Medical knowledge reengineering—converting major portions of the UMLS into a terminological knowledge base

Stefan Schulz a,*, Udo Hahn b

a Freiburg University Hospital, Department of Medical Informatics, Stefan-Meier-Straße 26, D-79104 Freiburg, Germany
b Freiburg University, Computational Linguistics Lab, Werthmannplatz 1, D-79085 Freiburg, Germany

Abstract

We describe a semi-automatic knowledge engineering approach for converting the human anatomy and pathology portion of the UMLS metathesaurus into a terminological knowledge base. Particular attention is paid to the proper representation of part-whole hierarchies, which complement taxonomic ones as a major hierarchy-forming principle for anatomical knowledge. Our approach consists of four steps. First, concept definitions are automatically generated from the metathesaurus, with LOOM as the target language. Second, integrity checking of the emerging taxonomic and partonomic hierarchies is automatically carried out by the terminological classifier. Third, terminological cycles and inconsistencies are manually eliminated and, in the last step, the knowledge base built this way is incrementally refined by a medical expert. Our experiments were run on a terminological knowledge base which is composed of 164,000 concepts and 76,000 relations. Empirical evidence for the lack of logical consistency, adequacy and improper granularity of the UMLS knowledge source is given, and finally, assessments of what kind of efforts are needed to render the formal target representation structures complete and empirically adequate. © 2001 Elsevier Science Ireland Ltd. All rights reserved.

Keywords: UMLS; Description logics; Anatomy; Pathology

1. Introduction

The health care domain and the biomedical sciences are somewhat unique compared with other scientific areas, since large portions of their terminological knowledge are already structured in terms of controlled terminologies, classification systems and thesauri. According to the different tasks they have been designed for, such as statistics, clinical communication, accounting or document indexing, they exhibit considerable variability both in terms of coverage and granularity. Also the way knowledge is organized differs between heterogeneous types of medical terminologies [1,2]. Classifications aim at providing exhaustive sets of mutually exclusive categories (or classes) such as the International Classification of Diseases (ICD) [3]. More complex systems such as nomenclatures (e.g.
SNOMED [4] or NHS clinical terms [5]) and thesauri (e.g. MeSH [6]) provide additional descriptive flexibility by way of compositionality of concepts, polyhierarchies and semantic links—often, however, at the price of increasing ambiguity and semantic vagueness. Although various kinds of medical terminologies are well adapted to different needs, the demand for homogeneous multi-purpose terminology servers has been increasingly expressed [7–11].

The ‘Unified Medical Language System’ (UMLS) [12] can be considered as a direct response to this request. It contains about 800,000 concepts from more than 60 different classifications, nomenclatures and thesauri, all of which have been merged into the UMLS Metathesaurus. Additional semantic structure can be imposed on concepts by using 134 semantic types, provided by the UMLS Semantic Network, together with 54 semantic relations. UMLS is accessible via http://umlsinfo.nlm.nih.gov. Given its size, evolutionary diversity and inherent heterogeneity, there is no surprise at all that the lack of a solid formal foundation leads to a bunch of inconsistencies, circular definitions, etc. [13,14]. This may not cause utterly severe problems when humans are in the loop and its use is limited to tasks such as those mentioned above. However, anticipating its use for more knowledge-intensive applications, such as natural language understanding of medical narratives [15] or medical decision support systems [16], those shortcomings might lead to an impasse.

As a consequence, formal models for dealing with medical knowledge have been proposed, using representation mechanisms based on conceptual graphs, semantic networks or description logics [17–19]. Not surprisingly, there is also a price to be paid for more expressiveness and formal rigor in terms of increasing modeling efforts and, hence, increasing maintenance costs. Therefore, concrete medical knowledge bases making full use of this rigid approach, especially those which employ high-end, KL-ONE-style knowledge representation languages (for a survey, cf. [20]), are usually restricted to rather small subdomains. Those systems developed within the framework of the above-mentioned formal approaches have all been designed from scratch—without making systematic use of the large body of knowledge contained in informal medical terminologies.

An intriguing approach would be to combine the massive coverage offered by informal medical terminologies with the high level of expressiveness supported by formally solid knowledge representation systems in order to develop sophisticated medical knowledge bases on a larger scale. This idea has already been fostered by Pisanelli et al. [10], who extracted knowledge from the UMLS semantic network as well as from parts of the metathesaurus and merged it with generic ontologies from other sources. In a similar way, Spackman and Campbell [21] describe how SNOMED [4] can be transformed from a multi-axial coding system into a formally founded ontology. Unfortunately, efforts up to now are entirely focused on taxonomic reasoning along generalization hierarchies (expressed by is-a relations) and lack a reasonable coverage of part-whole (i.e. part-of or has-part) relationships, a second major conceptual construct needed for reasoning in the anatomy domain, in particular.

