and instantiations<sup>4</sup>, space and time and granular partitions<sup>5</sup> are all included in the BFO theory. The main distinction in BFO is between the endurants (SNAP) and perdurants (SPAN). Endurants are those entities that endure through time, in contrast to perdurants, which unfold themselves through time and are never fully present at a given moment in time<sup>3</sup>. The LinKBase® hierarchy is integrated under and branches from the BFO upper level entities, representing general



**Figure 1 – Analysis of syntactical structure**

Syntactic analysis of the sentence “The patient’s mother was involved in a car accident and injured her hand”.

categories such as processes, properties and objects. By using the BFO theory<sup>1</sup>, LinKBase® is not only provided with a rigorous philosophical classification of all its entities, but is provided with the set of axioms that govern BFO entities and the relationships among them as well. These axioms are used to apply modeling restrictions and guidance to prevent erroneous editing and to maintain and improve the structure of LinKBase®. More important however, the BFO definitions of ontological entities can be used by reasoning applications, including applications designed for NLP, and aid to the filtering out of erroneous synonyms and the disambiguation of ontological structures that are inherent to the processing of free text<sup>6</sup>. To support correct and precise linguistic reasoning, the LinKBase® hierarchical structure is very strict and every child type is a subclass of its parent’s class. Thus, the application of BFO-driven philosophical knowledge and axioms offers several advantages that are not present in application ontologies lacking a philosophical backbone.

The structure of the LinKBase® mid-layer is partially structured according to the GUM<sup>2</sup>. The GUM is a

general task and domain independent linguistically motivated ontology intended for organizing information for expression in natural language. In LinKBase®, the “processes” are organized based on their linguistic properties. This allows us, by using a GUM-based grammatical analysis, to convert the syntactic structure of a given sentence into an ‘understandable’ structure of types and criteria. For example, we can determine that the actee (or object) and the actor (or subject) are identical in the sentences “The patient was treated by the doctor” and “The doctor treated the patient.” In the sentence “The patient’s mother was involved in a car accident and injured her hand”, we deduce that “injured her hand” refers to the mother and is not referring to the patient (figure 1). In addition, we use the semantics to disambiguate the syntax by relating specific processes to specific actors and actees, e.g. a “treatment process” is related to the actee “patient” and the actor “healthcare professional”. Using this strategy, LinKBase® has the capacity to support NLU applications.

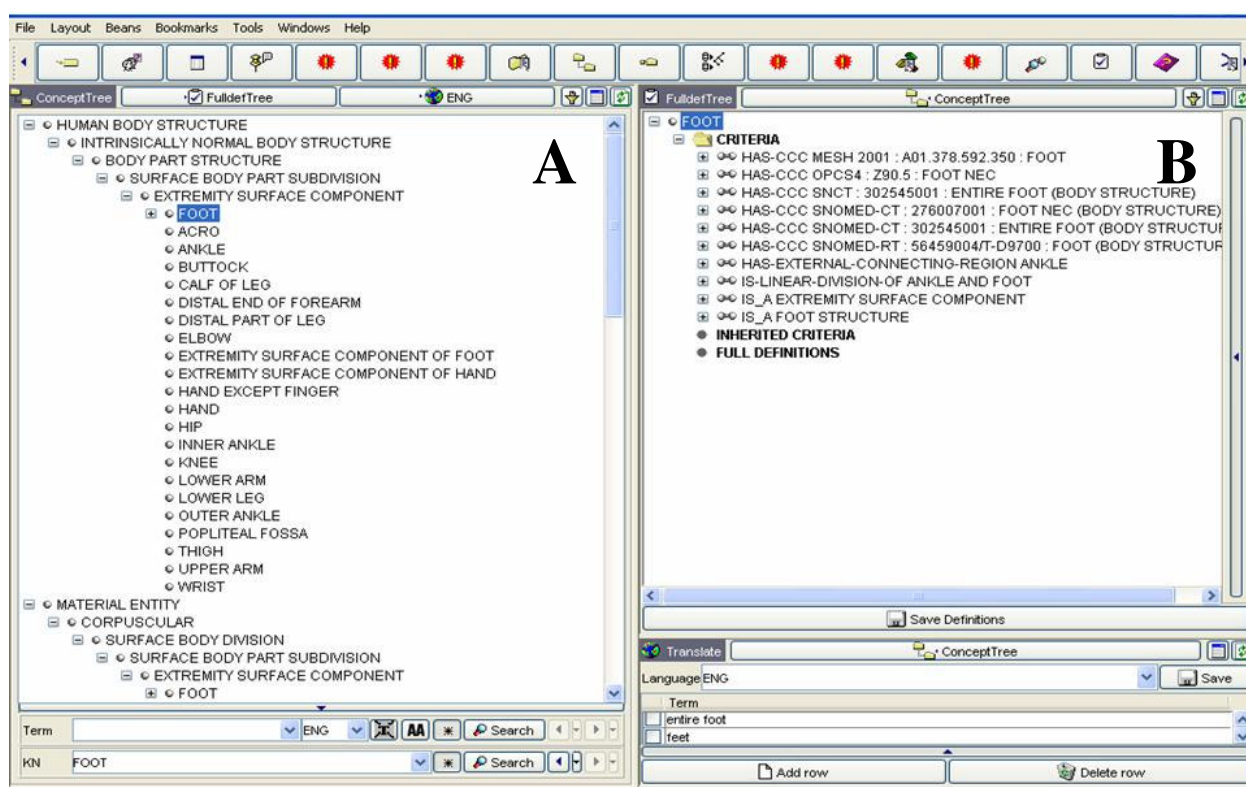
## TYPES

The more than 570,000 LinKBase® types represent real-world entities and not concepts in the mind of conscious beings that are abstractions of what these beings think the real-world entities are. To enable semantic reasoning, the types are hierarchically structured using a realist approach: child types

represent subclasses of a given parent for 100 % of the instances (figure 2A). Using this approach, the hierarchical relationships among LinKBase® types have a consistent meaning. In LinKBase®, for example, “rash” will never be a subclass of “allergic reaction” since it is not always allergic. However, in many classification systems that lack a strict hierarchical structure, such as ICD-9<sup>7</sup> or MEDCIN<sup>8</sup>, these situations do occur, hampering the use of algorithms in reasoning. Conflicts arise when analyzing the sentence “the patient was diagnosed with meningitis that was not due to infection” using an ontology in which “meningitis” is modeled as “a-

“meningitis”, namely “infective meningitis”, forms a solution to this problem. “Infective meningitis is-a meningitis” and “a-consequence-of infection”. However, “aseptic meningitis”, the illness of the above mentioned patient, is ‘only’ “meningitis” and does not have a relationship, direct or inherited, to “infection”. Thus, the principle of 100 % criteria allows LinKBase® to support NLU applications where other ontologies fail.

LinKBase® is a “living” ontology and types and subsequent relationships are added and edited on a daily basis by the modeling team. Although it is not required for types to be perfectly modeled from the



**Figure 2 - LinKBase® structure**

Screenshot of LinKBase® structure showing:

A) hierarchical structure; all child types are a representation of their parent(s), B) several types of relationships to 3rd party terminologies (the HAS-CCC relationships) as well as to other LinKBase® types and C) the terms that are assigned to, in this example, the type FOOT.

consequence-of infection”. Following the realist approach, the creation of an additional subclass of

beginning, the creation of new types and subsequent relationships is strictly regulated and new types can only be added if specific criteria are met<sup>9</sup>.

## RELATIONSHIPS

The types in LinKBase® are linked into a semantic network by a rich set of relationship types (figure 2B). Most relationships are based on theories, including BFO<sup>1</sup>, that deal with topics such as mereology and topology<sup>10,11</sup>, time and causality<sup>12</sup> and

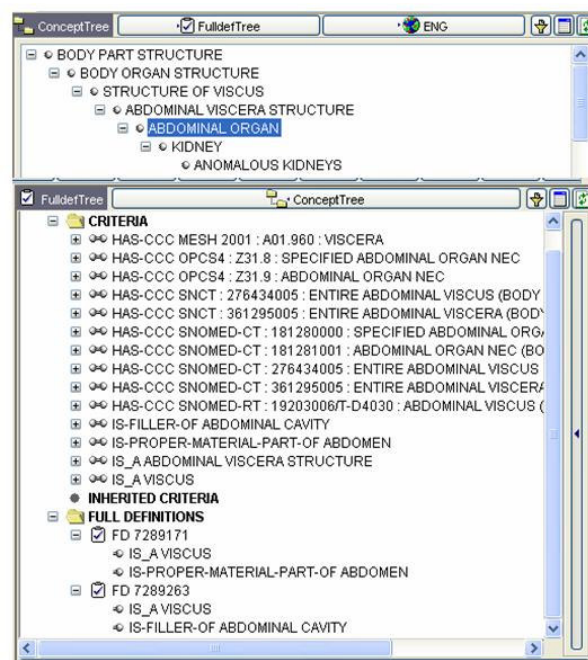
models for semantics driven natural language understanding<sup>13,14</sup>. In addition, LinKBase® contains relationships that fall out of any theory but are essential to express important notions in the medical world. One example is “absence of entity”, considered a lack of entity and not a real entity in most theories, but needed to represent types such as “anuria”, “absence of blood” or “noninvasive”. Since it is not possible to consider absences as processes<sup>6</sup>, absences are represented as a relation between the “absent entity” and the “entity from which the related entity is absent”. This avoids the creation of “absent processes” and keeps the distance between the related types to a minimum, which is relevant to many LinKBase applications<sup>15</sup>.

The LinKBase® relationship types are structured in a multi-parented hierarchy, taken into account both the formal realistic ontological implications and the linguistic aspects of the relationships. LinKBase® contains 383 different relationship types, covering many, often subtle, semantic differences; including spatial, temporal and process-related relationship types. New relationship types are added when the existing relationships are not capable to represent the semantics of new types or when new insights justify the creation of new relationship types that might provide a better quality assurance or are needed for certain applications. Within LinKBase®, we are currently revising the framework of “function” and “dysfunction”. New relationship types are needed to relate, for example, “function”, the function that the body part is supposed to perform, with “functioning process”, the body process that it is really performing at a given point in time. For this purpose, the relationship type “has-realisation” was created, going from “function” to the “functioning process”. The reverse relationship type is “is-realisation”.

Within LinKBase®, formal or full definitions are created by those criteria, whether direct or inherited, that are necessary and sufficient to uniquely define the type (figure 3).

The formal logic used by LinKBase® is an important prerequisite for an ontology with the ability to support reasoning applications<sup>16</sup>, since the system automatically infers that, if a real-world entity satisfies the full definition of a given domain-entity, it is an instance of that domain-entity.

Only around 15 % of the total number of relationships within LinKBase® is covered by formal subsumption relationships. As a consequence, the structure of LinKBase® is much richer compared to



**Figure 3 - Formal or full definitions in LinKBase®**  
Within LinKBase®, formal or full definitions are created by those criteria, which are necessary and sufficient to uniquely define the type. In this example, two full definitions are defined for the type ABDOMINAL ORGAN.

other ontologies and terminology systems, in which type-relationships are often expressed as “narrower” or “broader”, as is the case for the Unified Medical Language System (UMLS)<sup>16</sup>.

## TERMS

The LinKBase® ontology is connected to a lexicon of approximately 1.5 million terms. Terms are signs or symbols that are used to represent types in the real world. Terms can be synonyms, symbols, translations or, for example, singular or plural forms of the type name (figure 2C). In LinKBase®, the assignment of terms depends on the meaning of the types. Terms can only be assigned to types when they express exactly the same meaning in natural language. Bad synonym assignment often occurs because conditions are tightly connected in a medical cause-effect or symptom-disorder relation, as is the case for the SNOMED<sup>19, 20</sup> type “viral gastroenteritis (disorder)” that is linked to the terms “viral diarrhea”, “viral vomiting” and “viral gastroenteritis”. Although this

example of SNOMED term assignment might be correct from a medical point of view and is suited for terminology standardization/coding applications it is not compatible with NLP/NLU applications and is thus avoided in LinKBase®.

Next to types, criteria and relationships can receive terms as well; the criterion “has-happening-earlier-than systole” has the term presystolic and the relationship “is-part-of” has the German term “ist-ein-Teil”. Unlike types and relationships, terms can be stored in different languages. Thus, although LinKBase® itself is language independent, the assignment of multi-lingual terms and lexemes to its ontological elements allow the analysis of text in any European language.

### **EDITING/MODELING PROCESSES**

An accurate and consistent modeling is not always obvious when dealing with a large and complex ontology as LinKBase®. To overcome this problem and to guide and assist the modelers, several mechanisms have been developed. These tools include management issues, such as hierarchical user privileges and log file reviews, and modeling guidance in which the BFO theory<sup>1</sup> is used as automatic error detection. Both the BFO subsumption and disjoint axioms were implemented in LinKBase®. Of the BFO relationship axioms, only the domain-range restrictions were used. The axioms on the level of inference are not applied, but future work involves the application of these and other BFO axioms, to allow for further levels of inference. In addition, the BFO framework and the BFO partition theory are used as guidelines for the modelers to follow.

#### **hierarchical user privileges**

Hierarchical user privileges is a mechanism that assigns types to the modeler that created them. The users are organized in a hierarchical structure according to their skills and experience. Elements can only be modified by the ontologist who created the item or by a user at a higher level in the hierarchy. In this way, erroneous modeling of an already correctly modeled type is prevented as well as repetitive modeling of a certain type by different modelers.

