

Project Acronym:	<i>SemanticHEALTH</i>
Project Title:	Semantic Interoperability Deployment and Research Roadmap
Contract Number:	IST-27328-SSA
Starting date:	January 01, 2006
Ending date:	June 30, 2008

Deliverable Number:	D 6.1
Title of the Deliverable:	Barriers, approaches and research priorities for integrating biomedical ontologies
WP/Task related to Deliv.:	WP 6, Task 6.1
Type:	Restricted
Authors:	<i>Alan Rector</i>

Partners:	empirica Information & Communication Technology Research mbH Egészségügyi Stratégiai Kutatóintézet - National Institute for Strategic Health Research University College London Université Jean Monnet Saint Etienne Radboud University Nijmegen Medical Centre Uppsala Universitet World Health Organization University of Manchester
------------------	---

Project Co-ordinator

Organisation name: empirica Information & Communication Technology Research mbH
Name of representative: Simon Robinson
Address: Oxfordstrasse 2, D-53111 Bonn, Germany
Phone number: +49.228.98530.0
Fax number: +49.228.98530.12
E-mail: simon.robinson@empirica.com
Project WEB site: <http://www.semantichealth.org/>

Document History

Name	Date	Version	Description
Alan Rector	November 2007	0.10	Draft outline
Benjamin Jung	09 January 2008	0.10	Draft structure and bullet points from Alan Rector
Benjamin Jung	18 January 2008	0.12	Edits
Alan Rctor	28 Jan 2008	0.13	Rough draft for input from consortium
Alan Rector with input from Anand Kumar, JM Rodrigues, and other members of the consortium	8 Feb 2008	0.14	Partially consolidated draft with input from consortium members
Alan Rector with input from Anand Kumar, JM Rodrigues, and other members of the consortium	10 Feb 2008	0.15	Penultimate draft
Veli Stroetmann	11 Feb. 2008	Final	Final version, delivered to the EC

Contractual Date of Delivery to the EC: February 2008

Actual Date of Delivery to the EC: February 2008

Table of contents

0	Executive Summary	6
0.1	Major facts for terminologies and ontologies in healthcare in the next ten years.....	6
0.2	General Principles	6
0.3	Recommendations.....	6
1	Introduction	8
1.1	The need for clinical terminologies.....	8
1.2	Vocabulary.....	8
1.3	Interaction with Electronic Health Records and Clinical Decision Support.....	9
1.4	Interaction with molecular biology and translational medicine	10
1.5	The meaning of the word “Ontology”.....	10
1.5.1	Ontology sensu philosophy and Ontology sensu informatics	10
1.5.2	Ontologies and “compositionality”.....	11
1.5.3	The “Ontological Spectrum”.....	11
1.5.4	Logic based ontologies and description logics.....	12
1.5.5	Ontologies, Background Knowledge Resources, and Data Models	12
1.5.6	Summary of closely related notions.....	13
1.6	Issues in the development of clinical terminologies.....	14
1.6.1	Use cases	15
1.7	Visions for the future.....	17
1.7.1	Desirable outcomes.....	17
1.7.2	Inevitable outcomes	19
1.7.3	Outcomes to be avoided.....	20
2	Background	21
2.1	Historical perspective.....	21
2.2	Quality Indicators and Academic Studies.....	25
2.2.1	Internal measures.....	25
2.2.2	Human factors and Inter-rater reliability.....	26
2.2.3	Scaling and Zipf’s law usage	26
2.3	Technologies	27
2.3.1	Terminologies and Ontologies.....	27
2.3.2	Digital libraries and Scholarly Resources	29
2.3.3	Statistical and Social: Google, Web and text mining.....	29
2.3.4	Authoring Tool & Browsing sets.....	30
2.3.5	Infrastructure and software.....	31
2.3.6	Bioinformatics and Molecular Medicine	32
2.3.7	Social computing demonstrations from related fields.....	33

3	Analysis	34
3.1	Key determinants of adoption and success	34
3.1.1	Regulatory or remuneration requirements	34
3.1.2	Commercial requirements – de facto and de jure standards that are supported by industry to the extent that all systems must conform to be commercially viable.	34
3.1.3	Major international agreement and historical establishment	34
3.1.4	Community commitment and support	35
3.2	Key Players and “Facts on the Ground”	35
3.2.1	Clinical medicine and medical classifications	35
3.2.2	Bibliographic resources and cross referencing – US NLM, UMLS & PubMed	37
3.2.3	Drug Information: Commercial Drug Databases and National Standards	37
3.2.4	Translational medicine and Genomics	38
3.3	Critical barriers	40
3.3.1	Socio-political	40
3.3.2	SNOMED related barriers	41
3.3.3	Skills and Human Capacity	43
3.3.4	Tools and Software	43
3.3.5	Special technical issues	43
3.4	Issues critical to Europe	44
3.4.1	The limits of top-down development and danger of “hubris”	44
3.4.2	Multilingual and multicultural systems	44
3.4.3	Large scale participation and openness	45
3.4.4	Coping with the diversity of administrative and organisational systems	45
3.4.5	Rapid evolution and “just in time” deployment	45
3.4.6	The standardisation process	45
3.5	Open questions and research needs	46
3.5.1	Effective re-use	46
3.5.2	Binding of terminology and EHRs	47
3.5.3	Human factors	47
3.5.4	Natural language use for compiling ontologies	47
3.5.5	Integration of Semantic, Statistical Methods, and Social Computing Models: How best to exploit Google, Web 2.0 Text Mining and related tools	47
3.5.6	Integration of terminology and background information resources	47
3.5.7	Other technical issues	48
3.6	Gaps and development needs	48
3.6.1	Central services	48
3.6.2	Coordinated input to standards bodies and involvement of Providers and Vendors in the standards process	48
3.6.3	Collaborative development	48
3.6.4	Effective collaboration on translational medicine	49
3.6.5	Effective exploitation of the new technology flowing from Web 2.0 and the Semantic Web	49

4	Action plan and recommendations for the Semantic Interoperability Roadmap	50
4.1	Choices	50
4.1.1	Do Nothing	50
4.1.2	Focused areas only	50
4.1.3	Facilitation and support for common requirements	50
4.1.4	Full fledged support for common, possibly mandated standards	50
4.2	Principles	51
4.2.1	Purposes: What's it for	51
4.2.2	Priorities: What must be done now? What can be built on now?.....	51
4.2.3	Involvement of providers and vendors: Who will deliver the results?	51
4.2.4	Sustainability: Any action must be for the long term	51
4.2.5	Openness & collaboration: The terminologies that have succeeded in bioinformatics have all been developed on an open basis	52
4.2.6	E-Terminology and the Web:.....	52
4.3	Long term goals to achieve desirable outcomes	52
4.3.1	Target capabilities of new-generation ontologies and terminologies	52
4.3.2	Content	52
4.3.3	Tools	53
4.3.4	Process	53
4.4	Priorities for the next five years: Recommendations	54
4.4.1	Content	54
4.4.2	Tools and Capacity.....	54
4.4.3	Process	55
4.4.4	Clinical research and Translational Medicine.....	56
5	Glossary	57
6	Appendix	58

0 Executive Summary

0.1 Major facts for terminologies and ontologies in healthcare in the next ten years

- Most data about patient care are likely to be collected, if at all using idiosyncratic coding systems specific to each country or dictated by rules for remuneration or regulation.
- SNOMED-CT has now achieved sufficient momentum that there is unlikely to be an alternative for an international classification. However, SNOMED is predominantly an Anglophone organisation and is unlikely to respond to the requirements for multilingual or multicultural use without resources from Europe.
- HI7 will dominate the standards for messaging. The standard for EHRs is likely to be a combination or amalgam of the HL7 CDA and Archetype based standards from *OpenEHR*, CEN EN 13606,
- Terminologies from biomedicine, particularly the Gene Ontology and the associated ontologies in the Open Biomedical ontologies consortium will become of increasing importance to clinical medicine
- Web based initiatives and social computing will become increasingly important. Following the example of the bioinformatics community, open systems “owned” by their community are likely to make an increasing contribution.
- The lack of tools and human capacity in clinical terminology is a major limiting factor that should be addressed.

0.2 General Principles

- Terminologies and ontologies need to have well-established purposes and criteria for determining whether they are fit for those purposes. Development without clear criteria tends to become open ended and indefinitely expensive.
- Providers and vendors must be more heavily involved. They are the ones who must deploy the systems; they are the ones who must ‘interoperate’.
- Any action taken must lead to long-term institutions that can be sustained. The only constant in healthcare or biomedical research is change.
- Systems that do not involve key users are unlikely to succeed. Systems need to be open and “owned by those who use them. They need to be highly responsive.
- Multilingual and multicultural systems are of particular importance to Europe. They will not come into existence without significant resources being invested by European countries and/or the EC.

0.3 Recommendations

- Support for specific initiatives between ICD/WHO, ISO and SNOMED to develop ICD-11 along with mappings to and from SNOMED
- A feasibility study of a well curated and internationalised subset of SNOMED-CT not to exceed 25,000 concepts, to be followed by careful consideration of how to proceed, or not, on the next steps of collaboration and internationalisation.
- Investment in tools and capacity for collaborative terminology development, building on work already sponsored by the EC where possible, and collaborating with other international groups developing open and freely available software.

- Linking investment in tools with a programme of training and capacity building
- Encouraging user involvement and ownership through a series of centres and initiatives
- Providing a series of centres and web resources for collaborative involvement of end-users – different from those required for primary authors – and linked to mechanisms for highly responsive “just in time” delivery of material to encourage use of standard systems
- Establishment of specific collaborations with key organisations for translational medicine in the various member states and with in particular the European Bioinformatics Institute and the US National Cancer Institute.

1 Introduction

1.1 The need for clinical terminologies

Interoperability requires agreement on meanings and labels for those meanings – on ontology and lexicons, which together we label as terminology.

Development of standard terminologies for medicine dates back to the mid nineteenth century, but the major initiatives that concern us date from the beginning of wide spread use of computerised systems in the 1970s and 1980s, with major developments in the last decade.

The primary goal of ontologies and terminologies for interoperability is to enable the faithful exchange of meaning between machines and between machines and people. The test of whether or not they are successful is whether or not they perform this task effectively.

The goal of standard ontologies and terminologies is to make it easier to build systems that successfully exchange meanings. Currently, agreeing meanings is one of the most time consuming and resource intensive tasks in setting up clinical systems. Agreeing meanings between systems is a major barrier to system interoperability. The goal of improved standard terminologies is to reduce that barrier – to reduce the effort required to establish effective exchange of meaningful information between two systems.

Recently the use of the term ‘ontology’ has been steadily growing, and the question arises whether this constitutes any real advance or advantage? There is indeed good reason to cast at least some doubt on the claims made on ontology’s behalf: Too many recent publications, calls for research proposals and project descriptions have embodied what are in our view (and have sometimes already proven themselves to be) insupportable expectations. It is thus understandable that some have been tempted to see in ontology just one more new and flashy buzzword.

We therefore first lay out a vocabulary and issues and then realistic goals for the use of ontologies and terminologies as part of a strategy to achieve semantic interoperability.

1.2 Vocabulary

The vocabulary surrounding terminologies and ontologies is confusing, and different authors use the same words differently. We provide a glossary at the end of this deliverable. However, as an aide to understanding this paper we provide the following definitions for the terms as we shall use them.

- *Controlled Vocabulary* – a list of specified items to be used for some purpose, usually in an information system to reduce ambiguity, misspellings, etc.
- *System of identifiers (“codes”)* – Controlled vocabularies, and many lexicons, ontologies, and thesauri, are usually accompanied by systems of identifiers for their units, e.g. Typically, identifiers act as the primary unambiguous means of referring to the entities in the system for computational purposes with the text form being used for communication with users. Examples the “Concept Unique Identifiers (CUIs) from the UMLS, SNOMED Identifiers, etc. In many contexts, identifiers are known as “codes.”
- *Lexicon* – A list of linguistic units that may be attached to a controlled vocabulary or ontology, in a specific language or sublanguage, often including linguistic information such as synonyms, preferred terms, parts of speech, inflections and other grammatical

material. Example: Term terms and lexical material in UMLS identified by Lexical Unique Identifiers LUIs)

- *Ontology* (*sensu* information system) – a symbolic logical model of some part of the meanings of the notions used in a field, *i.e.* those things that are universally true or true by definition.¹ The key relationship in an ontology is “subsumption” or “kind-of”. Every instance of a subkind must be an instance of the kind, without exception. Typically, ontologies are implemented in logic languages such as Ontolog or OWL or frame systems such as Protégé-Frames. Examples: The GALEN Core Model, the stated form of SNOMED.
- *Classification* – an organisation of entities into classes for a specific purpose such as international reporting or remuneration. Examples ICD and Diagnosis Related Groups.
- *Thesaurus* – a system of terms organised for navigation with the primary relationship being “broader than”/“narrower than”. The “broader than”/“Narrower than” relation is explicitly not limited to subsumption/kind of relation. It is a general form of linguistic hyper/hyponymy aimed at assisting human navigation. However, it is explicitly not intended that it be used as the basis for logical inferences, *e.g.* in decision support. Examples MeSH, WordNet.
- *Knowledge Representation System / Background knowledge base* – the common knowledge to be assumed by the system, including both the ontology – what is universally true – and generalisations about what is typically true.
- *Terminology* – Any or all of the above in various combinations. Most health terminologies consist, at a minimum, of a controlled vocabulary and a system of identifiers. They may include extended lexicons, ontologies, thesauri or background knowledge base. This definition is deliberately broader and less specific than that in most of the standard references and intended to approximate common usage.
- *Coding system* – A terminology with attached identifiers or “codes”.

Some of the often fraught issues around the use of the word “ontology” are further discussed in Section 1.5.

1.3 Interaction with Electronic Health Records and Clinical Decision Support

Terminologies and ontologies interact intimately with Electronic Health Records and Clinical Decision Support Systems:

- Terminologies form the interface between human users and machines, and must support the requirements of both.
- Terminologies provide the key interface between Electronic Health Records (EHR) and Clinical Decision Support (CDS). Since the early days of clinical decision support, sharing terminology has, surprisingly, proved more difficult than sharing rules, algorithms, and heuristics.
- Medical record models define the structures in which information is carried. Terminologies define the meaning or content of the information carried by those structures.
- Decision support defines the use to be made of the meanings carried by the structures, and requires a coherent view of both. The Background Knowledge Resources or

¹ Different authors refer to the meanings as “concepts”, “universals,” “categories”. Note that the word “ontology” was borrowed from philosophy, and that there remain controversies concerning the extent to which the symbolic models referred to as ontologies used in information systems should conform to principles laid down by philosophers for ontologies understood as part of the philosophical study of being.

Knowledge Representation System used by Decision support system is usually built around the framework of an ontology.

- “Ontology based Architectures” may use the framework of the ontology plus background knowledge resources to generate software for information capture, decision support, and other purposes, as the next step beyond “model driven architectures”. “Ontology driven architectures” have the advantage of being able to use strongly logical frameworks and thus allow more extensive validation than typical model driven architectures.
- Different groups have, in general, been responsible for developing standards for Terminologies, Electronic Health Records (EHRs), and Clinical Decision Support Systems (CDSs). Although there is interaction between the groups, overall coordination between them is poor.

1.4 Interaction with molecular biology and translational medicine

A major, perhaps the major, feature of medicine in the first part of the twenty-first century is the application of new knowledge from molecular biology to clinical medicine. The “translation” of basic scientific understanding to clinical practice is happening at a faster pace than anyone could have imagined a decade ago. This is resulting in a rapid increase in the rate of clinical trials and “Bio-Banking” or various sorts – the prospective collection of samples of either normal or diseased tissue to be analysed later in the light of new analyses and knowledge not available at the time the material was collected. There are several consequences for ontologies and terminologies:

- Clinical terminologies must increasingly interact with terminologies from molecular biology, particularly those developing from the Gene Ontology (GO), now known as the Open Biomedical Ontologies (OBO) and its OBOFoundry.
- Clinical terminologies for patient care must increasingly interact with data standards – e.g. BRIDG –for clinical trials and currently a major controversy over the emergence of a standard ontology for clinical trials
- There is a rapid development of a parallel set of ontologies in the field of cancer at the US National Cancer Institute and an alternative technology for managing terminology in the NCI EVS.
- The US has funded a National Center for Biomedical Ontologies (NCBO)

These developments bring in a new set of players and stakeholders that will only be reconciled with significant effort.

1.5 The meaning of the word “Ontology”

1.5.1 Ontology sensu philosophy and Ontology sensu informatics

Knowledge Representation Systems have a long history in software development and artificial intelligence reaching back into the late 1960s. The word “ontology” was first borrowed for the part of the knowledge representation system represented in “frames” or logic in the early 1990’s by Gruber and his colleagues.² However, the meanings as used in philosophy and informatics are subtly different.

² Gruber TR. Toward Principles for the Design of Ontologies Used for Knowledge Sharing. *Journal of Human-Computer Studies*. 1993;43:907-928.

- Ontology (sensu philosophy): The study of what there is. Formal ontologies are theories that attempt to give precise formulations of the types of entities in reality, of their properties and of the relations between them.³
- Ontology (sensu informatics) Originally “The conceptualisation of the entities in a domain”. In a more modern formulation: “A symbolic logical model of some parts of the meanings of the notions used in a field, those things which are universally true or true by definition.” (See 1.2)

The difference is important. In philosophy, ontology is the study of what is; in informatics, ontology is the study of what is to be represented – and by extension the means of representation. The test of an informatics ontology is whether or not it is useful in information systems. This may, or may not, correspond to what any given school of philosophy considers to exist “in reality”.

There is continuing controversy concerning which principles from philosophical ontology it is appropriate to apply to ontologies as used in informatics. On the one hand, the fruit of thousands of years of thought since Aristotle often throws light on difficult issues; on the other hand, the distinctions and restrictions advocated by philosophical ontologists often seem at best irrelevant and on occasion actively counterproductive to the use of ontologies in information systems. (The most vociferous advocate of the philosophical approach is Smith⁴)

1.5.2 Ontologies and “compositionality”

Just as in human language, where a dictionary contains the basic (and relatively stable) building blocks for constructing complex phrases and sentences (but does not enumerate these phrases and sentences, an impossible task due to combinatorial explosion), a domain ontology contains the basic entities that form the building blocks of the representation of the domain of discourse.

A major issue in clinical systems is whether or not they use compositional ontologies/terminologies, and if so whether they allow “post-coordination”, *i.e.* whether they allow simpler entities to be composed into more complex entities rather than requiring all possible entities to be enumerated in advance.

1.5.3 The “Ontological Spectrum”

Since the word “ontology” was adopted into informatics by Gruber, it has been used in a wide range of different ways to cover a wide range of different information artefacts. Depending on one's point of view, “ontologically based” seems often to have been used as a synonym either for “good” or “bad”.

