Journal of Biomedical Informatics xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Formalizing MedDRA to support semantic reasoning on adverse drug reaction terms

Cédric Bousquet [a,b], Éric Sadou [a], Julien Souvignet [a,b], Marie-Christine Jaulent [a], Gunnar Declerck [a,*]

[a] INSERM, U1142, LIMICS, F-75006, Paris, France; Sorbonne Universités, UPMC Univ Paris 06, UMR_S 1142, LIMICS, F-75006, Paris, France; Université Paris 13, Sorbonne Paris Cité, LIMICS, (UMR_S 1142), F-93430, Villetaneuse, France

[b] University of Saint Etienne, Department of Public Health and Medical Informatics, Saint-Etienne, France

ARTICLE INFO

Article history:
Received 4 November 2013
Accepted 16 March 2014
Available online xxxx

Keywords:
MedDRA
SNOMED-CT
Ontology
Terminology
Semantic reasoning
Adverse drug reaction

ABSTRACT

Although MedDRA has obvious advantages over previous terminologies for coding adverse drug reactions and discovering potential signals using data mining techniques, its terminological organization constrains users to search terms according to predefined categories. Adding formal definitions to MedDRA would allow retrieval of terms according to a case definition that may correspond to novel categories that are not currently available in the terminology. To achieve semantic reasoning with MedDRA, we have associated formal definitions to MedDRA terms in an OWL file named OntoADR that is the result of our first step for providing an "ontologized" version of MedDRA. MedDRA five-levels original hierarchy was converted into a subsumption tree and formal definitions of MedDRA terms were designed using several methods: mappings to SNOMED-CT, semi-automatic definition algorithms or a fully manual way. This article presents the main steps of OntoADR conception process, its structure and content, and discusses problems and limits raised by this attempt to "ontologize" MedDRA.

1. Introduction

MedDRA[1] (Medical Dictionary for Drug Regulatory Activities) terminology is used to record and report adverse drug reactions (ADR) data for pre-marketing as well as post-marketing drug surveillance in most countries and is recommended by the ICH for the electronic transmission of individual case safety reports (ICSR) [1]. MedDRA includes most terms of common ADR dictionaries, e.g. WHO–ART (World Health Organization–Adverse Reactions Terminology), and terminologies such as the International Classification of Diseases (ICD-9) [2]. MedDRA generally allows a precise description of ADRs and related issues (*e.g.* investigations and surgical procedures carried out on the patient) [3,4]. MedDRA terms are organized in 5 levels: System Organ Class (SOC), High Level Group Terms (HLGT), High Level Terms (HLT), Preferred Terms (PT) and Low Level Terms (LLT). Each PT is linked to a primary SOC and can be linked to other secondary SOCs. LLTs are usually synonyms of PTs (including spelling or lexical variant) but can also be more precise terms. This structure and systematic hierarchy facilitates navigation and updating. MedDRA "multi-axiality" – *i.e.* the same PT can belong to several SOCs – allows to group terms in different ways, depending on context needs [5,6] and Standardised MedDRA Queries (SMQs) are available to aid in case identification.

However, MedDRA also suffers limitations and might be improved in several ways [7–11]. In our view, its main limitation comes from its standard terminological format, which restricts the possibility of accessing terms based on their semantics. MedDRA is a system halfway between first-generation systems (paper-based systems) and second generation systems (compositional systems) [12]. MedDRA is available in electronic format, but unlike second-generation systems, it does not enable the creation of new concepts by composition of preexisting atomic concepts and their meaning cannot be processed automatically by semantic tools. For example, *Gastric Ulcer* PT is part of MedDRA but neither of the atomic concepts, *Stomach* and *Ulcer*, necessary for an explicit representation of meaning are included. The only semantic information available in MedDRA derives from the terms' labels and to some extent from their hierarchical organization.

If this lack of compositionality would not interfere with the objectives of MedDRA, there would be no need to worry. But is this really the case? The primary stated objective of MedDRA is to

* Corresponding author.
E-mail address: gunnar.declerck@upmc.fr (G. Declerck).