#### **log files**

Every action performed by a modeler is stored in a log file. In the case of erroneous modeling, one can go back to the log files and check what went wrong, in order to be able to correct their mistake(s). In addition, the log files can be used for training purposes, in which the work of an ontologist is reviewed by an experienced ontologist and the performed actions are discussed.

#### **relationship type domain-range restrictions**

One method enforced by LinkFactory®<sup>21</sup>, the ontology management system used to edit, store and maintain LinKBase®, in order to limit the amount of modeling errors is domain-range restriction. A domain-range restriction on a relationship type limits the amount of types to which the relationship can refer, since that specific relationship type can only relate types that are located within its domain. These domain-range attributes have values corresponding to the SNAP and SPAN entities of BFO<sup>3</sup> between which they apply. In addition, the embedded GUM theory<sup>2</sup> and the linguistically structured processes allow the further refinement of domain-range restrictions to the mid-layer and linguistic layer of LinKBase® as well. For example, the relationship type “has-theme” holds between an endurant and a motion process and the theme is the entity that is displaced in the motion process (e.g. “excision of kidney has-theme kidney”). The source of the relationship type “has-actee”, an actee is someone or something that passively undergoes, is changed by, or is directly affected by a predicate, is always a-kind-of the linguistic process “directed action” (e.g. “treatment of acne has-actee acne”). Since both the relationship types and the types within LinKBase® are hierarchically structured, the relationship type domain-range restriction applies to the subtypes of the relationship type and type(s) in question as well. The relationship type “has-theme”, is a further refinement of the “has-participant” relationship type, valid between processes and endurants, of which it is a subclass. If a modeler tries to link a type to another type that is not within the domain of the specific relationship type used, the modeler receives a warning that a restriction is violated and has to revise his modeling.

#### **disjoint restrictions**

Another method enforced by LinkFactory®<sup>21</sup> to avoid modeling errors and to enhance the quality of LinKBase® is disjoint restriction. When two types are made disjoint, this implies that no type can be a subclass of both disjoint types. These checks are performed in real-time and the modeler receives a disjoint violation warning whenever he wants to make a type a subclass of both disjoint types. In addition, when (re)structuring the ontology, disjoint violations support the creation of a valid model of reality. Examples of disjoints in LinKBase® are the endurants (SNAP) and perdurants (SPAN) and the categories Corpuscular (e.g. organisms and organs) and Non-Corpuscular (e.g. tissues and liquids).

## THE META- AND DOMAIN-MAPPING FRAMEWORK

In LinKBase®, the domain-entity is defined as the set of types and their relationships that always have a consistent meaning. Outside this domain, in an area called the meta-entity, the 3<sup>rd</sup> party terminologies are located, standard classification systems such as ICD9<sup>7</sup> and SNOMED<sup>19, 20</sup>. The external ontologies are stored in their exact original style and structure and are linked to the LinKBase® domain-entity by specific formal relationship types<sup>22</sup>. This framework of a central domain-ontology linked to external (medical) information sources is called the “meta- and domain-mapping framework”. Table 1 contains an overview of some of the most important 3<sup>rd</sup> party terminologies that are linked to LinKBase®.

**Table 1** - Absolute number of meta-entity type names appended and subsequently processed within LinkBase®.

Meta-entity versus LinkBase®	METN <sup>a</sup> appended	LBKN <sup>b</sup> in support to METN	LBKN defined	English terms for LBKN	Non-English terms for LBKN
ICD-9-CM	27659	50105	20406	138186	223110
CPT-4	10202	8366	404	11121	2806
ICPC-2	746	3491	2297	14978	28823
MEDCIN	208981	208400	5450	243218	61432
MEDDRA 6.0	60616	51572	10401	111042	110221
MESH 2001	35780	19984	2997	54960	71349
SNCT	275222	322360	25121	539031	308786

<sup>a</sup> Meta-entity type name

<sup>b</sup> LinkBase® knowledge name

The “meta- and domain-mapping framework” has several advantages compared to a direct integration of external ontologies, such as the reusability of existing mappings, the ability to cross map several data sources and the ability to transpose divergent levels of granularity between external information sources. However, it also requires a careful mapping procedure to the central domain ontology LinKBase®, since the different information sources often have internally and mutually inconsistent structures<sup>22</sup>. Through the implementation of the “meta- and domain-mapping framework” LinKBase® becomes the ontology of choice to serve as a “translation-hub” between diverse 3<sup>rd</sup> party terminologies. Indeed, other ontologies that integrate several different 3<sup>rd</sup> party terminologies do exist, such

as the UMLS<sup>16</sup>. Why then, do we claim that LinKBase® is the preferred ontology for data integration? Is the UMLS®<sup>16</sup>, for this application, not a useful source? A comparison between LinKBase® and the UMLS®<sup>16</sup> will shed a light on the differences in structure and potential applications.

## LINKBASE® VERSUS THE UMLS®

Within the Metathesaurus of the UMLS®<sup>16</sup>, a large number of different source vocabularies and classification systems, e.g. ICD9<sup>7</sup>, Meddra<sup>23</sup> and SNOMED<sup>19, 20</sup>, are integrated with the purpose to facilitate the development of NLP/NLU computer systems and to overcome disparities in language, granularity and perspective. When integrating different vocabularies, it is important to respect the original structure and granularity of the source vocabularies. If not, circular hierarchical relationships might occur, as has been described in Bodenreider<sup>24</sup>. For example, in the UMLS® Metathesaurus, “maduromycosis” is related to “mycetoma of foot” in one vocabulary and to “eumycotic mycetoma” in another one. In LinKBase®, however, “eumycotic mycetoma” (mycetoma caused by fungi) and “mycetoma of foot” are child types of “mycetoma” (synonym of maduromycosis). The types are modeled according to their meaning and linked to their respective information sources, thus keeping a consistent and realistic view of the world (see figure 4).

A second distinction between the UMLS®<sup>16</sup> and LinKBase® are the relationship types and more specific the hierarchy within. Whereas LinKBase® follows a realist approach resulting in relationship types with a consistent meaning and child types that represent subclasses of a given parent for 100 % of the instances, this is not the case for the UMLS®. The hierarchical relationship types of the UMLS® can be both parent-child relationship types, comparable to the ones used in LinKBase®, or broader/narrower-than relationship types. An example of the latter is “toe is-a foot”. Although a toe is part of the foot, it certainly is not a kind-of foot and hence should not be placed as a subclass of “foot”.

Within LinKBase®, this problem is solved by creation of the type “foot structure” with the subclasses “foot”, referring to the extremity foot, and “foot part”. “Foot part”, in turn, contains the subclasses “toe part” and “toe”, which refers to the digit toe<sup>25</sup> (figure 5). This consistent class-subclass hierarchy of LinKBase® is a huge asset compared to the UMLS® hierarchy when considering NLP/NLU applications, since it avoids misclassification and allows clear and correct crossmapping.

**Concept:** Maduromycosis

**CUI:** C0024449

**Semantic Type:** Disease or Syndrome

**Definition:**

A disease caused by various fungi (Madurella swollen after infection. (MeSH)

**Synonyms:**

Maduromycosis

[X]Mycetoma, unspecified

[X]Mycetoma, unspecified (disorder)

E-430 MYCOTIC MYCETOMA

Eumycetoma

Eumycotic madura foot

Eumycotic mycetoma

Eumycotic mycetoma (disorder)

Eumycotic mycetoma of foot

Eumycotic mycetoma of foot (disorder)

Fungal mycetoma

Madura Foot

Maduramycosis

Maduromycosis NOS (disorder)

Mycetoma

Mycetoma, unspecified

mycetoma, infection, mycotic

mycetoma, madurae

mycetoma, madurae, mycotic

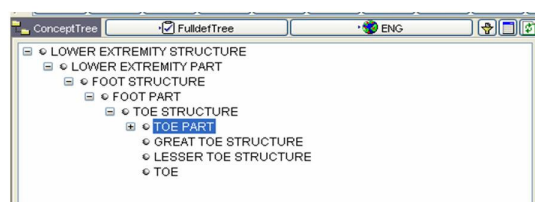
Mycetoma (disorder)

Mycetoma of foot (disorder)

Mycetoma, [unspecified] or [madura foot]



**Figure – 4 Comparison between LinkBase® (right panel) and the UMLS (left panel, see text for details)**



**Figure 5 - LinkBase® class-subclass structure**

When comparing LinkBase® to the UMLS®, we can conclude that LinkBase® is more suited for NLP/NLU applications. Conflicting relationships and the lack of a consistent hierarchy makes the mapping of free text to UMLS® a highly error-prone task. An example of a LinkBase®-based NLP/NLU

application is the development of an information extraction application for extraction of findings and procedures and their related context information, encoded into SNOMED according to the SNOMED Context Model guidelines<sup>26</sup>. Another example of a LinkBase®-based NLP/NLU application is the extraction of patient-related suicide- and self-harm behavior from medical reports that were generated during clinical trials. This aim of this project was to enhance data retrieval and to decrease manual review. In a first pilot study, based on 153 documents, the accuracy was more than 99 % (based on precision and recall against manual annotations).

## CONCLUSION

The novelty of LinkBase® compared to other terminologies is the LinkBase® “meta- and domain-mapping framework”. This framework of 3<sup>rd</sup> party terminologies, linked to the LinkBase® domain-entity, makes exchange, management and integration of data possible. The application-independency of LinkBase®, its strong framework based on established ontological theories, combined with a rich set of hierarchical relationship types, without any doubt, creates a flexible yet powerful ontology.

## References

1. B. Smith, Basic formal Ontology (BFO)  
<http://ontology.buffalo.edu/bfo/>
2. J.A. Bateman, R. Henschel, and F. Rinaldi. The generalized upper model 2.0. Technical report, GMD/Institut für Integrierte Publikations- und Informationssysteme, Darmstadt, Germany, 1995.  
<http://www.gmd.darmstadt.de/publish/komet/gen-um/newUM.html>
3. P. Grenon, B. Smith, SNAP and SPAN: Prolegomena to Geodynamic Ontology, Hornsby and Worboys Spatial Cognition and Computation, forthcoming,  
[http://www.ontology.buffalo.edu/smith/courses03/tb/SNAP\\_SPAN.pdf](http://www.ontology.buffalo.edu/smith/courses03/tb/SNAP_SPAN.pdf)
4. B. Smith, Westerstahl (ed), Invited Papers from the 10th Int Conf in Logic Methodology and the Philosophy of Science Oviedo, [http://ontology.buffalo.edu/bio/logic\\_of\\_classes.pdf](http://ontology.buffalo.edu/bio/logic_of_classes.pdf), 2003
5. Smith B. The Logic of Biological Classification and the Foundations of Biomedical Ontology. Westerstahl (ed), Invited Papers from the 10th International Conference in Logic Methodology and the Philosophy of Science, Oviedo, Spain, 2003.
6. M. Casella dos Santos, C., Dhaen, and J.M. Fielding, Philosophical scrutiny for Run-Time support of Application Ontologies Development, International Conference on Formal Ontology in Information Systems (FOIS), Torino, Italy, 2001.
7. International Classification of Diseases, Ninth Revision, with Clinical Modifications, Washington DC: U.S. National Center for Health Statistics, 1980
8. MEDCIN, <http://www.medicomp.com/>.
9. A. Flett, M. Casella dos Santos, W. Ceusters, Some Ontology Engineering processes and their Supporting technologies. Gomez-Perez, A, Benjamins VR (eds), Springer, 2002: 154-165.
10. B. Smith, A. C. Varzi., Fiat and Bona Fide Boundaries, Proc COSIT-97 Springer-Verlag, 1997: 103-119.
11. B. Smith, Data and Knowledge Engineering 20, <http://ontology.buffalo.edu/smith/articles/meretootology.htm>, 1996.
12. F. Buekens, W. Ceusters, G. De Moor, The Explanatory Role of Events in Causal and Temporal reasoning in Medicine, Met Inform Med 32, 1993: 274-278.
13. W. Ceusters, F. Buekens, T. Deray, A. Waagmeester, The distinction between linguistic and conceptual semantics in medical terminology and its implications for NLP-based knowledge acquisition, Met Inform Med 37, 1998: 327-333.
14. J. Bateman, Ontology construction and natural language. Proc International Workshop on Formal Ontology. Padua(Italy), 1993: 83-93.
15. W. Ceusters, B. Smith and J. Fielding. LinkSuite™: formally robust ontology-based data and information integration,. In DILS (Database Integration in the Life Sciences), Berlin: Springer, 2004.
16. B.L. Humphreys and D.A. Lindberg, The UMLS project: making the conceptual connection between users and the information they need, Bull Med Libr Assoc. 1993; 81(2): 170-177,
17. T.C. Rindflesch, B. Libbus, D. Hristovski, A.R. Aronson, and H. Kilicoglu, Semantic relations asserting the etiology of genetic diseases, AMIA Symposium Proceedings, 2003: 554-558.
18. G. Schadow and C.J. McDonald, Extracting structured information from free text pathology reports, AMIA Symposium Proceedings, 2003: 584-588.
19. R.A. Côté and S. Robboy, Progress in medical information management: Systematized Nomenclature of Medicine (SNOMED), JAMA, 1980, 243; 756-762
20. K.A. Spackman, K.E. Campbell, R.A. Cote, SNOMED RT: a reference terminology for healthcare, proc AMIA Annu Fall Symp, 1997; 640-644
21. W. Ceusters, P. Martens, C. Dhaen, B. Terzic, Linkfactory: an advanced formal ontology management system, Victoria, Canada K-CAP 2001, <http://www.landcglobal.com/images/linkfactory.pdf>, 2001.

