13

Medical Informatics

Alan Rector

Abstract

Description logics and related formalisms are being applied in at least five applications in medical informatics—terminology, intelligent user interfaces, decision support and semantic indexing, language technology, and systems integration. Important issues include size, complexity, connectivity, and the wide range of granularity required—medical terminologies require on the order of 250,000 concepts, some involving a dozen or more conjuncts with deep nesting; the nature of anatomy and physiology is that everything connects to everything else; and notions to be represented range from psychology to molecular biology. Technical issues for expressivity have focused on problems of part-whole relations and the need to provide "frame-like" functionality—i.e., the ability to determine efficiently what can sensibly be said about any particular concept and means of handling at least limited cases of defaults with exceptions. There are also significant problems with "semantic normalisation" and "clinical pragmatics" because understanding medical notions often depends on implicit knowledge and some notions defy easy logical formulation. The two best known efforts—*Open*GALEN and SNOMED-RT—both use idiosyncratic description logics with generally limited expressivity but specialised extensions to cope with issues around part-whole and other transitive relations. There is also a conflict between the needs for re-use and the requirement for easy understandability by domain expert authors. *Open*GALEN has coped with this conflict by introducing a layered architecture with a high level "Intermediate Representation" which insulates authors from the details of the description logic which is treated as an "assembly language" rather than the primary medium for expressing the ontology.

13.1 Background and history

13.1.1 Knowledge representation in medical applications

Description logics (DLs) and related frame-based and conceptual graph formalisms are being applied in at least five applications in Medical Informatics:

- Terminology development and, more broadly, the representation of information in health records.
- Intelligent user interfaces.
- Decision support and semantic indexing.
- Semantics oriented natural language processing.
- Semantic integration of information systems.

The seminal early work in the use of description logics in medical applications focused on the dilemma between expressiveness and tractability. Doyle and Patil [1991] attempted to apply NIKL to medical vocabulary and came to the firm conclusion that the NIKL TBox language was too restrictive to be useful for this purpose. More explicitly they despaired of users accepting the restrictions of minimally expressive TBox languages and predicted that users would find "work-arounds" which defeated the logical rigour which was their raison d'être. A first attempt at a more appropriate representation was made by Jang and Patil [1989].

However, as providing a standard controlled medical vocabulary came to be seen as one of the central issues of medical informatics, some researchers saw "compositional systems" as the only plausible route forward. The perceived urgency of the task motivated "pragmatic" approaches. Masarie *et al.* [1991] used a large frame based AI environment to produce an "interlingua" linking three of the then current terminologies in one of the exploratory projects to what became the Unified Medical Language System [Evans, 1987].

Although the National Library of Medicine chose to use lexical methods to cross map existing terminologies rather than to develop Masarie's approach to a logical interlingua, the project gave rise indirectly to the CANON group which became strong advocates of formal representations in medical terminologies [Cimino, 1994; Evans *et al.*, 1994]. A special issue of the American Journal of Medical Informatics (Volume 1, issue 3) summarised the material from its seminal workshop.

The CANON group brought together several other strands of then current work:

- The Medical Entities Dictionary developed by Cimino *et al.* [1989] as a large semantic network.
- The related GALEN [Rector *et al.*, 1993; Rector and Nowlan, 1994] and PEN&PAD [Nowlan *et al.*, 1991a; 1991b; Nowlan and Rector, 1991] programmes from Europe.

- A series of projects on the use of Sowa's conceptual graphs for representing medical vocabularies, of which the best known is the one by Campbell *et al.* [1994] but includes also work by Bell *et al.* [1994].

In addition, the group interacted with more linguistic work by Friedman *et al.* [1994] and Sager *et al.* [1994] which, along with Tuttle [1994], served as a contrast and a reality check.

There have been two large scale outcomes of this work:

- The SNOMED-Reference Terminology (SNOMED-RT) and SNOMED-Clinical Terms (SNOMED-CT) projects under the College of American Pathologists,[1] which seeks to produce a terminology, all of whose concepts are represented in a subset of KRSS and formally classified, which was released at the end of 2000 [Spackman *et al.*, 1997]. A further cooperation with the UK Clinical Terms project is to produce an international version to be released in 2002.[2]
- *Open*GALEN, which seeks to produce a reference ontology in a specialised description logic for use in developing and managing other terminologies and indexing knowledge required for decision support, user interfaces and other knowledge management tasks.[3]

In addition there have been a number of projects on language processing in medicine which have included significant work on formal knowledge representation, particularly the work by Hahn using LOOM [Hahn *et al.*, 1999a; 1999c], which has produced a range of large scale results in both language engineering and ontologies proper, and by Zweigenbaum using a specially restricted frame representation in a similar way [Zweigenbaum *et al.*, 1995]. Another important task is the indexing and retrieval of medical literature which has been addressed by McGuinness [1999].

Applications of ontologies within medicine, not based on description logics, include the work by Musen [1998] on re-usable problem solving methods and ontology driven knowledge acquisition in the PROTÉGÉ project which, at least so far, has specifically not used a description logic or other formal basis for its ontology, but rather based its ontologies around the OKBC and DAML standards. As these standards are converging with description logics in OIL and DAML+OIL [Fensel *et al.*, 2001; Horrocks and Patel-Schneider, 2001], convergence with PROTÉGÉ is under active discussion.