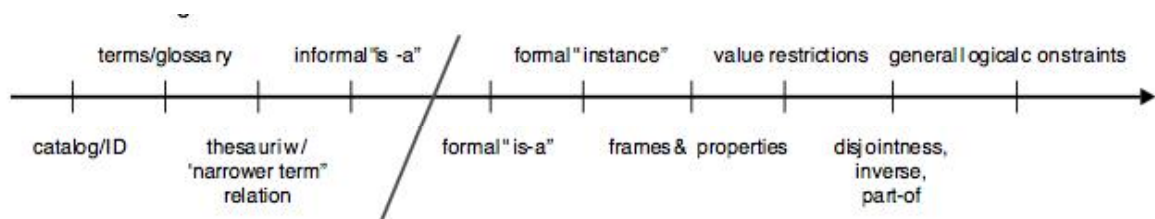
A useful notion is that of the “Ontological Spectrum” introduced by Welty and McGuinness and illustrated in Figure 1. It points out a range of artefacts with different levels of formality and commitment to different purposes and functions.

In this deliverable we shall use the term “ontology” in a restricted way, to refer to ontologies sensu informatics formulated using formal logical tools such as description logics (See Sections 1.5.4 2.3.1)

³ Quine WV. On what there is. The review of metaphysics. 1948;2:21

⁴ <http://ontology.buffalo.edu/smith/>

FIGURE 1: THE ONTOLOGICAL SPECTRUM OF WELTY, MCGUINNESS ET AL.⁵



1.5.4 Logic based ontologies and description logics

Ontologies sensu informatics are often identified with those things that can be represented in description logics which are specialized subsets of logic designed to be adequate for describing things but still relatively tractable with respect to inference.

What is meant by representing “meaning”? As a fundamental principle, in both philosophical and description logic ontologies, *all* assertions about a given type are true for *all* instances of this type. Thus, all instances of the type appendectomy are performed on some instance of appendix; all instances of water molecules contain oxygen and hydrogen. Interpreted strictly, this restricts the ability of an ontology to express seemingly obvious assertions such as “*hands have thumbs*” or “*aspirin alleviates headache*”, because there are hands without thumbs and not all aspirin tablets are used to alleviate any headache.

A further restriction is that probabilistic assertions cannot be expressed in an ontology in a simple way. For example, if a prevalence of 1 % is ascribed to lung cancer then this is not a property inhering in any instance of lung cancer. It is rather a factual statement about some given population with respect to the occurrence of this disease. However, the common consensus in science is in many areas based on probabilistic theories which describe results in terms of probabilistic states, processes and events. So is the assessment of risks (of signs, symptoms, and therapies for specific diseases) commonplace in medical practice. (E.g., arterial hypertension is considered a risk for stroke). Unsatisfactorily, the related entity types and relations cannot be straightforwardly represented in formal ontologies following the principles described above.

These fundamental constraints are corollaries of the fact that all assertions of relations between types in ontologies should be of the basic form of universal statements: “*for all instances of type T there is some...*” We could, of course, consider types and instances as two different ranges for our quantifiers. Then, however, we would have to accept some higher-order logic, which would cause problems for machine reasoning since such logics are known not to be computable in all circumstances using current algorithms. By contrast, languages from the family of Description Logics are computable, and are therefore frequently used in the development of ontologies.

1.5.5 Ontologies, Background Knowledge Resources, and Data Models

The restrictions on what can be said in a strictly logical ontology lead to the distinction between an “ontology” and a “background knowledge resource” or “Knowledge Representation”, although confusingly, the word “ontology” is often used to cover both.

The requirement for most information systems is either

⁵ This spectrum arose out of a conversation in preparation for an ontology panel at AAAI '99. The panelists (McGuinness, Ushold, and Welty), chosen because of their years of experience in ontologies found that they encountered many forms of specifications some people might call ontologies. McGuinness refined the picture.

- To represent the background knowledge that any user can be assumed to have and to be able to reason with it – e.g. that “pneumonia” affects the lung.
- To represent what it is “sensible to say” – e.g. that there is a “fracture of the femur” makes sense but “fracture of the blood” does not – i.e. to represent the rules for composing entities in compositional systems.
- To represent information that is commonly and widely known but which a clinician might need to look up – e.g. that “pneumonia” is commonly caused by “pneumococcus” which is usually sensitive to “penicillin”.

These restrictions lead to the distinction between an “ontology” and a “background knowledge reference resource”. There are very few interesting items of knowledge that are truly ontological in this strict sense. Much current work on informatics ontologies is aimed at integrating probabilistic and typical reasoning with universal “ontological” reasoning effectively.

Hence, the background information for a clinical system often goes well beyond the ontology, in this strict sense. Brachman introduced the notion of the ontology as a “conceptual coat rack” on which other information is held.

The relationship between knowledge representation and ontologies remains controversial and plagued by confusion of substance compounded by loose use of language.

A second closely related notion is that of an “information model” of “model of data structures”. Both Archetypes and HL7 V3 Messages are examples of data structures. Formalisms for data structures bear many resemblances to formalisms for ontologies. The confusion is made worse because the description logics are often used for both. However, there is a clear difference.

- Ontologies are about the things being represented – patients, their diseases. They are about what is always true, whether or not it is known to the clinician. For example, all patients have a body temperature (possibly ambient if they are dead); however, the body temperature may not be known or recorded. It makes no sense to talk about a patient with a “missing” body temperature.
- Data structures are about the artefacts in which information is recorded. Not every data structure about a patient need include a field for body temperature, and even if it does, that field may be missing for any given patient. It makes perfect sense to speak about a patient record with missing data for body temperature.

A key point is that “epistemological issues” – issues of what a given physician or the healthcare system knows – should be represented in the data structures rather than the ontology. This causes serious problems for terminologies coding systems, which often include notions such as “unspecified” or even “missing”. This practice is now widely deprecated but remains common.

“Classifications” – organisations for epidemiological or quality assurance purposes – almost inevitably require some epistemological notions. Hence, the rules for abstracting from an ontology or coding system to a classification system need to include the medical record structure. Patients cannot be classified solely on their pathophysiological state, but must often also be classified in terms of what is known about them, e.g., “survival rate of unknown patients after emergency admission for head injury.”

1.5.6 Summary of closely related notions

We therefore now supplement the definitions in

- *Ontology* – A representation of what is universally true, including what is true by definition
- *Knowledge Representation* or “*Background knowledge resource*”⁶ – a representation of what is generally true, or widely known to be true in some specific instance. In general, the knowledge representation is formulated in terms of and indexed by the Ontology.
- *Information model* or *Data model* a model of how information is structured in a given software system, message, or electronic health record. In general, the data structures carry codes for the ontology as their content.
- *Lexicon* – The linguistic artefacts attached to the entities in the ontology so that the ontology, knowledge representation and information model can be understood by human users.

1.6 Issues in the development of clinical terminologies

The problem of developing widely accepted and used clinical terminologies is made more difficult because it lies at the intersection of a series of issues

- *Disparate tasks* – and the requirements for different tasks are often confused. The entities and requirements for the different sorts of artefact described in the vocabulary in Section 0 are often confused and conflicting. For example, the best organisation for human navigation of a *thesaurus* may be completely unsuitable for use as an *ontology* in clinical decision support.
- *Usability* – terminologies must be usable for clinicians. They are not part of clinicians’ core task of patient care, and must not distract them from that task. While there are claims that issues of usability can be separated from issues of the structure of terminologies *per se*, this claim at best unproven in practice.
- *Scale* – Medicine is big and complicated. Knowledge is fractal. Any worthwhile medical terminology will be large ($\geq 20,000$ entities and perhaps orders of magnitude more), and grow continually. Terminologies are intrinsically combinatorial, and so tend to explode exponentially.
- *Maintenance, evolution, and responsiveness*. This means there are major ongoing as well as startup costs. None will ever be complete, both because medical knowledge changes and because no single enumeration of medical concepts can ever exhaust the combinatorial explosion of concepts implicit in any interesting subset of medical concepts.
- *Requirements to cope with multiple granularities* – from the molecular to the population and societal and with multiple different disciplines and expertise
- *Need to be Multilingual and Multicultural* –. Despite major international efforts, most development has been within national or linguistic groups. The influence of the US National Library of Medicine has provided an enormous boost to the field, but it has also reinforced the strong Anglophone bias of most terminologies in widespread international use.
- *The benefits to standard terminologies accrue to the healthcare system as a whole rather than to any one player*. Everyone wants them, but no one wants to be the one to pay for them.
- There are major controversies about where systems from various points on the ontological along the “ontological spectrum” are appropriate

⁶ “*ase*” is deprecated because of its multiple meanings and connotations.

1.6.1 Use cases

We organise use cases in three dimensions:

- By generic use cases for semantic interoperability
- By geographic scope of interoperability, since the impact of interoperability
- By task

Generic use cases for semantic interoperability – see Deliverable 7.1

- Patient care
- Public Health
- Research and translational Medicine
- Support for diverse Markets
- Cost reductions

Use cases by geographic scope of interoperability

- Locally within hospitals or groups of hospitals – inter-vendor interoperability

Within hospitals, the issues are likely to be care of individual patients and interoperability between vendors' systems. Hospitals would like to be able to "mix and match" "best of breed" systems. They can only do so of those systems interoperable. Vendors' interests are often directly in conflict, since they would rather supply a "unified service".

- Regionally and Nationally – issues of policy, planning, and research

The importance of regional and national interoperability for patient care can be divided into three issues:

- Transfer of information about individual patients. This case always receives much attention, but the amount, detail, and frequency with which information actually needs to be transferred, and on what time scale remains controversial. While the case for relatively low priority transfer of information between systems – e.g. between primary care physicians when patients move – and specific information between hospitals, health centres, and primary care – e.g. prescribing information and discharge instructions. The case for comprehensive national notes remains controversial, with many advocates but no clear consensus that the benefits justify the cost. The amount of information worth transferring is limited by the degree to which different clinical units trust and will re-use each others' information. On the one hand, patients should not be put at risk by ignorance in one institution of what is known at another; on the other, there is no point in transferring information from one institution if it will be ignored by the receiving institution.
- Aggregation of information for quality assurance, policy, and remuneration and research. (Much research is local, but interoperability issues are likely to come into play in multicentre studies at regional, national and international levels.)
- Large scale procurement policies, such as the English Connecting for Health Project.
- Internationally

Internationally the issues are likely to concern

- Scale of markets. The cost of developing systems is such that developing systems for a single country that operate only in that country may not be economic.

- Large scale collection of comparative data (e.g. WHO) and large scale multi-centred research studies and trials.
- Individual transfers across borders are significant but still represent a small fraction of total healthcare except in specific local border areas. Interoperability internationally is therefore likely to be of particular value in those areas where there is naturally significant cross border patient flow.

Use cases by Task for interoperability

Terminologies and ontologies have been widely advocated and/or used for many different purposes. Rather than attempt an exhaustive survey, we outline here the major use cases that will be used in the analysis of the different systems on offer and the current “facts on the ground”. For this purpose, we divide the use cases into two broad categories – the use of information about individual patients and the use of aggregated information about populations and groups of patients.

- *For individual patients – specific data*
 - Assisted Data capture
 - Information display
 - Access to the patient record
 - Access to pertinent knowledge resources
 - Ongoing clinical care
 - Automated support
 - Monitoring and warnings
 - Clinical decision support\
 - Remuneration
- *For populations – aggregations of data*
 - Reporting –
 - Remuneration
 - National and International reporting
 - Public Health – gathering of generic information to help in the formulation of hypotheses
 - Quality assurance
 - Surveillance
 - “Observatories” and “BioBanking”
 - Research – gathering of highly structured information for the testing of hypotheses
- *For communication*
 - *Between machines*
 - *Between machines and people*
- *For discovery*
 - Indexing of knowledge resources
 - Navigation of knowledge resources
 - Formulation of knowledge resources

Each of these use cases has different requirements and different constraints for interoperability.

For communication between machines the prime requirement is *constancy* – the receiving system must recognise what is being sent to be able to act on it. This requirement is fundamental to all others and dictates the need for unambiguous identifiers.

- For communication between machines and people and information display, *undertstandability and reproducibility* are critical.
 - For automated support, unambiguous and repeatable information capture is critical. However, there is a practical limit to the degree to which clinicians are willing to trust each other, let alone systems, so that there are fundamental limits to the degree of interoperability that can be achieved in practice, regardless of the precision of the terminology.
 - However, the relative lack of concern with reproducibility is peculiar. Consider the response of a funding body for an application for a new world wide study in which the reproducibility and inter-rater reliability of none of the data collection instruments had been, or were planned to be, validated.
- For aggregated data, the detail is limited by the procedures used to extract the information. The question of whether data captured from directly from clinicians (e.g. in SNOMED) can ever be sufficiently accurate for interoperability of international statistics using codes assigned by specialist coding staff (e.g. using ICD) is currently a matter of controversy and research. Concerns with reproducibility – see above – currently limit researchers' ability to trust and use routinely collected data.
- For information discovery, human *undertstandability* and navigation are critical. Humans can tolerate considerable ambiguity, but tend to focus too narrowly, so that the requirements are almost the reverse as for automated support.
- Because of the rapid rate of evolution, for nearly all these use cases, *responsiveness* and rapid change and correction is critical. Current systems response times vary from six months to decades. In the age of the Web, response times for almost all other purposes are measured in seconds or at most days.
- For decision support and aggregation of the ability to specify "*derived*" measures from data is essential, e.g. to specify the derivation of "body mass index" from height and weight.

1.7 Visions for the future

1.7.1 Desirable outcomes

Technical evolution

- Semantic operability will be achieved gradually beginning with applications with high benefit and modest cost. Given appropriate incentives, there will be a series of bottom up and top down measures that will achieve a level of interoperability that protects patient safety and supports common undertakings in public health, clinical research, and dissemination of best practice. Material for dissemination of best clinical practice will increasingly be linked to the structures and terminologies used for clinical care.
- It will be rare, and considered foolish, to develop terminologies or ontologies de novo. Most will be borrowed, adapted or mined, just as today most Java programs are developed out of re-usable packages widely available on the Web or in specialist repositories.
- There will be standard procedures for developing EHR, decision support, and vocabularies together, with rapid feedback amongst the participants.

- There will be a clear recognition of the limits to semantic interoperability. Effort will not be expended on attempting to achieve a degree of communication between systems not possible between the users of those systems.
- For terminologies, this will best be achieved by starting with areas where there is a high degree of consensus on both the content and the need. Key areas are likely to be sensitivities and adverse drug reactions, translational medicine, and large scale public health and population research initiatives such as “BioBanking”.
- The mechanisms that are successful will be open, collaborative and Web enabled, and much of the effort will be contributed by specialised communities standardising vocabularies for local purposes. These communities will “own” and take responsibility for their terminologies, helped by central servers and technologies which they will think of as part of their environment, just as much of the population today thinks of the Web, Google, Facebook, Flickr and related applications as just “there”. Feedback in most situations will be rapidly, often effectively instantaneous. It will be no more difficult to integrate new terminology or translations of terms than it is currently to make simple purchases on-line.
- The methods will become increasingly formal. The conflict between the scaling problems presented by pre-coordinated terminologies and the difficulty of maintaining consistency with post-coordinated terminologies will be overcome. To this end, the formal structure of SNOMED-CT and will be radically revised to take advantage of its purported underpinnings in description logic. HL7 v3 and/or Archetypes will likewise be reformulated to take advantage of modern technologies to ensure their mutual consistency and consistent binding to the new terminologies. Common links to terminologies from OBO and others used in molecular biology will be forged.
- Because of obvious usefulness, there will be serious involvement by clinical staff in medical terminologies as there is be bioinformatics and molecular biologists in bio-ontologies.
- Because of the recognition of limits, despite a high degree of interoperability and reliability for key systems, Semantic interoperability will not be complete. Much care will continue to be delivered locally using idiosyncratic systems or with minimal, or no, IT support
- Nor will there be complete harmonisation of either EHR models or terminologies. There will continue to be a major requirement for mappings and transformation services based on technologies analogous (or identical) to current data warehousing and mediation technologies.
-

Evolution with respect to use cases

- *Patient care* will improve dramatically with a significant reduction in avoidable errors and improvements in patient safety. Distributed care will become the dominant paradigm, with a rapid shift of care both to the community and to highly specialised centres expert in applying the latest techniques arising from accelerated clinical and translational research. Care in remote areas will be particularly affected. Patient’s will take increasing responsibility for their own care with the help of Web-enabled tools that link directly both to their own records and to the records held in the various institutions in which they seek care. The elapsed time to translate new findings into practice will be drastically reduced. The rise in the overall cost of care will be mitigated.
- *Public Health* will be facilitated by much faster and less costly collection of international statistics, as most statistics will be derived from data collected during patient care, although there will remain a need for experts to monitor and check data for critical measures. Surveillance for the emergence of new epidemic diseases and major health

problems will become more effective, and most outbreaks will be recognised early enough to be contained, although the increasing population and rate of travel will result in more small outbreaks.

- *Clinical and translational research* will advance very rapidly. Information sharing amongst researchers will be the norm. The lines between patient care and translational research and between translational research and basic research in molecular biology will become increasingly blurred. Most studies will be large scale, international. Many will re-use data from earlier studies to triage hypotheses and minimise the number of patients exposed to unsuccessful therapies. Research will depend increasingly on BioBanks and tissue banks which will have access to rich information on the lifelong outcome of large cohorts of patients collected in the course of their routine care. A uniform structure of privacy, consent and governance will manage data sharing for research in ways that are accepted by the vast majority of the population.
- *A balanced market will develop* with large suppliers managing hospitals as a whole but with innovative SMEs and specialist vendors supplying systems to address special functions and niche markets. The evolution between large and small, institutional and personal suppliers will be fluid, and European companies will play a major part in the overall commercial market. The time required to integrate a new specialised module or system into a hospital's infrastructure will drop from person years to person weeks, in some cases to person-hours. The difficulty of integrating systems will cease to be a barrier to adoption of best-of-breed solutions, and they will be embraced by central administrations and central IT directorates.

1.7.2 Inevitable outcomes

Technical evolution

- *Statistical text and Web mining technologies* will advance rapidly, and Google-like technologies will take over much of the burden of coarse-grained search for navigation information discovery. This will probably include linking of EHRs to text material for decision support such as the Map of Medicine⁷. The balance between semantic and statistical technologies will eventually be established, but where the balance will be remains to be seen. Cross language searching will improve rapidly, driven by general commercial imperatives, but is unlikely to eliminate the need for multilingual systems. Research on how best to use the two in concert is a major priority.
- *Direct encoding of free text into formal vocabularies* and EHR structures will improve radically, partly driven by voice recognition.
- Personal medical systems will proliferate. Whether they interact effectively with the local health care systems will depend on a combination of technical and commercial pressures. They may become a key driver for interoperability or may operate entirely outside it.

Evolution with respect to use cases

- Patient's will increasingly use web resources and take responsibility for their own care, with or without coordination with professional carers.
- Clinical medicine will advance and new treatments will inevitably be more expensive.