22. M. Casella dos Santos, C. Dhaen, D. Decraene, M. van Gorp, T. Deray, The methodology behind the military health system conceptual framework and core ontology, [http://www.landcglobal.com/images/TSB\\_Methodology.pdf](http://www.landcglobal.com/images/TSB_Methodology.pdf), 2005.
23. MEDDRA, <http://www.meddrasso.com/MSSOWeb/index.htm>.
24. O. Bodenreider, Circular hierarchical relationships in the UMLS: etiology, diagnosis, treatment, complication and prevention, Proc of AMIA Annual Symposium, 2001, 57-61.
25. U. Hahn, S. Schulz, M. Romacker, An ontological engineering methodology for part-whole reasoning in medicine, 1998, <http://citeseer.ist.psu.edu/hahn98ontological.htm>.
26. HL7 Term Info working group. Using Snomed CT in HL7 Version 3. <http://www.hl7.org/>



## The development of a schema for the annotation of terms in the BioCaster disease detecting/tracking system

**Ai Kawazoe<sup>\*1</sup>, Ph.D., Lihua Jin<sup>\*1</sup>, Ph.D., Mika Shigematsu<sup>\*3</sup>, M.D.,  
Roberto Barrero<sup>\*2</sup>, Ph.D., Kiyosu Taniguchi<sup>\*3</sup>, M.D., Nigel Collier<sup>\*1</sup>, Ph.D.**  
<sup>\*1</sup>National Institute of Informatics, Hitotsubashi 2-1-2 Chiyoda-ku Tokyo, JAPAN  
<sup>\*2</sup>National Institute of Genetics, Yata 1111 Mishima Shizuoka, JAPAN  
<sup>\*3</sup>National Institute of Infectious Diseases, Toyama 1-23-1 Shinjuku-ku Tokyo, JAPAN  
<sup>\*1</sup>{zoeai, lihua-jin, collier}@nii.ac.jp, <sup>\*2</sup>rbarrero@genes.nig.ac.jp,  
<sup>\*3</sup>{mikas, tanigk}@nih.go.jp

*Amid growing public concern about the spread of infectious diseases such as avian influenza and SARS, there is an increasing need for collecting timely and reliable information about disease outbreaks from natural language data such as online news articles. In this paper we introduce BioCaster, a text mining-based system for infectious disease detection and tracking currently being developed, and discuss the development of a domain ontology and schema for the annotation of terms. In particular we focus on the comparison between two approaches, 1) a traditional task-oriented approach with a simple schema that does not strictly follow ontological principles, and 2) a formal approach which is ontologically well-founded but adds extra requirements to the annotation schema. We report on several critical problems that were highlighted by an entity annotation experiment, attributable to the purely task-oriented ontology design. A second experiment based on a formally constructed ontology produced improved annotation results despite the apparent complexity of the annotation schema.*

### 1. INTRODUCTION

As shown by the recent outbreak of Severe Acute Respiratory Syndrome (SARS) and emerging cases of avian influenza, infectious diseases have the potential to spread rapidly through person-to-person transmission within densely populated areas and across country borders through international air travel. The first line of defense against rapidly spreading diseases is surveillance, led by the World Health Organization (WHO) and national health authorities. Catching an outbreak earlier has clear implications for both morbidity and mortality as well as the feasibility of containment [1]. However a lack of surveillance system infrastructure in Southeast Asia, which is currently the focus of an avian H5N1 epidemic is seen as hindering control efforts. In addition to traditional surrogate methods such as reporting notifiable diseases and over-the-counter (OTC) sales monitoring, public health experts are increasingly considering news and other reports available on the World Wide Web (Web) as a cost-effective means of helping to find and track early cluster cases, enabling a timely and appropriate response. Such *rumour-based* information may be of

particular value for assessing possible outbreaks in areas where formal reporting procedures are absent or not well established.

Several major challenges exist in locating Web-based information in a timely manner using traditional search methods: (1) the massively increasing volume of dynamically changing unstructured news data available on the Web makes it extremely difficult to obtain a clear picture of an outbreak in a timely manner, (2) the large-scale republication of reports from centralized news agencies requires redundancy to be identified and removed, (3) the initial reports of an outbreak are contained in only a few news articles which will usually be overlooked by traditional search engines which use keyword indexing, (4) the first reports of an infectious disease will often be reported in local news media which are only available in the local language. Experience has shown that this requires computer systems to have at least a partial understanding of the domain through ontologies, term lists and databases as well as specialized multilingual resources.

To address the information needs in the domain of infectious disease outbreaks, standard Information Extraction technology has been adapted for retrospective archive search [2] but only a few systems are currently actively deployed with the most prominent being the Global Public Health Intelligence Network (GPHIN) [3], a successful but semi-closed system used by the WHO. We are now developing BioCaster, a text mining system based on an openly available multilingual ontology for proactive notification about priority disease outbreaks. A key component of the BioCaster system is the use of automated learning methods to identify novel entities and events using features derived from annotated examples in a multilingual collection of news articles. The initial target languages are English, Japanese, Vietnamese and Thai.

In our early development of BioCaster it became clear that we needed a rigorous schema for markable entities. Since the system relies on high quality human annotated training data for constructing

named entity recognizers (NERs), any inconsistency introduced into the annotation schema by ontological inconsistencies should be harmful for annotation performance, both human and machine. Surprisingly while there have been several studies on the mapping problem between terms and coding systems such as the UMLS Metathesaurus [4] as well as biomedical annotation experiments [5] [6] [7] there have been to the best of our knowledge no studies conducted into the method by which new domain models suitable for biomedical text mining should be organized. We report here on our initial experience which showed that the task-oriented annotation schema based on a poorly-considered domain ontology can indeed be harmful to accuracy. Re-organizing this schema using well founded ontological principles produced better results, despite the added complexity.

## 2. USER NEEDS

Epidemiologists are concerned with the circumstances in which diseases occur in a population and the factors that influence their incidence, spread, recognition and control. Our initial discussions with domain experts at the National Institute of Infectious Diseases revealed several common scenarios for gathering information from Web news including cases involving the spread of a communicable disease across international borders and the contamination of blood products. From these initial discussions we collected examples of early outbreak news reports and compiled a list of significant entity classes which included DISEASE<sup>1</sup>, CASE, LOCATION SYMPTOM, TIME, DRUG, etc.

Subsequent follow up discussions and examination of the literature revealed that we can categorize these concepts according to the information needs of the scientists as shown in Table 1.

Genetic epidemiology adds another dimension to the information needs as the genetic makeup of the host plays a key role in determining susceptibility or resistance to pathogens. We therefore chose to add in a further level of detail about the host which includes genes and their products, identified with a §. Finally we had 19 categories of concepts which we want to identify in news texts (Table 2).

## 3. CONSIDERATION ON TWO APPROACHES

At this stage we were aware that some of the important concepts in Table 2 are contextually-dependent and intrinsically different from others. For example, CASE and TRANSMISSION represent roles (discussed in [8] [9] [10] [11] among others) which are dependent on the existence of events they

participate in, while most others, such as PERSON, BACTERIA, and NON\_HUMAN, represent types.

We had two options for constructing the ontology and annotation schema, according to how to deal with concepts of a different nature. The first approach is rather task-oriented. Here we do not make any distinction between context-dependent concepts and others. This results in a somewhat simpler ontology: all categories of concepts are represented as classes which follow a disjoint entity class principal that has been the underlying premise of NERs. The corresponding annotation schema will also be simpler, since instances of context-dependent classes are annotated in the same way as those of other classes, e.g.

```
<NAME cl="PERSON">Kofi Annan</NAME>
<NAME cl="CASE">a 12 year-old girl</NAME> infected
with H5N1
```

(The details of this schema will be given in the next section.) In this task-oriented approach, we can annotate exactly what the event frame needs to identify. For example, we can exclude from annotation non-named, non-case mentions, which we are not interested in. A defect of this approach is that it is not ontologically well-founded.

The alternative approach is a more formal one where we make a clear distinction between context-dependent concepts and others, based on well-founded ontological principles. The result is likely to be a more complex ontology in which context-dependent concepts have a different status from other concepts. The corresponding annotation schema will also be more complex as well, since roles are annotated in a different way from those of entity classes. In order to achieve ontological consistency we also need to annotate more mentions than the former approach, including those that will not instantiate event frames.

From the two approaches above, out of expediency we chose the former for the first annotation experiment. The reason being that it seemed easier for annotators and that we could find almost no precedent works in named entity annotation which dealt with formal analysis of entities and role concepts.

## 4. ANNOTATION EXPERIMENT 1

### 4.1 Method

Based on the list of categories of concepts in Table 2, we constructed the ontology shown in Figure 1. Note that CASE and TRANSMISSION, which represent

<sup>1</sup> We will adopt here the notation of using all upper case for domain entity classes.

Focus	Description	Example properties	Concept types
Agent	Pathogens	Infectivity, pathogenicity, virulence, incubation period, communicability	VIRUS, BACTERIA, PARASITE <sup>*</sup> , FUNGI <sup>*</sup>
Transmission	The delivery or dispersal method	Dermal, oral, respiratory	TRANSMISSION
Host	Persons carrying a disease	Age, gender, occupation,	CASE, SYMPTOM, DISEASE, ANATOMY, DNA <sup>§</sup> , RNA <sup>§</sup> , PROTEIN <sup>§</sup>
Environment	Location and climate	Large population centre, enclosed building, mass transport system, rural village	LOCATION, TIME
* Not included in the current schema			
§ Genetic level entities			

Table 1 Categorization of concepts

Classes	Examples	Description
ANATOMY	<i>liver, pancreas, nervous system, eLa cel,</i>	Body parts including tissues and cells
BACTERIA	<i>Escherichia coli O157, tubercle bacillus</i>	Eubacteria
CASE	<i>a 35-year-old woman, the third case</i>	Confirmed cases of diseases
NT_CHEMICAL	<i>beryllium, organophosphate pesticide</i>	Chemicals intended for non-therapeutic purposes <sup>*1</sup>
T_CHEMICAL	<i>Relenza, immunosuppressive drug, oseltamivir</i>	Chemicals intended for the treatment of diseases <sup>*1</sup>
CONTROL	<i>stamping out, screening, vaccination</i>	Control measures to lower the risk of transmission of a disease
DISEASE	<i>H5N1 avian influenza, SARS, cholera</i>	A deviation in the normal functioning of the host caused by a persistent agent (pathogen) or some environmental factor
DNA	<i>Sp1 site, triple-A, c-jun gene</i>	Includes the names of DNAs, groups, families, molecules, domains and regions <sup>*2</sup>
LOCATION	<i>Viet Nam, Jakarta, Sumatra Island, Asia</i>	A politically or geographically defined location <sup>*3</sup>
NON_HUMAN	<i>civet cats, poultry, flies</i>	Multi-cell organism other than humans, i.e. "animals"
ORGANIZATION	<i>the Ministry of Health, WHO, Pasteur Institute</i>	Corporate, governmental, or other organizational entity <sup>*3</sup>
PERSON	<i>Jean Chretien, Murray McQuigge</i>	A named person or family
PRODUCT	<i>botulism antitoxin, Influenza vaccine</i>	Biological product, (e.g. vaccines, immune sera)
PROTEIN	<i>STAT, RNA polymerase II alpha subunit</i>	Includes the names of proteins, groups, families, molecules, complexes and substructures <sup>*2</sup>
RNA	<i>IL-2R alpha transcripts, TNF mRNA</i>	Includes the names of RNAs, groups, families, molecules, domains and regions <sup>*2</sup>
SYMPTOM	<i>cough, fever, dehydration, convulsion</i>	Alterations in the appearance of a case due to a disease
TIME	<i>Tue Jan 3, winter, March, since October, 2003</i>	Temporal expressions that can be anchored on a timeline <sup>*4</sup>
TRANSMISSION	<i>HIV-tainted blood products, BSE-infected cows</i>	Source of infection
VIRUS	<i>Ebola virus, HIV</i>	Viruses such as HIV, HTLV, EBV <sup>*2</sup>
Descriptions marked with *1 , *2, *3, *4 are based on those in MeSH [12], GENIA ontology [13], MUC-7 [14], and HUB-4 [15], respectively.		

