

Exploiting thesauri knowledge in medical

guideline formalization

Radu SERBAN^{1,3} and Annette ten TEIJE^{1,2}

Vrije Universiteit, Faculty of Exact Sciences, Amsterdam, The Netherlands

Abstract. As in software product lifecycle, the effort spent in maintaining medical knowledge in guidelines can be reduced, if modularization, formalization and tracking of domain knowledge are employed across the guideline development phases. We propose to exploit and combine knowledge templates with medical background knowledge from existing thesauri in order to produce reusable building blocks used in guideline development. These templates enable easier guideline formalization, by describing how chunks of medical knowledge can be combined into more complex ones and how they are linked to a textual representation. By linking our ontology used in guideline formalization with existing thesauri, we can use compilations of thesauri knowledge as building blocks for modeling and maintaining the content of a medical guideline. Our paper investigates whether medical knowledge acquired from several medical thesauri can be molded on a guideline pattern, such that it supports building of executable models of guidelines.

Keywords: Linguistic and Control Patterns, Guideline Modelling and Formalization.

1. Objective

Evidence-based clinical guidelines, representing disseminated state-of-the-art medical practice, undergo frequent changes due to new research results, and require permanent maintenance, similar to that required in a software project. Existing guideline modeling languages and guideline formalization frameworks ([4, 16, 12]) only use to a limited extent the rich medical knowledge present in medical thesauri to facilitate formalization, i.e. producing an operational model of the guideline. Recent research ([11]) advocate a Software Engineering view of the problem of guideline formalization: the requirements, the literature sources and their supporting medical background knowledge are captured explicitly in a medical guideline project, in which several control components of the formalized guideline are linked to their corresponding textual representation in the guidelines and can be traced back to the medical knowledge they were derived from. This is based on the hypothesis that certain lexical similarities exist between all instances of a guideline component, and can be mapped to the same formalized representation, and that identifying lexical similarities between guideline texts is relatively easier and cheaper to perform than producing the formalized equivalent of the text from scratch.

Even though some guidelines contain program-like actions advising on the most effective and safe medical practice in complex clinical situations, they are much less structured than a software program and lack a structured method for change management.

¹ This work has been supported by the European Commission's IST program, under contract number IST-FP6-508794 Protocure-II: www.protocure.org

² Address: De Boelelaan 1081a, 1081HV Amsterdam, The Netherlands. E-mail:annette@few.vu.nl

³ Current affiliation: IBM Global Business Services, Amsterdam, The Netherlands

Updating existing knowledge and importing knowledge from other guidelines is also difficult, due to a lack of modularity and organization of the knowledge employed to produce the guideline. This, despite the fact that the vocabulary and the linguistic constructs used are rather compact, regular and supported by existing terminological systems, including medical thesauri.

This paper addresses the problem of identifying, using existing medical thesauri (MeSH [1] and NCI [2]), those categories of knowledge that can be acquired automatically. We use this knowledge to enrich and validate a model of the medical guideline, such that validation of its content with respect to medical terminologies and exchange of knowledge with other medical sources are possible. To obtain a consistent *guideline ontology*, knowledge from different thesauri needs to be integrated such that a *guideline domain model* is obtained, which is conceptually sound, formal, expressive, comprehensible and executable ([17]).

We show how such a guideline domain model can be produced, consisting of two components - an ontology of the medical domain targeted by the guideline, and an ontology containing knowledge building blocks for guidelines, particularly control relations between medical terms. Using this guideline domain model we seek to enrich frequently occurring lexical regularities that convey relevant medical and procedural knowledge, with knowledge from medical thesauri. The ultimate goal is to propose building blocks for authoring, refining and validating a medical guideline. When a syntactic regularity contains lexical elements allowed by a target guideline ontology and encodes medical or procedural knowledge, we call this a linguistic component. We have previously proposed a methodology for using linguistic components in guideline formalization (see [18]). Here, we exploit the knowledge present in medical thesauri to enrich and be able to recognize these linguistic components.

The structure of this paper is as follows. In Section 2 we sketch by means of examples our method for enriching the formal model of the guideline using medical thesauri such as UMLS [3], MeSH [1] and NCI [2], and for building knowledge components for guideline formalization. Section 3 shows how these components are used in guideline formalization. In Section 4 we argue why this method is feasible for an oncology scenario, in fact giving a set of prerequisites for applying this method. Section 5 summarizes our conclusions.