This article is organized as follows. In Section 2, we argue for the relevance of part-whole reasoning for the medical domain and introduce a representation model which is rooted in a description logics framework [20]. In particular, we propose a tripartite data structure for encoding anatomical concepts in
Fig. 1. SEP triplets: partitive relations within taxonomies.

order to emulate partonomic reasoning by taxonomic reasoning. Section 3 contains an in-depth description of a four-step knowledge engineering procedure for semi-automatically converting UMLS specifications into a terminological knowledge base. Throughout this procedure our emphasis is on maintaining the consistency of the emerging knowledge base. We conclude in Section 4 by discussing some implications of our approach and prospects of future work.

2. Part-whole reasoning

As far as medical knowledge is concerned, two main hierarchy-building relationships can be identified, namely taxonomic (is-a) and partonomic (part-whole) ones. Unlike taxonomic reasoning in concept hierarchies, no fully conclusive mechanism exists up to now for reasoning along partonomic hierarchies in description logic systems. As anatomical knowledge, a crucial portion of medical knowledge, is principally organized along part-whole hierarchies, any proper medical knowledge representation has to take account of both hierarchy types [22].

The outstanding importance of part-whole hierarchies for anatomy and, consequently, for clinical medicine has recently motivated the development of semantic networks of anatomical concepts [23,24]. Although they provide ontologically precise descriptions of partonomies, their granularity level is usually rather high. Also, these terminological resources do not provide a formally founded methodology for part-whole reasoning that underlies various object-centered representation approaches as discussed by Artale et al. [25]. In one of these branches, the description logics community, several language extensions for knowledge representation systems have been proposed which provide special constructors for part-whole reasoning [19,26].

Motivated by proposals from Schmolze and Mark [27], as well as by design principles underlying the Read Codes Version 3 [28], we advocate an alternative solution for part-whole reasoning, one that does not exceed the expressiveness of the well-understood, parsimonious concept language $ALC$ [29]. Unlike the constructor-based approaches mentioned before, our approach can easily cope with many of the exceptions to the transitivity of the part-of relation, which one encounters not only in medicine [30,31] but also in commonsense domains [32,33].

Instead of defining new operators with a built-in transitivity property, our proposal is centered around a particular data structure, so-called $SEP$ triplets, especially designed for
empirically adequate part-whole reasoning (cf. the structural description in Fig. 1). They define a characteristic pattern of is-a hierarchies, which support the emulation of inferences typical of transitive part-of relations, as well as exceptions to it. In this formalism, the relation anatomical-part-of describes the partitive relation between physical parts of an organism.

Each basic anatomical concept node is expanded to an SEP triplet. Such a triplet consists, first of all, of the anatomical concept itself, the so-called E-node (entity node). As an example, in Fig. 1, $H_E$ stands for the concept of the entire Hand. The second node of the triplet construct, the P-node (part node) is defined as the common subsumer of all concepts which have the role anatomical-part-of filled by the corresponding E-node. P-nodes can therefore be considered as a kind of reification of the relation anatomical-part-of. In Fig. 1, the P-node $H_P$ subsumes every concept which has $H_E$ (Hand) as a filler of the role anatomical-part-of, e.g. $F_E$ (Finger). Finally, both, the P- and E-node, have a common direct subsumer, the so-called S-node (structure node), $H_S$ (Hand-Structure) in Fig. 1. By definition, E-nodes and P-nodes are mutually disjoint, thus restricting anatomical-part-of to proper parthood, i.e. no anatomical concept can be anatomical-part-of itself (e.g. no object in the world can be considered a Hand and a Part of a Hand simultaneously. This constraint might be relaxed under certain circumstances [34].) The SEP triplet construct can then be used to emulate transitive part-of hierarchies by linking S-nodes to P-nodes (cf. the is-a link between $C_S$ and $D_P$ in Fig. 1), and to exclude transitivity by linking nontransitive properties to the corresponding E-node of a SEP triplet [30,31]. The solution we propose is computationally neutral insofar as we extend the number of concept nodes by a constant factor (viz. two additional nodes per concept at most).