Table 2 List of classes of markable concepts

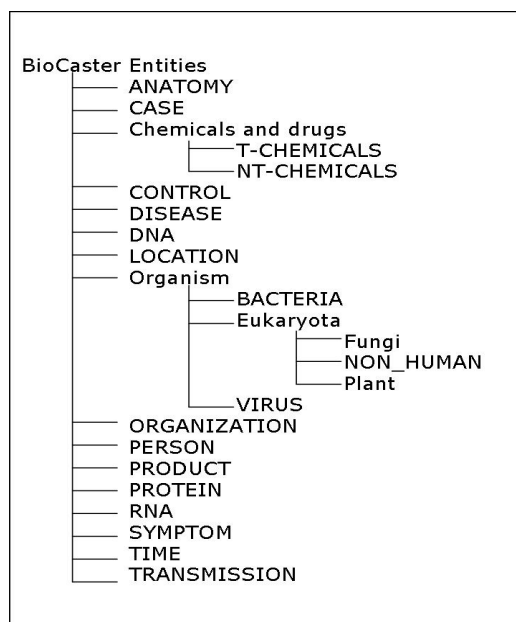


Figure 1 Initial domain ontology (simplified)

roles, have the same status as other classes since we adopted the task-oriented approach as discussed in the last section. We developed annotation guidelines to annotate non-overlapping mentions related to the classes in news articles and hired two PhD informatics students as annotators. After 1-week of training consisting of guideline review, case study discussions and test cases, we started the annotation process with 200 news articles taken from domain sources, including WHO epidemic reports, IRIN, and Reuter news.

In order to restrict the markable mentions to exactly those that we aimed to identify with the text mining system, we defined CASE as the class of confirmed cases which are unnamed, and PERSON as the class of named persons who are not cases. We considered this would narrow down the number of markable mentions since unnamed mentions for non-cases need not be annotated. We also instructed annotators to markup only the single most appropriate class, prohibited multiple classes. An example of annotated text is shown below:

The <NAME cl="ORGANIZATION">Ministry of Health</NAME> in <NAME cl="LOCATION">Indonesia</NAME> has today confirmed <NAME cl="CASE">a fatal human case</NAME> of <NAME cl="DISEASE">H5N1 avian influenza</NAME>. <NAME cl="CASE">A 27-year-old woman</NAME> from <NAME cl="LOCATION">Jakarta</NAME> developed symptoms on <NAME cl="TIME">17 September</NAME>. She contracted the virus from close contact with infected <NAME cl="TRANSMISSION">birds</NAME>.

In the annotation schema used in the example above, the attribute *cl* takes the entity class label as its value. For example "<NAME cl="PERSON">Kofi Annan</NAME>" means that the entity mentioned by "Kofi Annan" is *related* to the class PERSON. The reason for using this rather vague expression is to cover two relations between mentioned entities and the ontology we want to describe. The first is "is an instance of", and the other one is "is a subclass of". Some of the markable texts mention a particular and others mention a universal. For example, names of persons, locations and organizations are usually used to refer to a particular, whereas names of chemical substance, viruses and proteins are often used to refer to universals. This is one of the factors which makes ontology-based annotation a complicated process. It should be noted though that we intend to work towards a clear distinction between the two relations in future work.

#### 4.2 Annotation results and problems

During the first annotation experiment, we had many problem reports from annotators, and found a significant number of inconsistencies in the annotation results. Most of the problems could be traced back to poor design of the domain ontology and the annotation schema. Follow up analysis on the corpus yielded the following symptoms of error:

- Gaps in the annotation schema shown by the existence of mentions to entities which it is desirable to annotate but the annotation schema does not cover.
- Ambiguity between context-dependent concepts and context-independent ones
- Idiosyncratic annotations which are forced on annotators due to the disjoint entity class principal.

#### Gaps in the annotation schema

At the initial stage of our analysis we considered that distinguishing CASE (as confirmed cases of a disease which are unnamed humans) from PERSON (named persons who are not cases of a disease) was rather natural, since CASE entities are in general anonymous. However, in the news articles there were some examples where cases were mentioned by name as follows:

E1 Tests carried out in a UK laboratory confirmed that M.A and F died from the H5N1 strain<sup>2</sup>

In addition, we found that there were more frequent mentions of putative cases than we had expected.

<sup>2</sup> In this example we only show initials of the victims' names.

These mentions were often annotated as CASE by annotators although we restricted the scope of this class only to confirmed cases.

E2 a Taiwanese is suspected to have died of SARS

Follow up discussions with public health experts revealed that mentions of putative cases are important, especially in the early stages of disease outbreaks, and we concluded that they should be identified by the system. However, the existing framework made them difficult to capture.

#### Ambiguity caused by context-dependent concepts

One of the classes which confused annotators most was TRANSMISSION (source of infection). Below are typical examples of problematic cases.

- E3 Victims contract the virus from close contact with infected birds  
 E4 There is no known cure for Ebola, which is transmitted via infected body fluids  
 E5 An Irish woman infected with Hepatitis C by a contaminated blood product  
 E6 18 hospitalized after consuming chapattis

Annotators had a problem in annotating 'birds' in E3 since those can be classified as both TRANSMISSION and NON\_HUMAN (animals). 'Body fluid' in E4 is also ambiguous between TRANSMISSION and ANATOMY (body parts), and also 'blood product' in E5 is ambiguous between TRANSMISSION and PRODUCT (biological product). Most of the TRANSMISSION instances found in the text were those which could be categorized as NON\_HUMAN, and the cases which belonged only to TRANSMISSION, such as 'chapattis' in E6, were very few.

#### Idiosyncratic annotations due to the disjoint entity class principal

- E7 <NAME cl="PERSON">Hudd</NAME> has written several books on music hall and Variety...  
 E8 Doctors later diagnosed <NAME cl="CASE">Hudd</NAME> with a chest infection...

In the example above, it is clearly undesirable that the same entity is related to PERSON in E7 and CASE in E8. Although the annotator was aware of the choices the principal of disjoint classes forced a choice.

### 4.3 Empirical results from training an NER

We trained a support vector machine [13] (for details, see Takeuchi and Collier [14]) for named entity recognition based on the annotated corpus of 200 news articles. 10-fold cross validation experiments were performed using TinySVM<sup>3</sup>. A -2/+1 features window was used that included surface word, orthography, biomedical prefixes/suffixes, lemma, head noun and previous class predications. The F-score for the all classes in Table 2 was 76.96. Among the problematic classes were found to be PERSON, CASE and NON\_HUMAN (many instances of which had ambiguity with TRANSMISSION) which had F-scores below our expectation: PERSON (54.95), CASE (53.17), NON\_HUMAN (68.0).

## 5. ANNOTATION EXPERIMENT 2

### 5.1 Re-examination of the approach

Although we chose the task-oriented approach for its simplicity and ease of implementation the results from automatic NER and subsequent corpus analysis revealed that problems arose because we made no clear distinction between context-dependent and context-independent classes. We decided to take an alternative, formal and linguistically-sound approach, and distinguish context-dependent concepts from others in both the ontology and the annotation schema.

### 5.2 Classification of concepts

The first step was to use the classification method proposed by Guarino and Welty ([9] and [10]) which is based on meta-properties (rigidity, identity, dependency), in order to classify categories of concepts in Table 2. Definitions of the meta-properties we used are as follows:

<Rigidity> ([10], p.4)

**rigid property  $\phi$  (+R):**  $\forall x \phi(x) \rightarrow \Box \phi(x)$

**anti-rigid property  $\phi$  (~R):**  $\forall x \phi(x) \rightarrow \neg \Box \phi(x)$

<Identity> ([10], p.5)

**Identity Condition (IC):** An identity condition is a formula  $\Gamma$  that satisfies either of the followings<sup>4</sup>:

<sup>3</sup> Available from <http://cl.aist-nara.ac.jp/~taku-ku/software/TinySVM>

<sup>4</sup> In [9], further restrictions are added in order to avoid 1) the case where the necessary IC definition becomes trivially true regardless of the truth value of the formula  $x=y$  and 2) the case where  $\Gamma(x, y, t, t')$  is false and that makes the sufficient IC definition trivially true.

	rigidity	identity (supplying)	identity (carrying)	dependency	classification
ANATOMY	+R	+O	+I	-D	Type
BACTERIA	+R	+O	+I	-D	Type
CASE	~R	-O	+I	+D	Material Role
NT_CHEMICAL	~R	-O	+I	+D	Material Role
T_CHEMICAL	~R	-O	+I	+D	Material Role
CONTROL	~R <sup>*1</sup>	-O <sup>*2</sup>	+I	+D	Material Role
DISEASE	+R	+O <sup>*3</sup>	+I	+D	Type
DNA	+R	+O	+I	-D	Type
LOCATION	+R	+O	+I	-D	Type
NON_HUMAN	+R	+O	+I	-D	Type
ORGANIZATION	+R	+O	+I	-D	Type
PERSON	+R	+O	+I	-D	Type
PRODUCT	+R	+O	+I	+D	Type
PROTEIN	+R	+O	+I	-D	Type
RNA	+R	+O	+I	-D	Type
SYMPTOM	+R	+O	+I	+D	Type
TIME	+R	+O	+I	-D	Type
VIRUS	+R	+O	+I	-D	Type
TRANSMISSION	~R	-O	-I	+D	Formal Role

\*1 We consider that this class is anti-rigid, since it is possible that an action which is an instance of CONTROL in the current world is not an instance of CONTROL in some other accessible world. The same action may be conducted for different purposes in different worlds.

\*2 This class includes events. In DOLCE top level categories (Gangemi et al.[19]), Events are under the class of Perdurant/Occurrence. It seems to be controversial what the identity condition for events should be. Davidson [20] proposes a condition such that "events are identical if and only if they have exactly the same causes and effects". In any case it should be reasonable to assume that this class itself does not supply ICs but inherits them from the upper level classes.

\*3 What we consider ICs for this class is as follows: Two instances of diseases are identical iff the two are experienced by the same host at the same time, are caused by the same agent (e.g. H5N1 virus for "H5N1 avian influenza") and have the same set of characteristic alterations/symptoms (e.g. inflammation of the lung for "pneumonia").

**Table 3: Classification of concepts**

necessary IC:  $E(x, t) \wedge \phi(x, t) \wedge E(x, t') \wedge \phi(y, t') \wedge x=y \rightarrow \Gamma(x, y, t, t')$

sufficient IC:  $E(x, t) \wedge \phi(x, t) \wedge E(x, t') \wedge \phi(y, t') \wedge \Gamma(x, y, t, t') \rightarrow x=y$   
(E : "actually exist at time t")

**Any property  $\phi$  carries an IC (+I)** iff it is subsumed by a property supplying that IC.

**A property  $\phi$  supplies an IC (+O)** iff i) it is rigid; ii) there is a necessary or sufficient IC for it; and iii) the same IC is not carried by all the properties subsuming  $\phi$ .

**<Dependency>** ([10], p.7)

**externally dependent property  $\phi$  (+D):**

$\forall x \Box (\phi(x) \rightarrow \exists y \omega(y) \wedge \neg P(y, x) \wedge \neg C(y, x))$   
(P: "is a part of")  
(C: "is a constituent of")

Classification results are shown in Table 3. Most concepts such as ANATOMY, NON\_HUMAN, and PERSON are classified as Type, whereas the concepts which were problematic in the first

experiment were classified as Role: TRANSMISSION (Formal Role) and CASE (Material Role). According to the further classification of non-rigid concepts by Kaneiwa and Mizoguchi [18], these cases are classified as time-dependent concepts.

### 5.3 Modification of the schema

For some of the roles in Table 3, we modified their status in the annotation schema.

#### CASE

CASE and PERSON were problematic since we distinguished them according to the form of expression (unnamed/named), in addition to the case/non-case distinction. In order to cover the mentions which could not be annotated in the first experiment, we extended the scope of the PERSON class to include person instances in general, and eliminate the unnamed/named and case/non-case distinctions. We modified the annotation schema so that CASE is not the value of *cl* attribute, but is the *case* attribute which applies to the referred instance of PERSON. This attribute takes the value *true* when the mentioned instance is a confirmed case of disease,

*false* when the instance is not a case, and *putative* when the instance is a suspected case. Named case mentions and suspected case mentions are annotated as follows:

- E9 Tests carried out in a UK laboratory confirmed that <NAME cl="PERSON" case="true">M.A</NAME>...
- E10 <NAME cl="PERSON" case="putative">a Taiwanese</NAME> is suspected to have died of SARS

The meaning of *case* attribute-value pairs can be described in logical description and natural language as follows:

<...cl="PERSON" case="true">John</...>: **case(j)**  
 "It is true that the person **j** mentioned by "John" is an instance of the CASE class"

<...cl="PERSON" case="false">John</...>: **¬case(j)**  
 "It is false that the person **j** mentioned by "John" is an instance of the CASE class"

<...cl="PERSON" case="putative">John</...>:  
 ◇**case(j)**  
 "It is possible that the person **j** mentioned by "John" is an instance of the CASE class"

As shown above, the values of the *case* attribute correspond to logical operators such as  $\neg$  and  $\Diamond$ . The values of *case* attributes specify the modes of linkage between the referred concept and the CASE class. The formal basis we had in mind when formulating the *case* attribute are as follows: 1) every instance of a non-rigid class must be an instance of some rigid class, 2) the relations between a non-rigid class and its instance are often modified by modal/temporal operators. The first point drove us to create the case attribute which apply to instances of some rigid class, here, PERSON. The second point is the motivation for us to set values to include negative and modal operators. This schema can be extended if we allow a wider value range for the case attribute to include other modal/temporal operators, although currently we restrict the values to the three above.

It is worth noting that there is a trade-off between this revised schema and the former schema which is that we have increased the number of the markable entities, since we need to annotate unnamed, non-case mentions which are not directly related to the purpose of the system.

## TRANSMISSION

We defined the *transmission* attribute which applies to mentions of ANATOMY, PRODUCT, PERSON and NON\_HUMAN classes. As shown in the following examples, 'birds' are always related to NON\_HUMAN, and take a 'true' value only when they are mentioned as a source of infection. It can also take a 'putative' value to cover mentions to possible sources of infection.

