INTERNATIONAL JOURNAL OF MEDICAL INFORMATICS 81 (2012) 566–578

journal homepage: www.ijmijournal.com

Semantic similarity-based alignment between clinical archetypes and SNOMED CT: An application to observations

María Meizoso García[a,*], *José Luis Iglesias Allones*[a,1], *Diego Martínez Hernández*[b,2], *María Jesús Taboada Iglesias*[a,3]

[a] *Department of Electronics and Computer Science, Campus Vida, University of Santiago de Compostela, Spain*
[b] *Department of Applied Physics, Campus Vida, University of Santiago de Compostela, Spain*

ARTICLE INFO

Article history:
Received 29 July 2011
Received in revised form
13 February 2012
Accepted 15 February 2012

Keywords:
Terminology mapping
Electronic Health Records
Clinical archetypes
SNOMED CT
Semantic interoperability
Knowledge representation

ABSTRACT

Purpose: One of the main challenges of eHealth is semantic interoperability of health systems. But, this will only be possible if the capture, representation and access of patient data is standardized. Clinical data models, such as OpenEHR Archetypes, define data structures that are agreed by experts to ensure the accuracy of health information. In addition, they provide an option to normalize clinical data by means of binding terms used in the model definition to standard medical vocabularies. Nevertheless, the effort needed to establish the association between archetype terms and standard terminology concepts is considerable. Therefore, the purpose of this study is to provide an automated approach to bind OpenEHR archetypes terms to the external terminology SNOMED CT, with the capability to do it at a semantic level.

Methods: This research uses lexical techniques and external terminological tools in combination with context-based techniques, which use information about structural and semantic proximity to identify similarities between terms and so, to find alignments between them. The proposed approach exploits both the *structural context* of archetypes and the terminology context, in which concepts are logically defined through the relationships (hierarchical and definitional) to other concepts.

Results: A set of 25 *OBSERVATION* archetypes with 477 bound terms was used to test the method. Of these, 342 terms (74.6%) were linked with 96.1% precision, 71.7% recall and 1.23 SNOMED CT concepts on average for each mapping. It has been detected that about one third of the archetype clinical information is grouped logically. Context-based techniques take advantage of this to increase the recall and to validate a 30.4% of the bindings produced by lexical techniques.

Conclusions: This research shows that it is possible to automatically map archetype terms to a standard terminology with a high precision and recall, with the help of appropriate contextual and semantic information of both models. Moreover, the semantic-based methods provide a means of validating and disambiguating the resulting bindings. Therefore, this work is a step forward to reduce the human participation in the mapping process.

* *Corresponding author*. Tel.: +34 8818 13569.
E-mail addresses: maria.meizoso@usc.es (M. Meizoso García), joseluis.allones@usc.es (J.L. Iglesias Allones), diego.martinez@usc.es (D. Martínez Hernández), maria.taboada@usc.es (M.J. Taboada Iglesias).
[1] Tel.: +34 8818 13580.
[2] Tel.: +34 8818 23241.
[3] Tel.: +34 8818 13561.

1386-5056/$ – see front matter © 2012 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.ijmedinf.2012.02.007

1. Introduction

Electronic Health Records (EHRs) mainly involve clinical patient data including the personal and family history, the clinical state, the dispensed therapies, and other relevant information about the reached diagnostics and outcomes. Using natural language to describe this information provides a full expressivity level, but it hampers computational processing, interfering with one of the main challenges of eHealth [1], namely semantic interoperability of health systems. Achieving this would provide the possibility of automatically managing arbitrary fragments of the patient information from different EHRs, without the demand for particular mappings between them, at the same time that it would increase the quality of medical services. But, this will only be possible whenever health systems are endowed with the capability to share information in a meaningful, unambiguous and accurate way. Therefore, standardizing the capture, representation and access of the detailed and complete patient information is essential. Several EHRs proposals, such as Good electronic Health Record (GeHR) [2], OpenEHR [3] or CEN/ISO EN13606 [4], have started to pave the way for solving the problem. Aimed at supporting healthcare delivery in a correct, robust, reliable and unambiguously understood way, they establish the standardization of clinical data upon a three-layer architecture [1], including: (1) reference models; (2) clinical data structure definitions; and (3) clinical terminology systems.