⁷http://demo.mapofmedicine.com/demo/1/login_page.html?next=http%3A%2F%2Fdemo.mapofmedicine.com%2Fdemo%2F1%2Findex.html

1.7.3 Outcomes to be avoided

Technical evolution

- Little will be done and the status quo will be allowed to evolve without incentives to interoperability.
- Semantic interoperability will remain confined to special cases, with little advance on the current state where. National and specialist terminologies and EHR formats will remain silos. Virtually all records for patient care will remain in free text.
- Alternatively, enormous resources will be spent on over-ambitious plans for semantic interoperability that inevitably fail. In either case, communication will take place by going around rather than via the clinical information systems. In countries where it is mandated, SNOMED and HL7 V3 will become taxes on healthcare, absorbing significant resources while returning no, or in some cases negative, benefits.
- Terminologies will remain closed or partly closed. Most of the developing world and much of Europe will be excluded from their development, so they will neither be fit their purposes nor be owned by them.
- The revision time for major terminologies will remain years. The release time for “mandated” terminologies will remain months.
- The technologies will remain based on, often ill understood, techniques from the late 1980s and early 1990s, without formal validation and support. Tools will remain primitive. The defects in the resulting systems will large enough that no one will trust them.
- Serious resources will be spent on developing expertise in getting around these defects, so that the legacy becomes increasingly difficult to change.
- The profession will remain alienated from informatics in general and “coding” in particular.

Evolution with respect to use cases

- *Patient safety will improve only slowly* and the reduction of clinical errors will continue to be slow and sporadic. The current death toll of hundreds of thousands per annum will continue. Best practice will continue to require up to twenty years to be adopted by the profession as a whole, even where the evidence in its favour is unequivocal. Care will continue to be fragmented institutionally. Errors resulting from miscommunication will continue to occur and will account for significant morbidity. Well educated and informed patients may take matters into their own hands, but those with fewer resources or in less well developed areas will be left at the mercy of chance and hazard.
- *Public health* will continue to depend on specialist encoders and be limited by the cost of capturing the information and the accuracy of the information that can be captured post-hoc. The cycle for revising major terminologies. Biosurveillance will remain a specialist activity divorced from mainstream of clinical practice.
- Clinical and Translational research will continue to be conducted in silos. The cost of mounting multi-centre trials will become the dominant barrier to the application of basic biological knowledge to medical care.
- The market will continue to be dominated by a few large suppliers who supply “complete” one-size-fits-all solutions to entire hospitals or even entire countries. Innovation will become nearly impossible. Niche systems will be rare and will not interact with the main systems. The difficulty of integration will be the major barrier to the procurement of specialist systems and they will be resisted vigorously by administrative and central IT directorates.

2 Background

2.1 Historical perspective

Brief summary of systems evolution – Diagnoses and Findings

There have been many parallel developments in terminologies. An exhaustive survey would not advance a “roadmap” and is beyond the scope of this deliverable. Many of the major disease terminologies in use to day derive from a few sources:

- The International Classification of Diseases and its predecessors going back into the middle of the nineteenth century and now managed by WHO – developed for international reporting and public health, and fundamentally a “classification” ICD is now entering its eleventh revision and remains the standard for international recording and much public health information.
- The MeSH headings developed by the US National Library of Medicine for Index Medicus – and fundamentally a “thesaurus”
- The systematised nomenclature of pathology – SNOP – which later became the basis of SNOMED and provided much of the material for GALEN – developed to index pathology slides and a controlled vocabulary and identifier system⁸.

When computer systems began to become prominent in the 1980s, it became clear that none of these systems provided sufficient, or appropriate, detail for recording information about individual patients, either for care or for remuneration. There were two major Anglophone responses to this challenge.

- The “clinical modifications” of the ICD became the basis for reporting for insurance and remuneration in the US, with variants being used in Australia and Canada. Currently ICD9CM remains the standard for reporting in the US, although there are attempts to update to a corresponding modification of ICD10. Essentially the “Clinical Modification” added an extra fifth layer of codes to the four levels of coding in the ICD without changing the structure in any major way.
- In the UK, the Read Codes were developed. Although not explicitly based on ICD and containing much clinical information needed for UK general practice not included in ICD, their overall structure followed the main ICD headings and many of the rubrics can be traced directly to ICD. The success of the original “four digit” Read codes (now known as Read I) in UK primary care led to a major effort to develop a much larger set to include secondary care. This major effort consumed several millions of pounds without producing anything used outside of general practice, although a subset of Read 2 (“five digit read”) remains the standard in UK primary care. Subsequent to this an effort to develop a more “compositional” system, Clinical Terms Version 3 (CTv3) was progressing when it was decided that the cost of developing the system by the UK NHS on its own was prohibitive, and a decision was made to collaborate with the College of American Pathologists to produce what is now SNOMED-CT.
- SNOMED was fundamentally different from ICD and Read in that it had multiple axes and composed codes out of their several parts. This gave rise to the first compositional systems (See?), first a major effort by the College of American Pathologists to develop SNOMED-RT, and then the collaboration with the UK NHS to produce SNOMED-CT.

⁸ The structure of the original SNOP is actually more complex than this implies but not relevant to the material in this deliverable.

Both SNOMED RT/CT, and a major EU initiative, GALEN, was based on new “description logics” that allowed new terms to be composed out of old and had the potential for greater computer support for the development process.

In parallel with these developments were numerous other coding schemes for diseases, of which the most important is probably the Dutch based International Classification of Problems in Primary Care (ICHPPC), which was noted for its emphasis on careful quality assurance and studies of inter-rater reliability.

In addition, the EU funded a major initiative, GALEN, which developed a large and high structured compositional model for terminology that is still widely mined for its content but no longer actively maintained.

Current situation with relation to diagnoses and findings in countries where information is available

- The US – the primary system is ICD9/CM. There is a move towards an ICD10/CM but its precise status is not known at this time.
- UK –SNOMED-CT is mandated, but Clinical Terms (Read Codes) Version 2 continues to be used in primary care.
- Nordic Countries, collective and cooperative efforts – The Nordic countries have a tradition of cooperation in healthcare and welfare terminology work. This has been based on detailed definition work of key concepts in the field and has yielded only a low number of well-defined concepts. The people working in this area have had collaboration which has almost ceased due to lack of resources.

Nordic Centre for Classifications in Healthcare has had the terminology work on its work list for a number of years. However, almost no resources have been given to this work and thus the role of the centre has been to follow the situation in the field.

The SNOMED system for pathology has been used in all Nordic countries to register findings in medical pathology. There has been also interest to the modern versions of SNOMED but they have not been translated to any Nordic language before SNOMED CT.

- Denmark – Based on *National Strategy for Information Technology in Hospitals* (1999) and *National Strategy for Information Technology in Health Care Systems* (2003) Denmark started development of Electronic Patient Journal (EPJ), which is to be based on SNOMED CT. For this purpose, Denmark has had a testing license for SNOMED CT until they joined the IHT SDO collaboration in 2007. The translation is to be completed by mid 2008 and the EPJ has been developed parallel to the translation process.

Despite the high investments in the work, there is no decision of implementation of the EPJ system. The plans of pilot studies have been postponed and currently we do not have any information about such plans. The key person of this work Arne Kvaerneland has recently resigned from the National Board of Health and is now not working with SNOMED CT related matters. He is still responsible for the Danish post in the IHT SDO work.

- Sweden – has had testing licence for SNOMED but the work has been discontinued. This work seems to have been on county level and as an initiative of different individuals. However, during 2006 a national report on potential use of terminological systems was published (Nationell terminologi- och klassifikationsresurs med SNOMED CT). Based on this report a national strategy has been developed with 4000-5000 € yearly budget for years 2007-2010. As a part of this strategy, Sweden also joined the IHT SDO as one the

nine starting members. The plans include translation of SNOMED CT to Swedish, mapping and harmonisation with existing systems (in international collaboration) and linkage to common language. The project does not include the implementation in the health care systems that are expected to be part of the following National Strategy of Health.

- Finland – has passed a law for electronic health care record standardisation through a national archive that all patient record systems have to be able to communicate with by 1.4.2011. This means in practice that the systems will be able to communicate with each other either directly or through the national archive. The demand for paper based backup of the electronic patient records will be waved with this archive. The information will be categorized based on existing classifications and open text. The system will therefore be strictly national. Individual patients' information may be transferred to other health care facilities by the demand of the patient with the permission of the unit that has been treating the patient. These permission of the patient is always necessary and such permissions will be stored as part of the system. The system will not use SNOMED CT nor any other terminological system but the broadly used national classification systems will allow search for most key elements of the information.
- France is mainly following the ICD stream (since its initial start) and for the DRG/GHM systems for hospitals. For findings in pathology France is mainly using a French brand system ADICAP and not SNOMED. For biology the use of LOINC is increasing but a French brand system named NABM is broad use. The house passed an act in 2004 for a life long individual electronic record accessible to the patient named DMP starting in 2007. A recent audit has decided to postpone "sine die" the implementation which was unable to start. Within this program in 2006 the organisation GIP DMP has bought the rights of French speaking SNOMED 3.5 international from the French speaking Canadian co-founding father of SNOP and SNOMED. On the other hand the French department of health has decided not to participate in SNOMED IHTSDO. Contacted by the Canadian team in charge of translating SNOMED CT in French GIP DMP has decided to make the French version of SNOMED 3.5 available to them. In primary care there are some punctual implementations of ICPC and a French brand coding system develop by a scientific organisation of G. It is difficult to forecast the future but the implementations of either SNOMED 3.5 or SNOMED CT in the real world have a low probability when are considered the resistances of health care professionals to use any mandatory terminology as ICD 10 preferring their own point of care terminology or their specific college terminology.
- Norway – has earlier had test license of SNOMED, it is currently not interested in SNOMED CT work. There is now direct development work in this direction, but there is definite interest and need for further development of health care information systems.
- Iceland – There is no current development work in Iceland for starting using SNOMED or any other terminological system. Categorized information is collected using existing classifications mainly ICD-10 and NCSP (Nomesco Classification of Surgical Procedures) in the extended version adapted for use with NordDRG. There is local interest for SNOMED CT but resources for extensive projects like translation are not available.
- Australia – has joined the IHSDO and mandated SNOMED. Specific projects are known to be using SNOMED in practice, particularly in conjunction with Archetypes and the Ocean Informatics Software.

Surgical procedures and operations

Even more than findings and diagnoses, the coding of clinical procedures has been involved with remuneration. The systems have been specific to specific jurisdictions and often proprietary:

- Clinical Procedure Terminology (CPT) in the US owned by the American Medical Association
- Office of Population Census and Statistics 4 (OPCS4) in the UK, operated, in effect, cooperatively by the OPCS and the NHS
- In France a first attempt named CDAM was a kind of ICD 9CM volume 3 for DRG based on a French translation done in Belgium under the name HCIMO {A second generation named CCAM was developed for a more comprehensive approach .It was performed in 2 ways: a traditional expert consensus one and an advanced one based on GALEN ontology tools and Multilingual generation. As any “political system”, CCAM is now submitted to a reimbursement bias; nevertheless, several countries as Germany, Austria and Switzerland are trying to develop such ontology driven systems. More recently, the WHO -FIC has decided to propose an International Classification of Health Intervention (ICHI) based on ontology and on the CEN categorial structure standards.
- In The Netherlands there is a mixture of procedure classifications in use. The core structure is in CMSV, derived from the original WHO ICPM of 1978 by the Dutch WHO-FIC centre. Third parties have extended this classification with more axes. The Dutch CMSV has been translated and extended in German as ICPM, which now has developed into OPS that is used for the German DRG system.
- The GALEN project focused much of its efforts on providing resources for procedure classifications and was at the core of the methodology for the development of the French CCAM system. It demonstrated the effectiveness of compositional methods for describing complex procedures, a problem which had previously defied solution.

Laboratory procedures LOINC and HL7

In parallel with HL7’s development of messaging standards, initially for communication between hospital laboratory and other hospital systems, an open source terminology for laboratory procedures, Logical Observation Identifiers Names and Codes (LOINC) which is and remains the standard for naming of laboratory procedures in almost all systems using HL7. LOINC has an idiosyncratic structure which does not fit neatly into the vocabulary in Section 0, but for these purposes is best thought of as a controlled vocabulary plus system of identifiers. (Its structure is not intended to support clinical decision support or extensive human navigation).

As the major standards body for messaging, HL7 itself maintains an extensive set of small “structural vocabularies” through its vocabulary committee.

Images

As with HL7, the major standard for transmitting images, DICOM⁹, required has developed its own vocabulary that it maintains, with varying degrees of cooperation with other bodies.

Nursing

Nursing terminologies have been the subject of numerous attempts at standardisation, of which the two most important are probably

- The International National Classification for Nursing Practice (ICNP)¹⁰
- NANDA Taxonomy¹¹

⁹ <http://medical.nema.org/>

¹⁰ <http://www.icn.ch/icnp.htm>

Unified Medical Language System and Metathesaurus

Faced with the “tower of Babel” of terminologies developing in the early 1990s, the US National Library of Medicine extended its remit to the development of the Unified Medical Language System, which attempts to cross-reference the major terminologies that are actively deployed world wide. A key part of this process is that the “Meta-thesaurus” includes the Medical Subject Headings (MeSH) terms. The UMLS identifiers, known as “Concept Unique Identifiers” (CUIs) and “Lexical Unique Identifiers” (LUIs) have become the *de facto* standard identifiers for medical concepts and terms.

The UMLS also includes a range of lexical resources and knowledge sources, all freely available (subject to the licensing arrangements with the originating authors of some terminologies.)

2.2 Quality Indicators and Academic Studies

The issue of establishing quality of terminologies is a major topic of research.

2.2.1 Internal measures

There are two classic papers that touch on the issue, Cimino’s “Desiderata”¹² and Rector’s “Why it’s Hard”.¹³ More recently, there have been a rash of studies particularly by Cornet and Schulz on specific aspects of ontology quality.

In looking at issues of quality, it is worth distinguishing

- Issues of the use of the terminology / ontology in practice
- Issues of the internal structure of the ontology

Cimino’s desiderata, which relate to the use of the terminology/ontology are widely quoted and worth reviewing (slightly paraphrased):

- Sufficient content/coverage and scalability
- “Concept orientation” – it is the meaning that matters not the specific words or “terms” used to express that meaning. In current parlance, distinguish clearly between “Concept” – the underlying meaning – and “term” the linguistic or other label used to communicate that concept to human users.
- Concept permanence – Neither terms, nor concepts, nor identifiers can ever be deleted, because they will appear in records somewhere.
- Nonsemantic identifiers – the identifiers should not be derived from the structure of the terminology, otherwise when the structure changes the identifiers will have to be changed. (This is a known problem with ICD and the original Read codes)
- Polyhierarchy – concepts must be able to appear in more than one place. No single hierarchy will capture all meanings (NB. This has recently become controversial in some circles, primarily on philosophical rather than pragmatic grounds – see Barry Smith)
- Formal definitions – there should be sufficient information to support machine processing and to support human understanding. This would probably now be reformulated as two desiderata.

¹¹ <http://www.nanda.org/html/taxonomy.html>

¹² Cimino, 1998, Desiderata, *Methods of Information in Medicine*, 37, 394-403

¹³ Rector A., 1999, Why it is Hard, *Methods of Information in Medicine*, 38, 239-252

- Human understandable definitions
- Machine understandable definitions
- Reject “Not elsewhere classified” – and other “residual categories” because such categories are not stable; their meaning changes with changes in versions. Other residual terms include “Not otherwise specified” and “Other”
- Multiple granularities – a key requirement in an age of translational medicine in which terminologies need to span the range from the genetic and molecular to populations and clinical trials. This is a major challenge not yet dealt with in practice.
- Multiple consistent views – a key requirement given the multiple use cases and tasks to which the same terminology / ontology may be applied.

Many would now argue that this list gives too little emphasis to the internal structure of the ontology and does not distinguish sufficiently between the need for human readable text definitions and machine-readable formal definitions. Nonetheless, it provides a basic starting point for any discussion of quality.

2.2.2 Human factors and Inter-rater reliability

There has been much less study of the human factors and classic measures of inter-rater reliability with respect to terminologies than there has been of their internal and computational properties. Other than the early studies of Lamberts and ICHPPC and recent studies of Rogers on the use of SNOMED codes by British general practitioners¹⁴ and a few independent studies on a relatively small scale^{15 16}

2.2.3 Scaling and Zipf’s law usage

Building a small terminology or classification is deceptively easy. Scaling it up is unexpectedly difficult. For example, development of the original four digit Read Codes required at most five person years; attempts to scale them up to the five digit read codes required several person-centuries and involving over forty groups of clinicians. Development of SNOMED-RT/CT has been quoted as costing in excess of €40M.

Medical knowledge, like any significant body of knowledge is “fractal”. As much can be said about any subspecialty at the subspecialty’s level of granularity as can be said of all of medicine at the non-specialist level of granularity. The same goes for the sub-sub-specialist. As our knowledge increases, the total amount of terminology to be captured grows exponentially.

Furthermore, even within a given field our knowledge grows combinatorially. For example, if there are three degrees of burn, three classifications of the extent of the burn, three aetiologies for burns, and 130 body sites that might be burned, there are a total of nearly 10,000 potential codes for burns. Add another axis, e.g. for whether or not infected, and the number of potential codes increases to 30,000 (including for “unspecified”). This is despite the fact that underpinning this enormous coding system is a set of less than 150 concepts and relations.

¹⁴ Rogers JE, Wroe CJ, Roberts A, et al. Automated quality checks on repeat prescribing. *British Journal of General Practice*. 2003;53:838-844.

¹⁵ Chiang MF, Hwang JC, Yu AC, Casper DS, Cimino JJ; Reliability of SNOMED-CT Coding by Three Physicians using Two Terminology Browsers. 2006; AMIA Annual Symposium: 131-135

¹⁶ Andrews et al, Variation of SNOMED CT Coding of Clinical Research Concepts among Coding Experts Andrews et al. *J Am Med Inform Assoc*. 2007; 14: 497-506

Furthermore, it is difficult to know in advance, which concepts are going to be used in practice. Typically, in any language, the vast majority of usage is concentrated in a few notions, but the tail of usage is infinitely long. The result is that any attempt to enumerate all useful concepts risks running into the law of diminishing returns very early in its development. Practical experience with virtually all large terminology projects confirms this observation. The terminologies tend to grow much faster than the number of terms actually used in practice, and vastly faster than the number of terms used more than once.

2.3 Technologies

2.3.1 Terminologies and Ontologies

Terminology tools and traditions

There is a long tradition of developing terminologies, particularly strong in the Scandinavian countries, which formalised *ad hoc* methods developed over many years to develop what are, usually, thesauri or classifications. These were primarily disciplined and principled “pencil and paper” methods with minimal computer support. They are formalised most compactly in ISO 704, “*Terminology work – Principles and Methods*”.