- E11 Victims contract the virus from close contact with infected <NAME cl="NON\_HUMAN transmission="true">birds</NAME>

## T\_CHEMICAL /NT\_CHEMICAL

Concept classification revealed that T\_CHEMICAL and NT\_CHEMICAL have "the situation dependency obtained from extending types" discussed in [18] and have the same status as 'weapon' and 'table'. T\_CHEMICAL includes chemicals mentioned as drugs in any context and those regarded as drugs in some context. Here we removed the two classes and made the parent node CHEMICAL as a class for annotation.

We then defined *therapeutic* attribute which applies to mentions of CHEMICAL and takes the value *true* when the entity is intended for therapeutic use and *false* otherwise.

As a result of the modifications above, our revised ontology is shown in Figure 2. We also added new classes CONDITION (status of patients: 'hospitalized' 'died 'in critical condition', etc) and OUTBREAK (collective disease incident: 'outbreak', 'pandemic', etc). Information about CONDITION is important for experts to know the rate of hospitalization and death and determine the alert level. Mentions of OUTBREAK include expressions which are specific to disease outbreak news, increasing the specificity of our detection system. We located PERSON and NON\_HUMAN under metazoa, and added a *number* attribute (which takes *one* or *many* as its value) to be applied to PERSON instances.

With insights from the revised ontology we also changed the annotation method by dividing the process into two distinct stages as shown in Figure 3: 1) annotation of mentions to non-role (rigid) concepts and 2) annotation of role (non-rigid) concepts.

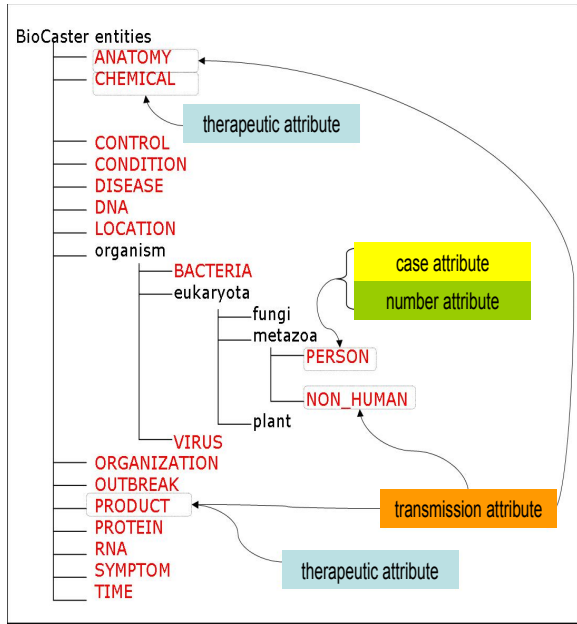


Figure 2 Current ontology (simplified)

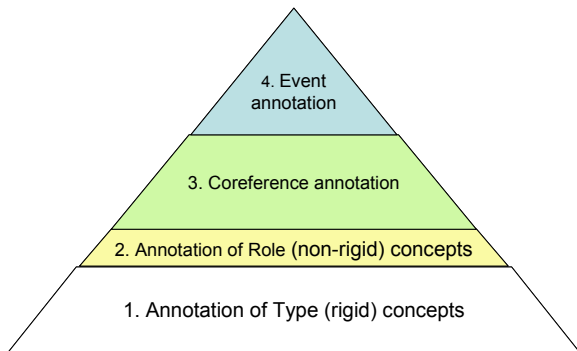


Figure 3 Annotation schedule

#### 5.4 Results of annotation and NE recognizer training

We asked three PhD students to annotate a further 300 news articles. This time we used the revised annotation method 1 and 2 shown in Figure 3.

As a result of distinguishing between Role concepts (case, transmission, therapeutic) from others in the annotation schema, problem reports on these classes were reduced, and the annotation results were also improved. Contrary to our expectations, the complexity of the new annotation schema and the increased number of markable mentions seemed to have no negative influence on the annotator's speed.

The improvement can be seen empirically in the NER results. We re-annotated the corpus used in the first experiment using the revised annotation schema. This time the F-score for all classes rose to 79.96 (+3 compared to the previous result). Especially,

significant increases of the F score were observed in the classes for PERSON (66.28; +11.33 compared to the previous result), case mentions among PERSON (65.63; +12.46), and NON\_HUMAN (73.21; +5.21).

#### 5.5 Remaining issues

Some of the problems reported in this second experiment were related to context dependency (anti-rigidity, situation dependency) discussed in Section 6.2.

The most difficult class seemed to be CONTROL (control measures to lower the risk of diseases). As shown in Table 3, we consider this class is also non-rigid, and it includes mentions which refer to subclasses of the CONTROL class regardless of situation ("quarantine" "vaccination"), and others which can be a control measure depending on the situation ("warning" "blockade"). This characteristic seems to cause the difficulty.

So far we have resolved the complexity of non-rigid concepts by defining attributes which apply to instances of rigid classes (e.g. the *case* attribute for the class PERSON). This strategy, however does not seem to be effective for CONTROL since it is not easy to identify a rigid superclass for CONTROL which can be realistically annotated in the text. For example, EVENT can be considered as a rigid class subsuming CONTROL, but currently it is not realistic to manually annotate every mention of an event. Currently we are seeking for a way to deal with this problem.

#### 6. CONCLUSION

The study in this paper was motivated by our need for a high quality annotation schema to support detection of novel entities in the infectious disease outbreak domain. We discussed two experiments based on alternative approaches for constructing an ontology-based annotation schema. The amount of data in our study is relatively small but empirical results indicate support for our view that there is a positive effect in adopting well founded ontological principals over an ad-hoc task-based approach. Although this study is not a formal evaluation of ontologies, it is still an evaluation from the viewpoint of ontology application to the task of natural language annotation. The classification method of Guarino and Welty ([9], [10]) which was originally proposed to achieve consistency in the configurational structure of ontologies, was adapted and found to be useful for improving annotation performance.

An alternative possibility exists which we have not addressed in this paper which is to reformulate the tradition NER task to allow for overlapping (nested) and multi-class entities. This however introduces

significant additional complications in both the recognizer models and in the annotation schema so we have adopted a less radical formulation in this work.

As the next step in this study, we are now extending our simple taxonomy to a multi-lingual ontology; enriching the current taxonomic structure with domain-sensitive relations. The resulting ontology will be freely available for re-use. At the initial stage we are focusing on English, Japanese, Vietnamese, Thai, Chinese (standard) and Korean. We hope to add other Asia-Pacific languages in the future.

#### Acknowledgements

We gratefully acknowledge partial funding support from the Japan Society for the Promotion of Science (grant no. 18049071). We also thank the anonymous reviewers for helpful comments.

#### References

1. Ferguson NM, Cummings DA, Cauchemez S, Fraser C, Riley S, et al. Strategies for containing an emerging influenza pandemic in Southeast Asia. *Nature* 437: 209–214. 2005.
2. Grishman R, Huttunen S, and Yangarber R. Information extraction for enhanced access to disease outbreak reports. *Journal of Biomedical Informatics*, Vol. 35, No. 4, 236 - 246, 2002.
3. Public Health Agency of Canada. GPHIN system. [http://www.phac-aspc.gc.ca/media/nr-rp/2004/2004\\_gphin-rmispbk\\_e.html](http://www.phac-aspc.gc.ca/media/nr-rp/2004/2004_gphin-rmispbk_e.html)
4. Aronson A.R. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proceedings of AMIA Symposium*, 17–21, 2001.
5. Rindflesch T.C., Tanabe L, Weinstein J.N. and Hunter L. EDGAR: extraction of drugs, genes and relations from the biomedical literature. *Proceedings of Pacific Symposium on Biocomputing* 5:514-525, 2000.
6. Kim J.D., Ohta T, Tsuruoka Y, Tateishi Y, Collier N. Introduction to the Bio-entity Recognition Task of the JNLPBA workshop. *Proceedings of the JNPBA*, 70-76, 2004.
7. Yeh A, Morgan A, Colosimo M, Hirschman L. BioCreAtIvE task 1A: gene mention finding evaluation. *BMC Bioinformatics* 2005, 6(Suppl 1):S2.
8. Sowa J.F. *Conceptual structures: Information processing in mind and machine*. Addison-Wesley, New York; 1984.
9. Guarino N, Welty C. A formal ontology of properties. Dieng R, Corby O (eds.) *Proceedings of EKAW-2000: The 12th International Conference on Knowledge Engineering and Knowledge Management*, volume 1937: 97-112.
10. Guarino N, Welty C. Ontological analysis of taxonomic relations. Lander A, Storey V (eds.) *Proceedings of ER-2000: The International Conference on Conceptual Modeling*, vol. 1920, 210-224, Springer Verlag LNCS, Berlin, Germany.
11. Steimann F. On the representation of roles in object-oriented and conceptual modelling. *Data and Knowledge Engineering* 35, 1: 83-106. 2000.
12. U.S. National Library of Medicine. *Medical Subject Headings (MeSH)*, 2006.
13. Kim J.D., Ohta T, Tateishi Y, Tsujii J. GENIA corpus - a semantically annotated corpus for bio-textmining. *Bioinformatics* 19(suppl. 1), pp. i180-i182, Oxford University Press, 2003.
14. Hirschman L, Chinchor N. MUC-7 named entity task definition. *Proceedings of the 7th Message Understanding Conference (MUC-7)*.
15. Hirschman L, Chinchor N, Grishman R, Sundheim B. Hub-4 Event Guidelines Version 2.6. [http://www-nlpir.nist.gov/related\\_projects/muc/proceedings/hub4/guidelines.html](http://www-nlpir.nist.gov/related_projects/muc/proceedings/hub4/guidelines.html)
16. Vapnik, V. N. *The Nature of Statistical Learning Theory*, Springer-Verlag, New York, 1995.
17. Takeuchi, K and Collier, N. "Bio-medical entity extraction using support vector machines", in vol. 33, no.2, *Artificial Intelligence in Medicine*, Elsevier, pp. 125-137, 2005.
18. Kaneiwa K, Mizoguchi, R. An order-sorted quantified modal logic for meta-ontology. *Proc. of the International Conference on Automated Reasoning with Analytic Tableaux and Related Methods (TABLEAUX 2005)*, Koblenz, Germany: 169-184, 2005.
19. Gangemi A, Guarino N, Masolo C, Oltramari A, Schneider L. Sweetening ontologies with DOLCE. Benjamins et al. (eds.), *Proceedings of the 13th European Conference on Knowledge Engineering and Knowledge Management (EKAW2002)*, 166-181, Sigüenza, Spain, 2002.
20. Davidson D. The Individuation of events. Rescher N (ed) *Essays in Honor of Carl G. Hempel*: 216-234, 1969, D. Reidel.



## BioFrameNet: A Domain-specific FrameNet Extension with Links to Biomedical Ontologies

Andrew Dolbey<sup>1,2</sup>, Michael Ellsworth<sup>3</sup>, and Jan Scheffczyk<sup>3</sup>

<sup>1</sup>University of Colorado, Center for Computational Pharmacology, Aurora, Colorado

<sup>2</sup>University of California Berkeley, Linguistics Dept., Berkeley, California  
[andy.dolbey@gmail.com](mailto:andy.dolbey@gmail.com)

<sup>3</sup>International Computer Science Institute, Berkeley, California  
[{infinity,jan}@ICSI.Berkeley.EDU](mailto:{infinity,jan}@ICSI.Berkeley.EDU)

*Biomedical domain ontologies could be better put to use for automatic semantic linguistic processing if we could map them to lexical resources that model the linguistic phenomena encountered in this domain, e.g., complex noun phrase structures that reference specific biological entity names and processes. In this paper, we introduce BioFrameNet – a domain-specific FrameNet extension. BioFrameNet uses Frame semantics to express the meaning of natural language, is augmented with domain-specific semantic relations, and links to biomedical ontologies like the Gene Ontology – all of which are expressed in the Description Logic (DL) variant of OWL. Thus, BioFrameNet annotations of natural-language text precisely map to biomedical ontologies, which in turn facilitates inference using DL reasoners.*

### INTRODUCTION

Many currently available Natural Language Processing (NLP) tools limit language processing to levels of linguistic detail that involve form, e.g. Part of Speech tagging and syntactic parsing (Stanford Parser<sup>1</sup>). In this endeavor, they are quite successful. What is missing is, however, an automated analysis of meaning. With the vast amount of knowledge expressed via textual resources publicly available, we see an increasing demand to include automated meaning analysis in our NLP toolkits. We intend to develop tools that provide users with fast access to what is being discussed in a large set of documents of potential interest. This will include tasks like entity recognition, question answering, thread discovery, and summarization.

At the same time, there has been a rapid emergence of a great number of ontological resources including the Gene Ontology and Entrez Gene Database. This is particularly true in the domains of molecular biology and biomedicine. This emergence offers opportunities to achieve new levels of success in

Natural Language Understanding (NLU), the task of automatically determining and extracting meaning from texts. But for this to happen, the *interface between form and meaning* must also be modeled.

We propose to model this interface by combining Frame semantics [1] with links to domain-specific biomedical ontologies, all of which we express in the Description Logic (DL) variant of OWL in order to facilitate inference by means of DL reasoners like Racer [2] or FaCT++ [3]. The primary goal of *BioFrameNet* (BioFN), a resource currently being developed, is to model the mapping of form and meaning in the linguistic structures that occur in biomedical texts.