2. Guideline Formalization Examples

This section is an illustration of our approach to guideline formalization by means of an example taken from our case study on breast cancer guidelines.

The following fragment from a guideline for treatment of breast-cancer [7]: [...]Special attention should be paid to those side-effects which are more likely to occur as the result of combining treatment modalities, such as the possibility of <u>more severe</u> skin reactions to radiotherapy following anthracycline-containing chemotherapy and the <u>increased</u> risk of lymphoedema and shoulder problems in the longer term after combining axillary radiotherapy and surgery.

It contains several linguistic templates, identifiable by the lexical markers in bold face. For guideline modeling, we are interested in linguistic templates expressing intentions and procedural knowledge, such as:

"{**Attention should be paid to**} [side-effects:med_effect], such as [skin-reactions:med_effect]." The intentional content of this linguistic template is:

"Avoid actions known to produce as side-effects 'skin reactions'."

Such intentional templates are found rather frequently in guidelines, combined into more complex procedural knowledge (or control) templates, such as:

DO treatment modalities: med action AND AVOID [skin reactions : med effect] == severe.

Such background knowledge is necessary for recognizing procedural knowledge in the text. For instance, the fragment:

I1 = [radiotherapy + following + anthracycline_containing_chemotherapy] (where "+" denotes concatenation) represents a composed action (by operational sequencing SEQ):

O1 = DO("anthracycline containing chemotherapy") SEQ DO("radiotherapy"). To recognize medical terms and control information in the text, we annotate the text using knowledge of the following form:

- radiotherapy subClassOf med_action
- anthracycline_containing_chemotherapy subClassOf med_action M1 = following subClassOf seg act op

This type of knowledge is similar to the one present in existing thesauri. By using thesauri as background knowledge, we can detect ambiguous or inconsistent usage of medical terms. Also, we can define control knowledge transformations, such as: Any instance, such as I1, associated with the control template:

CT = [med_action1 + seq_act_op + med_action2] has an equivalent operational representation (i.e., the formalized and executable

procedural part of the guideline) described by the template: FT = {DO(action1) SEQ DO(action2)}.

CT is a control template and FT is an operational template. The translation rule CT \rightarrow FT is called a control pattern (CP), and is used in guideline formalization, to map two equivalent guideline representation elements at different levels of abstraction: CK=

I1 instantiates CT using_ontology_mapping M1 CP: CT formalized_as FT using_formalization_mapping M2

11 translated_to O1 using_pattern CP

where M2 is the following mapping:

instanceOf(med_action1) \rightarrow DO(action1); M2 =

instanceOf(med_action2) \rightarrow DO(action2);

instanceOf(seq_act_op) → SEQ

By using patterns that map linguistic and control templates and capture the knowledge transformations of guideline fragments in the process of guideline formalization, a more direct translation of the text into a formal representation is possible, as depicted in Figure 1 and as suggested by our experiments using macro-rules in the guideline modelling language ASBRU [12].

Several sources of background knowledge can help us to recognize intentions, actions and other elements relevant in modelling the procedural part of a guideline:

A. information present in a guideline ontology, such as: med_action has_result med_effect therapy(subClassOf action) has_result(inherited) side_effects(subClassOf effect) chemotherapy(subClassOf therapy) skin reactions(subClassOf side effects)

B. information extracted from medical thesauri such as MeSH: chemotherapy subClassOf "MeSH/Analytical, Diagnostic and Therapeutic Techniques and Equipment Category/Therapeutics" skin_reactions subClassOf "MeSH/Diseases-Category/Pathological Conditions, Signs and Symptoms/Signs and Symptoms/ Skin Manifestations"

C. mappings from thesauri terms to guideline ontology terms, realized manually or by using semantic distance algorithms:

med action superClassOf "MeSH/Analytical, Diagnostic and

Therapeutic Techniques and Equipment Category/Therapeutics"

side_effect **superClassOf** "MeSH/Diseases-Category/ Pathological Conditions, Signs and Symptoms/Signs and Symptoms/Skin Manifestations"

Integrating knowledge from several sources as indicated above, facilitates validation of guideline ontology, allowing answering of questions such as: What instances of medical actions can represent legitimate instantiations of the concept "treatment modalities"? Which of these can influence the parameter "skin reactions"? Knowing that "radiotherapy" and "anthracycline-containing chemotherapy" are instantiations of the class "therapy", which is part of the definition of the term "treatment modalities", and any instance of class "chemotherapy" can produce as effect "skin reaction", the procedural core of the guideline can easier be reconstructed and even modified such that it maintains its consistency.