There have been various efforts at specialising and standardising this work and related developments in the medical field, including work in both CEN TC251 and ISO TC 215. The most recent standard in this tradition is ISO 17115:2007 “*Health informatics – Vocabulary for terminological systems*”

Today, this work remains important, although primarily as the preliminary stages in compiling ontologies for use by various more formal, computer-based tools.

Semantic path based classifications (ICD, Read I and II)

Early terminologies were characterised by mono-axial classification systems in which the identifier was actually the path to the code. Although convenient for some purposes, this approach has come to be deprecated for most applications in favour of “non-semantic identifiers” (See 2.2.1) because a) they cannot be easily adapted to poly-hierarchies, and b) changing the organisation requires changing the identifier. Most such systems have been managed by hand or by very simple outlining tools – indeed one of their advantages is that tooling is simple.

(Note the use of paths in ICD and Read codes should not be confused with that in the MeSH codes, despite superficial similarity and the fact that MeSH codes are often referred to by their path. However, used correctly, the MeSH identifier is actually the text term, and the same term can appear at the end of many different paths.)

Description logics, Ontolog, GRAIL, and OWL

In the late 1980s, previous knowledge representation systems evolved into what we now refer to as “description logics.” Description logics are subsets of first order logic that allow classes (corresponding to codes) to be described and defined rigorously. Description logics provide a formal method for stating the universal truths of what must be so in any possible model of the world consistent with their axioms. Because of their strong logical structure, they provide two key capabilities.

- *Composition* – new classes can be defined by logical combinations of existing codes, e.g. a “fracture of the femur” can be composed out of the classes for “fracture and femur”

and the property for the relation between them (usually termed “site” or “has_locus”). Composition allows a choice between

- *Pre-coordination* – enumeration of all concepts before use
- *Post-coordination* – composing of concepts as used
- *Automatic classification* and consistency checking – based on the definitions and descriptions, “classifiers” (sometimes called “reasoners”) can construct the classification automatically and check that all descriptions and definitions are consistent with each other.

These capabilities address six out of the key desiderata enumerated by Cimino as cited in Section 2.2.1. For brevity, we consider them under three headings:

- *Composition* addresses the problem of *scaling and combinatorial explosion*. If compositions can be created as required, then only the base 150 entities in the example of burns in Section 2.2.3 need be enumerated as opposed to the 30,000 plus potential combinations.
- *Formal definitions* are intrinsic to the approach.
- *Polyhierarchy, Multiple consistent views, and multiple granularities* can all be determined algorithmically given a suitable set of formal definitions and descriptions. Since experience has shown that manual development of polyhierarchies reliably is almost impossible, this is a major advantage.

Given these advantages, why have not description logic solutions become universal? They have become widespread and underpin both SNOMED-CT and GALEN and parts of ICNP. However, there are several drawbacks.

- They are difficult to understand, and existing tools are not intuitive to users. The major achievement of GALEN was to reduce the time required for training before being able to participate effectively from three months to three days. Unfortunately, GALEN’s techniques for “intermediate representations” (roughly SNOMED’s “close to user forms”) have not been widely adopted.
- In order to be tractable, description logics impose limitations on expressiveness. Description logics are all subsets of first order logic with some restrictions. In the past decade there have been major advances so that many of the awkward constructs required in GALEN (GRAIL) and SNOMED-CT (Ontolog) are no longer necessary (See appendix), but learning to work around the limitations adds to the difficulty of using them.
- Scaling description logic classifiers to very large terminologies such as SNOMED has proved challenging. Until recently, this was a major barrier to using all but the simplest description logics for very large systems. However, the combination of the increase in computing power (A “standard laptop is now between 100 and 10000 times the power of the most powerful servers available at the start of GALEN and SNOMED) and improved algorithms, this problem is rapidly being overcome.
- Until recently there was no standard syntax or API, so that every tool was limited to a single reasoner and experimentation with alternatives was difficult. The emergence of the W3C standard ontology language, OWL, has changed this situation radically. A large and growing community are now developing tools and underlying software to a single standard.

Given the major advances, the remaining barrier is tooling which has not kept pace with technical advance. In fact, in many ways tooling for modern description logics and OWL is far behind that available for GRAIL and earlier languages of a decade, or even two decades ago. Resolving this situation should be a high priority for development.

It should be noted that OWL comes in several flavours. OWL DL and its successor OWL 1.1 are fully based on description logics and can be reasoned over using standard classifiers. OWL lite is a subset of OWL DL and likewise can be reasoned over. OWL full includes constructs from higher order logic and other constructs not supported by description logic, and can be queried but not reasoned over. Unless otherwise noted in this document, “OWL” refers to “OWL DL” or “OWL 1.1”.

2.3.2 Digital libraries and Scholarly Resources

MeSH, UMLS and PubMed

The Medical Subject Headings (MeSH) are maintained by the US National Library of medicine using their internal procedures and used to index their major bibliographic resource PubMed, which is the *de facto* standard for indexing the biomedical literature.

2.3.3 Statistical and Social: Google, Web and text mining

Web 2.0 and social computing

A major unanticipated development of the Web has been the advent of social computing often summarised under the heading of “Web 2.0”. Social computing attacks large tasks by making them available to many developers, often volunteers, via the Web. The best known example is Wikipedia, but it is less well known that the taxonomies that underpin Yahoo and Google evolved from a social computing effort, Open Directory¹⁷, which outperformed directed efforts dramatically.

Social computing tools have been surprisingly little used in clinical terminology development, although the ICD-11 effort is experimenting with Semantic Wiki. We would expect to see this become one of the important areas for development in the near future.

Google: Statistical text and web mining

The major development of the past decade in computing has been the rise and rise of search engine technology now dominated by Google. Google has proved remarkably successful in being able to locate and present information based on a combination of lexical and web-mining techniques. It also makes many tools widely available for software development. The Web 2.0 development to no small part owes its development to a combination of these tools.

The question as to the optimal balance between statistical “google-like” techniques and semantic techniques remains a key issue for interoperability.

Google’s pioneering techniques for “mash-ups” using tools such as Google earth offer alternative software paradigms for interoperability and are a major force behind the move towards “restful computing”.

¹⁷ <http://www.dmoz.org/>

2.3.4 Authoring Tool & Browsing sets

Because of its size and specialised structure, SNOMED presents special challenges. Also, because the IHTSDO has so far failed to make the description logic “stated form” generally available, there are issues with exactly what is required to cope with SNOMED in its original form.

Tools may also be divided into purely browsing tools, such as CLUE and more general authoring environments, all of which include a browsing capability.

SNOMED Browsers without authoring capability

- CliniCLue¹⁸ from the Clinical Information Consultancy headed by David Markwell is the *de facto* standard browser for SNOMED used by much of the community. It is freely available, although the SNOMED databases and licenses must be obtained separately. It provides extensive searching and browsing facilities, but no authoring facilities

SNOMED oriented authoring environments

- *Apelon*¹⁹ - the toolset used in the development of SNOMED. The Apelon Distributed Terminology Server has now been made open source – see Chapter 0 – but the authoring environment remains proprietary and expensive (Pricing is commercial in confidence but reported to be on the order of several €10K/annum/seat)
- *SNOB*²⁰, a powerful freeware SNOMED browser and editing tool. Partially based on GALEN technologies. SNOB is probably the only well developed openly available editor that supports SNOMED in its native form at full scale with full use of classification.

Generic SNOMED Capable

- *Protégé-OWL*²¹ has become the *de facto* standard editing environment for OWL DL, and is increasingly used in biomedical applications, including by both the US National Cancer Institute (NCI) and many of the OBO. Protégé supports both the more traditional frame representation and the new Web Ontology Language, OWL. The latest version of the OWL editor also reads the standard format for the SNOMED Stated Form (KRSS). With the advent of 64 bit architectures, standard Protégé is able to handle the full SNOMED in OWL form, although optimisations would make the system easier and quicker to use. Protégé is committed to develop a distributed collaborative environment for ontology development but has not so far delivered. The tools are open source and based on the open source and standard OWL API.

Other Generic Ontology Development

- *NEON*²² – a current EU project that seeks to produce an open-source collaborative ontology development environment. As far as can be determined, it has not so far been applied on a large scale to well known biomedical ontologies.
- *CEL*²³ – The CEL classifier is specifically designed for the minimal description logic sufficient to describe SNOMED as currently expressed (*EL+*). It has the great advantage

¹⁸ <http://www.cliniclue.com/>

¹⁹ <http://www.apelon.com>

²⁰ <http://snob.eggbird.eu/>

²¹ <http://protege.stanford.edu>

²² <http://www.neon-project.org/web-content/>

²³ <http://lat.inf.tu-dresden.de/systems/cel/>

over other classifiers that it is both provably complete and runs in polynomial time – *i.e.* it can be guaranteed to scale smoothly to the size of SNOMED and beyond. CEL is being used as part of several environments, but it is in itself only one component – the classifier/reasoner – of an ontology development environment, albeit a critical one.

- KAON2²⁴ is the result of the predecessor project to NEON and used by it and provides a range of features and alternative ontology development resources

Terminology Binding: The interaction of terminologies and EHRs

- *Ocean Informatics Archetype editor* – Ocean Informatics has demonstrated a terminology query language for SNOMED as an adjunct to its Archetype Editor, but no official release is available for testing at the time of writing.
- *Linköping University Archetype editor and MoST*²⁵ – The alternative archetype editor developing by Linköping University has been used as the basis for the development of the MoST binding system by a PhD student funded by the Semantic Mining project. MoST is a highly promising technology which combines many different natural language and semantic resources, but which would require significant further development for widespread deployment.

2.3.5 Infrastructure and software

Terminology Servers and Services

No one terminology is ever going to be suitable for all tasks. Most applications will need to “serve up” a variety of terminologies, both for configuration and at run time. A clear distinction needs to be made between systems for authoring and maintaining terminologies – development environments – and systems for delivering them – Terminology Servers. In many cases a terminology will be developed in one form – *e.g.* a description logic – and then delivered in another – *e.g.* as a set of relational tables. (This is, in fact, the case with SNOMED-CT).

Three related projects are developing large scale “terminology servers” in addition to at least two commercially available systems.

- LexGrid – an open source terminology server being developed at Mayo Clinic under the auspices of the NIH
- LexBig – the National Cancer Institute’s adaptation of LexGrid to their specific purposes and linking it to their metadata repository and Enterprise Vocabulary Services²⁶ (see ISO 11179 below).
- The Apelon²⁷ Distributed Terminology System (DTS), which is now open source (although many of the associated tools remain proprietary)
- Ocean Informatics²⁸ Terminology Server, which is specialised to SNOMED and includes a specialised query language and search engine.
- Health Language Terminology Server²⁹, which forms part of its broader language and terminology system.

²⁴ <http://kaon2.semanticweb.org/>

²⁵ <http://www.imt.liu.se/mi/ehr/>

²⁶ <http://www.library.ucsf.edu/db/record.html?idrecord=82>

²⁷ <http://www.apelon.com>

²⁸ <http://oceaninformatics.biz/CMS/index.php>

²⁹ <http://www.healthlanguage.com/>

In addition, there is a novel system of “just in time” terminology services developed as part of the SAGE project.

NCBO and BioPortal

The NIH has funded the National Center for Biomedical Ontologies (NCBO or NCBO)³⁰ which provides an alternative form of ontology service, BioPortal, a tool for browsing, searching and downloading ontologies from many sites, including OBOFoundry.

The NCBO has a number of collaborating centres and can be expected to play an increasing role in the development of ontologies in the future.

Versioning systems and identifier tracking

Managing versioning and tracking identifiers is a major part of any system that includes a controlled vocabulary. Good identifier tracking is one of the major features of successful terminology systems including both SNOMED and the Gene Ontology. Currently there are rumours of a new system developed by Informatics, Inc³¹ that is under consideration by the SNOMED IHTSDO and the Australian government. No further information is available at this time.

ISO 11179 and the NCI Formal Metadata repositories and EVS

The US National Cancer Institute is developing a major platform for translational medicine CaBIG³² one of whose key components is an enterprise vocabulary service (EVS) based on the ISO 11179 Metadata standard. This represents an important alternative technology for the interaction between vocabulary and software whose outcome cannot yet be properly assessed.

2.3.6 Bioinformatics and Molecular Medicine

GO, OBO, OBOFoundry and OBOEdit

The Gene Ontology (GO)³³ has been a key resource for the molecular biology community. Starting from a relatively simple structure and a single resource, it has developed into the Open Biomedical Ontologies consortium. The most visible face is the OBOFoundry³⁴, a major effort at collaborative ontology development.

The Gene Ontology (GO) itself is authored using its own tools, OBO Edit with its own flat file format. Significant efforts are under way to provide translations to and from OWL from the OBO format.

The two formats have complementary strengths and weaknesses. OBO format provides excellent metadata and user interface features but relatively weak semantics; OWL provides strong semantics but weak features for metadata and user interfaces.

³⁰ <http://www.bioontology.org/>

³¹ <http://www.informatics.com/>

³² <https://cabig.nci.nih.gov/>

³³ <http://www.geneontology.org/>

³⁴ <http://obofoundry.org/>

2.3.7 Social computing demonstrations from related fields

myExperiment and Taverna

Another Bioinformatics initiative that should be examined is the experiments in managing bioinformatics workflows for the widely used Taverna workflow engine using the social computing site myExperiment³⁵. There are many analogies between workflows and ontologies – indeed workflows require and depend, to some degree, on strong semantic underpinnings. Workflows are complex; the underlying technology is difficult and unfamiliar to biological scientists and most bioinformaticians; but it is only biological scientists and bioinformaticians who have the expertise to formulate them. The goal is that they should be re-usable and re-used, and that they should be built from re-usable parts. MyExperiment is taking a bold step towards cooperative development and re-use of such workflows using social computing. It is a lead that should be investigated as to whether it could be followed within the terminology world.

Semantic Wiki's

There is a major effort to combine the power of the Semantic Web with the hugely successful Wiki technology.³⁶ It is hoped to use this technology as part of the development process for the latest version of ICD. The technology remains unproven but highly promising, and a major area for research.

³⁵ www.myExperiment.org

³⁶ <http://www.semwiki.org/>

3 Analysis

3.1 Key determinants of adoption and success

Major coding systems have been adopted for general use only for a few reasons.

3.1.1 Regulatory or remuneration requirements

You don't get paid unless you use them.

Submission, and an eventual reimbursement, of clinical activities carried out in hospitals is allowed only when specific codes are used for that purpose.

Examples:

- ICD9/CM and CPT in the US;
- Read/CTV-2 in the UK.
- MEDDRA for adverse drug reaction reporting
- CCAM in France
- DRGs in Germany
- ICD codes for disease outbreaks to be submitted to WHO

3.1.2 Commercial requirements – de facto and de jure standards that are supported by industry to the extent that all systems must conform to be commercially viable.

You can't sell a system that doesn't conform to them.

It is hard to sell a system that is not conformant to the standards. The customers typically look for a heterogeneous IT solution – meaning they buy different components of software from different vendors. In such an environment, it is evident that the different systems should share a basic level of interoperability. Since most of the vendors support certain standards, it becomes necessary for a new product to support those existing standards.

Examples:

- LOINC and HL7v2 for laboratory communication
- DICOM for Images
- ASTM standards for biosignals
- MeSH/UMLS for bibliographic searching

3.1.3 Major international agreement and historical establishment

Long-standing commitments require that you use them

Example

- ICD from WHO, which in its various forms has been the standard for international public health reporting for over a century.

3.1.4 Community commitment and support

The bioinformatics community, in contrast to the medical informatics community, has been more cooperative towards building de facto standards which is led by a broad research community and supported, both in content and budget, by the commercial organizations. The prime example is:

- The Gene Ontology and the OBO consortium

However, there are hundreds of bioinformatics databases and terminologies, (some claim that a comprehensive workbench must have access to over 700 resources, with the number growing at the rate of several per months)

These resources are interesting in that they have several characteristics

- Openness – anyone, or almost anyone, can participate
- Curation – there is a core group with an intense commitment to “quality”, defined on the basis of good principles of ontology. OBO Foundry is one such example where a subset of OBO ontologies have agreed to follow not only a retrospective, but also a prospective policy of curation based on formalized class definitions and relations. There are serious controversies over the standards adopted for curation, but they are at least explicit, open to critique, and a genuine attempt is made to adhere to them.
- Web-based Access – Most of these resources have grown up in the era of the Web and provide web-enabled access through a variety of tools built around these resources.

There are, of course, many analogues in other fields of which the most important are the open source movement and the development of, amongst many other tools, Linux and the development of Wikipedia. All of these applications show that community mechanisms can work and produce high quality results under proper circumstances and with proper support.

3.2 Key Players and “Facts on the Ground”

A number of key facts about the landscape can now be seen and predictions made with reasonable confidence.

3.2.1 Clinical medicine and medical classifications

National Coding Schemes and Specialised Requirements

- National coding schemes will continue to be important, particularly in surgical procedures, e.g. CPT in the US and OPCS in the UK.
- The ICD-9/10-CM will continue to be used for reporting for remuneration purposes in the US. There is little prospect of its being supplanted for remuneration purposes by SNOMED.
- LOINC and HL7v2 will continue to be used for communication of laboratory and related results for the foreseeable future. An HL7 v2.6 is now planned, and LOINC continues to be developed in parallel.
- The US FDA will continue to play a key role in drug certification and require MEDDRA for various purposes. It is also moving to extend the use of SNOMED for some other purposes that may provide a critical boost to SNOMED within certain fields.
- The UK will continue to pursue its mandate for HL7 v3 and SNOMED, and will continue to be a major player in both organisations. However, the rate of practical uptake by

clinicians in the UK is limited by a) the cost of migration from Clinical Terms (Read Codes) version 3 in primary care, and b) difficulties with the overall Connecting for Health Project. Nonetheless, the largest set, and *de facto* standard of SNOMED codes specified for use with HL7 v3 messages comes from the UK.

- Currently, outside the UK, as far as can be determined, the only large users of SNOMED and HL7 v3 for clinical systems are the US Veterans Administration and Kaiser Permanente (Epic Systems³⁷), and Australia. Even in the UK, practical implementation and acceptance is currently limited (2008).

LOINC and HL7v2

- LOINC will remain the standard for communication of laboratory values within HL7, at least as long as HL7 version 2 remains the *de facto* standard for communication with laboratory systems, which appears likely to be for the foreseeable future. The two standards have “grown up together”, and there is a massive legacy that would have to be modified. Even in the UK, where SNOMED and HL7 v3 are mandated, *de facto*, HL7 v2 and LOINC are being used in most laboratory applications.