Stefannelli and Schreiber likewise have produced a body of work based around adaptations of the KADS architecture using ontologies as the basis for intelligent

[1] `http://www.snomed.org/`
[2] `http://www.coding.nhsia.nhs.uk/`
[3] `http://www.opengalen.org`

systems and agent architectures [Schreiber *et al.*, 1993; Vanheijst *et al.*, 1995; Falasconi *et al.*, 1997].

Another major effort on knowledge representation in medicine is the Digital Anatomist project [Rosse *et al.*, 1998; Agoncillo *et al.*, 1999; Mejino and Rosse, 1999], which currently does not use a description logic but which represents a benchmark for a comprehensive, carefully curated and validated knowledge base based on carefully analysed ontological commitments and distinctions manifest in a meticulously defined hierarchy of high level concepts such as "organ", "tissue", etc. It poses a challenge to any system purporting to a comprehensive representation of medical knowledge.

13.1.2 The medical environment

Behind most of these applications is the aspiration to re-use clinical data—either to integrate systems, to link patient records to decision support and knowledge management, or to re-use information collected in the course of patient care for management, remuneration, quality assurance or research.

There has been a widespread move to greater integration and to "Electronic Patient Records" (EPRs), also known variously as "Computer based Patient Records" (CPRs) or (CBPRs). The goal behind these moves is three-fold:

- To improve patient care through providing better information on current patients, warnings, and decision support to healthcare professionals—e.g., to be able to identify patients' known problems and treatments, warn of potential drug interactions and contraindications, or suggest management based on established guidelines.
- To capture improved information for planning and management within healthcare institutions by re-using information collected at the point of care for all secondary functions—e.g., to re-use diagnosis and treatment information collected during patient care for statistical reporting, quality assurance, and remuneration.
- To integrate the disparate information systems typical of most healthcare institutions.

Major reports justifying electronic patient records have been issued, amongst others, by the Institute of Medicine [Dick and Steen, 1991], the Computer based Patient Record Institute (CPRI), and the UK National Health Service [NHS National Health Service Executive, 1998]. This pressure is increasing with moves to greater clinical accountability and concern with clinical errors [Kohn *et al.*, 2000]. That every patient should have an electronic medical record is now government policy in a number of western countries including, the UK and US.

Despite the widespread use of management, billing, and laboratory systems in

medicine, the vast majority of the information required for such medical records currently exists only as unstructured narrative text. Capturing more of this information in structured form is a central task of medical informatics. The absence of a standard "controlled vocabulary" or "coding system" is seen as a major barrier to this task [Sittig, 1994] and a key to its success [Rossi Mori and Consorti, 1999]. Hence several countries have mandated, or will soon mandate, standard terminologies for use in medical records.

However, most existing terminologies or "coding systems" are mono-hierarchical classifications developed either for public health reporting (the International Classification of Diseases "ICD") or bibliographic retrieval (the Medical Subject Headings—MeSH). They are much too coarse grained for recording care of individual patients. Attempts to extend them to make them finer grained have run into combinatorial explosions with some systems now running to over 250,000 "terms" which are beyond manual maintenance. Their structure is largely implicit, and writing software to use them is therefore problematic. An alternative faceted system, SNOMED-International, has existed for some time, but has no strong semantics defining the relationships amongst the facets and has always been considered difficult to use outside its origin in Pathology—both because of its unfamiliar structure and an organisation which reflects its origins in pathology and often does not cater for the needs of other medical specialities.

The US National Library of Medicine has mounted a major programme to tame this chaos in its Unified Medical Language System (UMLS) which cross maps, insofar as possible, all of the general and special purpose vocabularies [Lindberg *et al.*, 1993]. It has developed into a massive (15 Gbyte) cross reference and cataloguing system.[1] However, although cross referenced, the Unified Medical Language System is fundamentally limited by the nature of the underlying systems which it cross maps. It itself provides only a minimal amount of additional semantic information—less than 200 categories in a loose semantic network.

Hence the hope by various researchers that description logic based ontologies can provide a better solution for at least some of the problems of terminology, decision support, language processing and integration.

13.2 Example applications

13.2.1 Description Logics in terminology development and "coding"

13.2.1.1 SNOMED-RT *: tightly coupled development and pre-coordination*

SNOMED-RT is a cooperative enterprise between the College of American Pathologists and Kaiser Permanente, a large health maintenance organisation. It has

[1] `http://umlsks.nlm.nih.gov/`

re-represented in a subset of KRSS the information in the SNOMED-International. In a first approximation, the SNOMED facets for anatomy, morphology, function, etc. have been turned into roles, hasTopography, hasMorphology, etc. [Campbell *et al.*, 1998]. The initial mechanical translation has then been re-modelled in place by domain experts using a set of tools with a highly developed change management mechanism [Campbell, 1998]. The development methodology has placed a high emphasis on achieving repeatability of domain experts' results, and made extensive use of lexical tools to suggest additional relationships which are implied by the rubrics but may not be explicitly present in the faceted representation, for example the term "retinal vasculitis" was correctly related to "eye" but not to "vasculitis" (inflammation of the blood vessels) in early versions of SNOMED-International [Campbell *et al.*, 1996]

The first released version consists of a pre-enumerated set of 180,000 or more disease and procedure codes, each defined in an ontology represented in KRSS and classified accordingly into an acyclic directed graph. The intention appears to be a standard pre-coordinated (i.e., pre-defined) set of concepts and associated terms to be presented and used in a form analogous that of traditional hierarchical coding schemes.