3. Semi-automatic transformation of an informal knowledge repository into a formal terminological knowledge base

Our goal is to extract conceptual knowledge from two highly relevant subdomains of the UMLS, anatomy and pathology, and to map it (semi-)automatically into a formally sound medical knowledge base. We use LOOM [35,36], a KL-ONE-style terminological knowledge representation language, as our implementation platform (for alternatives, cf. [37]), though our approach does in no way depend on particular features of that language.\(^2\) The knowledge transformation task is divided into four steps: (1) the automatic generation of terminological assertions, (2) their submission to a terminological classifier\(^3\) for consistency checking, (3) the manual restitution of formal consistency in case of inconsistencies, and, finally, (4) the manual rectification and refinement of the resulting knowledge base. These four steps are illustrated by the workflow diagram depicted in Fig. 2.

3.1. Step 1: automatic generation of terminological assertions

Sources for concepts and relations were the UMLS semantic network and the mrrel, mr-con and mrsty tables of the 1999 release of

\(^2\) Cf. also the work of Carenini and Moore [38] who have already suggested a graphical interactive tool for mapping UMLS concepts semi-automatically into a loom knowledge base environment.\(^3\) The description classifier of a terminological knowledge representation system [36] is the inference engine that computes subsumption relations between concepts, i.e. the generalization hierarchies that can be derived from is-a relations.
Fig. 2. Workflow diagram for the construction of a terminological knowledge base from the UMLS.

<table>
<thead>
<tr>
<th>UMLS relation</th>
<th>Step 1</th>
<th>Step 2</th>
<th>Step 3</th>
<th>Step 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>sibling_of</td>
<td>267,218</td>
<td>SIB</td>
<td>add negations in order to express taxonomic or partitive disjointness</td>
<td></td>
</tr>
<tr>
<td>child_of</td>
<td>59,004</td>
<td>CHDRN</td>
<td>include related concepts into primitive or part-of clause where plausible</td>
<td></td>
</tr>
<tr>
<td>narrower_term</td>
<td>24,223</td>
<td>CHDRN</td>
<td>substitute primitive links by non-primitive ones where possible</td>
<td></td>
</tr>
<tr>
<td>isa</td>
<td>9,755</td>
<td></td>
<td>check for definitional cycles</td>
<td>remove taxonomic parent concepts</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>substitute primitive links by non-primitive ones where possible</td>
<td></td>
</tr>
<tr>
<td>location_of</td>
<td>4,820</td>
<td>LOCATION_OF</td>
<td>include related concepts into has-part clause where plausible</td>
<td></td>
</tr>
<tr>
<td>has_location</td>
<td>4,820</td>
<td>HAS_LOCATION</td>
<td>include related concepts into part-of clause, where plausible</td>
<td></td>
</tr>
<tr>
<td>has_part</td>
<td>4,321</td>
<td>has-part</td>
<td>check whether this part is mandatory (under “real-anatomy” assumption)</td>
<td></td>
</tr>
<tr>
<td>has_conceptual_part_of</td>
<td>126</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>part_of</td>
<td>4,321</td>
<td>part-of</td>
<td></td>
<td></td>
</tr>
<tr>
<td>conceptual_part_of</td>
<td>126</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>parent</td>
<td>59,004</td>
<td>PARRB</td>
<td>include related concepts into has-part clause where plausible</td>
<td></td>
</tr>
<tr>
<td>broader_term</td>
<td>24,223</td>
<td>PARRB</td>
<td></td>
<td></td>
</tr>
<tr>
<td>inverse_isa</td>
<td>9,755</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>associated_with</td>
<td>13</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mapped_from</td>
<td>2,643</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>other_relation</td>
<td>18,906</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>qualified_by</td>
<td>1,864</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>allowedQualifier</td>
<td>1,864</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mapped_to</td>
<td>2643</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;other named relations&gt;</td>
<td>11,888</td>
<td>&lt;some x&gt;</td>
<td>check for inherited constraints</td>
<td>remove constraints</td>
</tr>
<tr>
<td>sibling_of</td>
<td>457,549</td>
<td>SIB</td>
<td>add negations in order to express taxonomic or partitive disjointness</td>
<td></td>
</tr>
<tr>
<td>child_of</td>
<td>72,429</td>
<td>PARRB</td>
<td>include related concepts into has-part clause where plausible</td>
<td></td>
</tr>
<tr>
<td>narrower_term</td>
<td>26,973</td>
<td>PARRB</td>
<td>include related concepts into has-part clause where plausible</td>
<td></td>
</tr>
<tr>
<td>isa</td>
<td>3,635</td>
<td></td>
<td>substitute primitive links by non-primitive ones whenever possible</td>
<td></td>
</tr>
<tr>
<td>inverse_isa</td>
<td>3,635</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>associated_with</td>
<td>12,905</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mapped_to</td>
<td>15,023</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>other_relation</td>
<td>25,786</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>qualified_by</td>
<td>6,235</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>allowedQualifier</td>
<td>6,235</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;other named relations&gt;</td>
<td>4,162</td>
<td>&lt;some x&gt;</td>
<td>check for inherited constraints</td>
<td>remove constraints</td>
</tr>
</tbody>
</table>