IHTSDO and SNOMED-CT

- The rights and control of SNOMED have been successfully transferred to the *International Health Terminology Standards Development Organisation (IHTSDO)*³⁸, which now has at least ten national members – effectively the Anglophone community plus the Netherlands, Denmark, and Sweden, with negotiations with several other countries reportedly nearing completion. Participation by non-anglophone countries remains limited, both in Europe and in the wider world.
- SNOMED is gaining key support, most notably from the UK National Health Service, in recognition by the FDA, and in use in Australia.
- Despite several attempts, there is now little prospect of a serious alternative to SNOMED as an international standard for coding for diseases, signs, and symptoms. Many therefore argue that despite its many flaws, the issue is therefore how to manage the improvement and use of SNOMED rather than its replacement.
- The practical impact of SNOMED remains limited. Acceptance and implementation by vendors, even within the Anglophone world, remains extremely limited. The legacy of systems using codes and the continuing requirement in the US to report for remuneration and legal purposes using ICD9/CM and CPT, remain serious barriers to the further implementation of SNOMED. Whether the level of acceptance will increase in future remains a matter on which opinions differ.

HL7 v3 and Terminfo

Conversion to HL7 v3 has received a major boost from its adoption by the NHS Connecting for Health programme in the UK. Nonetheless, take up by vendors remains patchy. However, the Terminfo project has produced a comprehensive manual for the use of SNOMED within HL7 which removes many of the previous ambiguities in the standard, although it has not produced, or even attempted to produce, any automatic tools for conformance testing.

³⁷ <http://www.epicsystems.com/>

³⁸ <http://www.ihtsdo.org/>

WHO and ICD-11 and ISO TC215

- ICD9/10/11 will continue to be used for international reporting.
- The ICD-11 initiative is being conducted in conjunction with ISO (and possibly SNOMED). It is experimenting with social computing mechanisms, and this presents a major opportunity to produce an innovative international classification scheme in a truly cooperative manner that would benefit both Europe and other countries not currently. However, this is a major and innovative undertaking. It risks failing if not provided with sufficient resources and support.
- Collaboration between WHO and the SNOMED IHTSDO will be a key nexus for development and influence. There have been extensive discussions on the “aggregation logic” for use in population health and public health informatics.

DICOM and Images

DICOM is both a *de jure* and *de facto* standard for image information. There are efforts at convergence with SNOMED and HL7, but it remains largely independent. Although there is a significant research effort in the field of radiology terminology, no major innovations have been identified as immediately in prospect. Radiology Society of North America (RSNA) is supporting a new terminology called RADLex and it might

3.2.2 Bibliographic resources and cross referencing – US NLM, UMLS & PubMed

The US National Library of Medicine’s (NLM’s) Unified Medical Language System (UMLS) and its PubMed indexing system can be expected to remain the dominant sources of indexing of scientific publications and scholarly resources. They will also remain the primary source for cross-referencing of existing major terminologies. The use of CUIs (Concept Unique Identifiers) and LUIs (Lexical Unique Identifiers) as the common currency for coarse-grained interoperability and access to scholarly resources will continue for the foreseeable future.

3.2.3 Drug Information: Commercial Drug Databases and National Standards

Storing information on drugs is a challenge due to the huge number of available medicines. Drug manufacturers may sell similar active ingredients using different trade names. Moreover it is necessary to consider different formulations, dosage forms, routes of administration and therapeutic and pharmacological classes.

A first strategy is to consider only the active ingredients, and therapeutic and pharmacological classes.

- The WHO maintains a list of International Nonproprietary Names (INN) that facilitates the identification of pharmaceutical substances or active pharmaceutical ingredients. Each INN is a unique name that is globally recognized and is public property. Some countries recommend drug prescription using the INN instead of trade names.
- The Anatomical Therapeutic Chemical Classification System is used for the classification of drugs. It is controlled by the WHO Collaborating Centre for Drug Statistics Methodology, and was first published in 1976. In the ATC classification system, the drugs are divided into different groups according to their site of therapeutic effect, therapeutic indications and pharmacological nature. Drugs are classified into five different levels.

The second strategy consists in providing entries for every trade name and combinations.

- The WHO Drug Dictionary is the world's most comprehensive dictionary of medicinal product information. It is used by pharmaceutical companies, clinical research organisations and drug regulatory authorities for identifying drug names, their active ingredients and therapeutic use, in the course of their drug safety surveillance. Drugs are classified according to ATC codes. Drugs containing the same active ingredient are referred by the INN that acts as a preferred name.
- The US National Library of Medicine (NLM) / Unified Medical Language System (UMLS) provides a standard drug vocabulary, RxNorm³⁹. In addition, the US FDA⁴⁰ has its own listing of standard names largely based on RxNorm which drive a variety of resources and are required for reporting.
- In the UK the DM+D⁴¹ – Drugs, Medications, and Devices – database provides a standard dictionary of identifiers and textual descriptions.

The third strategy is to benefit from vocabularies included in drug knowledge bases. Although not strictly terminological, the provision of drug information, codes, and databases is a major business. Companies and organisations with a major presence include First Data Bank (FDB)⁴² and Micromedex⁴³. Many of these resources are more than a vocabulary. They provide information models for describing drugs, medical products, devices, etc. However, they all have major vocabulary components, and their identifiers are widely used in commercial systems that subscribe to them. This includes applications for computerized physician order entry (CPOE), electronic prescribing (e-Prescribing), pharmacy dispensing, electronic medication.

Furthermore, the update and liability requirements attached to drug information are stringent. Monthly update cycles are standard. It is unlikely that any non-commercial source will compete, or wish to compete, with the comprehensive product lists provided by First Data Bank and its analogues in other countries.

The vocabularies also have strong regulatory backing, making them effectively mandatory for many purposes. Mandating by the FDA provides an effective *de facto* standard, at least throughout the developed world.

One special issue concerns the representation of drug information in the Summary of Product characteristics (SPC). The European commission has produced guidelines in the Rules Governing Medicinal Products in the European Union to recommend presentation of the adverse reaction section of the SPC using MedDRA (Medical Dictionary for drug regulatory Activities). This includes a single table of adverse reactions according to the MedDRA system organ class. On the other hand the US Food and Drug Administration recommends the use of SNOMED for description of adverse drug reactions in the SPC.

3.2.4 Translational medicine and Genomics

Gene Ontology, OBO, and OBO Foundry

The Gene Ontology is the most successful standard vocabulary in the biomedical area, and has played a significant role in the success of the Human Genome Project. The related ontologies, brought together under the heading of the OBO Foundry⁴⁴ have raised controversy because of the philosophical commitments of some of those directing the effort.

³⁹ <http://www.nlm.nih.gov/research/umls/rxnorm/>

⁴⁰ <http://www.accessdata.fda.gov/scripts/cder/drugsatfda/>

⁴¹ <http://www.dmd.nhs.uk/>

⁴² <http://www.firstdatabank.com/>

⁴³ <http://www.library.ucsf.edu/db/record.html?idrecord=82>

⁴⁴ <http://obofoundry.org/>

These can be expected to be the standard resources for terminologies related to molecular biology.

Some of the OBO ontologies, notably PATO, the ontology of phenotypes, and the OBI – the ontology of biological investigations, overlap into areas traditionally the province of translational medicine and even general clinical medicine. Their impact upon the general arena of biomedical terminologies remains to be seen.

US National Cancer Institutes (CaBig)

The US National Institute of Health (NIH), National Cancer Institute (NCI) and its CaBig project and associated Enterprise Vocabulary Service (See 2.3.6) will remain a major force in translational medicine both in cancer and beyond. NCI Metathesaurus tends to use its counterpart from UMLS, is being widely used in cancer trials. Improvements to NCI terminology are being proposed and will be carried out in the near future.

It is also likely to be a major sponsor of research in this area, particularly in the area of clinical trials ontologies.

US National Center For BioOntologies (BioPortal)

The NCBio and its BioPortal are likely to play an increasingly important role, although they remain too new and immature to assess properly at the time of writing. If successful, they will become the primary point of access to a large range of ontologies that are maintained in the common. A European link might well develop to build on and harmonise development.

Clinical Trials Ontologies and BRIDG

There is a major effort to converge on standards for clinical trials. The BRIDG⁴⁵ initiative brings together an industry-based initiative, CDISC⁴⁶, and HL7. In addition, the Ontology for Clinical Research (OCRe) was recently formed out of activities around an NIH sponsored workshop. A first release version of OCRe is available from BioPortal⁴⁷.

The interaction between the various translational medicine resources, the NCI EVS, and Clinical Systems, and BioBanking remains an important open question on both sides of the Atlantic.

Healthcare and Life Sciences Special Interest Group of the W3

The Healthcare and Life Sciences Special Interest Group⁴⁸ has been involved in a series of efforts around data standards and ontologies in the life sciences and translational medicine.

BioBanking

There are initiatives in several European countries to create “BioBanks” – collections of data and tissue samples and large segments of the population that will then be followed for long periods of time. The hope is to be able to perform what have been termed “pro-retrospective” studies so as to identify the genetic and molecular predictors and precursors of disease and response to therapy without having to wait quickly when new hypotheses are and to screen for early indicators of disease as tests improve. For example, to be able to

⁴⁵ Biomedical Research Integrated Domain Group, <http://www.cdisc.org/standards/bridg.html>

⁴⁶ Clinical Data Interchange Standards Consortium, <http://www.cdisc.org/>

⁴⁷ http://www.bioontology.org/ncbo/faces/pages/ontology_list.xhtml and click on OCRe.

⁴⁸ <http://www.w3.org/2001/sw/hcls/>

look at samples from patients who responded well or poorly to a particular preventative treatment in the light of genetic and tissue tests not available at the time that the samples were taken.

A key aspect of BioBanking is about looking at large numbers of patients so that international collaborative studies are critical. Describing these patients in a consistent way so that data can be compared across BioBanks is a major challenge. One such project is regarding the collection of certain genotypic and phenotypic data of the people living close to the banks of the Danube. At some future instance, such data would need to be integrated with similar efforts and a standardized approach needs to be developed. Achieving agreement on an appropriate level of reproducible clinical information for BioBanking internationally would claim a high priority for further advances in the field of translational medicine.

At least one major project in this area is known to be in the development stages. Achieving agreement on an appropriate level of reproducible clinical information for BioBanking internationally would seem a high priority for advancement of translational medicine.

Other known EU Initiatives

There are a large number of EU initiatives touching on translational medicine and genomics. Two with major emphases on terminology are:

- Bootstrep⁴⁹ - concerned with the interaction of text mining and terminologies
- "ACGT - Advancing Clinico-Genomic Trials on Cancer"⁵⁰ - concerned with coordinating trials on cancer

In addition, there are known to be major efforts to put together a joint European project on BioBanking. (See above.)

3.3 Critical barriers

3.3.1 Socio-political

Who is responsible? Who will pay? Who will accrue the benefits?

The total expenditure on healthcare in the EU has approached almost ten percent of GDP. The costs of healthcare IT programmes are vast, probably exceeding ten billion Euros per annum. (The cost in the UK alone over ten years is estimated to be on the order of €30B / £20B). However, the economic system has yet to associate benefits of interoperability in general, let alone terminology in specific, with those who must pay for it.

Standard interoperable solutions always cost more

It is almost always easier and cheaper to build a stand-alone system from scratch than to use standard interoperable systems. While those who pay are not those who benefit, this will remain a problem unless affordable technical solutions that allow the re-use of technology can be found.

⁴⁹ <http://www.bootstrep.org/bin/view/Extern/WebHome>

⁵⁰ <http://www.eu-acgt.org>

Legacy and the cost of migration

Healthcare systems are no longer “green fields”. Systems, of some sort, exist in virtually all hospitals and most primary care settings. Most have their own vocabularies, some use national standards such as ICD9/CM and the Clinical Terms (Read Codes) V2; many are entirely local applications. Myriad *ad hoc* solutions have grown up to link such systems, some successfully, some less so. However, the cost of adaptation of existing systems, to the extent that they are convergent, is potentially massive. It requires a convincing demonstration of benefit to those who will pay to justify the expense. For vendors, it requires a major economic incentive or savings in cost. Neither is currently foreseeable, particularly for key components such as communication with laboratory systems, imaging modalities, material management systems, billing and administration systems, and so on.

Diversity of healthcare systems and fragmentation of market

The diversity of healthcare systems throughout the EU and the world means that systems are highly specialised. The use of specialised vocabularies is merely one aspect of that specialisation. Different kinds of healthcare system cater to different market systems. Some of the major segments that are addressed by a diversity of healthcare systems include:

- Market segmentation according to language
- Market segmentation according to clinical speciality
- Market segmentation according to maturity of economic development in various countries
- Market segmentation according to acceptance of Information Technology by the society in general
- Subsegmentation of the above markets according to the level of care provided by the healthcare institutions – primary, secondary and tertiary.

Dispersion of costs of terminology and centralisation of costs of change

The total costs of lack of standard terminology may be very large. The total expenditure by an organisation on *ad hoc* terminologies may be very large. However, that spending is rarely done through a single central agency and can rarely be aggregated for accounting purposes. Therefore, net saving is difficult to measure, whereas the costs of a major modification to any central standardised system are highly visible.

Lack of large vendor commitment

Despite the participation of numerous vendors in both HL7 and various SNOMED activities, the commitment of large vendors to interoperable solutions, except to HL7v2 and LOINC for laboratory systems and DICOM for images, has been limited, and large vendors are notable by their limited presence in the SNOMED IHTSDO.

3.3.2 SNOMED related barriers

Organisational

SNOMED is now a sufficient force on the scene and raises unique barriers, so that no discussion of barriers to semantic interoperability can avoid treating issues relating specifically to SNOMED.

- The fact that SNOMED is licensed only to certain countries rather than being truly open is a major barrier to large-scale international development. It effectively precludes many social computing solutions that would otherwise be promising and excludes most of the developing world that include large emerging markets.
- SNOMED is effectively an Anglophone consortium. Unless there is major input from non-anglophone countries, there will be no effective translation, except possibly into Spanish.
- There are serious questions about the effectiveness of the IHTSDO and its structures for maintaining, quality assuring, and developing SNOMED. There is little evidence that the issues that have led to the unsatisfactory technical status described below have been addressed or can be addressed.

Technical

- SNOMED is only fit for purpose as a controlled vocabulary and system of managed identifiers
- The structure and implementation of SNOMED remain deeply flawed in terms of the definition of classes, the class hierarchy, and the relationships between classes.
 - It cannot safely be used as a source of semantics.
 - The potential for using its semantics for reliable binding to EHRs cannot be realised.
 - Semantic equivalence, or other relations, between different formulations of terms or different ways of expressing terms cannot be used reliably.
 - Post coordination cannot be performed reliably.
 - The errors are so egregious, that few clinicians will take it seriously.
- The scale of SNOMED is a major issue. SNOMED currently consists of over 400,000 concepts and is regularly issued. Any action on SNOMED as a whole is therefore a major undertaking. However, studies suggest that only a small fraction of this material has ever been used, and that no group of applications are known to use more than 5% of the total. Managing an overhead of nearly 500,000 terms for an application that needs at most 10,000 in an enterprise that needs at most 25,000 (the size of the VA/Kaiser subset), is clearly a major barrier to progress.
- The large scale and poor quality of the SNOMED relations and hierarchies make it questionable whether they are of any value. The cost of reconstructing them *de novo* on a scale suitable for high value applications may exceed the cost of repair.
- Developing a reliable, quality assured core subset of SNOMED is a feasible task that has so far failed to attract sufficient resources to succeed. The experience of the Scandinavian countries in translating SNOMED testifies to the difficulties of coping with the current flawed structure.

Reference Terminology and Subsets

The standard description of SNOMED-CT is as a “reference ontology”. However, given the sheer size and many faults in SNOMED, much of the effort is going into the creation of specialised subsets for specific tasks. Informal presentations and discussions indicate that the creation of subsets can be a major task, often involving several person years of effort and involving difficult choices to resolve the many ambiguities and stick correctly to the rules of SNOMED. The question then arises as to whether it is the central terminology or the subset that is the “reference”.

The SNOMED identifier system, however, has become a standard that is likely to remain a reference point. Which other aspects of the SNOMED corpus will serve as a reference, and

which will serve as something from which a reference may, or may not, be developed, remains to be seen.

3.3.3 Skills and Human Capacity

The range of skills needed to develop large-scale terminologies – clinical, computational, logical and linguistic – is large. The number of people with those skills in useful combination able to work on the issue is extremely limited. Many practitioners are self-taught consultants. At least until recently, the links to draw on wider expertise in logic, linguistics, and systems design was largely missing.

More seriously, there is no clear consensus on what needs to be done and which disciplines should be prioritized is unclear

With the greatly increased interest in ontologies in both the Semantic Web and Bioinformatics community, the number of potential staff with the required skills is beginning to improve. Now the required skills can even be recruited from outside the health informatics community. However, whenever a potential solution is proposed, the immediate question remains: “Who will do it?”

3.3.4 Tools and Software

The range of tools available to develop terminologies remains extremely limited for authoring, maintaining, and deploying, and deploying terminologies, let alone for building applications that use terminologies. (See 2.3.4 above.)

Few of the tools have yet caught up with the advances in Web technology, particularly not with the Web 2.0 developments. The result

Lack of such tools that are easy to use and still be able to deal with the complexities of semantic interoperability across the Web hampers any practical use of such a system.

3.3.5 Special technical issues

Because terminologies and ontologies serve both humans and machines, their specification needs to be clear for both humans and machines. This requires that the content and specification of each terminology provide both kinds of information:

Natural language definitions

Curiously, despite the enormous technical effort, few medical terminologies, except ICPC, contain extensive natural language definitions. Volume 2 of ICD contains rules and instructions that are operational equivalents,

Most contain “preferred terms” (SNOMED) or their equivalent, which are intended to be “context free”, but are often not sufficient to disambiguate entities clearly or explain meanings to the uninitiated.

One of the major improvements brought by the involvement of “philosophical ontologists” in the OBO ontologies has been the insistence on clearly formulated text definitions for every item. While there may be a few medical terms for which such definitions would seem superfluous – e.g. “Femur” – for most it is necessary.

In addition, if formal definitions are to be used, a formal “paraphrase” of exactly what information is to be represented formally is often required. However, “paraphrases” and “text definitions” serve different purposes, and should not be confused. “Text definitions” are there for end users; “paraphrases” are there for developers.

Extent of computability and “truth in advertising”

The extent to which the semantics of a terminology or ontology is explicit in a form that can be used directly by a computer without reference to the term names or text definitions should be explicitly specified.

The specification of the computational properties of terminologies, ontologies, and thesauri is often vague or missing. This was perhaps acceptable in the early 1990s when the technologies were immature and ill defined, but does not provide a foundation for progress in the twenty-first century when these issues are increasingly well understood, standards exist, and there is an extensive comparative literature on which to draw to make description precise.

There is a particular problem with SNOMED-CT. It claims to be based on a description logic but does not make the “stated form” available routinely nor is the precise semantics of the description logic publicly posted. Both can be obtained informally, but serious questions such as the degree to which SNOMED is “post-edited” after classification so that it does not conform in all respects to the description logic definitions cannot be answered independently.