Recently a collaboration has been formed between SNOMED-RT and the UK Clinical Terms (Read Codes) project to produce a combined product which is aimed at being a standard English controlled vocabulary for medicine. Details have not yet been announced, but it is assumed that the form will be closely related to that of SNOMED-RT.

The ontology used is relatively shallow, including under ten roles in its pre-release version, and avoiding embedded expressions wherever possible. However, the standard semantics of KRSS have been enhanced by the inclusion of right-identities to cater for part-whole relations (see Section 13.3.2).

SNOMED-RT itself includes no tools or transformations for data entry or for other applications involving dynamic post-coordination. However, a range of tools based on SNOMED-RT, including the authoring suite, is available from the company that supplies the development tools (Apelon,[1]), which are descended in part from K-REP, a DL style KR system used in many of the early experiments which led up to the project [Mays *et al.*, 1991a; 1996].

13.2.1.2 GALEN *: loosely coupled development and post-coordination*

GALEN is the result of a series of European Commission funded projects and its ontologies and specifications as well as some of the tools are available in open source form from `http://www.opengalen.org/`.

[1] `http://www.apelon.com/`

The GALEN tools are designed for loosely coupled development, and the ontology is aimed primarily at post-coordinated applications such as, intelligent user interfaces, and tools to empower users to adapt core terminologies to their specific needs. It is based around the idea of a dynamic "terminology server" rather than enumerated table of pre-coordinated terms [Nowlan *et al.*, 1994; Rector *et al.*, 1995a], although there is a limited set of common concepts predefined.

An important feature of GALEN is the clean separation of functions within the server architecture:

- logical representation in the description logic;
- language generation and text recognition;
- mapping to and from existing coding systems;
- indexing of non-terminological information;
- additional calculations such as unit and coordinate transformations.

GALEN's ontology was created de novo but with close reference to the standard classifications particularly the International Classification of Diseases. It uses the GRAIL description logic [Rector *et al.*, 1997] whose core includes the subset of operations of the KRSS used by SNOMED-RT including transitive roles, with the addition of inverse roles and role subsumption. (See Section 13.3.2.2 for a further discussion of transitive roles and related issues.) In addition GRAIL provides an additional construct, "sanctioning", analogous to slot definitions in frame systems or function signatures in object oriented systems, which supports answering queries of the form "what can be said about this". GRAIL is implemented using a graph comparison algorithm which, although known to be incomplete, has still proved to be extremely useful in practice.

GALEN's most distinctive feature is the use in authoring tools for domain experts of a much simplified "intermediate representation" which is then translated into the description logic which is relegated to the status of an "assembly language" (see Section 13.5.1 below).

The GALEN project has also devoted much effort to mapping to existing coding systems—a more complex task than is at first apparent because of the idiosyncratic construction of the target schemes. Each code in such schemes is mapped to the disjunction of one or more GALEN concepts. A GALEN concept is taken as being mapped to the most specific code mapped to a subsuming concept, and conversely, a code is mapped to all those GALEN concepts subsumed by its mapping except those subsumed by a more specific mapping. This mechanism deals with almost all of the complex sets of exclusions and inclusions in the International Classification of Diseases (ICD)—e.g., "Hypertension excluding hypertension in pregnancy" is coped

with automatically simply by mapping to the general concept "Hypertension", because there is a mapping to a specific concept "Hypertension in pregnancy" which will cause it, and its descendants, to be excluded automatically. In the very few cases where conflicts occur they are resolved by separate exception handling tables.

A similar mechanism provides a surrogate for inheritance with exceptions as a means of indexing information ranging from triggers for decision support rules to data entry forms and user interface specifications. Any information may be labelled and attached to the ontology, and the server provides operations to retrieve the set of all the values "inherited". The GALEN server makes no attempt to reduce the set to a single value; if required this is a matter for the client application.

13.2.2 Description Logics and language processing

13.2.2.1 Language analysis and information extraction

Most medical information originates and is stored as natural language text Medical texts present classic "sublanguages" with peculiarities of vocabulary and syntax. Many utterances are telegraphic or highly elliptical which cannot be easily parsed without semantic knowledge. These features seem natural to combine with lexicalised grammars in which most or all syntactic information is stored with the lexical item rather than in a separate grammar, e.g., Tree-Adjoining Grammars (TAG) [Joshi, 1994], Lexical-Functional Grammar, and Combinatory Categorical Grammar (CCG) [Steedman, 1996].[1]

Hahn's work on medSYNDICATE [Hahn *et al.*, 1999a], provides a detailed example using a specially constructed ontology in LOOM. The medSYNDICATE architecture features close coupling of the ontology ("Domain knowledge base") with the parser and extensive use of learning techniques to deepen and extend both the ontology and the grammar. It uses the integrity conditions, and conceptual constraints, and cardinality restrictions in the ontology to reduce ambiguity and select plausible interpretations. It makes use of knowledge within the ontology to complete ellipses within the original text—e.g., to know that the connection between a gland and its product is "secretes". It also makes extensive use of partonomic information using a unique approach discussed in Section 13.3.2.3 below.

Rassinoux and Baud have used the GALEN ontology to augment a strongly semantic approach likewise to constrain ambiguous or incomplete parsings [Baud *et al.*, 1993; Rassinoux, 1998]. Zweigenbaum has used a restricted application specific ontology to similar purpose [Zweigenbaum *et al.*, 1995].