3. Method for guideline formalization using reusable guideline components

We view a medical guideline text from a Knowledge Engineering perspective, searching reusable knowledge components that convey process information, i.e. correlations condition-action, action-goal, repeated actions, functional decomposition and sequencing of actions. We assume the guideline text is a combination of explanations (definitions and descriptions of medical terms), procedural knowledge (process information, e.g. how to implement a treatment plan), goal statements (what is to be achieved) and argumentations (why it has to be achieved). From this text, we produce executable models of the procedural content of the guideline in two steps:

Step 1: Use of thesauri to enrich domain and guideline ontology. Input: guideline text, initial guideline ontology, thesauri.

Output: guideline domain model.

In this step, we exploit control and linguistic templates, by marking up semantic categories in the text and mapping text fragments to a set of control templates allowed by the guideline representation language. For this, we acquire the most commonly used medical categories, and the medical thesauri play an essential role in identifying these categories. The resulting guideline domain model can be used to produce a representation of the guideline text in several guideline representation languages.

Step 2: Transformation from guideline model to operational model Input: guideline domain model, guideline representation language (such as Asbru), control patterns (mapping guideline domain model onto guideline representation language modules) Output: operational model of the guideline (for the specific guideline representation language selected)

We generate an executable translation for each control template allowed by the guideline modelling language, then use these generic translations/mappings to produce an executable representation for the instances of each control template.

Knowledge templates can be found at several levels of abstraction, depicted in Figure 1: in the lexical, medical, guideline and operational representation of a guideline.

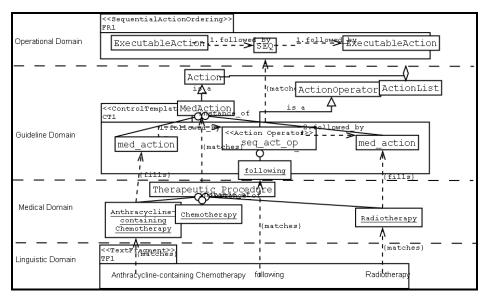


Figure 1 Abstraction levels for a semantic pattern.

The figure shows the relation between the terms used in the guideline, their relation to descriptors from the medical and guideline domain and the correspondence of guideline terms in the operational/procedural domain (the "semantics" of guideline control templates). A linguistic template (expressing intentional and procedural aspects) is mapped to a control template, and to a guideline ontology fragment (which covers the semantic categories and control categories present in the template). With the help of a control pattern, which uses as much as possible background knowledge from thesauri, the control template is mapped to a formalized executable representation of the guideline, the so-called operational model.

<u>Constructing the guideline ontology</u>. We have built a custom guideline ontology to cover medical relations often encountered in the guideline text, which are useful in guideline modeling, such as:

Therapy A helps against disease B. Treatment A consists of therapies B,C,D.

Drug A helps against disease B. Therapy A uses drug A.

This (*formalization-driven*) guideline ontology can be obtained from a simplified medical domain model similar to that of the semantic network of a meta-thesaurus like UMLS, by selecting only semantic categories and relations that are relevant to the guideline modeling language used. The procedural fragment of the guideline can be annotated, manually or semi-automatically, using UMLS semantic network classes. Then the initial guideline model is established, containing solely the UMLS classes and relations mentioned in the procedural fragment analyzed, which can be mapped to elements in the operational domain. Subsequently, this initial ontology is enriched with relations from existing medical thesauri, this process being driven by the goal of supporting guideline formalization. This gradual enrichment of the guideline ontology, in which new terms and relations are added based on common terms/relations in the guideline in an earlier phase of the guideline modeling process than when no background knowledge is used. The

use of thesauri is in this case beneficial to the modelling and validation of medical guidelines.