Furthermore

- all but the simplest post-coordination depends on access to the stated form
- all but the simplest mechanisms of quality assurance and potential reformulation depend on access to the stated form.

There must be serious questions about the “truth in advertising” of any product which claims to be based on a formalism but publishes neither the formalism nor the product as formulated in that formalism.

3.4 Issues critical to Europe

3.4.1 The limits of top-down development and danger of “hubris”

SNOMED has more than 450,000 codes and yet, in many applications, achieves only 25% to 50% coverage. On the other hand, most studies suggest that in practice, less than 25,000 codes are actually used, in some cases very much less. SNOMED is not alone in suffering from this problem; the same is true of almost every terminology developed, top-down, with the possible exception of those that are mandated in such a way that codes must be made to fit cases for remuneration.

The problem is that any coding system, however, large, can only enumerate a tiny fraction of all possible expressions in medical language. (Based on the GALEN experience, reasonable estimate is that, at the lowest estimate, there are over 10 billion meaningful compositions of medical concepts, just within signs, symptoms, diseases, and procedures.) Which ones will be critical is extremely difficult to determine in advance. Therefore, any system to be comprehensive must have a significant element of bottom up development, whether through feedback, natural language processing, or social computer/Web 2.0 techniques.

3.4.2 Multilingual and multicultural systems

Interoperability across Europe requires interoperability across languages and cultures. SNOMED, in particular, has been developed largely in a monolingual Anglophone environment. Given the current members of the IHTSDO, there is little motivation for change without a major input from continental Europe or other cultures with requirements for

multilingual multicultural systems. While, on the one hand, clinical medicine and clinical research become ever more international, practical care delivery at the bedside remains largely national, or more properly, local.

3.4.3 Large scale participation and openness

It is barely plausible that the translations and adaptations required for a truly multicultural and multilingual system can be managed by traditional means. The Web and the examples of open development whether of software or terminologies (e.g. Open Directory) bring the possibility of much wider participation. They also bring the possibility of spreading the involvement much wider than just Europe. However, this requires that the systems be truly open and that all those with a potential interest be not just free but encouraged and given incentives to participate.

3.4.4 Coping with the diversity of administrative and organisational systems

Each European country has its own mechanisms for managing and reimbursing healthcare. It is not realistic to expect a harmonisation of these mechanisms, let alone a single system that can be tuned to all of them. However, identification of common elements and translations should be attempted, so that the efforts done in one country or at an international basis can be used to a large extent for local needs. A good ontological structure and logically correct definition of unambiguous terms – are some of such common elements.

3.4.5 Rapid evolution and “just in time” deployment

The rate of change of medicine dictates rapid change. Large standardised systems are usually slow to change and adapt. Systems will not long remain interoperable unless the standards and terminologies evolve quickly.

In the age of the Web, other groups – e.g. the Gene Ontology and related – have achieved very rapid turn around for identifiers and much greater responsiveness for significant changes through widespread collaborative development.

Compositional systems that allow “post coordination” bring with them the possibility of generating most novel terms “just in time” by constrained composition of simpler terms. GALEN achieved much of this a decade or more ago, although without the benefit of the Web to deliver⁵¹ The Sage project in the US has made a first step in this direction with SNOMED.⁵²

3.4.6 The standardisation process

There is much concern about the standardisation process itself. HL7’s executive held a meeting in October 2007 concerned with a process which has taken 15 years to produce a “version 3” that is not yet in routine use. The SNOMED process has been closed, and there remain serious concerns about the IHTSDO. The WHO ICD revision process has likewise been extremely slow – 15-20 years per revision.

⁵¹ Rector AL, Zanstra PE, Solomon WD, et al. Reconciling users' needs and formal requirements: Issues in developing a reusable ontology for medicine. IEEE Transactions on Information Technology in BioMedicine. 1999;2:229-242

⁵² The SAGE Guideline Model: Achievements and Overview, Samson W. Tu MS1*, James R. Campbell MD2, Julie Glasgow MD3, Mark A. Nyman MD4, Robert McClure MD5, James McClay MD2, Craig Parker MD, MS6, Karen M. Hrabak MSN, RNC2, David Berg, Tony Weida PhD5, James G. Mansfield PhD7, Mark A. Musen MD, PhD1, and Robert M. Abarbanel MD, PhD8 Journal of the American Medical Informatics Association 2007;14(5):589-598

The standardisation process is also unusual in the relative paucity of serious commitment from the major players in the healthcare industry – both systems vendors and large providers. Although some vendors and providers participate in both HL7 and SNOMED, the number is limited, and many of the largest are noticeable by their absence. The exception is in the HL7 v2 and DICOM, both of which are seen as vital to the interests of industry and both of which attract vigorous industry participation and controversy amongst industrial participants – a sign that the standards matter to the participants.

The largely self-selected nature of the standardisation bodies also leads to a lack of self-criticism. Challenges to the foundational structure, whether or SNOMED, HL7 or CEN come primarily from the academic community. Unfortunately, the nature of academic incentives and funding mean that academic participation is limited. The result is that those who attend are too often those who expect to profit from attending and have a vested interest in the status quo. This situation is beginning to change, at least in HL7, with the recognition that the organisation is losing credibility because version three has been in development for nearly fifteen years without significant deployment. (By contrast, version 2 was developed in less than three years and rapidly reached near universal deployment.)

If standards are to affect interoperability on a realistic scale, then a drastically streamlined and more agile standardisation process is required. In this respect, there may be much to learn from other organisations, particularly the W3C⁵³ (the organisation that manages the World Wide Web and its related standards such as HTML, XML, RDF, SPARQL, OWL, SKOS, SML, etc. In the relatively short time since its inception W3C has managed to produce a host of standards that have become the basis of much of the IT industry. It has done so through a combination of web based resources, funded staff, seconded staff from industry.

Another organisation which has grown at a much faster rate using Web-based technologies is the Gene Ontology and OBO, as already mentioned above (See 2.3.6)

Although the healthcare IT industry and community are smaller, at least than the Web Community, the experience of these groups suggests that large standardisation processes, including vocabularies and terminologies, can be developed at a much faster rate than is currently the experience in healthcare.

3.5 Open questions and research needs

3.5.1 Effective re-use

The goal of re-use of ontologies and terminologies remains almost as elusive today as software re-use was two decades ago. Some progress is being made, and modularisation and re-use is a ‘hot topic’ for research.

Few applications are likely to require entire terminologies/ontologies as large as SNOMED. Few may even need the entire NCI thesaurus or even perhaps the Gene Ontology. However, it is also true that few applications will be able to be built using just one terminology or ontology, even of these broad terminologies large reference resources. Therefore, the issues of re-use, segmentation and modularisation are of paramount interest.

The author has recently raised the issues of Ontology Binding Interfaces and patterns of modular use and of ontology mining as two paradigms that need to be explored.

⁵³ <http://www.w3.org>

3.5.2 Binding of terminology and EHRs

Given two complex standards that have been developed semi-independently with minimal coordination, specifying which codes can be used where in EHR standards is a major problem. The Archetype framework takes a first step in this direction by providing a mechanism of indirection and local placeholder codes. Recently, specification using expressions in a “terminology query language” have been suggested informally by Ocean Informatics⁵⁴. This approach seems promising, but remains to be tested.

The Terminofo group of HL7 has produced a large guide, but there is no formal means of validating or verifying it and it is not supported by tools.

A more formal approach, also not tested on a large scale has been proposed by Rector in two recent papers^{55 56}. The difficulties of developing tools to assist in this process is documented in a further recent paper that derives from the experience in the Semantic Mining Network of Excellence.⁵⁷[[Qamar R, Rector A; Unambiguous data modelling to ensure higher accuracy term binding to clinical terminologies. 2007; Medinfo 2007: Brisbane, Australia: 675-678.]]

3.5.3 Human factors

The importance of reproducibility, inter-rater reliability, and human factors more broadly is seriously under researched. Although some work has been done in this area (see 2.2.2), the disproportion between technical efforts and human factors efforts in the whole area of terminologies remains striking.

3.5.4 Natural language use for compiling ontologies

It is a peculiar fact that linguistics and ontologists often find it difficult to work with each other. Recently there is beginning to be some useful developments in using natural language methods to supplement ontologies, particularly in the BootStrep project (See 3.2.3) and elsewhere in Bioinformatics where “Text Mining” is taking an increasingly influential role

3.5.5 Integration of Semantic, Statistical Methods, and Social Computing Models: How best to exploit Google, Web 2.0 Text Mining and related tools

The rise of statistical methods, link mining, search engines and Google presents a major new paradigm. How best to integrate it with semantic techniques remains a major question. (See also research challenges.)

3.5.6 Integration of terminology and background information resources

One of the major successes of the GALEN project was the generation of natural language from its description logic framework. This was particularly used by the French in the development of their procedure terminology.

⁵⁴ <http://oceaninformatics.biz/CMS/index.php>

⁵⁵ Rector AL; What's in a code: Towards a formal account of the relation of ontologies and coding systems. 2007; Medinfo 2007: Brisbane, Australia: IOS Press; 730-734

⁵⁶ Rector A, Qamar R, Marley T; Binding ontologies & coding systems to electronic health records and messages. 2006; Formal Biomedical Knowledge Representation (KR-MED 2006) (To appear in J Applied Ontologies) CEUR Workshop Proceedings 222: Baltimore: CEUR; 11-19

⁵⁷ Qamar R, Rector A; Unambiguous data modeling to ensure higher accuracy term binding to clinical terminologies. 2007; Medinfo 2007: Brisbane, Australia: 675-678.

More recently, there has been extensive work on natural language generation for presenting information, most notably by Scott's group.

Given that Natural Language Generation is now a rapidly developing field (see e.g. the ACM group on natural language generation⁵⁸) and that the use of description representations suitable for natural language generation is becoming more widespread with SNOMED, it would seem opportune to extend this work.⁵⁹

3.5.7 Other technical issues

- Epistemics, data structures, and domain knowledge
- Uncertainty
- Generalities and normative modelling
- Metamodeling and the relationship between data structures and ontologies.
- Formal concept analysis

3.6 Gaps and development needs

3.6.1 Central services

The only central resources for terminology are the US NLM related services, the NCI, and the US National Center for BioOntologies. None of these provides strong access to multilingual resources. None have a strong coverage of European national terminologies.

There are no central services for collaborative ontology authoring and maintenance, except through the Open Biomedical Ontologies Consortium, although several groups aspire to provide such a platform.

3.6.2 Coordinated input to standards bodies and involvement of Providers and Vendors in the standards process

The shortcomings of the standards process are noted in 3.4.6. The lack of an effective forum in which standards that matter to their users – vendors and service providers – can be openly discussed and developed is a major gap in the system.

3.6.3 Collaborative development

Collaborative open development of terminologies remains a dream rather than a reality. Collaborative tools are a major goal of both the NCBO in the US and the NEON project in Europe.

Progress growing out of the GALEN tools has begun to be part of the SNOB editing environment and related Terminologies EU effort⁶⁰.

⁵⁸ <http://www.siggen.org/>

⁵⁹ Hallett, C., Power, R., and Scott, D.(2006). Summarisation and Visualisation of e-Health Data Repositories. UK E-Science All-Hands Meeting, Nottingham, UK

⁶⁰ See <http://www.terminologies.eu>

3.6.4 Effective collaboration on translational medicine

Although there are many initiatives, the gap between healthcare IT and translational medicine remains broad. A major threat is that incompatible standards will develop in the two areas. This is already happening to some extent with the use of the ISO 11179 metadata standard in NCI efforts, which is largely unknown otherwise in the healthcare informatics field.

3.6.5 Effective exploitation of the new technology flowing from Web 2.0 and the Semantic Web

There have been massive changes in technology since the major healthcare terminology (and other standards) were established. However, by and large the technological base remains firmly rooted in a previous generation of technologies. With the exception of the widespread use of XML, other aspects of Web technology for collaboration or of the new tools for knowledge representation – RDF(S), SKOS, and OWL – have remained woefully limited. While there is an active W3C Healthcare and Life Science Special Interest Group, there is only limited interaction with the mainstream healthcare IT community.

4 Action plan and recommendations for the Semantic Interoperability Roadmap

4.1 Choices

The choices below represent a spectrum. They are not mutually exclusive.

4.1.1 Do Nothing

If nothing is done, the individual EU states will continue to go their own way, which will make it impossible to realise potential synergies in research and public health and make market fragmentation inevitable.

4.1.2 Focused areas only

Focus on areas where there is maximum benefit to international collaboration. We would aim these as key enabling technologies. Prime targets would be

- Public health
- Clinical and translational research
- Patient safety

Concentration on focused areas might naturally evolve into a policy of gradual convergence.

4.1.3 Facilitation and support for common requirements

Facilitation can be of two types

- Facilitation of initiatives

There are a number of initiatives, notable the ICD-11/ISO/SNOMED Subset cluster that would bring major benefits that require collective support.

Facilitation of member states engagement with SNOMED and ISO, particularly over efforts to make them multilingual and multicultural, also fall into this category

- Facilitation of tools, facilities, and human capacity

There is a desperate shortage of good, freely available and/or open source tools for terminology development and deployment. The absence of such tools severely limits participation in the standards process and the development of new initiatives. Support for the development of such tools would have a major impact. Failure to support the development of such tools will make progress all but impossible.

Along with support for tools must go the support for the training in their use.

4.1.4 Full fledged support for common, possibly mandated standards

Start now to move towards an EU-wide set of standards that would allow full interoperability. Any such effort must show that it is cost effective, and faces serious barriers. Support for

fully common standards across medicine may come, but it seems likely that it will best be built out of more modest initiatives.

4.2 Principles

4.2.1 Purposes: What's it for

Any terminology supported should have clearly defined purposes and tests that it is fit for those purposes. It may be hoped that the terminology will be useful beyond the purposes for which it was originally developed; it may be that the design deliberately emphasises broad use across a range of functions. However, experience shows that at least one, at preferably two, key purposes and use cases need to be defined for any terminology or ontology effort to be successful.

4.2.2 Priorities: What must be done now? What can be built on now?

Some things must be done before others.

Relatively coarse grained standards for international collaboration on tasks in translational medicine and public health achieved quickly may be of more value than fine grained ones achieved much more slowly.

In many cases, tools must be built before they can be used, or at least before their use can be expanded beyond a limited group of users intimately involved in their development.

4.2.3 Involvement of providers and vendors: Who will deliver the results?

Interoperability in terminology, as in everything else, must be deployed by the providers. Both incentives and sanctions may be required. However, it will rarely be in any one providers' interest, on their own, to opt for interoperability. In most cases, only small providers attempting to work in niche markets have natural benefits. For large providers the balance may be on the side of avoiding interoperability; at best, the added value is unlikely to outweigh the cost of dealing with legacy applications.

The cost of converting legacy systems must be addressed in these incentives and sanctions. Health informatics is in the awkward position of having poor penetration, but still having a large enough installed base that the cost of conversion from legacy systems is a major issue to be addressed

4.2.4 Sustainability: Any action must be for the long term

Change and evolution are inevitable. The scale of the task suggests that any institutions will have to be sustained for decades, probably indefinitely. The best example is that of the US National Library of Medicine, MeSH and PubMed whose influence is based, above all, on a century's development and the confidence of all concerned that it will continue to be sustained into the future.

In the face of the total cost of healthcare at between 8% and 15% of GDP, and even of healthcare informatics at billions of Euros per country across the EU the total costs of sustaining such facilities is small.

4.2.5 Openness & collaboration: The terminologies that have succeeded in bioinformatics have all been developed on an open basis

The example of bioinformatics and other fields makes it clear that only open processes will work for most large scale development. However, in a biomedical, area, there are special challenges to governance and editorial policy to be addressed.

4.2.6 E-Terminology and the Web:

An important part of open participatory processes will be the new Web technologies. Any initiative that ignores the Web is almost bound to be overtaken by one that embraces it.

4.3 Long term goals to achieve desirable outcomes

The vision and desirable and undesirable outcomes are set out in 1.7: "Visions for the Future".

The goals desirable outcomes can be divided into Capabilities – what it should be able to do – content – what is actually in the terminologies and resources – tools – what software is needed to make it possible to work with them – and process – what is required to sustain them.

4.3.1 Target capabilities of new-generation ontologies and terminologies

- Clear articulation of purpose, scope, limitations and criteria for quality assurance.
- Clear technical specifications formal foundations coupled to clear text based resources
- Integration with other Web Resources and Digital Libraries.
- Reliable recognition of semantic equivalence between pre-coordinated and post-coordinated forms
- Methods and tools for formal specification and automatic validation of the binding of terminologies to EHRs and Decision Support Systems.
- Flexibility, rapid adaptation, and responsiveness
- Scalability – the terminologies, ontologies, and knowledge representations themselves should not grow exponentially, and they should contribute to controlling the combinatorial explosion of special cases and special functionality in the software that uses them.
- Collaborative development – which requires that they be understandable and intuitive for the clinicians and scientists who must build and maintain them.

4.3.2 Content

- (i) a set of interoperable and computable biomedical terminologies responsive to the needs of the communities that need them.
- (ii) inclusion of unambiguous formal and natural language definitions for all terms in (i).
- (iii) availability of lexical and linguistic to support the use of (i) and (ii) in all important European languages.

- (iv) availability of knowledge representation resources and formalisms using (i-iii) sufficient to support the needs of patient care, public health, clinical research, and health service management.
- (v) Comprehensive maps between (i-iv) and major classification systems, including the successors to the ICD.
- (vi) Availability of a library of adaptable information models along with bindings to (i-v).
- (vii) Availability of multilingual language generation based on above
- (viii) Availability of multilingual automatic encoding from text to support above.(to be coupled with voice recognition which is expected to be available from independent sources.)

4.3.3 Tools

- (viii)Widespread availability of methodologies and tools for implementing the content into practical applications at less cost than developing new content from scratch coupled with
- tools that use the content to generate applications quickly and at far less cost than currently.
- (ix) Wide spread availability and use of tools exploiting the Web for collaborative development and maintenance of (i-vi) by the communities involved.
- (x) Widespread availability of tools, often based on Web technologies, for deploying and disseminating (i-vi) including mediators and repositories so that specialist ontologies and terminologies can exchange information wherever possible.
- (xi) Widespread availability of tools for systematic open quality assurance and feedback on all content
- Tools to allow above to synergise with Web Mining and Text mining approaches

4.3.4 Process

- (xii) A highly responsive sustainable process for maintaining and updating the content by clinical experts. This process will need to be radically different from current standardisation processes if it is to succeed. It should be world-wide and engage with the developing world.
- Sustainable development of tools, central repositories and distribution resources. This almost certainly requires government involvement analogous to the resources devoted in the US to MeSH, PubMed, and the other common resources provided by the National Library of Medicine.
- (xiii) Clear scoping statements limiting the scope of terminologies and ontologies to be interoperable and when investment in interoperability has a positive cost/benefit ratio.
- (xiv) Existence of sustainable mechanisms that engage communities and result in effective QA
- (xv) Close integration of Clinical, Research, Best Practice, and Management communities in their relevant areas of content.