Ceusters, by contrast, attempted to use natural language processing to under-

[1] However, it should be noted that the classic medical natural language work, the Linguistic String Project [Sager *et al.*, 1987; 1994], while it makes extensive use of semantics, makes no use of ontologies or related mechanisms.

stand the text attached to codes (the "rubrics") to build and make mappings to the GALEN ontology. Ceusters' work was based on a range of pre-existing tools and experienced significant difficulty because of serious differences in the information processing oriented ontology developed by GALEN and the language oriented ontologies which underlay his tools. For example, the distinctions between location and part-whole relations and the distinctions amongst different part-whole relations have no direct linguistic counterpart. An adaptation of the GALEN Intermediate representation was used to bridge this gap, but with only partial success [Ceusters and Spyns, 1997; Ceusters, 1998; Ceusters *et al.*, 1999].

13.2.2.2 Language generation, user interfaces, and quality assurance

Any ontology intended for use by domain experts presents a problem quality assurance, or curation, by those experts. Any post-coordinated use of an ontology also presents a serious problem for the user interface—standard DL expressions are not acceptable for most uses by most domain experts. Even if they are simplified to an "intermediate representation" or transformed to conceptual graphs, the complexity is too great for most domain experts to take in quickly.

One way to make such expressions accessible to users is to generate language expressions from them. Not only are the language expressions more readable, they are usually much more compact. GALEN has found language generation to be essential in virtually all applications involving post-coordination including most approaches to independent quality assurance of the ontology.

Curiously, one of the major applications of GALEN technology has been by the French government to produce unambiguous definitions for their new national classification of surgical procedures. Curiously, in this application, the usual language generation goals of concise idiomatic expression do not apply. The value of the technique is its pedantic, but completely unambiguous, presentation of the underlying formal definitions. Once the definitions are agreed and quality assured, idiomatic "preferred terms" can be composed manually where required [Baud *et al.*, 1997; Rodrigues *et al.*, 1997].

13.2.3 Decision support, indexing, and re-usable ontologies for problem solving

Many decision support methodologies, notably Musen's PROTÉGÉ and AEON [Tu *et al.*, 1995; Musen *et al.*, 1996; Musen, 1998; Grosso *et al.*, 1999] and Stefanelli's GAMES [Schreiber *et al.*, 1993; Vanheijst *et al.*, 1995; Falasconi *et al.*, 1997], are based around the existence of a domain ontology, but in general the ontologies are constructed specifically for one application and have proved less re-usable than the

problem solving methods they support. Both use ontologies primarily as frame systems

A more specific use of the classification reasoning in description logics is provided by GALEN's work on drug ontologies carried out in collaboration with the PRODIGY project on computerised guidelines for prescribing in UK general practice [Johnson *et al.*, 2000]. Traditional classifications for diseases and drugs have only a single axis of generalisation which conflates several different criteria. For example, standard drug classifications conflate indication (e.g., for "treatment of asthma"), molecular-effects (e.g., "stimulates alpha adrenergic receptors"), physiological effect (e.g., "dilates the airways") and chemical structure. As result, even simple generalisations such as "steroids reduce inflammation" are difficult to operationalise using the classification because various steroids may be classified in many different ways—under antiasthmatic drugs, topical skin preparations, anti-rheumatic drugs, etc.

Separating the conflated axes and then using them as the basis of formal descriptions which can be classified by a DL offers a potential solution. After early prototype demonstrations [Solomon and Heathfield, 1994], GALEN is now being used to construct an ontology of drugs and related conditions to be used as part of the PRODIGY project, a system of protocols for prescribing for patients with chronic diseases which being developed by the UK Department of Health [Solomon *et al.*, 1999; Wroe *et al.*, 2000]. Experience to date suggests that the ontology provides efficiently precise indexing at the varying levels of granularity required and can provide a framework for the necessary default reasoning via the mechanisms described in Section 13.2.1.1 for coding. Further evaluation awaits the next phase of the project.

13.2.4 Intelligent data entry

Data capture is the largest single barrier to greater information use in healthcare. GALEN developed from a project in user centred design to improve user interfaces for health care professionals with particular emphasis on data entry, PEN&PAD [Nowlan *et al.*, 1991a; 1991b], i.e., to construct forms which would capture most, if not all, of the information currently recorded as narrative text.

The ontology provides two services in PEN&PAD—both related to the question "What can be sensibly said in this situation?":

- Indicating how a given concept could be refined by modifiers.
- Indexing the form associated with each starting concept—often a disease or a symptom. Each such form may contain numerous subforms allowing further refinement of a concept or inclusion of further less common signs and symptoms.

The total number of forms required to provide a clinical interface is very large—certainly hundreds of thousands and possibly more. The goal of the system is

to assemble forms dynamically from the indexed "recipes" in such a way that it would fail soft—i.e., that forms for important frequently encountered situations could be highly tailored at a very fine granularity whereas rarely encountered areas could be served by a form related only to the broad class of condition. In its commercial version, Clinergy™, a knowledge base of under 10,000 concepts and a similar number of auxiliary facts and forms specifications covered essentially all data entry for British general practice—a task requiring several hundreds of thousands of forms.[1]

Related systems were developed by Poon and Fagan [1994] and Lussier *et al.* [1992] using conceptual graph representations of SNOMED-International.

13.2.5 Integration

A major ostensible goal for common terminologies in medicine is system integration [Evans *et al.*, 1994; Rector *et al.*, 1995b; Spackman *et al.*, 1997]. While specialised terminology systems are being used in a few places as part of an enterprise wide effort at integration [Rocha *et al.*, 1993; 1994; Cimino *et al.*, 1998], ontologies based on description logics have yet to be demonstrated convincingly in this context. Much of the reason for this is the sheer scale and coverage required for such mediation tasks.