We organize this knowledge into compositional components that can be re-used for testing and change management of the guideline. A knowledge component is obtained by mapping control templates and engineered linguistic templates, as suggested by Amaral et al ([10]). We established mappings between several existing medical thesauri - MeSH, UMLS, NCI and our guideline ontology. For instance, we mapped manually classes from our guideline ontology into one or more UMLS classes, and imported several UMLS relations into our guideline ontology:

Mappings Guideline Ontology Classes --> UMLS Classes:

TargetGroup --> {Age_Group, Patient_or_Disabled_Group, Population_Group} Medication --> {Clinical_Drug} BodyPart --> {Body_Location_or_Region} MedAction --> {Diagnostic_Procedure, Therapeutic_or_Preventive_Procedure} Disease --> {Disease_or_Syndrome} MedContext --> {Sign_or_Symptom} Guideline Ontology Relations reusing UMLS class relations: MedAction **treats** Disease **uses** Medication **affects** BodyPart MedAction **part_of** MedAction **achieves** MedGoal **produces** MedEffect Based on such mappings, UMLS relations such as ClinicalDrug **affects** Body Location or Region

can be transformed into relations in our guideline ontology, such as

Medication affects BodyPart.

Several MeSH and NCI classes were associated with the same guideline concept. When two classes were mapped to a guideline ontology concept, their corresponding relations were also imported as relations in our guideline ontology.

To investigate what type of knowledge present in thesauri can be used in guideline formalization, we have done a case study: we identified the medical knowledge conveyed by sentences containing more than 2 thesauri terms in 5 oncology guidelines (2 for treatment of lung cancer, 3 for treatment of breast cancer). We reverse engineered a part of the guideline domain model which contains solely the medically relevant sentences in the guideline, proposing a skeleton of the knowledge used to produce the guideline. This guideline domain model needs to be validated and populated using classifications from existing thesauri. Apart from medical terms, it includes concepts such as ActionGroups, which are compositions of MedActions linked with special ActionOperators. Similar methods for extending existing medical terminologies using lexical and terminological knowledge and for mapping medical text to medical thesauri have been proposed by Rindflesch et al ([15, 6, 5]).

4. Prerequisites for formalizing medical guidelines using thesauri-based guideline ontology

To be able to extract an operational model for a specific guideline using existing medical thesauri, we have to verify that the guideline satisfies several requirements. If the following hypotheses can be confirmed for a class of guidelines, our method to produce an operational model of these guidelines is feasible.

Requirement 1: *Medical terms used by guidelines are covered by a controlled vocabulary.*

In our case study targeting oncology guidelines, we selected three procedural guidelines with the same field of study, breast cancer: CBO [7], SIGN [13], RCR [14].

We evaluated the size of the medical vocabulary shared by these guidelines, and checked their overlap with the terms of two thesauri - MeSH [1] and NCI [2].

We used the Text2Onto tool ([8]) to extract the most frequently used medical terms in each of the three guideline texts. The numbers of the most relevant terms, according to a common relevance threshold defined for the three documents were 174, 267 and 190 terms for SIGN, CBO and RCR guidelines, respectively. From a total of 394 unique relevant terms used by all three guidelines, 202 terms (i.e., more than 50%) were present in MeSH, 144 terms (about 36%) were present in NCI, and 120 - in both of them. Due to a different writing and organization style of the three guidelines, and different levels of abstraction addressed by them, a much smaller number of terms were shared by all three guidelines: 70 terms (about 18% from the union set of the three guidelines). However, 60 of these terms were present in MeSH, 45 of them were present in NCI and 42 of them in both, which suggests a good coverage of the shared vocabulary by existing controlled vocabularies. In fact, this simple experiment reveals that even for randomly selected guidelines, between 60% and 85% of the terms shared by these guidelines are present in the controlled vocabulary (even above 90% if the UMLS metathesaurus is used). **Requirement 2**: Semantic tagging of guidelines with medical domain knowledge from multiple thesauri is feasible.

We checked whether sufficient mappings exist between our guideline ontology and the intersection of NCI and MeSH semantic categories, such that the majority of the medical terms shared by the analyzed guidelines can be assigned a semantic category. The structure of MeSH and NCI indicates that merging of medical terms from the two thesauri are possible, based on their common fields (UMLS semantic categories and NCI classes). For instance, synonymous medical terms "Radiotherapy" and "Radiation therapy" are found in MeSH and NCI, assigned to the same medical category: "Therapeutic or Preventive procedure":

MeSH: DescriptorName="Radiotherapy" AllowableQualifierName={"adverse effects","contraindications"} ConceptList: {ConceptName, ConceptUMLSUI} SemanticTypeList: {SemanticTypeName="Therapeutic or Preventive procedure"} NCI: owl:Class ID="Radiation_Therapy" rdfs:subClassOf rdfs:resource="#Cancer_Treatment" Semantic Type: Therapeutic or Preventive Procedure Preferred Name,Synonyms={"Radiotherapy"},UMLS_CUI