4.4 Priorities for the next five years: Recommendations

4.4.1 Content

Open Collaborative development of ICD-11 using Web based technologies

This is needed both for its own sake and could serve as a demonstration for other open collaborative developments. Where possible this should be mapped to the semantically sound subset of SNOMED, or if necessary, to SNOMED itself. (Mapping it to the existing structure may be necessary as a political exercise but will squander resources in terms of technical achievement)⁶¹.

Demonstration of a well curated, internationalised version of a subset of SNOMED

This is required both in its own right and to provide evidence for long term decisions on the role of SNOMED in semantic interoperability in Europe.

Demonstration of a reformulation of one or more suitable subsets of SNOMED using modern tools, with clear QA, with the resources to make it multilingual and multicultural. The VA/KP subset is one possible target, Various European subsets are others. The action should be undertaken in cooperation with the IHTSDO but should *not* be under its control or limited by its strictures.

Those states that are members of the SNOMED IHTSDO, perhaps with negotiation of others, should be encouraged to lead if possible. The effort should focus on a particular application – e.g. sensitivities and allergies – and be of a size to make it representative but manageable – e.g. not greater than 25,000 concepts. Given the evidence from this effort, the remaining states, and the EC itself, would be in a position to make informed decisions about how to interact with SNOMED in the future. One possible target for such an action would could be based on the VA/Kaiser Permanente Subset of SNMED-CT, which would have the advantage of engaging international collaboration. Alternatively a subset relevant to the ICD-!! effort might be chosen. The timescale for this effort is seen as not more than three years.

Clear policy on linkage to UMLS CUIs and LUIs

Adoption of a clear policy that, insofar as possible, all terminologies should be mapped to UMLS CUIs and LUIs, either by their originators or in collaboration with the US National Library of Medicine.

4.4.2 Tools and Capacity

Development of one or more freely available, preferably open-source, authoring environments and toolsets to support terminologies and ontologies.

These should be capable of handling compositional terminologies in general and SNOMED in particular. A combination of the capabilities of the SNOB⁶² browser growing out of

⁶¹ sometimes called “aggregation logic”, but this phrase is avoided here because of its possible confusion of aggregations of codes with aggregation of patients.

⁶² <http://snob.eggbird.eu/>

GALEN and Protégé-OWL⁶³ environment linked to UK projects and the US National Center for BioOntologies should be encouraged. There is an urgent need for collaborative open tools for terminology development to be developed and made widely available.

Linking of tools to developments in Web 2.0, Social Computing, the Semantic Web, Text Mining, and related disciplines – probably in conjunction with development of ICD-11.

In satisfying the need for open collaborative tools, developers should be given incentives to look beyond traditional approaches to new methods drawing on Web 2.0 technologies and text and web mining should be encouraged. In this respect, the myExperiment⁶⁴ development in Bioinformatics might provide a model of a lightweight process that is having a clear impact. myExperiment is aimed at those developing workflows; we need a similar site devoted to those developing terminologies.

The development of the ICD-11 within a collaborative framework forms a near ideal vehicle for this task.

Development of environments for co-ordinated development of terminologies and medical record standards, starting with the addition of facilities for linking terminologies to editors for the Archetype standard and CEN EN 13606.

This is a high priority task. See also deliverable 4.

Development of a prototype network of terminology and archetype servers for European countries.

The development of tools for coordinated authoring of terminologies and archetypes should lead, in the medium term to a network of “just in time” centres/web sites where users can get quick responses to their needs. Ongoing support for such centres will be required.

Research on combined approaches to knowledge representation, ontologies, and web technologies.

How to achieve the optimal balance between the various technologies for clinical applications is an open question. Most clinical applications are currently focused on technologies first developed in the 1980s and 1990s and have largely ignored the last decades explosion in both semantic, logical, web-based, and language tools. A major effort to rectify this is essential.

4.4.3 Process

User involvement and “ownership” of resources

A process should be started to involve the prospective primary user community i.e. the health care workers to an extent that they feel they are the ‘owners’ and ‘custodians’ of this resource. The health care workers have the burden of entering the data. If there is no dividend for this burden, all other actions are futile!

⁶³ <http://protégé.stanford.edu>; <http://www.co-ode.org>

⁶⁴ <http://www.myexperiment.org/>

An honest and public evaluation by independent evaluators of what SNOMED-CT can and cannot be used effectively and safely in healthcare systems software

The current state of SNOMED is a matter of controversy and little sound evidence on several important questions:

- The relative cost of implementation using different techniques.
- The cost of migration from proprietary terminologies
- The reproducibility and inter-rater reliability of various mechanisms of entering SNOMED terms
- The comparative accuracy
- The potential benefits and hazards to patient safety of using SNOMED in decision support.

If SNOMED is to be seriously considered as a standard, whether *de facto* or even more *de jure*, these questions need answers urgently.

Formal links and a European centre of expertise for collaboration with the US National Center for BioMedical Ontologies (NCBO) and collaboration with them on their tools and kits and with the National Cancer Institute on its CaBig and EVS systems

Collaboration requires resources. It takes real expertise to use software developed elsewhere. The NCI core technologies are particularly large. It is unrealistic to expect each project that might wish to use them to take on this effort on its own without special support. The CancerGrid project⁶⁵ in the UK has already undertaken much of this task and provides a local host. NCBO material is already provided and hosted by the European Bioinformatics Institute. These, and similar, initiatives should be supported and provided with the additional resources to enable them to make their expertise more widely available.

4.4.4 Clinical research and Translational Medicine

Convergence on a common terminology for clinical trials and longitudinal studies: Convergence on a common terminology for BioBanking

Clinical research and BioBanking are both world-wide efforts. Major initiatives in translational medicine involving both are in progress in many countries. A specific initiative to investigate:

- The UK BioBank initiative
- Other UK BioBank initiatives, in progress or planning
- The European Bioinformatics Institute, which is the home of the OBO consortium
- The US National Cancer Institute
- The Ontology for Clinical Research (OCRe)

⁶⁵ <http://www.cancergrid.org/>

5 Glossary

- *Background knowledge base* – the common knowledge to be assumed by the system, including both the ontology – what is universally true – and generalisations about what is typically true.
- *Classification* – an organisation of entities into classes for a specific purpose such as international reporting or remuneration. Examples ICD and Diagnosis Related Groups.
- *Controlled Vocabulary* – a list of specified items to be used for some purpose, usually in an information system to reduce ambiguity, misspellings, etc.
- *Lexicon* – A list of linguistic units that may be attached to a controlled vocabulary or ontology, in a specific language or sublanguage, often including linguistic information such as synonyms, preferred terms, parts of speech, inflections and other grammatical material. Example: Term terms and lexical material in UMLS identified by Lexical Unique Identifiers LUIs)
- *Ontology (sensu information system)* – a symbolic logical model of some part of the meanings of the notions used in a field, *i.e.* those things which are universally true or true by definition.⁶⁶ The key relationship in an ontology is “subsumption” or “kind-of”. Every instance of a subkind must be an instance of the kind, without exception. Typically, ontologies are implemented in logic languages such as Ontylog or OWL or frame systems such as Protégé-Frames. Examples: The GALEN Core Model, the stated form of SNOMED.
- *Paraphrase* – the nearest statement in natural language to a definition in some formally computable language.
- *Post-coordination* – composing of concepts as used
- *Pre-coordination* – enumeration of all concepts before use
- *System of identifiers (“codes”)* – Controlled vocabularies, and many lexicons, ontologies, and thesauri, are usually accompanied by systems of identifiers for their units, *e.g.* Typically, identifiers act as the primary unambiguous means of referring to the entities in the system for computational purposes with the text form being used for communication with users. Examples the “Concept Unique Identifiers (CUIs) from the UMLS, SNOMED Identifiers, etc. In many contexts, identifiers are known as “codes.”
- *Terminology* – Any or all of the above in various combinations. Most health terminologies consist, at a minimum, of a controlled vocabulary and a system of identifiers. They may include extended lexicons, ontologies, thesauri or background knowledge base. This definition is deliberately broader and less specific than that in most of the standard references and intended to approximate common usage.
- *Text Definition* – a user oriented unambiguous of the meaning of an entity.
- *Thesaurus* – a system of terms organised for navigation with the primary relationship being “broader than”/“narrower than”. The “broader than”/“Narrower than” relation is explicitly not limited to subsumption/kind of relation. It is a general form of linguistic hyper/hyponymy aimed at assisting human navigation. However, it is explicitly not intended that it be used as the basis for logical inferences, *e.g.* in decision support. Examples MeSH, WordNet.

⁶⁶ Different authors refer to the meanings as “concepts”, “universals,” “categories”. Note that the word “ontology” was borrowed from philosophy, and that there remain controversies concerning the extent to which the symbolic models referred to as ontologies used in information systems should conform to principles laid down by philosophers for ontologies understood as part of the philosophical study of being.

6 Appendix

Why do it the hard way? The Case for an Expressive Description Logic for SNOMED

Alan Rector, Sebastian Brandt, Jay Kola
School of Computer Science, University of Manchester, Manchester M13 9PL,
(rector | brandt | jay.kola)[@cs.manchester.ac.uk](mailto:rector | brandt | jay.kola@cs.manchester.ac.uk)

Abstract

Considerable effort has been invested in suggestions for a modernization of SNOMED-CT. In the present paper, we argue that many of the difficulties identified could be more easily dealt with using schemas in a more expressive language than that in which SNOMED was originally formulated. There has been major progress both on description logics and associated classifiers since SNOMED was originally developed. The emergence of the standard Web Ontology language and its latest revision, OWL 1.1 are leading to a rapid proliferation of tools. Combined with the increase in computing power in the past two decades, these developments mean that many of the restrictions that limited SNOMED's original formulation no longer need apply. We contend that future development of SNOMED could be made much easier by adopting a more expressive language and more modern tools.

Introduction

Since its first release in 2002, several countries around the world have embraced SNOMED-CT (SNOMED for short) as a reference terminology for their national health-care institutions. Apart from national extensions for content, however, the structure of SNOMED, neither the structure of SNOMED nor the expressiveness of the underlying formalism, Ontylog, has changed significantly since their initial development in early to mid 1990s.

Since then, there have been major developments in both logic-based formalisms for knowledge representation

and formal ontology design. SNOMED has been reviewed in the light of these advances, and several proposals for improvements have been made. For example, Bodenreider (1) examined the specialization hierarchy of SNOMED classes and suggested that thousands of classes are apparently defined at variance with basic ontological principles. Schulz (2) discussed 'relationship groups', a construct unique to SNOMED's representation language, most probably a relic from concerns with performance in early implementations combined with the inheritance from the structure of older SNOMED International. He concludes that relationship groups can be replaced by mereological relations using syntactic constructs available in all modern ontology languages, including Ontology, and that this would significantly improve the clarity of the definitions. In a separate publication, by Schulz and his colleagues (3) a broad range of ontological problems in SNOMED are identified and a comprehensive set of remedies are proposed, affecting not only the definitions of classes in SNOMED. In that paper, the authors argue for a modest extension of the logical formalism underlying SNOMED, permitting simpler and clearer definitions of classes and especially of the relations used to define relationships between classes.

In this paper we review these problems and further focus on the issues around SNOMED's "context model" and the notion of "Situations involving specific context." We also further discuss issues around part-whole relations and the determining the equivalence between findings and observables. Following on from the theoretical arguments in previous

papers (4)(5), we argue for a schema for that integrates context with other concepts, so that all related concepts appear in the same hierarchy. Formulating such an integrated schema requires a further extension of the formalism beyond that proposed by Schulz to one that includes negation, disjunction and “general inclusion axioms”, the obvious candidate being the W3C standard Web Ontology Language (OWL 1.1). Although in the past the scale of SNOMED has been a barrier to the use of OWL and related formalisms, this is no longer the case.

We argue that reformulation using such a schema and a more expressive language would have major advantages:

- A uniform, clear and understandable schema for all concepts used in clinical records.
- Elimination of the need for special mechanisms to deal with context, partonomy, and role groups.
- More effective leveraging of the underlying logical representation to organise and quality assure the SNOMED hierarchies
- Improved ability to recognise semantic equivalence between post-coordinated and pre-coordinated expressions and between “observables” with “values” and the corresponding “findings”.
- Improved ability to modularise and segment SNOMED for specific purposes
- Access to the tools and techniques being developed by the wider Semantic Web and OWL communities.
- In outline, the proposals are:
- To represent all concepts used in clinical re-cords (findings, observables, and procedures) uniformly as fully defined “situations” that include any context required and that deal with negation explicitly and formally.
- To represent all sites explicitly whether as the whole or as the disjunction of the whole and its parts.
- To define observables and related findings in such a way that the

classifier can be used to recognise the equivalence between a situation involving an observable with a given value and the corresponding finding of the observable with that value – e.g., between an observable of “blood pressure” qualified by “elevated” and a finding of “elevated blood pressure”.

- To organise the stated form as a set of modules that can be easily separated for specific applications.

The remainder of the present paper is organized as follows. In Section 2, we begin by highlighting features of SNOMED for which improvements might be gained by utilizing standard OWL modelling patterns. In Section 3, we discuss potential obstacles to the changes advocated in Section 2, the general feasibility of structural changes to SNOMED in the future and the potential benefits gained by them.

SNOMED from an OWL perspective

As the present section deals with the extension of Ontylog by constructs from OWL 1.1⁶⁷ and as OWL is strictly more expressive than Ontylog, we begin by viewing SNOMED as an OWL ontology. Hence, we present SNOMED definitions using the Manchester OWL Syntax (6). Moreover, we refer interchangeably to SNOMED *concepts* or their equivalent OWL *classes*.

Situations with explicit context

Within SNOMED, findings and conditions can either appear as plain subconcepts of clinical finding, condition, etc., or can be embedded within a construct previously called “context dependent concept”, now “situation with explicit context”. Situations can be used to specify additional information relating to:

- The presence or absence of the phenomena under consideration;

⁶⁷ <http://www.webont.org/owl/1.1/>

- “Modalities” such as “risk”, etc.;
- “Temporal” positioning such as “past”, “present actual”, etc.; and
- “Subject of care”, such as “subject of record”, “fetus”, etc.

For instance, history of vertebral fracture is defined in SNOMED as follows:

391094005 H/O: vertebral fracture *equivTo*:

243796009|Situation with explicit context|:

```
(246090004|Associated finding|=64572001|Disease|
(116676008|Associated morphology|=72704001|Fracture|
363698007|Finding site|=51282000|Bone structure of spine)),
408729009|Finding context|=410515003|Known present|,
408732007|Subject relationship context|=410604004|Subject of record|,
408731000|Temporal context|=410513005|Past))
```

The class does not merely define vertebral fracture, but a fracture known to have been present in the bone structure of the spine of the subject of the record in the past. The way these situations are defined in SNOMED has three major drawbacks.

Firstly, a finding of interest can either occur on its own, e.g., 195826005|Nasal obstruction, or as a situation: 267100006|Nasal obstruction present (situation). This means that any query for the notion of nasal obstruction must look in both places and that the distinctions between the two hierarchies are often unclear.

Secondly, the use of negation to express the absence of, e.g., a disorder in the definition of a situation is restricted and requires special processing, making it difficult to verify that situations involving negation are classified correctly.

Thirdly, some of the additional qualifiers available for the definition of situations conflate categories that should be kept apart. The overall model is complex as indicated in the detailed documentation required for the use of SNOMED in HL7 (7).

Moreover, a careful review suggests that no situation with explicit context defined in SNOMED includes the absence or presence of more than one proper clinical finding. Apparent exceptions are cases where the associated finding also carries temporal information or is altogether redundant. For instance, 407553003|History of glandular fever has the associated findings 40733004|Infectious disease (disorder) and 307294006|Personal history finding, the latter rather addressing the temporal context. In 161496006|History of chronic ear infection, the findings associated are 129127001|Infection of ear and 118236001|Ear and auditory finding, the latter being redundant due to the definition of the former.

Given the flexibility of class definitions in OWL, the above drawbacks can be overcome easily. Firstly, the distinction between findings within and without situations can be removed by separating codes/classes into “kernel codes/classes” that represent the entity itself and “recordable codes/classes” for the class of situations in which the entity either is, or is not, present. Kernel classes would then be purely internal. Only recordable codes would then be used in the EHR.

Each recordable code would represent a class expression for a “Clinical situation”, at its simplest:

```
Situation THAT
  [includes|NOT includes] SOME (Kernel_entity THAT
    ... values and qualifiers)
...
and has_temporal_qualifier SOME Temporal_qualifier
```

For example:

```
Situation THAT
  includes SOME Diabetes_type_2
  and has_temporal_qualifier VALUE Current_actual
```

```
Situation THAT
  NOT includes SOME Diabetes_type_2
  and has_temporal_qualifier VALUE Current_actual
```

Presence or absence is dealt with by the choice of either “includes” or “NOT includes”. Note that “NOT includes” is a formal logical statement and dealt with automatically by the classifier without further special treatment.

Qualifiers such as “risk” or “family history” that “modify the axis” would be in the

situation differently from other qualifiers, as “prefixes”, e.g.,

```
Situation THAT
  includes SOME (Family_history THAT
    is_of SOME Diabetes_type_2)
```

Family history, risks, etc. can be either included or not included – *i.e.* be stated explicitly to be either present or absent just as any other kernel concept. Because of their placement as “prefixes”, all danger of confusion of axes is eliminated.

Using a related mechanism, subjects other than the patient can be dealt with by nesting situations, e.g. by nesting a situation about the fetus within a situation about the mother:

```
Situation THAT
  includes SOME (Situation THAT
    clinical_subject SOME Fetus
    and includes SOME (heart_rate THAT quant VALUE 120))
```

Classes and their definitions

In addition to merely managing manually constructed hierarchies of terms, logic-based ontology languages such as Ontylog and OWL allow classes to be defined by complex class expressions built from atomic objects contained in the ontology and constructors provided by the ontology language. The benefits of this approach are that (1) the meanings behind the classes are made explicit, and (2) the actual hierarchy of classes can be computed automatically on the basis of their definitions in the ontology and (3) multiple hierarchies for different purposes.

In SNOMED, only about 11% of the classes are fully defined. Furthermore, the vast majority of the remaining “primitive” classes have only ‘trivial’ information asserted about them. (bodenreider2004). The information is usually insufficient even to permit consistency checking, let alone automatic classification.

One reason for the limited use of fully defined classes in SNOMED is that in Ontylog, once a class is fully defined, it cannot be further qualified, which limits the use of fully defined classes.