Pathology Concepts Linked to Anatomy Concepts

<table>
<thead>
<tr>
<th>CUIpat</th>
<th>CULana</th>
<th>2,247</th>
<th>&lt;some has_anatomic_correlate&gt;(some has_anatomic_correlate)</th>
<th>plausibility check of concept “duplication” (assignment to both domains)</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;missing&gt;</td>
<td>&lt;missing&gt;</td>
<td>&lt;missing&gt;</td>
<td>&lt;missing&gt;</td>
<td>add pathology-anatomy links</td>
</tr>
<tr>
<td>associated_with</td>
<td>2,714</td>
<td>&lt;some associated_with &lt;anatomy_concept&gt;_S&gt;</td>
<td>check for consistency</td>
<td>render links complete, link to E-node instead of S-node when role propagation has to be disabled</td>
</tr>
<tr>
<td>has_location</td>
<td>9,238</td>
<td>&lt;some has_location &lt;anatomy_concept&gt;_S&gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;other&gt;</td>
<td>&lt;do nothing&gt;</td>
<td>&lt;do nothing&gt;</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
the UMLS metathesaurus. The mrrel table contains roughly 7.5 million records and exhibits the semantic links between two concept unique identifiers (CUIs)\(^4\), the mrcon table contains the concept names and mrsty keeps the semantic type(s) assigned to each CUI. These tables (cf. Fig. 3 for a fragment), available as ASCII files, were imported into a Microsoft Access relational database and manipulated using SQL embedded in the VBA programming language. For each CUI in the mrrel subset its alphanumeric code was substituted by the English preferred term given in mrcon.

From a total of 85899 concepts, we extracted 38059 anatomy and 50087 pathology concepts from the metathesaurus. Each concept was included in this set, which belonged to a set of predefined anatomy\(^5\) and pathology\(^6\) types given in the UMLS semantic network. 2247 concepts were included in both sets, anatomy and pathology. This finding can easily be justified by the observation that these hybrid concepts exhibit, indeed, multiple meanings.\(^7\) As we wanted to keep

<table>
<thead>
<tr>
<th>CUI</th>
<th>MRREL</th>
<th>MRCON</th>
<th>RELA</th>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>C005847</td>
<td>CHD</td>
<td>C0014261</td>
<td>part_of</td>
<td>CSP98</td>
<td>CSP98</td>
</tr>
<tr>
<td>C005847</td>
<td>CHD</td>
<td>C0014261</td>
<td>isa</td>
<td>MSH99</td>
<td>MSH99</td>
</tr>
<tr>
<td>C005847</td>
<td>CHD</td>
<td>C0025862</td>
<td>part_of</td>
<td>MSH99</td>
<td>MSH99</td>
</tr>
<tr>
<td>C005847</td>
<td>CHD</td>
<td>C0026844</td>
<td>part_of</td>
<td>MSH99</td>
<td>MSH99</td>
</tr>
<tr>
<td>C005847</td>
<td>CHD</td>
<td>C0034092</td>
<td>isa</td>
<td>SMNI98</td>
<td>SMNI98</td>
</tr>
<tr>
<td>C005847</td>
<td>CHD</td>
<td>C0035330</td>
<td>isa</td>
<td>MSH99</td>
<td>MSH99</td>
</tr>
<tr>
<td>C005847</td>
<td>CHD</td>
<td>C0042366</td>
<td>part_of</td>
<td>MSH99</td>
<td>MSH99</td>
</tr>
<tr>
<td>C005847</td>
<td>CHD</td>
<td>C0042367</td>
<td>part_of</td>
<td>MSH99</td>
<td>MSH99</td>
</tr>
<tr>
<td>C005847</td>
<td>CHD</td>
<td>C0042449</td>
<td>isa</td>
<td>MSH99</td>
<td>MSH99</td>
</tr>
</tbody>
</table>

Fig. 3. Semantic relations in the UMLS metathesaurus.