BioFN is the dissertation project of the first author. It extends and refines FrameNet (FN) [4] – a lexicon for English, which is based on Frame semantics [1]. A semantic Frame (hereafter simply Frame) represents a set of concepts associated with an event or a state, ranging from simple (Bringing, Placing) to complex (Revenge, Criminal\_process). For each Frame, a set of roles (or arguments), called Frame Elements (FEs), is defined, about 10 per Frame. We say that a word can evoke a Frame, and its syntactic dependents can fill the FE slots. Semantic types (STs) constrain the types of FE fillers. Semantic relations between Frames are captured in Frame relations, each with corresponding FE-to-FE mappings. Syntactic-semantic mapping in FN and BioFN is captured by means of defining sets of valence patterns, where triples of FE, grammatical function, and phrase types observed in natural language text are enumerated for each Lexical Unit (LU) = word sense. FN currently contains more than 780 Frames, covering roughly 10,000 LUs; these are supported by more than 135,000 FrameNet-annotated example sentences.<sup>2</sup>

<sup>1</sup> See  
<http://www-nlp.stanford.edu/software/lex-parser.shtml>.

<sup>2</sup> For further information on FrameNet, see  
<http://Framenet.icsi.berkeley.edu>.

This paper proceeds as follows: First, we briefly discuss related work. Second, we introduce BioFN. We then propose mappings to biomedical ontologies and show our technique for creating these mappings, which will use OWL DL. This is followed by a description of how biomedical natural-language text can be annotated using BioFN and how these annotations can be put to work for reasoning by expressing them in OWL DL. Finally, we discuss lessons learned and show how others can benefit from our approach.

## RELATED WORK

The HunterLab<sup>3</sup> transport ontology has also been developed to model transport processes [5], and shares certain properties with BioFN. However, by using the explicit semantics provided in (Bio)FrameNet, we get, for free, a more inclusive formal analysis of the semantics of a transport event. Therefore, we would not need to produce and specify separate axioms with systems such as PAL. We model this semantics directly with BioFN.

BioFN uses our OWL DL translation of FrameNet [6] and augments it with domain-specific semantic relations between FEs and links to GO, the Entrez Gene database, and the protein transport knowledge representation created by the HunterLab<sup>4</sup>. Thereby, BioFN leverages on our experiences with linking FrameNet to the Standard Upper Merged Ontology (SUMO) [7], which, so far, are not domain specific.

PASBio [8] is a project that aims to produce definitions of Predicate Argument Structure (PAS) frames, similar in spirit to PropBank [9], but focusing on the domain of molecular biology. Although the PAS frames have much in common with BioFN valence patterns, it does not offer a direct linking of the predicates or their arguments to domain or general merged ontologies. The work of Korhonen et. al. [10] reports on the automatic induction of lexical verb classes for the domain of biomedicine, where the classes link together syntactic and semantic properties of groups of verbs, much like the work of Levin [11] and Kipper [12]. Providing syntax-semantic linking at the level of lexical class helps compensate for missing individual lexical entries, but runs the risk of error for individual predicates that share most of the semantics of the class, but nevertheless show divergent linking behavior [13].

"Kicktionary"<sup>5</sup> is a multi-lingual application of the FrameNet methodology to the domain of soccer. The kicktionary structure can be brought into accordance with ontological principles [14] and thus be mapped to soccer ontologies, e.g. [15]. BioFN can be extended to a multi-lingual lexicon based on the principles shown in [14]. Additional domain-specific semantic relations between FEs distinguish BioFN from the kicktionary.

## BIOFRAMENET

BioFN is a lexical resource modeled after FrameNet (FN) proper [4]. Indeed, it is an extension of FrameNet, one that builds on – i.e., includes and links to – the general FN frames. The primary data of the project is a collection of text data items (discussed later in the paper) annotated by biologists associated with the HunterLab of the University of Colorado Health Sciences Center.<sup>6</sup> The text data has a primary focus on the domain concept of intracellular transport. The annotations were carried out with a reported consistency score of over 90%. For purposes of this work, the annotations provide reliable indications of the locations of the spans of text that correspond to FE values.

The primary additions to FN proper consist of semantic frames relevant to the domain of molecular biology. As is the case elsewhere in FrameNet, these frames are linked with other frames in a set of clearly defined ways. For each Frame, there is a definition of Frame elements – the “arguments” or “slots” that the Frame licenses. Each Frame is also associated with a list of predicators, the lexical units that evoke the Frame.

For example, BioFN includes the domain-specific Frame “Transport\_intracellular”, which describes the biological process of intracellular transport of molecular entities. The Frame elements for this Frame are Cargo (the transported entity), Carrier (the transporting entity), Origin (the start point of transport), and Destination (the end point of transport). The following predicators, with part of speech appended to the name, are among the more frequently occurring lexical units that evoke this Frame:

translocate.v, translocation.n, transport.v,  
transport.n, shift.v, shuttle.v, export.v

<sup>3</sup> Center for Computational Pharmacology, of University of Colorado Health Sciences Center, directed by Dr. Lawrence Hunter.

<sup>4</sup> See <http://compbio.uchsc.edu/grifs/transport/schema.shtml>.

<sup>5</sup> See <http://www.kicktionary.de>.

<sup>6</sup> See <http://compbio.uchsc.edu>.

In many cases, new Frames added are related to other Frames that already exist in FN proper. For example, the `Transport_intracellular` Frame is included as a subtype of the `Brining` Frame, a Frame concerning the movement of a Theme and an Agent and/or Carrier.<sup>7</sup> It should be noted that the focus of the texts in the HunterLab corpus data will place a limit on the number and coverage of biomedical Frames included in the initial version of BioFN.

An important question that arises when incorporating new Frames in FN is whether or not a new Frame is warranted. This ties in to a general lumping vs. splitting decision the FN team often faces [4]. When the Frame under consideration is for domain specific semantics, there are special pros and cons to splitting with a new Frame. One disadvantage is an increase in the complexity of the network of Frames. We believe this is outweighed by the advantage of being able to specify richer information and constraints specific to the particular domain. Thus it will be possible to elaborate and constrain the general semantics of bringing with meaning, entailments, and domain knowledge particular to the event of intracellular transport. This shows up most clearly in the linking of Frames and FEs to domain specific ontologies. Maintaining close relations with more general Frames allows access to the more general semantics as well, thus simplifying the task of connecting the Bio-specific Frame to related Frames, since many of the connections will already be modeled in the general vocabulary.

### MAPPING BIOFRAMENET TO DOMAIN ONTOLOGIES

The domain ontologies we used for BioFN's mappings are GO, Entrez Gene, and a small transport knowledge representation schema of the HunterLab (HL) [5]. These were chosen for three reasons. First, and foremost, the community consensus is that GO and Entrez Gene are reliable, trusted, and actively updated. Second, all three are free and publicly available. And third, the HunterLab transport schema is currently under active development, and itself makes use of the other two domain resources.

There are two levels of mappings that must be formalized. On one level, the `Transport_intracellular` Frame and its Frame Elements are described. This frame is mapped to a

node in the GO `biological_process` tree, "protein transport". The FEs "Origin" and "Destination" are mapped to nodes in the `cellular_component` tree. The FEs "Cargo" and "Carrier" are disjunctively mapped to either an Entrez Gene element, or otherwise to the HL items "molecule or molecular complex" or "molecular part". This is shown in Fig. 1. On another level, we also need to map `SemanticType` (ST) filler constraints to the same (or related) ontologies<sup>8</sup>.

We have developed an approach that automatically translates a crucial portion of FrameNet (and its specializations) and annotations into OWL DL [6]. Fig. 1 shows the OWL DL translation of the `Transport_intracellular` Frame.

Frames, STs, and FEs are represented as OWL classes, where an FE class represents the type of the FE fillers. Frame and FE relations are modeled as existential restrictions on these classes; inheritance is represented via OWL subclassing. This way the generated ontology stays OWL DL – a crucial precondition for automated reasoning. The connection between a Frame and an FE filler is represented by the "hasFE" relation. We do so because in OWL relations are not first-class objects.<sup>9</sup> For example, the FE filler for `Origin_relation` is in fact a relation but we represent it as an OWL class in order to connect spans of text to it and to have the possibility of specifying relations to other FEs (like the `Origin` FE, which fills `Origin_relation`).

BioFN also uses the FrameNet STs, which are linked to the Standard Upper Merged Ontology (SUMO) [7]. Thereby, BioFN immediately benefits from SUMO's rich axiomatization.

We augment the OWL translation of BioFN with links to the Gene Ontology (GO), the Entrez Gene Ontology (EG), the HunterLab transport ontology (HL), and Smith's Relation Ontology (RO) [16]. These links are represented via subclass relationships and appear as bold arrows in Fig. 1.<sup>10</sup>

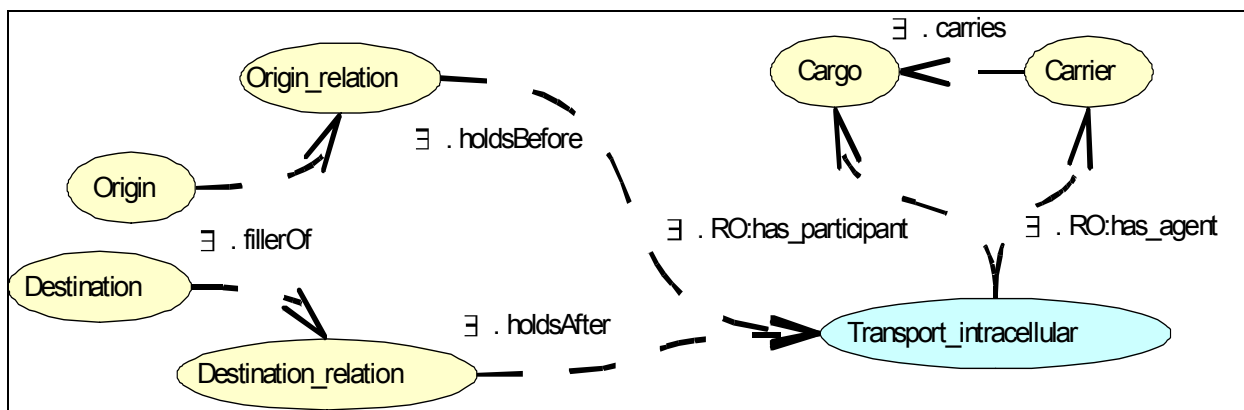
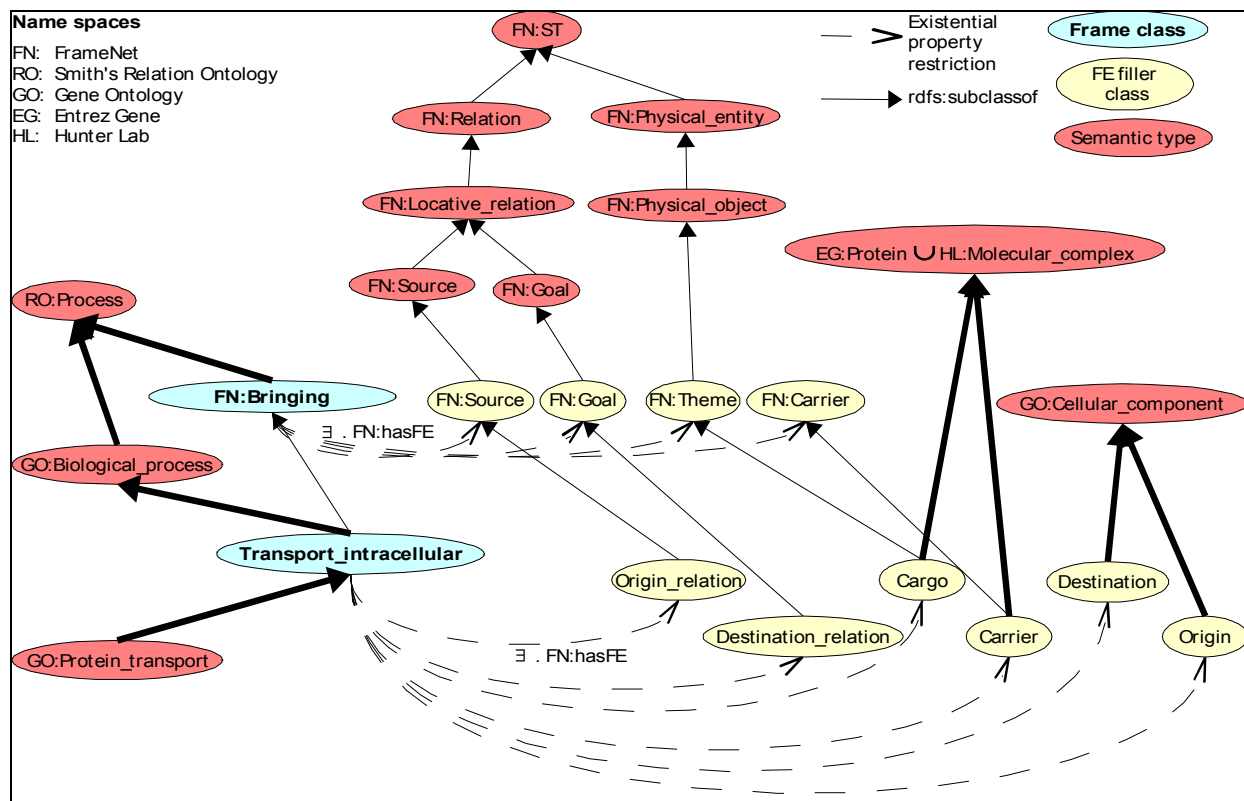
For example, the Frame class `Transport_intracellular` is a subclass of `GO:Biological_process`. Our way of modeling supports the use of OWL's expressive class language, e.g., to create anonymous union classes. For example, the class `Cargo` is a subclass of the

<sup>7</sup> See [http://framenet.icsi.berkeley.edu/index.php?option=com\\_wrapper&Itemid=118&frame=Brining&](http://framenet.icsi.berkeley.edu/index.php?option=com_wrapper&Itemid=118&frame=Brining&).