More technically, in Ontylog, *fully defined* classes are introduced by expressions using the keyword “defconcept” and specify set of necessary and sufficient conditions. Expressions using “defconcept” correspond to OWL “equivalentClasses” axioms. *Primitive classes* are introduced by expressions using the keyword “defprimclass” and introduce a set of necessary but not sufficient conditions. Expressions introduced by “defconcept” correspond to OWL subclass axioms. In Ontylog, the same class cannot be the subject (left-hand side) of both a *defconcept* statement (equivalence class axiom) and *defprimconcept* statement (subclass axiom). In OWL, there is no such restriction. This means that in OWL additional information can be added to fully defined classes. In fact this just a special case of OWL’s support for what are termed “general inclusion axioms” – *i.e.* axioms that allow any class, primitive or defined, to be asserted to be a subclass of any other class including the “restrictions” that correspond to SNOMED qualifiers.⁶⁸

Since in our proposed schemas, all “recordable codes” correspond to fully defined classes, OWL’s support for “general inclusion axioms” is essential. Given that, the proposed schema leads straightforwardly to a single classification hierarchy that integrates context with the hierarchy as a whole.

Representing parts and wholes

Any clinical representation needs to support the pattern that, in general but not always, a disorder of the part is a disorder of the whole. For example, a disorder of a heart valve is a disorder of the heart, a fracture of the neck of the femur is a fracture of the femur, a procedure on a lobe of the lung is a procedure on the lung, etc. SNOMED originally dealt with

⁶⁸ Technically, qualifiers in SNOMED correspond to “restrictions” in OWL. A “restriction” in OWL is just a special kind of class, the class of all those entities that satisfy the restriction.

this issue using “right identities”, a special type of axiom for properties. For example, the right-identity `site o is_part_of a site` expresses purely on the level of properties that it is always the case that anything that has a site that is a part of a whole also has the whole as a site. Right-identities are equivalent to the construct that GALEN called “refinement” (8) and can be expressed in OWL 1.1 by a more general construct called “property chains”.

Using property chains (or “right identities”) Property chains to represent this pattern, however, requires great care. For instance, the valves are part the heart, yet failure of a heart valve does not imply heart failure, although it may lead to it. Similarly, we may want to represent the notion of “Lung operation” to include “removal of a lobe of the lung”, we do not want “pulmonectomy” (removal of the lung) to include “removal of a lobe of the lung”. These problems originate in imprecise function or procedure definitions that do not distinguish between its site (failure *at* the heart, removal operation *at* the lung) and its object (failure *of* the valve, removal *of* the lobe).

Recent versions of SNOMED have experimented with an approach suggested in (9) involving so-called SEP-Triples -- triples of the thing (E), its parts (P), and the disjunction of the thing and its parts (S). This approach was found cumbersome when implemented literally because it necessitated enumerating three nodes for every anatomical structure.⁶⁹

As OWL supports disjunctions, the distinction between whether a property applies to just the thing itself, just its parts, or both is easy to express. For example:

- Operation THAT site SOME (Lung OR is_part_of SOME Lung)
- Removal THAT site SOME Lung

- Removal THAT site SOME (Lobe THAT is_part_of SOME Lung)

The first includes all operations on a lung as a whole or any of its parts, the second includes the removal of a lung as a whole, the third, the removal of a lobe of a lung, which since it is a part of a lung, and will be classified under the first automatically. The same mechanism deals naturally with “partial” and “total”, e.g. “Removal THAT site SOME Kidney” represents a total nephrectomy; “Removal THAT site SOME (is_part_of SOME Kidney)” represents a partial nephrectomy. The disjunction `Removal THAT site SOME (is_part_of SOME (Kidney OR is_part_of some Kideny))` includes both total and partial nephrectomies.

Observables and findings

One of the persistent issues in clinical information systems is the distinction between observables and findings. Although there exists no universal consensus for both, the term observable generally refers to an aspect of the patient that can be quantified or qualified, e.g., “blood pressure”, “skin colour”, “body-mass index”, etc. A finding, on the other hand, usually refers to something which is either present or absent, possibly with additional qualification, e.g. “diabetes”, “fractures” etc, or to the state of some observable such as “increased blood pressure” which likewise may be present or absent.

SNOMED distinguishes between the classes “finding” and “observable entity” but the relationship between them is not always easy to understand. As an example, consider the finding of increased blood pressure defined in SNOMED as follows.

```
24184005|Finding of increased blood pressure (finding) □
38936003|Abnormal blood pressure (finding) AND
roleGroup SOME
(363714003|Interprets (attribute) SOME
75367002|Blood pressure (observable entity))
```

Hence, the finding of increased blood pressure implies a finding of abnormal blood pressure that interprets the observable entity blood pressure. The fact that a finding of an increased blood pressure qualifies the blood pressure as abnormally *high* as opposed to abnormally

⁶⁹ The only remaining right-identity is unrelated to procedures, stating that the active ingredient of a direct substance is an active ingredient of the whole.

low is not at all reflected in the expression. This is a common phenomenon. In many cases, most of the intended meaning behind concepts such as `finding of increased blood pressure` remains in the term name and is not reflected in a definition. This is even more obvious when comparing SNOMED's (primitive) definition of a decreased blood pressure:

```
12763006|Finding of decreased blood pressure (finding)␣
  392570002|Blood pressure finding (finding) AND
  roleGroup SOME
  (363714003|Interprets (attribute) SOME
   75367002|Blood pressure (observable entity))
```

A comparison shows that there is no distinction between the definition of increased and decreased blood pressure, except that `decreased blood pressure` does not even imply an abnormal blood pressure. (Whether or not this is intentional cannot be determined from the information available.)

Furthermore, in many models of the medical record, it is possible to express elevated blood pressure by the combination of the code for blood pressure and a code for elevated or elevated, which would naturally be translated into SNOMED as a qualifier for "elevated". However, because this qualifier is not represented in the SNOMED definition of elevated blood pressure, it is not possible to recognise the equivalence between the two formulations.

Using GCIs available in OWL ontologies, it is easy to represent the relation between observables and findings more faithfully. If the definition of increased blood pressure finding were to be:

```
Finding of increased blood pressure ==:
  Blood pressure THAT hasJudgedLevel SOME elevated
```

Then whether or not the named finding were used, the underlying meaning in the logic representation for the recordable code would be the same:

```
Situation THAT includes
  Blood pressure THAT hasjudgedLevel SOME elevated
```

Hence, we have a unified representation in which the finding corresponds to a

qualified observable and conversely. (Note that the qualifier `hasJudgedLevel` indicates a clinical judgement as opposed to a simple value reading. The issue of representing thresholds belongs to another paper.) The finding can be coded either as the code `blood pressure` (an observable) plus the qualifier `elevated` or, if used frequently, as a single code for increased blood pressure.

The notion of "finding" and "observable", properly understood are "meta" to the ontology proper. "Findings" are things whose mere presence carries information; "observables" are things that must be qualified or given a value to convey information. There are large sections of the ontology where the distinction between findings and observables follows the natural hierarchies. For example, disorders are generally findings and laboratory tests (or the physiological parameters underlying them) typically observables. However, there are cases where more complex representations may be appropriate. By these definitions some physical signs are "findings"— e.g. the presence of a lump — others are observables — e.g. pulse rate or body temperature.

Modularization

The national versions of SNOMED comprise an international core and national extension. Nevertheless, as an ontology it is currently classified and managed as a whole.

OWL ontologies have an import mechanism by virtue of which it is easy to create ontologies comprising several modules together with an import graph specifying the way in which the modules are combined. Modern ontology editors, such as Protégé, support the user in the maintenance of module structures when editing the ontology as a whole.

The reasons for breaking up the contents of an ontology into modules are numerous. Firstly, just as with chapters in books, modules allow one to partition the

contents of an ontology into thematically related sub-units that are easier to maintain and exchange. Moreover, the modules mechanism of OWL affords a natural way to import content from other sources – e.g. specialised vocabularies for different realms or termsets for alternative languages.

Modules might be of interest for SNOMED for a third reason related to the highly imbalanced frequency distribution of classes (or codes) used for documentation in the health-care domain. Statistical records about READ codes published by The Health Improvement Network⁷⁰ in the UK show that 0.9% of the READ codes account for 80% of all information ever coded by them – and 9.2% of the codes for 96% of it. Hence, the vast majority of references refer to a tiny subset of the content. A similar imbalance can be expected of usage patterns of SNOMED classes in health-care systems, not only because READ codes form a part of SNOMED but also because SNOMED is intended for use as a broad reference terminology just as READ codes are. This imbalance motivates the question whether it might not be possible to identify a small subset of SNOMED that could satisfy the users' needs in 95% of the cases. Such a subset could be used for performance improvements of systems using SNOMED by first trying the small subset and returning to the full ontology only when necessary. Moreover, techniques have been developed and are being implemented that allow to classify the whole ontology without re-classifying those modules that have already been classified.⁷¹)

Future extensions of SNOMED could benefit from the modular structure of OWL ontologies in several ways. In particular, small local or specialized extensions could be formulated as modules of new class definitions built from the existing vocabulary of SNOMED. Such so-called “pre-post-coordinated” extensions offer the advantage of limiting the number of

concepts that have to be enumerated and named in advance while still tailoring the system to the needs of particular circumstances

For example, it is neither practical nor necessary to enumerate named concepts for the family history of all possible diseases. In the modern world of “omics” and translational medicine, the family history of many more conditions is becoming relevant, and indeed the range of “post coordinated” concepts is likely to grow rapidly. A special module of pre-coordinated family history concepts appropriate to a given genomics clinic or project could speed coding without affecting the logical content of the overall coding system

Discussion

How much OWL is required?

The previous sections have suggested various extensions to the SNOMED schemas that are possible using OWL and would address identified weaknesses within SNOMED's current representation. Features of OWL required include conjunction, disjunction, full negation, existential restrictions, property chains, and general inclusion axioms. In addition, it is suggested that the ease with which OWL can be broken into modules would greatly simplify managing extensions, subsets, and localisation. Note that we are not advocating the use of numerous other constructs in OWL, in particular universal restrictions and cardinality constraints, which are known to reduce performance.

Schulz (3) suggested using a fragment of OWL with attractive computational properties, EL++ (10) to achieve many of the reformulations suggested here. EL++ meets the above requirements with the exception of full negation and disjunction. However, we argue that these features are needed to address context and partonomy cleanly. While we recognise that a move to EL++ would allow important steps forward, we argue that the

⁷⁰ THIN: <http://www.thin-uk.com/>

⁷¹ Personal communication, Ulrike Sattler, October 2007.

absence of negation and disjunction are serious disadvantages.

Is OWL ready for SNOMED?

The main argument for using a less expressive representation such as EL++ is performance and scaling. The key issue is, therefore, whether it is plausible that a more radical reformulation using a larger subset of OWL would be practical. Do the reasoners and tools available for OWL 1.1 have the computational power and maturity to deal with an ontology of about 434,000 classes and more than a million relationships between them?

SNOMED can be loaded into Protégé⁷², the standard editor for OWL ontologies, and classified by FaCT++⁷³ without any extraordinary hardware requirements, provided a 64bit Java implementation is available. On a machine with a 2Ghz dual core CPU and 2 GB memory, classification takes 30 minutes; on a larger quad-core machine with 16GB memory something less than half this.

This situation is likely to improve rapidly due to new developments in three directions: faster reasoners, module-aware reasoners, and incremental reasoners. Hermit (11), a novel hypertableaux-reasoner supports the required expressivity and shows promise to deliver significantly more performance than FaCT++ today. As long as the rephrasing of SNOMED stays within the EL++ fragment, the specialized EL++-reasoner CEL (10) might be another promising alternative. IBM's SHER project⁷⁴ has produced a modularisation strategy for the OWL-reasoner Pellet to classify all of SNOMED-CT effectively, and to classify several million individuals against the classified ontology (12). Finally, incremental classifiers are being developed for both FaCT++ and Pellet, so that the need to reclassify the ontology from scratch can be minimised.

Note also that several purposes for which SNOMED is likely to be used do not require reclassifying. One such example is classifying post-coordinated class expressions against SNOMED. Moreover, there exist small subsets of SNOMED that satisfies the vast majority its uses and which can be expected to be classifiable much more quickly. One candidate for such a subset might be one collected by the UK Connecting for Health project; another is the VA/Kaiser Permanente (VA/KP) Problem List subset⁷⁵ of SNOMED terms proposed as part of the US FDA Structured Product Labelling Resources (SPL)⁷⁶.

Altogether, we obtain the picture that OWL and its infrastructure is reasonably well in shape for SNOMED already, with significant improvements to expect in the near future.

Is SNOMED Ready for OWL?

The need for significant quality assurance and development of SNOMED is widely accepted. The issues are increasingly well documented, and the complexity of using SNOMED in practice with HL7 (e.g.(7)) or Archetypes (e.g.(13)) increasingly well documented.

Migration and legacy

The pragmatics of migration of any large software artefact should not be underestimated. In the case of SNOMED, it includes not only the ontology itself, but also the applications already based on its current structure.

However, from the perspective of health-care institutions or third-party software developers, the basic set of codes, class names, and delivery file structures for the pre-coordinated terms would remain largely unchanged. Hence, the changes might be compared to any regular update.

⁷² Available from protégé.stanford.edu.

⁷³ Personal communication, Dmitry Tarkov, 2008.

⁷⁴ http://domino.research.ibm.com/comm/research_project.s.nsf/pages/iaa.index.html

⁷⁵ <http://www.fda.gov/oc/datacouncil/term.html#med>

⁷⁶ <http://www.fda.gov/oc/datacouncil/spl.html>

Outside the context dependent branch of the hierarchies, the changes advocated in Section 2.1 can be accomplished by wrapping the relevant definitions into the expression “Situation that includes ...”, an operation that can be automated easily. The changes to the handling of partonomy would have little affect on applications and end users, although they would greatly simplify the work of developers.

The changes to the context dependent branch are significant. A significant number of concepts would be redefined to the extent that a new code was felt to be required. The guidelines for using context dependent codes would have to be revised. However, despite recent efforts, the context dependent codes are known to be problematic and are modest in number. A thorough revision that produced a uniform structure integrated with the rest of the SNOMED structure rather than a piecemeal revision with multiple variants should be attractive.

The exact costs of implementing all the recommendations expressed in the present paper cannot be determined with precision because of the irregularities of SNOMED. Where the structure of SNOMED is sound and regular, it could be done largely via scripting. Where the structure is flawed, the transformations would make those flaws more obvious but would require manual revision. Unfortunately, the correct semantics of many terms can be determined neither from the lexical structure of the identifiers nor from the semantics of the qualifiers.

Summary

In this paper we argue the case for a reformulation of the definitions of classes in SNOMED's stated form, utilizing constructors available in the ontology standard OWL 1.1, a considerable extension of SNOMED's current underlying formalism Ontylog. The main advantages of this reformulation are (1) a simpler, uniform representation of situations with explicit context, (2) a more flexible way of handling definitions of

classes, (3) a simple and uniform way to specify partonomic information for procedures and findings, (4) a clearer and more principled relation between observable entities and findings, and (5) the chance to modularize SNOMED.

The relationship between codes and EHR models such as the HL7 RIM, Archetypes, and CEN 13606 has been discussed in a previous paper (4) and is touched on in (5). These patterns also integrate much more easily with the OBO family of ontologies and other ontologies being used in molecular biology.

The price to pay for this reformulation is twofold. Firstly, the representation would need to be transformed to OWL (or an equivalent extension of Ontylog), which is an entirely automatic process already catered for by existing tools. Much more seriously, to take full advantage of the proposed schemas, the explicit content of SNOMED's definitions would need to be extended. However, that could be done progressively.

However, A number of modest sized subsets of SNOMED exist that could be used for feasibility tests. The transfer of responsibility to an international standards body (the IHTSDO) provides a natural point to consider new developments. Although increasing, use of SNOMED is still in its early stages, even in the United Kingdom. Changes made now will be much easier to implement than changes made in the future when the legacy is greater. The result of the suggested changes would be a simpler and clearer representation. Why do it the hard way? At a minimum, the feasibility of alternative schemas should be tested.

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), the EU funded Semantic Mining Network of Excellence and SemanticHEALTH FP6 Specific Support Action, IST-27328-SSA, <http://www.semantichealth.org>. The HL7 Terminology working group stimulated and contributed to many of the ideas presented here.

References

1. Bodenreider O, Smith B, A K, Burgun A. Investigating subsumption in SNOMED CT: An exploration into large description logic-based biomedical terminologies. *Artificial Intelligence in Medicine*. 2007;39:183-195.
2. Schulz S, Hanser S, Hahn U, Rogers J. The semantics of procedures and diseases in SNOMED CT. *Methods of Information in Medicine*. 2006;45:354-358.
3. Schulz S, Suntisrivaraporn B, Baader F; SNOMED CT's Problem List: Ontologists' and logician's therapy suggestions. 2007; *Medinfo 2007*: IOS Press; 802-806.
4. Rector AL; What's in a code: Towards a formal account of the relation of ontologies and coding systems. 2007; *Medinfo 2007*: Brisbane, Australia: IOS Press; 730-734.
5. Rector A, Qamar R, Marley T; Binding ontologies & coding systems to electronic health records and messages. 2006; *Formal Biomedical Knowledge Representation (KR-MED 2006)* (To appear in *J Applied Ontologies*) CEUR Workshop Proceedings 222: Baltimore: CEUR; 11-19.
6. Horridge M, Drummond N, Goodwin J, Rector A, Stevens R, Wang H; The Manchester OWL syntax. 2006; *OWL: Experiences and Directions (OWLED 06)*: Athens, Georgia: CEUR; http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS//Vol-216/submission_9.pdf.
7. Cheetham E, Markwell D, Dolin R. Using SNOMED CT in HL7 Version 3: Implementation Guide, Release 1.3. HL7 Ballot Document. 2006
8. Rector AL, Bechhofer S, Goble CA, Horrocks I, Nowlan WA, Solomon WD. The GRAIL concept modelling language for medical terminology. *Artificial Intelligence in Medicine*. 1997;9:139-171.
9. Schulz S, Hahn U; Mereotopological reasoning about parts and (w)holes in bio-ontologies. 2001; *Formal Ontology in Information Systems (FOIS-2001)*: Ogunquit, ME: ACM; 210-221.
10. Baader F, Lutz C, Suntisrivaraporn B; CEL - A polynomial time reasoner for life science ontologies. 2006; *3rd International Joint Conference on Automated Reasoning (IJCAR'06)*: Springer-Verlag (LNI 4130); 287-291.
11. Motik B, Shearer R, Horrocks I; Optimized reasoning in description logics using hypertableaux. 2007; *21st Conference on Automated Deduction (CADE-21)*: Bremen, Germany: Springer LNAI; 67-83.
12. Patel C, Cimino J, Dolby J, et al.; Matching patient records to clinical trials using ontologies. 2007; *International Semantic Web Conference (ISWC 2007)*:
13. Qamar R, Rector A; Unambiguous data modeling to ensure higher accuracy term binding to clinical terminologies. 2007; *Medinfo 2007*: Brisbane, Australia: 675-678.
