User Centred Networked Health Care
A. Moen et al. (Eds.)
IOS Press, 2011

doi:10.3233/978-1-60750-806-9-517

Automatic Definition of the Oncologic EHR Data Elements from NCIT in OWL

Marc CUGGIA[a], Annabel BOURDÉ[a], Bruno TURLIN[b], Sebastien VINCENDEAU[b], Valerie BERTAUD[a], Catherine BOHEC[c], and Régis DUVAUFERRIER[a]
[a]*UMR 936 Inserm, Faculté de médicine de Rennes. France*
[b]*CHU Pontchaillou, Rennes, France*
[c]*Réseau ONCOBRETAGNE, Rennes -France*

Abstract. Semantic interoperability based on ontologies allows systems to combine their information and process them automatically. The ability to extract meaningful fragments from ontology is a key for the ontology re-use and the construction of a subset will help to structure clinical data entries. The aim of this work is to provide a method for extracting a set of concepts for a specific domain, in order to help to define data elements of an oncologic EHR. Method: a generic extraction algorithm was developed to extract, from the NCIT and for a specific disease (i.e. prostate neoplasm), all the concepts of interest into a sub-ontology. We compared all the concepts extracted to the concepts encoded manually contained into the multi-disciplinary meeting report form (MDMRF). Results: We extracted two sub-ontologies: sub-ontology 1 by using a single key concept and sub-ontology 2 by using 5 additional keywords. The coverage of sub-ontology 2 to the MDMRF concepts was 51%. The low rate of coverage is due to the lack of definition or mis-classification of the NCIT concepts. By providing a subset of concepts focused on a particular domain, this extraction method helps at optimizing the binding process of data elements and at maintaining and enriching a domain ontology.

Keywords. Semantic interoperability, information system, ontology modularization, date elements, value-set

1. Introduction

The development of health information systems, including EHR (Electronic Health Record), implies to define information models that capture increasingly complex patient data. Standardization efforts of these models are in process either through HL7 (version 3) templates [1] or archetypes (Open EHR / EN13606) [2]. These models define and organize data elements i.e. a basic unit of information built on standard structures having a unique meaning and distinct units or values [3]. These data elements are used in the forms, messages or documents in order to capture or to transmit patient data in an interoperable way. The process of defining an information model contains two steps. Firstly, we define all the data elements necessary and sufficient to capture information from a domain. Each data element contains a label and a value-set based on an interface terminology (end-user oriented) [4,5]. Secondly, in order to ensure a semantic interoperability, we bind this interface terminology with a pivot terminology i.e. controlled vocabularies or reference ontologies. This “bottom-

up" approach, beginning from a consensus of experts and leading to formalization of the semantic of an information model, is particularly cumbersome whereas biomedical ontologies are precisely providing the domain knowledge. The objective of this work is to show the interest of using an ontology in a top-down approach, to automatically extract, from a few key concepts (or "seed concept") , a sub-ontology to define data elements and their value sets. We extracted from the National Cancer Institute's Thesaurus (NCIT), a sub-ontology representing the data elements of the MDMRF (Multi-Disciplinary Meeting Report Form) related to prostate neoplasm. In recent years, many ontolgies' fragments extraction techniques, starting with a search criterion, have been developed. LexValueSets [6] defines an approach for extracting data elements (value-sets) from SNOMED CT. This technique uses two complementary processes. The extensional one where extraction is conducted using a set of concepts chosen by experts. The intentional process where extraction is done from a semantic definition of a concept. The modularization techniques have also been explored by several studies [7]. It consists of methods for the extraction of ontological modules from an original ontology. The objective is to enable the reuse of ontology, but also to facilitate their development, their management and their use [8]. Our work was conducted under the research project ANR ASTEC, whose objective is to automatically determine the eligibility of patients to be included in clinical trials, from the MDMRF data.

2. Methods

Extraction algorithm: The goal of this work is to provide a consistent sub-ontology with the domain concepts i.e. a subset of the NCIT with all the semantic relations between the concepts. *Our work is based on the NCIT (version 10.07)* because of its availability in OWL format, its free use, its specificity in the oncology domain, and because it is both a terminology for encoding and a reference ontology internationally recognized. Using Protege-OWL API to access Ontology model, the extraction algorithm takes some parameters as input: an ontology with OWL format, a list of key concepts from which extraction should begin, the directions in which it searches (i.e., towards parents, children and restrictions) and the list of restriction types to be followed (i.e, relations between concepts except the subsumption). It searches for semantically related concepts of interest and adds them in an empty ontology, which will progressively grow to create the sub-ontology.

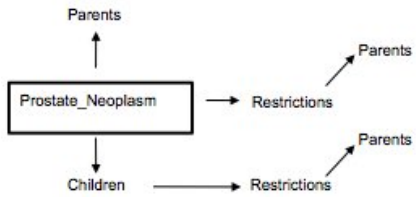

Figure 1. Extraction from Prostate_Neoplam key concept.