Reference model provides the building blocks to be reused in the specification of particular facets of clinical information. Using these building blocks, formal definitions of clinical data can be agreed by experts to ensure the accuracy of health information. The main global candidates for data structure definitions are OpenEHR and ISO/EN 13606 archetypes, and HL7 templates. OpenEHR archetypes provide an option to standardize the clinical data content which captures patient-related information by means of binding terms used in the model definition to standard medical vocabularies. To date, several relevant institutions have participated in the development of on-line archetype repositories [5–7]: the OpenEHR, the UK National Health Service Connecting for Health (NHS CFH) [8] and the National E-Health Transition Authority (NEHTA) [9]. The archetypes of these repositories are updated and maintained by a variety of cooperating groups of expert in different domain. With the exception of a small number of cases, bindings to standard vocabularies are infrequent in the archetypes openly accessible in the repositories. However, aligning archetypes and terminology systems is a crucial step to get semantic interoperability among different health information systems.

When a relevant set of archetypes have already been developed and are stable, the effort needed to bind archetype fragments to a standard terminology will be considerable. So far, the International Health Terminology Standards Development Organisation (IHTSDO) [10] and the openEHR Foundation have shown their interest in collaborating in order to explore how the SNOMED CT terminology and openEHR archetypes can best be aligned to support EHRs [11]. Additionally, once the bindings are obtained, a relevant human labour will be needed to suitably interpret them and guarantee the validity of the resulting bindings. Therefore, automated methods are needed in order to simplify the mapping process and then, to interpret and evaluate the resulting alignments in a more effective way than manual revision.

The present study proposes an automated approach to binding terms from archetypes to the external terminology SNOMED CT. This approach applies a combination of two basic matching methods (lexical and context-based techniques) in order to produce and then validate the alignment. The lexical techniques find alignments among concepts handling the synonymy, whereas the context-based techniques identify semantic similarities between the structure of the archetype and the relations in SNOMED CT.

2. Background

2.1. *Archetype Definition Language (ADL)*

The Archetype Definition Language is a formal language to express standard archetypes for any reference model. The three sections of an archetype – header, body and ontology – can be seen in Fig. 1. The header section includes meta data about the archetype; the body section, which is hierarchically organized, involves the structure and restrictions associated with the clinical concepts required to record a particular clinical statement; and the ontology section includes term definitions and bindings, which map archetype terms to standard terminology concepts.

2.2. *SNOMED CT*

SNOMED CT is a comprehensive clinical terminology managed by the IHTSDO that provides a standard for clinical information [12]. In July 2011, it contained 295,708 active unique concepts, each one described by a preferred term and one or more additional terms called synonyms. Each concept is logically described through the relationships to other concepts. There are four types of relationships: defining, qualifying, historical and additional. The defining relationships specify the meaning of the SNOMED CT concepts. They include the 'is a' and 'defining attribute' relationships. The 'is a' relationship logically represents a concept by linking it with other concept in a subsumption hierarchy, whereas the 'defining attribute' models the difference with other concepts in the hierarchy.

2.3. *Techniques to find bindings*

Some disciplines (e.g. information science [13,14], databases [15,16] or ontologies [17–20]) have been actively working on improving techniques to find equivalences between concepts across sources, techniques that can be now applied to automatically create bindings. The most principal *name-based techniques* use the lexical properties of term names to find correspondences between concepts. The *structure-based techniques* use structural properties (such as, shared relationships across sources) to find alignments between concepts. Generally, these techniques are used in combination with name-based techniques, as it is demonstrated