\(^4\) As a coding convention in UMLS, any two CUIs must be connected by at least a shallow relation (in Fig. 3, Child relations in the column REL are assumed between CUIs). These shallow relations may be refined in the column RELA, if a thesaurus is available which contains more specific information. Some CUIs are linked either by part-of or is-a. In any case, the source thesaurus for the relations and the CUIs involved is specified in the columns X and Y (e.g. MeSH 1999 (MSH99), SNOMED International 1998 (SNMI98).

\(^5\) Anatomical Structure, Embryonic Structure, Congenital Abnormality, Acquired Abnormality, Fully Formed Anatomical Structure, Body System, Body Part Organ or Organ Component, Tissue, Cell, Cell Component, Gene or Genome, Body Location or Region, Body Space or Junction, Anatomical Abnormality.

\(^6\) Pathologic Function, Disease or Syndrome, Mental or Behavioral Dysfunction, Cellular or Molecular Dysfunction, Experimental Model of Disease, Neoplastic Process.

\(^7\) For instance, Tumor has the meaning of a malignant disease on the one hand, and of an anatomical structure on the other hand. The same applies to congenital and acquired malformations, e.g. Claw Foot.

(dsdeftriplet HEART
:is-primitive HOLLOW-VISCUS
:has-part (:p-and
ANATOMICAL-FEATURE-OF-HEART
FIBROUS-SKELETON-OF-HEART
WALL-OF-HEART
CAVITY-OF-HEART
CARDIAC-CHAMBER-NOS
LEFT-CORONARY-SULCUS
RIGHT-CORONARY-SULCUS
SURFACE-OF-HEART-NOS
LEFT-SIDE-OF-HEART
RIGHT-SIDE-OF-HEART
AORTIC-VALVE
TRICUSPID-VALVE
PULMONARY-VALVE
MITRAL-VALVE
HEART-VALVES-100))
the two subdomains strictly disjoint, we duplicated these hybrids and prefixed them with ‘ana-’ or ‘pat-’ according to their respective subdomain. The a priori assignment to the above-mentioned semantic types in the UMLS is the only selection criterion; we refrained from any manual interference at this processing stage.

Anatomy and pathology concepts received a different formal treatment, however. As target structures for the anatomy domain we chose SEP triplets. These were expressed in the terminological language LOOM, which we had previously extended by a special DEFTRIPLET macro (cf. Table 1 for an example).

Only part-of, has-part and is-a relation attributes (RELA fields in the mrrel table) from the UMLS were considered for the construction of taxonomic and partonomic hierarchies (cf. Fig. 2). Hence, for each anatomy concept, one SEP triplet is created. The result is a mixed is-a/part-whole hierarchy (cf. Fig. 4).

For the pathology domain, we assumed the values CHD (child), ISA and RN (narrower relation) from the mrrel REL field as indicators of taxonomic links. For all anatomy concepts referred to in the definitive state-

dentification of pathology concepts, the ‘S-node’ is the default concept to which they are linked, thus enabling the propagation of roles across the part-whole hierarchy (see below).

In both subdomains, shallow relations, such as the extremely frequent SIB (sibling) relation, were included as comments into the code to give some heuristic guidance for the manual refinement phase (cf. Fig. 2).

3.2. Step 2: consistency checking by the description classifier

The import of UMLS anatomy concepts resulted in 38 059 DEFTRIPLET expressions for anatomical concepts and 50 087 DEFCONCEPT expressions for pathological concepts. Each DEFTRIPLET was expanded into three DEF-

CONCEPT (S-, E-, and P-nodes), and two DEFRELATION (anatomical-part-of-x, inv-anatomical-part-of-x) expressions, summing up to 114 177 concepts and 76 118 relations. Thus we obtained (together with 382 concepts from the semantic network) a total of 240 764 definitory LOOM expressions.

From 38 059 anatomy triplets, 1219 DEFTRIPLET statements exhibited a :has-part clause followed by a list of a variable number of triplets, containing more than one argument in 823 cases (average cardinality: 3.3). 4043 DEFTRIPLET statements contained a :part-of clause, only in 332 cases followed by more than one argument (average cardinality:
1.1. The obtained knowledge base was then submitted to the terminological classifier and automatically checked for terminological cycles and consistency. A terminological cycle is given when $A$ subsumes $B$ and $A$ is subsumed by $B$, as well. Inconsistencies occur when constraints (e.g. role restrictions) are violated. In the anatomy subdomain, one terminological cycle and 2328 inconsistent concept definitions were identified; in the pathology subdomain 355 terminological cycles were determined though no inconsistent concept definition at all was found (cf. Table 2).