By integrating the knowledge in these two thesauri entries, we can obtain a richer description of Radiotherapy: Radiotherapy equivalentOf Radiation_Therapy subClassOf CancerTreatment subClassOf TherapeuticOrPreventiveProcedure. All medical terms in MeSH and NCI can be assigned at least one UMLS semantic type. We found shallow equivalent classes between UMLS and our guideline ontology (med_action equivalentOf TherapeuticOrPreventiveProcedure). Therefore, we conclude that shallow semantic tagging of medical guidelines using medical terms from MeSH, NCI or UMLS and medical categories from our guideline ontology is feasible. Search of knowledge components at linguistic level are then possible, as medical terms can be replaced by their corresponding medical category, and the linguistic relations between these categories can be analyzed. MeSH is a comprehensive categorization of general medical terms, with more than 23,000 descriptors and more than 151000 subject headings and medical terms corresponding to semantic categories at various levels of abstraction. MeSH is a vocabulary source for UMLS, and it was preferred to UMLS for size considerations. From the 202 most relevant medical terms in any of the three guidelines analyzed, which were present in MESH, for 142 of them a category could automatically be assigned, based on mappings between categories in our guideline ontology and UMLS categories. For 40 others a correct semantic category could be learned (using a learning feature of the Text2Onto tool). For the remaining medical terms we had to assign manually a category. The results can be improved if we extend our guideline ontology and find new mappings to the UMLS semantic categories. Nonetheless, we conclude that for about 70% of the relevant medical terms, automatic semantic tagging is feasible.

Requirement 3: Control knowledge is abundant in the guideline text and can be recognized using linguistic regularities.

To verify that sufficient procedural constructs useful in an operational representation of a guideline can be identified, we have semantically tagged one guideline ([7]), built its operational model, then checked how many of the sentences used in the (semi)formal model contained control templates, i.e. had a clear control structure.

As result of this experiment, described in [18], it can be concluded that action sequencing [med_action+"following"+med_action] is the most frequently used specialization of template CT (Control Template) = [med_action+seq_act_op+ med_action], as it occurs in 40% of all occurences of CT in the text. In many cases, overlapping control templates can be identified only by having a categorization of lexical markers: sequencing action operator (seq_act_op) has instances such as: "after", "then", "followed by"; action composition relations: "combination of", "addition of"; or therapy effect markers: "results in", "improves", "expected to produce".

5. Conclusions

This paper proposes an approach to formalization of clinical guidelines by integrating knowledge from existing medical thesauri, such as UMLS, MeSH and NCI, in order to reduce the effort spent in keeping guidelines up-to-date and reuse control knowledge in subsequent formalizations. An ontology with two components was built, one representing medical domain knowledge reusing semantic categories and relations from thesauri, and one containing building blocks for guidelines (called control templates). The terms and relations present in a guideline model are similar to the terms and relations between semantic categories retrieved from medical thesauri. Shallow semantic tagging of the guideline text, using categories mapped from existing thesauri, enables identification of control templates that have an operational translation. This leads to a more structured medical guideline formalization process that supports frequently changing guidelines, guaranteeing their adherence to medical terminologies and improving their maintainability. To this end, we build a guideline ontology containing classes and relations from medical thesauri and control constructs allowed by a class of guideline representation language. The quality of the formalization depends on the quality of this guideline ontology, and in turn is sensitive to the quality of the thesauri used as source. Two vocabulary sources for UMLS (MeSH and NCI) are considered for enriching our guideline ontology. NCI focuses on non-operational knowledge (see detailed statistics in [9]), therefore it is less suitable for acquisition of control knowledge and more suitable for isolating oncology specific information. MeSH is a more comprehensive source of control knowledge, but it contains relatively few types of medical relations that can be used in formalizing guidelines.