<sup>8</sup> Mappings of the ST filler constraints are not shown in Fig. 1.

<sup>9</sup> OWL does not support relations between relations other than inheritance.

<sup>10</sup> The subclass relationships were added by hand in the OWL representation, they are not expressible in FrameNet itself.



union of EG:Protein and HL:Molecular complex.

In order to aid reasoning we specify further semantic relations between FE filler classes of the same Frame (see Fig. 2).

Wherever possible we use relations and constraints defined in Smith's Relation Ontology in order to



and connect this instance by the relation R. Also, for FE mappings (including inheritance) we generate owl:sameAs relations between the generated FE instances, which aid reasoning [6]. Thus we generate a new instance of the FrameNet:Bringing Frame because the Transport\_intracellular Frame inherits from FrameNet:Bringing. We also express that the connected FE instances are the same. Therefore, the span "its" in the example GRIF actually evokes three FEs, all of which have an identical filler: Cargo (in Transport\_intracellular), FrameNet:Figure (in FrameNet:Goal), and FrameNet:Theme (in FrameNet:Bringing).

Generation of BioFN-specific semantic relations between FEs and Frames is straightforward. Fig. 4 shows the additional semantic relations generated for the Transport\_intracellular Frame instance.

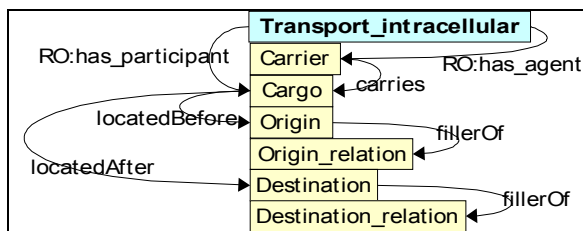


Figure 4 –Transport\_intracellular relations.

In Fig. 5, we represent an instance of the Dimension Frame bound (via the Cargo FE) to an instance of the Transport\_intracellular Frame.

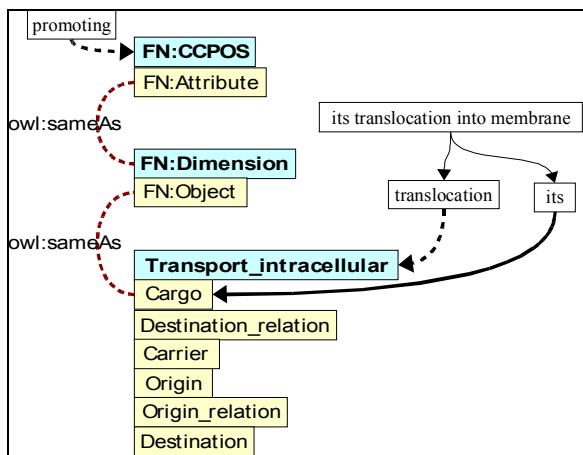


Figure 5 – Dimension Frame : an instance of metonymy.

This interpretation arises through a metonymic relation between events and quantities which is beyond the scope of the current paper; the interpretation with Transport\_intracellular filling the

Attribute role of Cause\_change\_of\_position\_on\_a\_scale ought to be discarded since Attributes and Events are disjoint.

## LESSONS LEARNED

### Changing FrameNet

Even during our preliminary investigation of annotation for BioFN, we have discovered new LUs (e.g. *promote.v* and *enhance.v*) for the Cause\_change\_of\_position\_on\_a\_scale Frame. This is despite FrameNet having studied this concept in some detail, showing definitively that domain-specific annotation will be necessary to capture the vocabulary of the biological domain.

This elaboration of FN is similar in spirit to other current efforts to link FN with other similar resources like VerbNet, PropBank, and Cyc [17]. These resources will be used for comparison and evaluation, when appropriate, as BioFN work proceeds.

### Changing biomedical ontologies

The lack of reference to GO in many entries of the HunterLab ontology will make integrated processing very difficult. The ultimate usefulness of BioFN will rely on a merged ontology and knowledge-base, with seamless references to FrameNet, SUMO, the HunterLab ontology, GO, and Entrez-Gene. The cross-reference between the ontologies required by BioFN will reveal errors and unnecessary points of difference between these ontologies, thus enabling their improvement.

### The impact of our approach for reasoning

We have already demonstrated elsewhere [6] that our OWL DL model of FrameNet is usable for the kind of reasoning needed for question answering, using queries in Racer. With some loss of power, the method could be made more efficient by implementation as a graph-traversal or querying of an SQL version of the ontology.

However, since the approach was not integrated with a large-scale ontology, it has so far been hampered by variations in the linguistic form of objects not captured in FrameNet or even in WordNet. Since BioFN will be integrated with the appropriate ontologies from its inception, the same approach should be much more powerful using the BioFN resource (together with its associated ontologies) than it is with FrameNet resources alone. In

addition, applications built with BioFN or FrameNet will make use of other NLP tools such as stemmers and lemmatizers for handling variation in linguistic form. We predict having similar success with BioFN in carrying out Question Answering and a variety of other NLU tasks.

*How can others benefit from our approach?*

Current biological ontologies have very few relations and events, and considerably less experience with modeling language than FrameNet. The work demonstrated here shows that FrameNet-style ontological descriptions of language can be integrated with information from biological ontologies using the expressive power of Description Logic.

*How can our technique be applied to other problems/domains?*

Since FrameNet provides a general-domain (if limited) ontology, it seems promising to apply our methodology to other domains that have associated ontologies and a need for textual processing. One area in which some work has already proceeded is event tracking in the terrorism domain [18].

## CONCLUSIONS, FUTURE WORK

In this paper we introduced BioFN – a domain-specific FrameNet extension. BioFN bridges form and meaning of natural-language biomedical texts by (1) new domain-specific Frames, (2) links to established biomedical ontologies like GO and Entrez Gene, and (3) domain-specific semantic relations between FEs. We model BioFN as an OWL DL ontology, which we populate with BioFN annotations of biomedical texts. Thus, natural-language biomedical texts become available for DL-based reasoning.

Since the BioFN project is dissertation work currently in progress, we are not yet able to provide full numbers and statistics for coverage of the data under consideration and counts and definitions of all the new Frames that need to be created. This is indeed one of the primary goals of the dissertation: a complete analysis of the collection of GRIFs in the HunterLab corpus. An analysis of coverage of WMD-related<sup>13</sup> text by the FN project shows that analyzing texts in a particular domain does yield

significantly greater coverage of new texts in the same genre.<sup>14</sup>

In the future, we will enhance BioFN with more biomedical Frames and richer semantic relations. Also, we aim at an (OWL DL + SWRL) axiomatization of domain-specific relations much in the fashion of [16]. We will conduct experiments in automatic parsing using the Shalmaneser Frame parser [19]. GO and Entrez Gene classes provide narrow semantic types, which can significantly aid automatic Frame recognition and role (i.e., FE) labeling.

Finally, we envision operationalizing the generation of ontology instances of metonymy by unpacking types of metonymy in the ontology itself. Currently, to the best of our knowledge, no ontology includes the explicit indications of metonymy that this would require, but ongoing work [7] is moving in this direction.

We are confident that the technique we use for BioFN scales well to other domains. Domain-specific lexical resources that are linked to domain-specific ontologies – under the roof of an upper lexical resource (like FrameNet), an upper ontology (like SUMO), and modeled using a common formal language (like OWL DL) – seem to be a reasonable approach to natural-language understanding. Thus, in the long run, we see FrameNet as a backbone of several domain-specific FrameNets that in turn are linked to domain-specific ontologies.

## Acknowledgments

This work is supported in part by grant #5R01-LM008111-02 from the National Library of Medicine.

## References

1. C. J. Fillmore. Frame semantics and the nature of language. *Annals of the New York Academy of Sciences*, (280):20-32, 1976.
2. M. Wessel and R. Möller. A high performance semantic web query answering engine. In *Proc. International Workshop on Description Logics*, 2005.

<sup>13</sup> WMD = Weapons of Mass Destruction.

<sup>14</sup> See updated FN FAQ for further discussion of this point, at <http://framenet.icsi.berkeley.edu/index.php?option=content&task=catalogorv&sectionid=11&id=86&Itemid=49>.

3. I. Horrocks. The FaCT system. In H. de Swart, editor, *Automated Reasoning with Analytic Tableaux and Related Methods: International Conference Tableaux'98*, number 1397 in *Lecture Notes in Artificial Intelligence*, pages 307-312. Springer-Verlag, May 1998.
4. J. Ruppenhofer, M. Ellsworth, M. R. Petruck, and C. R. Johnson. *FrameNet: Theory and Practice*. ICSI Berkeley, 2005. <http://framenet.icsi.berkeley.edu>.
5. Z. Lu. Mining protein transport data from GeneRIFs. University of Colorado, Center for Computational Pharmacology presentation, 2006.
6. J. Scheffczyk, C. F. Baker, and S. Narayanan. Ontology-based reasoning about lexical resources. In *Proc. of OntoLex 2006: Interfacing Ontologies and Lexical Resources for Semantic Web Technologies*, pages 1-8, Genoa, Italy, 2006.
7. J. Scheffczyk, A. Pease, and M. Ellsworth. Linking FrameNet to the suggested upper merged ontology. In *Proc. of FOIS 2006*, Baltimore, MD, 2006. to appear.
8. T. Wattarujeekrit, P. K. Shah, and N. Collier. PASBio: predicate-argument structures for event extraction in molecular biology. *BMC Bioinformatics*, 5:155, 2004.
9. P. Kingsbury and M. Palmer. From Treebank to PropBank. In *Proceedings of 3rd International Conference on Language Resources and Evaluation (LREC2002)*, 1989-1993. Las Palmas, Spain, 2002.
10. A. Korhonen, Y. Krymolowski, and N. Collier. Automatic Classification of Verbs in Biomedical Texts. To appear in *Proceedings of ACL-COLING 2006*. Sydney, Australia, 2006.
11. B. Levin. *English Verb Classes and Alternations*. Chicago University Press, Chicago. 1993.
12. K. Kipper, H. T. Dang, M. Palmer. Class-Based Construction of a Verb Lexicon. *AAAI/IAAI 2000*: 691-696. North Falmouth, MA, 2000.
13. C. F. Baker and J. Ruppenhofer. FrameNet's Frames vs. Levin's Verb Classes. In J. Larson and M. Paster (Eds.), *Proceedings of the 28th Annual Meeting of the Berkeley Linguistics Society*. 27-38. 2002.
14. T. Schmidt. Interfacing lexical and ontological information in a multilingual soccer FrameNet. In *Proc. of OntoLex 2006: Interfacing Ontologies and Lexical Resources for Semantic Web Technologies*, pages 75-81, Genoa, Italy, 2006.
15. P. Buitelaar et al. Generating and visualizing a soccer knowledge base. In *Proc. of the EACL'06 Demo Session*, Trento, Italy, 2006.
16. B. Smith, W. Ceusters, B. Klagges, J. Köhler, A. Kumar, J. Lomax, C. Mungall, F. Neuhaus, A. Rector, and C. Rosse. Relations in biomedical ontologies. *Genome Biology*, 6(5), 2005.
17. A. Meyers, A. C. Fang, L. Ferro, et. al.. *Annotation Compatibility Working Group Report*, Coling/ACL, Sydney, Australia, 2006.
18. S. Sinha and S. Narayanan. Model based answer selection. In *Textual Inference in Question Answering Workshop, AAAI 2005*, Pittsburgh, PA, 2005.
19. K. Erk and S. Padó. Shalmaneser – a flexible toolbox for semantic role assignment. In *Proc. of LREC 2006*, Genoa, Italy, 2006, to appear.

# Posters



## **A CPG-Based Ontology Driven Clinical Decision Support System for Breast Cancer Follow-up**

**Samina Raza Abidi Health Informatics Laboratory,  
Dalhousie University, Halifax, Nova Scotia.**

abidi @cs.dal.ca

### **Abstract:**

Breast cancer is the most common cancer among women in Canada. Due to recent advancements in treatment and diagnosis, more women are surviving breast cancer than ever before and breast cancer survivors are the most prevalent female cancer survivor group in Nova Scotia. As a consequence, the delivery of long-term follow-up care, which has traditionally been provided at the specialized cancer clinics, places a strain on specialist resources. However, there is evidence that family physician follow-up of women with breast cancer who are in remission is a safe and viable alternative to follow-up in the cancer centers. Therefore, there is an incentive to the transfer of breast cancer follow-up care to family physicians after primary treatment is completed by specialists. Notwithstanding, the benefits of such a transfer of services from the tertiary to the primary care centers the main issue is the transfer of specialized breast cancer follow-up care knowledge to family physicians expertise. In this regard, as a first step, Cancer Care Nova Scotia has developed a Breast cancer follow-up clinical practice guideline for use by the family physicians. Yet, the adoption of the said CPG is a challenge in the clinical setting.