3.3. Step 3: manual restitution of consistency

The inconsistencies of the anatomy part of the knowledge base identified by the classifier could be traced back to the simultaneous linkage of two triplets by both is-a and part-of links, an encoding that raises a conflict due to the disjointness required for corresponding P- and E-nodes. In most of these cases the affected parents belong to a class of concepts that obviously cannot be appropriately modeled as SEP triplets, e.g. Subdivision-Of-Ascending-Aorta, Organ-Part. The meaning of these concepts almost paraphrases that of a P-node, so that in these cases the violation of the SEP-internal disjointness condition could be accounted for by substituting the involved triplets with simple LOOM concepts, by matching them with already existing P-nodes, by relaxing the disjointness constraint, or by disabling is-a or part-of links.

In the pathology part of the knowledge base, we expected a large number of terminological cycles to arise as a consequence of interpreting the notoriously weak, thesaurus-style RN (narrower) and CHilD relations through taxonomic subsumption (is-a). Bearing in mind the size of the knowledge base, we consider 355 cycles a tolerable amount of noise. Those cycles were primarily due to very similar concepts, e.g. Arteriosclerosis vs. Atherosclerosis, Amaurosis vs. Blindness, and residual categories (‘other’, ‘NOS’ = not otherwise specified). These were directly inherited from the source terminologies and are always difficult to interpret out of their definitional context, e.g. Other-Malignant-Neoplasm-of-Skin vs. Malignant-Neoplasm-of-Skin-NOS. The cycles were analyzed and a negative list which consisted of 630 concept pairs was manually derived. In a subsequent extraction cycle, we incorporated this list in the automated construction of the LOOM concept definitions. By adding these new constraints a fully consistent knowledge base was generated.

3.4. Step 4: manual rectification and refinement of the knowledge base

Adding value to a consistent though possibly underspecified or even misspecified knowledge base is an extremely time-consuming job and requires broad and in-depth medical expertise. In order to roughly assess the potential workload for future knowledge base finishing, we extracted two random samples ($n = 100$ each) from both the anatomy and pathology part of the knowledge base; the samples were then analyzed by a medical student and a physician. From the experience we gained in both subdomains so far, the following workflow can be derived:

<table>
<thead>
<tr>
<th>Table 2</th>
<th>Classification results for anatomy and pathology concepts</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Anatomy</td>
</tr>
<tr>
<td>Triplets</td>
<td>38 059</td>
</tr>
<tr>
<td>DEFCONCEPT statements</td>
<td>114 177</td>
</tr>
<tr>
<td>Cycles</td>
<td>1</td>
</tr>
<tr>
<td>Inconsistencies</td>
<td>2328</td>
</tr>
</tbody>
</table>
3.4.1. Checking the correctness and completeness of both the taxonomic and partitive hierarchies

Taxonomic and partitive links are manually added or removed in order to eliminate inadequate concept descriptions and to increase the completeness and to deepen the granularity of concept descriptions. Primitive subsumption (where necessary conditions for a specialization relation between concepts are specified only) is substituted by a nonprimitive one (where necessary and sufficient conditions for a specialization relation between concepts are specified) whenever possible. This is a crucial point, because the automatically generated hierarchies contain only information about the parent concepts and necessary conditions. As an example, the automatically generated definition of Dermatitis includes the information that it is an Inflammation and that the role has-location must be filled by the concept Skin. An Inflammation that has-location Skin, however, cannot be classified automatically as Dermatitis.

3.4.2. Results

In the anatomy sample, only 76 concepts out of 100 could be unequivocally classified as belonging to 'canonical' anatomy. (The remainder, e.g. ana-Phalanx-of-Supernumerary-Digit-of-Hand, referring to pathological anatomy was immediately excluded from analysis.) Besides the assignment to the UMLS semantic types, only 27 (direct) taxonomic links were found. Another 83 UMLS relations (mostly CHilD or RN (narrower) relations) were manually upgraded to taxonomic links. 12 (direct) part-of and 19 has-part relations were found. Four part-of relations and one has-part relation had to be removed, since we considered them as implausible. 51 UMLS relations (mostly CHilD or RN (narrower) relations) were manually upgraded to part-of relations, and 94 UMLS relations (mostly PARRB, i.e. parent and broader relations) were upgraded to has-part relations. After this workup and upgrade of shallow UMLS relations to semantically more specific relations, the sample was checked for completeness again. As a result, 14 is-a and 37 part-of relations were still considered missing.