Initially, the key concept to apply the extraction algorithm was `Prostate_Neoplasm` (figure 1). We searched all its parents, children, and all the target concepts related to either `Prostate_Neoplasm` or its children with a restriction. Then we searched all the parents of these target concepts until ontology's root to have a

consistent ontology. It's the sub-ontology 1. To have a better coverage of the domain, we added in a second time 5 concepts (given by two experts as concepts that best represent the domain) to `Prostate_Neoplasm` as key concepts. Then we apply again the extraction algorithm with these 6 concepts. It's the sub-ontology 2.

Method for evaluating the algorithm: To evaluate our method, we used, as a target to reach, the MDMRF about prostate cancer created by the clinicians. We manually encoded all data elements of this MDMRF in NCIT concepts to compare them to those of our sub-ontologies. To analyze the sub-ontologies, we grouped the MDMRF concepts and those of each sub-ontology in 5 subsets (figure 2). **Set A**: manually encoded MDMRF concepts. **Set B**: concepts of the sub-ontology that exactly match those of A. The ratio B/A gives the proportion of ontology concepts that were strictly the same as those in MDMRF. Then to analyze more precisely the sub-ontology 2, two physicians evaluated all concepts of this sub-ontology and classified them in subsets. **Set B'**: concepts that are strictly the same as those in the MDMRF (Set B) and the concepts that experts defined as semantically close to concepts in Set A. These concepts could substitute for Set A concepts or could complete the data elements of the MDMRF. It's an extension of Set B. **Set C**: concepts that could be present in an EHR but the expert didn't keep for MDMRF. **Set D**: other concepts that couldn't be present in the MDMRF nor in the EHR but that are necessary to formally define the sub-ontology concepts. We also analysed the MDMRF concepts that were not extracted in the two sub-ontologies (Set A minus Set B) in order to determine the reason.

3. Results

The NCIT ontology contains 83143 concepts. The manual encoding of the 36 data elements of MDMRF produced 82 NCIT concepts. 11 concepts were not found in the NCIT, such as hip replacement notion, hepatic or renal chronic insufficiency: recall is 86,5%. The extraction algorithm performance is compared to these 82 concepts.

The extraction from a single key concept (`Prostate_Neoplasm`), produced the sub-ontology 1 containing 434 concepts. Among these concepts, only 16 concepts (set B) matched exactly with the 82 MDMRF concepts (set A: recall B/A is 19,5%). When we checked the sub-ontology 1, we highlighted the absence of some concepts like those about the TNM (international classification of the extension of malignant tumors).

The extraction from 5 key concepts in addition to the `Prostate_Neoplasm` concept (`Prostate_Adenocarcinoma`,`Prostate_Cancer_TNM_Finding`,`Biopsy_of_Prostate`,`PSA_Assay`,`Total_Gleason_Score_for_Prostate_Cancer`) produced the sub-ontology 2 that contained more concepts (483). However, the recall B/A (51%) was better since 42 concepts (set B) matched exactly with the 82 MDMRF concepts (set A). The precision (proportion of B in the sub-ontology 2) was lower (9%). However, sub-ontology 2 contained 140 concepts that experts classified in the set B' (precision is 27%). These concepts were either semantically close to MDMRF concepts and could be substituted to them (e.g., `recurrent_prostate_neoplasm` instead of `recurrent_disease`). or completed the list of possible value as well. For example we find many missing histologic types that experts didn't integrate in the MDMRF (e.g., `Prostate_Adenosquamous_Carcinoma` or `Prostate_Basal_Cell_Carcinoma`). We also found concepts about the tumor staging nut no longer used in favor of TNM classification (e.g., `Stage_I_Prostate_Adenocarcinoma`).

The set C contains 34 concepts that were not present in the MDMRF but that could be in the EHR. The set C contained essentially semiological concepts like `Bone_Pain` or `Urinary_Retention` that were not used in the MDM record because they were not taken into account for the MDM decision. Moreover, we found concepts about benign tumor forms like `Prostate_Adenoma` that can co-exist with a cancerous transformation.

The set D contained 309 concepts that were not essential for encoding MDMRF and EHR. These concepts represented intermediate concepts that structured the sub-ontology and made the definition of concepts of interest. These concepts could be used for reasoning and automatic classification. For example, they may be parents of `Prostate_Neoplasm` concept like `Reproductive_System_Neoplasm` or `Disorder_by_Site`, which were too general to characterize the disease but necessary for the reasoning. We found concepts about genomic (e.g., `Gain_of_Chromosome_2p`) that had no interest to be used routinely but participated to the definition of the disease. 49% of MDMRF concepts were not found in the sub-ontology 2. These concepts were essentially either about Medical antecedents (e.g., `Ischemic_Heart_Disease`) or about prostate neoplasm treatments (e.g., `Adjuvant_Therapy`), or findings elements (e.g., `birth_date` or `age` or `Zubrod_Performance_Status`).