We have developed a Clinical Practice Guideline (CPG) based interactive decision support system for the family practice setting to guide family physicians conducting breast cancer follow-ups. The idea is to computerize the breast cancer CPG and then operationalize it using patient data to

assist the practitioner to make CPG mediated decisions, recommendations and referrals. In order to achieve the above functionality we have seamlessly integrated the CPG with patient data through electronic interface for collecting patient information. The implementation of the system is achieved in three main steps. In the first step we have converted the breast cancer follow-up CPC in electronic format using the Guideline Element Model (GEM). In the second step we use the logic in the conditional statements of the CPG, to develop a domain ontology using Protégé. Finally in the third step the guideline is executed using the execution engine developed in Health Informatics Laboratory at Dalhousie University. The rule authoring and execution modules of the execution engine is used to develop IF and THEN forward rules with a list of decision variables followed by IF part and list of action variables followed by THEN part of the rule and executing them using the patient specific data.

The next steps are the deployment of the clinical decision support system within two clinics in Nova Scotia, followed by an evaluation study to measure the efficacy of the CDSS in terms of providing point-of-care support to family physicians conducting breast cancer follow-up.

**Keywords:** Breast cancer follow-up, clinical practice guideline, domain ontology, Guideline Element Model, clinical decision support system.

## Inferring Gene Ontology category membership via gene expression and sequence similarity data analysis

**Murilo Saraiva Queiroz, Francisco Prosdocimi, Izabela Freire Goertzel, Francisco Pereira Lobo, Cassio Pennachin and Ben Goertzel, Ph.D,**  
**Biomind LLC, Rockville, MD/USA**

The Gene Ontology (GO) database annotates a large number of genes according to their functions (the biological processes, molecular functions and cellular components in which they are involved). However, it is far from complete, and so there is a need for techniques that automatically assign GO functional categories to genes based on integration of available data. The present work describes one such technique, that uses a combination of sequence similarity and a similarity measure based on mutual information applied to cross-experiment microarray gene expression analysis.

First of all, in order to test the relevance of sequence similarity for gene function inference, similarity searches of genes belonging to the same GO (from here on we will use "GO" as a shorthand for "GO category", as well as for the "Gene Ontology" as a whole) were done across the human genome. A BLAST attachment value (BAV) for each GO was defined as the sum of the e-value exponents found between pairs of genes in the GO, divided by the sum of all e-value exponents found between genes in the GO and genes outside the GO.

Next, to assess the "expression based similarity" of human genes, we used a dataset (GDS181) from GEO, a gene expression and molecular data repository maintained by the NCBI, providing gene expression profiles from 85 different tissues, organs, and cell lines in the normal physiological state. The dataset contains 12,625 probes, and we used 9,725 of them associated to genes with identifiable GO relationships. For each gene in the dataset, we calculated the Mutual Information (MI) between its expression values measured across all tissues and the corresponding values for the other genes. In order to calculate MI, the gene expression values were discretized, meaning that each one was replaced by one of K symbols. The symbol replacing an expression value was calculated by first normalizing the values into [0,1], and then partitioning this interval into K equally sized subintervals. The normalization was done on a per-gene basis. After experimenting with several different values of K, a value of K=3 was chosen for all further experiments. Using a similar procedure to the one used for calculating the BAV, a MI attachment value (MiAV) was obtained. For each GO, the MiAV was defined as the sum of the MI expression values found for all

pairs of genes in the GO, divided by the sum of all MI values between genes in the GO and genes outside the GO.

Then, our gene function inference (GFI) process proceeds as follows. Given a gene for which one wants to know the function, one begins by comparing it with all other genes, using both BLAST and expression data. Then, given a GO, one may calculate the values Bs and Ms, representing the maximum similarity found between the query gene and any gene inside the GO, using BLAST or MI, respectively. Those values plus attachments are used in the following equation for estimating the pertinence of a gene to a GO:

$$(I) \quad f(Bs, BAV, Ms, MiAV) = \frac{x_1(Bs^{y_1} BAV^{y_2}) + x_2(Ms^{y_3} MiAV^{y_4})}{z}$$

Here,  $x_1, y_1, y_2, x_2, y_3, y_4$ , and  $z$  represent the weights of the equation. A gene should be classified as belonging to a GO if the equation above gives a value greater than zero when fed similarity and attachment values derived from the GO. A genetic algorithm was used to optimize the weights of this formula, based on assessing the performance of each weight-vector at predicting GO category membership over a training set. The objective function used by the GA was based on the F-measure, which takes into account both precision and recall.

A rigorous testing methodology was utilized. We set aside a subset of the GO and a subset of our overall human genome dataset to train our the genetic algorithm involved in our GFI model. Another subset of the GO and of the human genome dataset was used for testing; and further validation was obtained by applying the parameters learned with human data to the yeast genome. Yeast expression data was composed of the familiar Spellman dataset, and corresponding sequence data from SGD.

In our computational experiments, 2,386 new links were predicted between human genes and GO categories; and 1,111 links between yeast genes and GO categories, spanning the biological process, molecular function and cellular component ontologies. According to tests using the method to replicate already-known GO category assignments, the results are estimated to have precision bounded below by 73% for human data, and 83% for yeast.

# Experience with an Ontology of Pediatric Electrolyte Disorders in a developing country

Vorapong Chaichanamongkol, B.Sc., M.D., Wanwipa Titthasiri, M.Sc., Ed.D.

Department of Information Technology, Rangsit University, Bangkok, Thailand  
Email: vorapongch@yahoo.com

## Motivation

There are few pediatric nephrologists in Thailand and physicians in rural Thailand have limited access to up-to-date biomedical information such as biomedical journals. Moreover, biomedical information is often compiled in developed countries and may not be appropriate for use in developing countries. For example, the guidelines established by the World Health Organization (WHO) for the treatment of acute diarrhea in children are not always applicable in the case of rotavirus gastroenteritis, because the concentration of sodium in oral saline solutions is too high for infants. In this context, we believe ontologies can play an important role in patient management. Semantic Web ontologies foster sharing and reuse of knowledge and facilitate collaboration between pediatricians and consultant pediatric nephrologists. Such ontologies can be part of the telemedicine arsenal and help physicians in rural areas of Thailand to better manage difficult cases. In this paper, we report our experience in developing and using an ontology of pediatric electrolyte disorders.

## Developing and publishing the ontology

In the knowledge elicitation phase, we used concept maps to formalize the knowledge of a small group of eleven experts. Knowledge was contributed by pediatric nephrologists, pediatricians, general practitioners, as well as extracted from clinical practice guidelines, text books and the medical pediatric literature. Some 500 concepts were identified in the domain of pediatric electrolyte disorders.

These concepts were then organized into an ontology and related to other concepts. Textual definitions were created. For example, the concept *severe hyponatremia* is defined as “Sodium concentration is below 125 mEq/l” and is a subclass of the concept *disease*. In addition to subclass relations, we use the relationship “look for” between diseases and symptoms. Another example is the concept *Urine Sodium concentration*, subclass of *Urine test*, and for which an important property is “more than or less than 20 mEq/l”. The Web Ontology Language OWL-DL was selected for representing the ontology. In this phase, we used Protégé-OWL (<http://protege.stanford.edu/>, Stanford University and University of Manchester) and SWOOP (<http://www.mindswap.org/2004/-SWOOP/>, University of Maryland) for building a

prototype of the Pediatric Electrolyte Disorder Ontology. There is a reliable ADSL network in the capital Bangkok. Protégé and Swoop installed on a web server are used to publish and share the ontology. The same applications are also installed on client computers and can be used both online and offline.

## Usability study

In order to evaluate the ontology, questionnaires were sent to the 25 end users, i.e., general pediatricians and family physicians in rural areas. 24 responses were received and analyzed. Most physicians involved with the study were young (age 30-40).

The Jambalaya plug-in in Protégé was found to provide good visualization support, displaying 2D interactive representations of the domain of the electrolyte disorders. Some pediatricians liked Swoop publishing, because it is easy to understand, especially in Text mode.

Speed was sometimes an issue in those areas with a large number of ADSL users. Even lower connectivity was available in rural areas. In some cases, the ontology had to be sent by email in several pieces. However, once downloaded, the OWL ontology can be exploited offline, using Swoop or Protégé.

## Conclusions

The development of our ontology of pediatric electrolyte disorders took more than one year. It was motivated by the need for providing up-to-date therapeutic to general practitioners in this specialized domain, and to tailor this information to the particular patient population. Preliminary results show that the ontology has helped physicians better manage pediatric patients, especially in the rural areas of Thailand. Despite limited connectivity in some areas and limited performance of computer systems, the experience was globally successful, in both creating the ontology from expert knowledge and making it available to physicians in rural areas. Ontologies such as the one we created for pediatric electrolyte disorders will play an increasing role in telemedicine.

In future work, we plan to build a larger Semantic Web Ontology for Pediatric Nephrology. Rule languages such as SWRL – the Semantic Web Rule Language – may be used in addition to OWL in order to represent clinical guidelines.

## Issues in Representing Biological and Clinical Phenotypes Using Formal Models

Ying Tao<sup>1,\*</sup>, M.D., Chintan Patel<sup>1,\*</sup>, M.S, Carol Friedman<sup>1,†</sup>, PhD, Yves A. Lussier<sup>1,2,†</sup>, M.D.

1- Dept. of Biomedical Informatics, Columbia University, New York, NY, USA

2- Center for Biomedical Informatics and Dept. of Medicine, The University of Chicago, Chicago, IL, USA

*Representing phenotypes in a structured and standardized manner across different biological species poses significant challenges. We performed a modeling experiment to compare a model called the Canon model, and the PATO for representing a range of biological and clinical phenotypes. The formal nature of Canon model allows for complex representations, but lacks the simplicity offered by PATO. A phenotype model allowing flexible representation with unique semantic interpretation is desired.*

### BACKGROUND

The Phenotype Attribute and Value Ontology (1) (PATO) is an emerging standard to annotate assayed phenotypes in a structured and coherent manner across different biological species. Canon group (2) developed a model for the formal (canonical) representation of clinical information for data exchange and medical applications.

### METHODS

We selected a diverse set of phenotypes from Wormbase, OMIM and chest radiology report (radiographic findings/phenotypes). We then evaluated the PATO and Canon models by encoding the phenotypes into each model.

### RESULTS AND DISCUSSION

Examples of the phenotype modeling experiment are described in table 1. The flexibility of choosing the

entity from an external ontology in PATO can lead to multiple representations, for example, *vulval\_differentiation* (mammalian phenotype ontology) or *vulva* (anatomy ontology); it is not clear how semantic equivalences can be inferred from such representations. Developing a symbolic model that can represent and reason with complex concepts such as 'penetrance' is challenging. Furthermore, concepts having deep nested structures need a more formal representation framework to capture the knowledge at finer granularity (e.g. *slight interval decrease*). The Canon model with its logic based representation allows for formal and complex representations but the familiarity and acceptance of such a model among end-users remains an open issue. We conclude that using PATO with a formal description logic language, as the one provided in Canon, would provide a more expressive and less ambiguous framework for representing clinical and biological phenotypes, however additional studies are required to evaluate the usability aspects of the combined model.

### References

1. [www.bioontology.org/wiki/index.php/PATO:Main\\_Page](http://www.bioontology.org/wiki/index.php/PATO:Main_Page)
2. Friedman C, Huff SM, Hersh WR, Pattison-Gordon E, Cimino JJ: The Canon Group's effort: working toward a merged model. J Am Med Inform Assoc 2:4-18 (1995)..

**Table 1.** Modeling biological and clinical phenotypes using the PATO and the CANON model

Phenotype	PATO (observable entity   attribute   value) Note: the latest version of PATO does not have notion of 'attributes' (1)	CANON Model (conceptual graph)
<i>negatively regulates vulval differentiation (WormBase)</i>	Vulva Differentiation   regulation   negative	[phenotype: #ark1Fun] - (has-observation) → [differentiation] (has-location) → [vulval] (has-process*) → [negatively regulated]
<i>Cystic Fibrosis with pancreatic insufficiency in 80% (OMIM)</i>	Cystic Fibrosis Pancreas   enzyme_function   Insufficient*	[phenotype: MIM:219700]- (has-observation) → [enzyme function] (has-location) → [pancreas] (has-degree) → [insufficient] (has-penetrance) → [80%]
<i>Slight interval decrease in left pleural effusion (Radiology Report)</i>	Pleural effusion   local_qualifier   left Pleural effusion   temporal   decrease Left Pleural Cavity   pathological change   pleural effusion Left Pleural Cavity   temporal   decrease	[phenotype: # BWH22.09] - (has_observation) → [pleural_effusion] (has_location_qualifier) → [left] (has_temporal) → [decrease_in] - (has_degree) → [slight] (has_temporal) → [interval]

\* represents concepts not present in the model

\* These authors contributed equally to the work    † Corresponding authors who have contributed equally

# Notes

---

# Notes

---

# Notes

---

# Notes

---

# Notes

---