13.3 Technical issues in medical ontologies

13.3.1 Issues of scaling

13.3.1.1 Size

The fundamental issue in any medical ontology intended to capture clinical terminology is scale. The smallest useful medical terminologies contain on the order of 10,000 concepts; "comprehensive" terminologies require on the order of 250,000 or more concepts. The *Open*GALEN model of basic anatomy alone contains over 5000 concepts, the model of surgical procedures some 15,000. SNOMED-RT currently has some 180,000 concepts, and the combined Clinical Terms (Read Codes) SNOMED-CT expects to have substantially more. The Unified Medical Language System has issued nearly a million "Unique Concept Identifiers" (UCDs) with over a million lexical variants.

13.3.1.2 Connectivity

Medical ontologies are notoriously highly connected. Most medical concepts depend on anatomy, and every anatomical structure is ultimately connected to every

[1] See `http://www.galen-organisation.com/furthertut.html` for further information.

other, at least trivially, by virtue of being part of the body. The causal and functional interrelationships are of similar density. SNOMED-RT reduces connectivity by omitting inverses. GRAIL supports role inverses and transitive roles, but GALEN's ontology explicitly avoids expressions of the form "A which is part of B which has part C", for which the classifier is known to be incomplete. It is not known whether complete and decidable reasoning for a DL including role transitivity and inverses is practical for a large scale comprehensive medical ontology: some form of heuristic constraint on the depth or computational resources used for individual inferences may prove necessary.

13.3.1.3 Range of granularity or organisation

Common medical notions span the range from the molecular to the physiological to the behavioural. To form a truly re-usable skeleton for medical knowledge representation, the ontology needs to encompass concepts such as "substances which cause mood change and tremor by binding to specific receptor sites". If the promise of "genomics" is to be realised, this may soon need to be extended to include concepts which add "... by stimulating the expression of a genetic sequence homologous to some specified allele in some reference source".

13.3.1.4 Complexity of concepts to be represented

The areas of medicine most resistant to traditional manual terminologies and therefore most ripe for formal representation tend to include very complicated concepts. For example, a not untypical surgical procedure rubric to be represented might be "Removal of the gall bladder using an endoscope inserted via an abdominal incision" or "Fixation of fracture of the femur by means of insertion of pins" More complex rubrics may go on for several lines in their natural language formulation. The full expansion in a description logic may include several dozen conjuncts nested five or six levels deep. This complexity is not an academic artifact; these are the categories used to determine payment, quality of outcome, and prognosis.

13.3.1.5 How much to represent—detail of the ontology

SNOMED-RT has a relatively simple ontology with less than ten roles. The GALEN ontology is relatively complex, with some fifty roles, including seven different partonomic roles, and sharp distinctions between two-dimensional and three-dimensional objects. The Digital Anatomist appears to be a representation of similar complexity to GALEN's anatomical representation. At the extreme, Gangemi *et al.* [1996] have produced a high level ontology which claims strong philosophical grounding but is yet more elaborate. How much of this complexity is required for which purposes is still not established.

13.3.2 Issues of expressivity: part-whole relations

13.3.2.1 Transitivity and anatomy

A large fraction of all medical terminology is based on anatomy and dependent on part-whole relations. "Fracture of foot" must be classified as "Trauma to lower extremity", "Repair of the aortic valve" must be classified as an "Operation on heart", etc.

Conflation of part-whole and IS-A relations is ubiquitous in informal clinical classifications and thesauri [Rector, 1998]. In general this works because for the key locative attributes it is, in general true, that a disease of the part is a disease of the whole and a procedure on a part is a procedure on the whole. This is closely related to CYC's TRANSFERS-THRO notion and to some frame systems notion of inheritance of certain slots via relations other than IS-A.

13.3.2.2 GALEN *'s specialisedBy axioms and* SNOMED-RT *'s right identity axioms*

All medical ontologies must face this problem in one way or another. GALEN allows axioms equivalent to $\mathsf{R} \circ \mathsf{S} \sqsubseteq \mathsf{R}$ (R `specialisedBy` S in GRAIL notation). SNOMED-RT allows the declaration that S is a right identity for R, which appears to be equivalent [Spackman, 2000].

Hence if R is hasLocation and S is isPartOf, then

$$\exists \mathsf{hasLocation}.(\exists \mathsf{isPartOf}.\mathsf{Heart}) \sqsubseteq \exists \mathsf{hasLocation}.\mathsf{Heart}$$

where hasLocation is the relation used to link lesions and diseases to anatomy. Given axioms such as that

$$\mathsf{AorticValve} \sqsubseteq \exists \mathsf{isPartOf}.\mathsf{Heart},$$

the required inferences that lesions of the aortic valve are lesions of the heart follows, i.e., it can be inferred that

$$\exists \mathsf{hasLocation}.\mathsf{AorticValve} \sqsubseteq \exists \mathsf{hasLocation}.\mathsf{Heart}.$$

There are, in practice, a variety of other situations in which this construct seems essential, for example to say that the "risk of a syndrome involving a disease" is subsumed by a "risk of the disease itself".