We state three prerequisites for using medical thesauri in guideline formalization and show that in our case study on oncology guidelines these prerequisites are verified. We argue that, based on a shared controlled vocabulary between existing guidelines, and based on semantic categories shared by several medical thesauri, it is possible to enrich the medical domain model targeted by guidelines and to bring it closer to the constructs allowed by a guideline domain model. By using a list of semantic patterns (i.e., a text fragment matching linguistic and control templates) built on top of an enriched guideline domain model, and populating these patterns with knowledge from medical thesauri, we are able to refine and modularize procedural knowledge in medical guidelines. This facilitates integration of knowledge from external sources, validation of guideline content consistency against medical vocabularies, and reduces the effort spent in formalization, maintenance and authoring of medical guidelines.

References

- MeSH (Medical Subject Headings). URL: <u>http://www.nlm.nih.gov/mesh/meshhome.html</u> (last accessed June 15, 2008).
- [2] National Cancer Institute (NCI) Ontology. URL: <u>http://www.mindswap.org/2003/CancerOntology/</u> (last accessed June 15, 2008).
- [3] Unified medical language system. URL: <u>http://www.nlm.nih.gov/research/umls/</u> (last accessed June 15, 2008)
- [4] Shahar Y et al. DEGEL: A hybrid, multiple-ontology framework for specification and retrieval of clinical guidelines. In Proc. of 9th Conference on AI in Medicine - Europe (AIME) 2003, Protaras, Cyprus, 2003.
- [5] Aronson AR.. Effective mapping of biomedical text to the UMLS metathesaurus: the MetaMap program. In Proc. of American Medical Informatics Association Symposium, pages 17–21. National Library of Medicine, 2001.
- [6] Bodenreider O, Rindflesch TC and Burgun A. Unsupervised, corpus-based method for extending a biomedical terminology. In Proc. of the Workshop on Natural Language Processing in the Biomedical Domain, Philadelphia, pages 53–60. Association for Computational Linguistics, 2002.
- [7] CBO. Guideline for the Treatment of Breast Carcinoma. Dutch Institute for Healthcare improvement, 2002. PMID: 12474555.
- [8] Cimiano P and Volker J. Text2onto: A framework for ontology learning and data-driven change discovery. In E.Metais, A.Montoyo, R.Munoz, eds, NLP and Information Systems: 10th International Conference on Applications of Natural Language to Information Systems, NLDB 2005, Alicante, Spain, June 15-17, 2005, pages 227–238. Springer, 2005. LNCS Vol.3513.
- [9] de Coronado S et al. NCI thesaurus: using science-based terminology to integrate cancer research results. Medinfo, 11(1):33–37, 2004.
- [10] do Amaral MB, Roberts A and Rector AL. NLP techniques associated with the OpenGALEN ontology for semi-automatic textual extraction of medical knowledge: abstracting and mapping equivalent linguistic and logical constructs. In Overhage JM, editor, Proceedings of the 2000 American Medical Informatics Association Annual Symposium (AMIA 2000), Los Angeles, pages 76–80. American Medical Informatics Association, 2000.
- [11] ten Teije A et al. Protocure: Supporting the development of medical protocols through formal methods. In Proc. of the Symposium on Computerized Guidelines and Protocols (CGP-04), April 2004.
- [12] Johnson P, Miksch S and Shahar Y. ASBRU: A task-specific, intention-based, and time-oriented language for representing skeletal plans. In Pierret-Golbreich C, Filby I, Wijngaards N, Motta E, van Harmelen F(eds), Proc. of the 7th Workshop on Knowledge Engineering: Methods and Languages (KEML-97), 1997.
- [13] Scottish Intercollegiate Guidelines Network. Breast Cancer in Women, SIGN#29. SIGN, 1998.
- [14] Royal College of Radiologists. Guidelines on the non-surgical management of lung cancer. Springer-Verlag, 2001.
- [15] Rindflesch TC and Aronson AR. Semantic processing for enhanced access to biomedical knowledge. In Real World Semantic Web Applications, pages 157–172. IOS Press, 2001.
- [16] Boxwala A, Peleg M and Tu S. GLIF3: a representation format for sharable computer-interpretable clinical practice guidelines. J.of Biomedical Informatics, 37(3):147–161, 2004.
- [17] van der Maas A, ter Hofstede A, and ten Hoopen AJ. Requirements for medical modeling languages. Journal of the American Medical Informatics Association, 8(2):146–162, 2001.
- [18] Serban R et al. Ontology-driven extraction of linguistic patterns for modelling clinical guidelines. In S.Miksch et al., editor, Proc. 10th Conference on AI in Medicine (AIME 05), LNAI 3581, Aberdeen, Scotland, pages 194–203. Springer, 2005.