In the pathology sample, the assignment to the pathology subdomain was considered plausible for 99 of 100 concepts. A total of 15 false is-a relations were identified in 12 concept definitions, while 24 is-a relations were considered to be missing.

3.4.3. Checking has-part arguments assuming 'real anatomy'

In the UMLS sources part-of and has-part are considered symmetric. According to our transformation rules, the attachment of a role has-anatomical-part to an E-node \( B_E \), with its range restricted to \( A_E \), implies the existence of a concept \( A_E \) for the definition of concept \( B_E \). On the other hand, the classification of \( A_E \) as being subsumed by the P-node \( B_P \), the latter being defined via the role anatomical-part-of restricted to \( B_E \), implies the existence of \( B_E \) given the existence of \( A_E \) (cf. Fig. 5, left). This assumption does not always match ‘real’ anatomy, i.e. anatomical concepts that may exhibit pathological modifications. Fig. 5 (left part) sketches a concept \( A_E \) that is necessarily anatomical-part-of a concept \( B_E \), but whose existence is not required for the definition of \( B_E \). This is typical of the results of surgical interventions, e.g. a large intestine without an appendix, or an oral cavity without teeth, etc.

3.4.4. Results

All 112 has-part relations obtained by the automatic import and the manual workup of our sample were checked. The analysis revealed that more than half of them (62) should be eliminated in order not to obviate
Fig. 5. Patterns for partonomic reasoning using SEP triplets: anatomical-part-of without has-anatomical-part (left), has-anatomical-part without anatomical-part-of (right).

a coherent classification of pathologically modified anatomical objects. As an example, most instances of Ileum do not contain a Meckel’s Diverticulum, whereas all instances of Meckel’s Diverticulum are necessarily anatomical-part-of Ileum. Many surgical interventions that remove anatomical structures (appendix, gallbladder, etc.) produce similar patterns. In our formalism, this corresponds to a single taxonomic link between a P-node and a S-node (cf. Fig. 5, left part). The non-linkage situation is also possible (cf. Fig. 5, right part). The definition of $A_E$ does not imply the role anatomical-part-of to be filled by $B_E$, but $B_E$ does imply that the inverse role be filled by $A_E$. As an example, a Lymph-Node necessarily contains Lymph-Follicles, but there exist Lymph-Follicles that are not part of a Lymph-Node. This pattern is characteristic of mereological relations between macroscopic (countable) objects, such as organs, and multiple uniform microscopic objects [34].

3.4.5. Analysis of the sibling relations and defining concepts as being disjoint

In the UMLS mrrel table, the SIB(ling) relation targets at concepts which share the same parent in a taxonomic or partonomic hierarchy. Pairs of sibling concepts may either have common descendants or not. If not, they constitute the root of two disjoint sub-trees. In a taxonomic hierarchy, this means that one concept implies the negation of the other (e.g. a benign tumor cannot be a malignant one, et vice versa). In a partitive hierarchy, this corresponds to two topologically disconnected objects, $C_E$ and $D_E$, with the following interpretation: There are no common parts shared by any instance of $C_E$ with any instance of $D_E$. In our triplet formalism this can be expressed as follows (for a formalization and further discussion of topological aspects in anatomical ontologies, cf. [39,40]): topological disconnectedness refers to a pair of concepts, $C_E$ and $D_E$, whose S-nodes, $C_E$ and $D_E$, belong to two disjoint subgraphs (i.e. $C_S$ implies the negation of $D_S$). As a consequence there is no instance that is both anatomical-part-of an instance of $C_E$ and anatomical-part-of of an instance of $D_E$ (cf. Fig. 6). As an example, the concepts Right Hand and Left Hand are topologically disconnected, whereas Right Hand and Right Forearm are not (there are instances which share a common boundary structure).
3.4.6. Results

We found, on average, 6.8 siblings per concept in the anatomy domain, and 8.8 in the pathology domain. So far, the analysis of sibling relations has been performed only for the anatomy domain. From a total of 521 sibling relations, 9 were identified as is-a, 14 as part-of, and 17 as has-part, whereas 404 referred to topologically disconnected concepts.