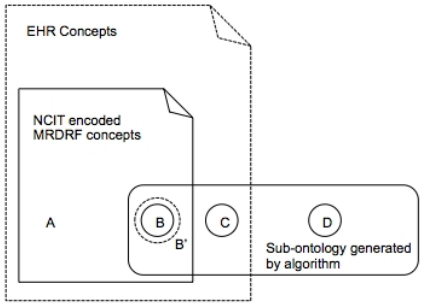

Figure 2. Sets used to evaluate sub-ontologies concepts.

4. Discussion

Our extraction method allows a modularization of the NCIT in domain specific sub-ontologies. This extraction makes available a limited number of concepts (those of interest in a domain) and facilitates the process of defining the data elements of an information model. This approach has been already defined as relevant and is being integrated into terminology server. This study was carried out about one single neoplasm type. Other neoplasms could be better defined in the NCIT both by concepts number and richness of the relations. Performing the same study on other cancers should assess this selection bias. Moreover, if we targeted the data elements of prostate medical record, instead of the MDMRF data elements (that contains only the crucial elements for the therapeutic decision), we would have a better coverage with the sub-ontology (through concepts of Set C), especially with finding concepts. Our coverage of relevant terms (51%) is higher than LexValueSets project (35%) if we consider a strict comparison. We reviewed exhaustively the concepts of our sub-ontology whereas

it was done only on a sample in the LexValuesSets project. The precision of the Set B is quite low because our sub-ontology contains many concepts that don't belong to the MDMRF but are essential for automatic reasoning. Our algorithm takes as input parameter a list of restrictions that we followed during the extraction. If we don't follow the restrictions that are not interesting for clinical data collection like the "omics" domain, the precision of the algorithm will be increased. Furthermore, our approach allows extracting concepts of interest (Set B') that can be used to complete the MDMRF during its creation. The relatively low coverage of the sub-ontology 1 was increased by the multi-concepts extraction (sub-ontology 2). This is explained by a lack of concept definitions and misclassified concepts. Firstly, concepts are well represented in the NCIT (86,5%), but with absence of relationships between concepts, they cannot be extracted by the algorithm. Secondly, some concepts were misclassified in NCIT, e.g., the extractor cannot get back primary and secondary Gleason scores concepts whereas we used `Total_Gleason_Socre_for_Prostate_Cancer` concept as key concept. These three concepts are not in the same hierarchy (same parents) and are not related to the `Prostate_Adenocarcinoma` concept. Therefore, another contribution of our work is that this modularization allows identifying more effectively misclassified or insufficiently defined concepts. The task of maintenance and enrichment of the ontology is easier by working on reasonable sized ontology and with an approach by domain. Thus we describe a virtuous cycle where the ontology is optimized for encoding patient data, and in return, extraction algorithm has optimal results due to enrichment and best expression of the ontology. Compared to bottom-up approach used for designing templates or archetype, we propose to use a top-down approach, starting from the ontology, to get back the semantics of a domain. Although, this approach supposes that the ontology contains concepts coming from the "reality" (e.g. MDM), and definitions that are sometimes non-formal and so, non-computable (e.g. `may_have`, is extensively used in NCIT).

Acknowledgement: We would like to thank the ANR-TECSAN for its financial support and Sahar BAYAT for her help.

References

[1] *HL7 Template*, http://www.hl7.org/Special/committees/template/index.cfm
[2] Kalra D, Beale T, Heard S. "The openEHR Foundation," *Studies in Health Technology and Informatics*, vol. 115, 2005, p. 153-173.
[3] *Data element* - Wikipedia, the free encyclopedia. http://en.wikipedia.org/wiki/Data_element
[4] Rosenbloom ST, Miller RA, Johnson KB, Elkin PL, Brown SH. "Interface Terminologies: Facilitating Direct Entry of Clinical Data into Electronic Health Record Systems," *JAMIA*, 13, 2006, pp. 277-288.
[5] Daniel C, Buemi A, Mazuel L, Ouagne D, Charlet J. Functional Requirements of Terminology Services for Coupling Interface Terminologies to Reference Terminologies. *MIE 2009*:205-209
[6] Pathak J, Jiang G, Dwarkanath SO, Buntrock JD, Chute CG, Chute C. "LexValueSets: an approach for context-driven value sets extraction," AMIA. Annual Symposium Proceedings / AMIA Symposium. *AMIA Symposium, 2008*, p. 556-560.
[7] Seidenberg J, Rector A. "Web ontology segmentation: analysis, classification and use," *Proceedings of the 15th international conference on World Wide Web*, ACM, 2006, p. 13-22.
[8] Rector A, Napoli A, Stamou G, et al. 'Report on modularization of ontologies', *Technical report, Knowledge Web Deliverable D2.1.3.1*, (2005)