GALEN also makes extensive use of the implication of such axioms for the inverse roles, i.e., $\mathsf{S}^- \circ \mathsf{R}^- \sqsubseteq \mathsf{R}^-$. For example, let S be isSubProcessOf and R be isActedOnBy, then S^- and R^- are hasSubprocess and actsOn respectively. The implication of such an axiom for the inverse roles then allows us to express the rule that surgical procedures can be said to act on all those structures acted on by their subprocedures, e.g.:

$$\exists \mathsf{hasSubprocess}.(\exists \mathsf{actsOn}.\mathsf{FemoralArtery}) \sqsubseteq \exists \mathsf{actsOn}.\mathsf{FemoralArtery}.$$

This is a practical example. The Femoral Artery is the usual route by which the heart is catheterised. Without such inferred subsumptions, cardiac catheterisation would not be found as a target for the procedure—e.g., by a decision support system seeking to identify possible causes of damage to the femoral artery. Numerous parts of the classification of surgical procedures depend on such inferences.

The GRAIL language allows chains of such axioms which can imply complex paths. Such axioms also interact strongly with the role hierarchy. Re-representing these paths as regular expressions of roles taking into account the role hierarchy is a current topic of research.

13.3.2.3 The "triples" approach

Hahn *et al.* [1999c; 1999b] have developed an alternative representation for partonomic relations based on what they have termed "SEP-triples," which captures much partonomic reasoning within a framework compatible with the standard $\mathcal{ALC}$ description logic. In the SEP triple formulation, each anatomic part X is represented by a parent concept X_s, and two subsumed concepts X_e and X_p. X_e represents the entity as a whole, and X_p the concept of its parts. For all parts Y of X, X_p subsumes Y_s, and since Y_s subsumes both Y_e and Y_p, both the entire part Y_e and all of its parts Y_p are subsumed by the parts of X.

$$Y_p \sqsubseteq Y_s \sqsubseteq X_p \sqsubseteq X_s$$
$$X_p \sqsubseteq \exists \mathsf{anatomicalPartOf}.X_e$$

This captures the transitive relation, i.e., that any part of Y is a part of X.

For invariant anatomic relations, a separate existentially qualified role called hasAnatomicalPart links X_e to Y_e.

$$X_e \sqsubseteq \exists \mathsf{hasAnatomicalPart}.Y_e$$

This scheme allows Hahn to capture the notion that something is always part of the whole if it is present, but that it may not necessarily be present (e.g., that it may have been removed or be congenitally absent)—this is achieved by omitting the third axiom.

This allows inferences such as that a diseases of a part must be a disease of the whole structure (s) node, but not of the whole taken as in its entirety (e) node. By careful selection of which of the three members of an SEP triplet is used in an assertion, it appears to be possible to be selective about which properties are "inherited". For example: "diseases of parts are diseases of the whole", but "surfaces of parts are not surfaces of the whole". Hence in Hahn's schema, "surface of" should always refer to an entity (e) node representing the entire object, whereas diseases should refer to the structure (s) node representing the complex of the entire object and all of its parts.

Detailed comparison of the expressiveness of SEP triples with SNOMED-RT's right identities and GALEN's `specialisedBy` axioms is not yet known. However, the scheme presents a number of advantages and is relatively easy to implement with existing classifier technology.

13.3.2.4 Construct not implemented in any major medical ontology

Padgham and Lambrix [1994] point out a number of other potential patterns for relationships between parts and wholes of which at least one is potentially important for anatomical reasoning but not implemented in any current DL. This formalises the pattern that from "the hand is part of the arm" we may infer that "the skin of the hand is a part of the skin of the arm". One way to capture the essence of this notion formally would be to allow axioms of the form, $\mathsf{R} \circ \mathsf{S} \sqsubseteq \mathsf{S} \circ \mathsf{R}$ so that we have:

$$\mathsf{isLayerOf} \circ \mathsf{isPartOf} \sqsubseteq \mathsf{isPartOf} \circ \mathsf{isLayerOf},$$

from which may be inferred, for example,

$$\exists \mathsf{isLayerOf}.(\exists \mathsf{isPartOf}.\mathsf{Arm}) \sqsubseteq \exists \mathsf{isPartOf}.(\exists \mathsf{isLayerOf}.\mathsf{Arm}).$$

The GALEN ontology makes the necessary distinctions between different partonomic relations but the GRAIL language does not implement this inference.

13.3.3 Other issues of expressivity

Both GALEN and SNOMED-RT use description logics with a very limited range of core constructors—usually only existential quantification and conjunction. Both even exclude conjunctions of primitives. Neither uses universal quantification in its constructors, although GRAIL's "sanctioning" mechanism provides constraints which serve some of the same functions [Rector *et al.*, 1997]. (Hahn uses LOOM, but exploits only a limited subset of the concept language.) On the other hand, both include constructs for transitive relations as described above. Two other issues deserve mention.

13.3.3.1 Negation

Neither GALEN nor SNOMED-RT use negation, at least in the subset of the DL used in the ontology itself. This reflects real questions about the appropriate interpretation of negative statements in clinical records. In the context of medical records, there needs to be a clear differentiation at all levels between "false" and "not done" or "unknown". GALEN simulates some of the effects in the ontology by the use of "modalities" such as "presence/absence" and "done/not-done" [Rector and Rogers, 2000; Rector *et al.*, 2000].

13.3.3.2 General inclusion axioms

GALEN makes extensive use of a subset of general inclusion axioms—i.e., axioms which state that one defined concept is classified under another concept. In GALEN the subsuming term is restricted to be a conjunction of existentially qualified constructed concepts. GALEN uses such expressions for two purposes:

- To indicate which structures, states and processes are normal, abnormal but harmless, or pathological, i.e., to be treated as "diseases". In many cases it is the presence of specific modifiers which implies that a structure of process is "pathological".
- To bridge levels of granularity and add implied meaning, e.g., to indicate that "ulcer of stomach" really occurs in the "lining of the stomach" or to cope with normalisation as discussed in Section 13.4.2.2.