3.4.7. Completion and modification of anatomy–pathology relations

For each pathology concept (such as determined by the LOOM system after classification) it has to be checked whether the anatomy–pathology links are correct and complete. Incorrect constraints have to be removed from a concept definition itself or from one of the subsuming concepts. For each correct anatomy–pathology relation the decision must be taken whether the E-node or the S-node has to be addressed as the target concept for modification. In the first case, the propagation of roles across part-whole hierarchies is disabled. As an example (cf. Fig. 7), Enteritis implies has-location Intestine. The range of the relation has-location is restricted to the E-node of Intestine, $I_E$. This precludes, for instance, the computation of an is-a relation between Appendicitis and Enteritis, though Appendicitis is related to Intestine via an anatomical-part-of relation. In the second case, the target is the S-node of the anatomical triplet, and, thus, the propagation of roles is enabled. Glomerulonephritis (has-location Glomerulum) is therefore subsumed by Nephritis (has-location Kidney), since Glomerulum is defined as an anatomical-part-of Kidney. In the same way, Perforation-of-Appendix is generalized as Intestinal-Perforation (cf. [30,31] for a comprehensive analysis and formal specification of these phenomena).

3.4.8. Results

In our random sample we found 522 anatomy–pathology relations, from which 358 (i.e. 69%) were judged as incorrect by the domain experts. In 36 cases an adequate anatomy–pathology relation was missing. All 164 has-location roles were analyzed as to whether they were to be filled by an S-node or an E-node of an anatomical triplet. In 153 cases, the S-node (which allows propagation across the part-whole hierarchy) was considered to be adequate; in 11 cases the E-node was preferred. The analysis of the random sample of 100 pathology concepts revealed that only 17 of them were to be linked with an anatomy concept. In 15 cases, the default linkage to the S-node was considered to be correct, in one case the linkage to the E-node was preferred, in another case a given linkage was considered to be false.

The high number of implausible constraints points to the lightweight semantics of has-location links in the UMLS sources. While we interpreted them in terms of a conjunction for the import routine, a disjunctive meaning seems to prevail implicitly in many definitions of top-level concepts such as Tuberculosis. In this example, we find all anatomical concepts that can be affected by this disease, linked by has-location. All these
4. Conclusions

There is a growing demand for high-quality terminology services and their embedding in functionally advanced health information systems. Among the desiderata that have to be fulfilled is the need to make consistent, conceptually rich knowledge bases available so that their inference engines can derive valid results. While there is a long tradition of developing medical knowledge bases from scratch, we here propose a conservative approach—reuse existing large-scale repositories, but refine the data from these resources so that advanced requirements imposed by more expressive knowledge representation languages are met. Consistency checking comes almost for free, once the informal knowledge sources are embedded in a formal reasoning framework (cf., e.g. the work of Mejino and Rosse who recognized inconsis-
tencies in the UMLS based on formal representation structures in the Digital Anatomist model [41]). The resulting knowledge bases can then be used for sophisticated applications requiring sound medical reasoning.

The knowledge engineering approach we have proposed in this paper does exactly this. It provides a formally solid description logics framework with a modeling extension which supports not only taxonomic reasoning, but also incorporates partonomic reasoning adapted to the requirements of anatomy as the foundation of medical terminology. In spite of their evident weaknesses, the subsets of the UMLS we analyzed proved to be useful as a source of terminological knowledge on a large scale. Whereas the restitution of logical consistency could be achieved in a straightforward way, the cleansing of the resulting knowledge base from inadequate concept definitions and specification gaps implies a high degree of manual involvement, which requires enormous efforts when it has to be performed on the knowledge base as a whole. A realistic setting would be to eliminate inadequacies once and for all, but to remedy specification gaps only when required by concrete applications.

For anatomy and pathology, the domains under analysis, this study sheds light on the conditioned usability of the conceptual ‘raw material’ the UMLS metathesaurus provides for knowledge engineering. For macroscopic anatomy the existing resources proved fruitful due to the inclusion of the UWDA (University of Washington Digital Anatomist, cf. [42]) knowledge base which delivers semantically precise relationships. Severe weaknesses and underspecification arise in the pathology portion where the necessary linkage to anatomy proved to be entirely insufficient.

While plain automatic conversion from semi-formal to formal environments causes problems of adequacy of the emerging representation structures, the step-wise refinement methodology we propose already inherits its power from the terminological reasoning framework. In our concrete work, we found the implications of using the terminological classifier, the inference engine which computes subsumption relations, of utmost importance and of outstanding heuristic value. Hence, the knowledge refinement cycles are truly semi-automatic, fed by medical expertise on the side of the human knowledge engineer, but also driven by the reasoning system which makes explicit the consequences of (im)proper concept definitions.

**Acknowledgements**

Stefan Schulz was supported by a grant from DFG (Ha 2097/5-2).

**References**