Many DLs have explicitly disallowed general inclusion axioms because of the difficulty of devising suitable algorithms and worries about intractability. However, motivated by GALEN, Horrocks has shown effective optimisations for DLs including general inclusion axioms. Furthermore, he has shown that all such axioms in GALEN are of a particular form which can be transformed so as to be "absorbed" within term definitions, and therefore reasoned with relatively efficiently [Horrocks and Rector, 1996; Horrocks *et al.*, 1996; Horrocks, 1997b; 1998b].

13.3.4 Frame-like behaviour

The use of description logics in both decision support and data entry systems stemmed from the use of frame systems to manage default inheritance and identify the slots relevant to a particular object. Neither are easy to implement directly in description logics. Both are particularly important in medical applications. Because of their size and variability, exhaustive manual enumeration of cases is neither practical initially nor maintainable.

13.3.4.1 Defaults and indexing

A major function of an ontology in a decision support system is to index information. However, the natural representation for a domain expert of this indexing is usually in terms of generalisations with exceptions. For examples drug indications, interactions, and side effects are all almost invariably expressed as general principles plus exceptions (chemical structure, biochemical and physiological actions can usually be treated as being indefeasible). To require all statements to be indefeasible in the domain users' environment drastically limits its usability and usefulness.

GALEN's approach is to attach "extrinsic" statements to the ontology and provide operations in the server which deliver all potential most specific candidates

as described in Section 13.2.1.2. Experience has shown that if the ontology is well constructed, the incidence of conflict is small and almost always represents a real requirement for additional information. Often this information is application specific—how seriously a drug's side effects should be viewed in a given situation, for example, or which of several minor variant codes matches the World Health Organisation's detailed coding criteria—and not appropriate to a re-usable ontology.

It has been suggested that similar behaviour could be achieved by "compiling" all defaults at the user level to explicit exclusions in the underlying description logic. A practical demonstration of this approach on a large scale in the medical field has yet to be demonstrated.

13.3.4.2 Available "slots": "what is it reasonable to say?"

GALEN's original approach was to represent "all and only what it is medically sensible to say". PEN&PAD (as well as non-medical uses of GRAIL such as the BioInformatics project TAMBIS [Baker *et al.*, 1998]), depends on assembling data entry forms and queries dynamically. The total number of potential forms is vastly greater than could be enumerated individually. Both applications depend on being able to determine which roles are "sensibly" applicable to a particular concept. GRAIL's sanctioning mechanism provides this information directly, but there is no direct way to form such a query within a standard DL framework. How best to address this issue remains an issue for research.

A key part of the GALEN experience in this regard is that only part of this "sanctioning" information is re-usable. In the original PEN&PAD application, changes to the user interface were made by changing the underlying ontology. In GALEN, and in the commercial version of PEN&PAD, Clinergy™ changing the re-usable ontology to fit an application specific requirement was unacceptable, so an additional layer of "perspectives" was interposed between the ontology itself and applications. This layered architecture now seems essential to many applications of ontologies which aspire to be re-usable.

13.4 Ontological issues in medical ontologies

13.4.1 Normative statements and abnormalities

Congenital and other deformities present a major difficulty to clinical knowledge representations, because they require that statements which would otherwise be absolute be made somehow contingent and that an extremely wide variety of statements be permitted in exceptional circumstances. They also require drawing distinctions that seem odd. Even in a Thalidomide patient with an absent left arm, we still need to be able to make statements about the left arm. Hence physical and potential presence must somehow be distinguished.

Likewise, in determining what it is "sensible" to say, congenital anomalies make a nonsense of the usual constraints. For example, most patients have their heart on their left side, three lobes to their right lung, and two lobes to their left. Most patients have a "right middle lobe" but no "left middle lobe" of the lung. However, a small percentage of patients reverse the pattern. The anomaly is not always complete, so many combinations of abnormalities are possible. Doctors tend to be highly intolerant of being presented with options such as "left middle lobe" in normal circumstances. Unfortunately, they are equally intolerant of the inability to express the notion of a "left middle lobe" in that small number ($\ll 1\%$) of cases where it is needed. Taken individually, such anomalies are rare. Taken collectively, they are surprisingly common, i.e., a significant percentage of all patients are atypical in one respect or another.

13.4.2 Clinical pragmatics

13.4.2.1 Conventional idioms

As in any language, many terms or phrases have conventional meanings different from their literal interpretation. Such differences are not always immediately obvious. A typical example is "endocrine surgery" which it might seem natural to define as "surgery on an endocrine organ". However, procedures on both the male and female reproductive organs are normally excluded, even though no doctor would dispute that they are endocrine organs. Similarly, "Heart valve", might naively be defined as a "structure in the heart with valvular function", but this includes numerous embryonic and sometimes congenitally deformed structures as well as the four "major valves" which serve the four "great vessels" entering and leaving the heart. Much of the effort of formulating a satisfactory medical ontology goes into reconciling such conventional usages with their apparent meaning.

13.4.2.2 Normalisation and implied information

Many medical notions, particularly of actions and procedures, carry strong implications about their purpose. O'Neil's classic example illustrates this problem [O'Neil *et al.*, 1995; Brown *et al.*, 1998]. A common procedure to treat hip fractures is "Insertion of pins in the femur". The only reason to insert pins in the femur is to "fixate" a fracture, and the operation is expected to be classified under both "insertion of pins" and "procedures to fixate fractures of long bones". Should the ontology contain axioms to extend the procedure definition automatically by adding "... to fixate fracture of femur"? If so, should the procedure be "Fixation of fracture of femur by means of insertion of pins in the femur" or "Insertion of pins in order to fixate fracture of femur". Ordinarily such "qua-induced" duals are distinct—e.g., the "infection caused by a virus" is very different from the "virus caused by an

infection". In these cases, at least two or more logically distinct possible representations are clinically equivalent. Most systems cope with this situation by imposing external "guidelines" on domain expert authors to normalise such expressions to one form or the other, but the problem is far from solved.

13.4.3 Semantic normalisation and level of intent

Consider the problem of what constitutes a "surgical procedure". It is easy to agree that all surgical procedure are constituted by an "act" on some "thing" which either is, or is located in, an anatomical structure. It is less easy to agree on what constitutes an "act" when there is a hierarchy of motivations: for example, "inserting pins to fixate a fracture of a long bone" or "destruction of a polyp by cautery" or "removal of a polyp (by excision)". Furthermore, important classifications hang on notions of motivation such as "palliative surgery" versus "corrective surgery". In addition, some systems wish to be able to record operations just as "correction of X" without describing the exact "act", while others wish to record "insertion of pins in fractured bone" without recording that the purpose is fixation. To address this problem within GALEN, Rossi Mori *et al.* [1997] proposed a classification into four levels:

L4 clinical goal (palliation, cure);

L3 physiologic goal: (correction, destruction, ...);

L2 primary surgical method (excision, insertion, lysis, ...);

L1 low level surgical act (cutting, cautery, ...).

It is tempting to believe that a list of concepts in each category could be agreed, so that resolution could be done automatically. However, at least within the GALEN project, intuitions and requirements clashed sufficiently to make this difficult. For example, "cautery" can sometimes be a low level act or sometimes a primary method. This ambiguity is dealt with in the formal ontology by having separate concepts for "simple cautery" and "removal by cauterisation", and by care in formulating the intermediate representation (see Section 13.5.1). However, achieving consistent usage amongst a range of authors with different applications requires vigilance and careful quality assurance.

13.5 Architectures: terminology servers, views, and change management

13.5.1 Intermediate representations and views: GALEN 's layered architecture

There is an inevitable conflict between the need for an ontology to be re-usable and the requirement that it be easily understood by the domain experts who must author and maintain it. SNOMED-RT addresses this problem by keeping the ontology relatively simple. GALEN addresses these problems by placing an "intermediate representation" and views ("perspectives") between the re-usable ontology and users oriented applications [Rector *et al.*, 1999; 2001]. The intermediate representation and perspective layers in the architecture hide complexities irrelevant to the current application from domain experts and other users. It also allows for variations amongst domain experts in the vocabulary, structure, and—critically for an international project—language. In this layered architecture, the description logic ontology is effectively reduced to a role analogous to that of an assembly language program. Using an intermediate representation both allows loose coupling amongst authors and simplifies the authoring task.

Within the GALEN project, use of an intermediate representation reduced training time for new authors from months to days. It also drastically reduced the time required centrally to harmonise the work of different authors so that the resulting classification would pass an agreed quality assurance. Prior to the introduction of the intermediate representation, central harmonisation had consumed over fifty percent of the effort; following introduction of the intermediate representation this dropped to less than ten percent. This is a major saving given that the knowledge engineers required for central harmonisation take a year or more to train fully. The experience of developing the drug ontology in Prodigy (See Section 13.2.3) has been roughly comparable. In addition, in the drug ontology, the use of the intermediate representation has allowed the quality assurance experts to participate directly in correcting the authored ontology—something which would be entirely impractical in its expanded formulation in the description logic.

13.5.2 Learning versus building

Given the scale of medical ontologies, it would obviously be attractive to use learning techniques for at least some of their construction. Hahn *et al.* [1999a] are focusing on using language plus the structure of the Unified Medical Language System as a major source for inducing their ontology. Campbell *et al.* [1998] have outlined a strategy which makes use of lexical "suggestions" to guide manual modelling as part

of the SNOMED-RT methodology. GALEN has experimented with various linguistic techniques but so far with limited success [Ceusters *et al.*, 1999].

13.5.3 Version and change management

Any medical ontology for general use must be a living developing structure. There are both clinical and technical issues to be dealt with. Campbell *et al.* [1996] have developed a tightly coupled methodology for change management in conjunction with SNOMED-RT, while Oliver *et al.* [1999] and Cimino [1996] have discussed the issues of changes in medical vocabulary.

13.6 Discussion: key lessons from medical ontologies

Medicine is big and complicated. It has a long tradition of controlled vocabularies and coding systems. Developing re-usable medical ontologies presents at least three major classes of issue to the description logic community:

- Developing implementations which scale.
- Developing architectures which reconcile the needs of users for simplicity with the formal constraints required for tractability and the ontological richness required for re-use.
- Developing formalisms expressive enough to cope with constructs of particular concern to medicine, particularly part-whole relations but also other spatio-temporal constructs such as adjacency.

Perhaps most critically, medicine presents the challenge of presenting description logic notations in forms which users can use to meet real problems—whether in representation of medical records, indexing of information for decision support, or supporting user interfaces and natural language processing.