[1] MedDRA® is a registered trademark of the International Federation of Pharmaceutical Manufacturers and Associations.

http://dx.doi.org/10.1016/j.jbi.2014.03.012
1532-0464/

provide an internationally approved classification for efficient reporting and communication of ADR data between countries [13]. To that aim, MedDRA must ensure accurate and consistent term selection[2]. However, the large number of terms in MedDRA makes this goal difficult to achieve. Several studies have observed that MedDRA users often experience difficulties in selecting appropriate terms for coding due to their higher specificity or generality compared to their verbatim [15,16] (see [3,7,17] for similar issues). MedDRA provides more precise terms than its predecessors (WHO–ART or COSTART), but a precise and consistent selection of terms (i.e. homogenous from one report to another) remains difficult due to the absence of univocal definitions, or the impossibility to assist the coding activity with querying tools able to process their meaning.

We can hypothesize that converting MedDRA into a third generation system – *i.e.* its "ontologization" – could help to overcome those difficulties: a formal representation of MedDRA terms' meaning might improve the coding accuracy and contribute to harmonizing the coding strategies by making possible the use of semantic processing tools. Computational ontologies and associated semantic web techniques are known to favor interoperability in the medical data sharing process [18–22]. Contrary to traditional medical classifications where the meaning of concepts relies on the implicit knowledge of the user, ontologies provide an explicit and logical (thus univocal) description of the semantics, which reduces the risk of misinterpretation of the terms and ensures the possibility of reliable information sharing [18,20]. MedDRA ontologization could also benefit data mining techniques used for post-market drug surveillance. Studies of the impact of the MedDRA hierarchical organization on automated signal detection – i.e. measures of drug-reaction causal relatedness based on statistical comparison between observed and expected cases – have pointed out that taxonomic limitations decrease the sensitivity and specificity of the signals computed [7,23–25]. Conversely, the possibility to perform semantic reasoning on MedDRA terms meaning has been shown to increase the performances of signal detection algorithms [26,27].

We previously experimented with providing formal definitions to WHO–ART terms using an alignment with the SNOMED-CT clinical terminology [28]. We assume that SNOMED-CT is still the best candidate for providing formal definitions to MedDRA because SNOMED-CT terms are defined using a Description Logic (DL) formalism and because a fair number of alignments between MedDRA and SNOMED-CT is described in the UMLS (Unified Medical Language System) metathesaurus [29].[3]

We report here our first step for providing an "ontologized" version of MedDRA: adding formal definitions to MedDRA terms using SNOMED-CT. We first describe our alignment techniques between MedDRA and SNOMED-CT terms and present the resulting formal definitions of MedDRA terms that are given in an OWL file named OntoADR. The primary objective of OntoADR is to define the PT level of MedDRA, and only secondarily the terms of upper levels. Focusing on the PTs was motivated by our intended use for grouping ADRs reported at the PT level because this level is recommended for analysis of pharmacovigilance data. Moreover, ADRs are described using PTs in the public version of the FDA (U.S. *Food and Drug Administration*) pharmacovigilance database that we use for signal detection. After presenting OntoADR, we discuss why formal definitions are not sufficient for an ontological version of MedDRA when keeping the original hierarchy. We finally highlight limits and problems related to this choice and propose additional work that would be necessary to achieve an improved and more advanced version of OntoADR.

Before the version of OntoADR presented in this paper, different preliminary versions were designed, the first one in 2003. We used them in several studies to perform grouping of case reports with different methods of terminological reasoning: mainly subsumption and approximate matching. Our studies showed that grouping of medically related conditions using terminological reasoning significantly improves signal generation performances [27]. More occurrences of drug-ADR associations could be identified with that method than by using the MedDRA hierarchy.

2. Methods

2.1. Providing MedDRA terms with computable formal definitions

Two main strategies are possible for providing MedDRA terms with formal definitions and thus achieving the advantages of third generation systems: (1) building an OWL-DL (Web Ontology Language – Description Logics) representation of MedDRA; (2) keeping the terminological format of MedDRA and mapping MedDRA terms with concepts from other existing ontologies in order to use their semantic representations and make possible indirect reasoning. The second strategy is adopted by the UMLS metathesaurus and the Bioportal Website [31], or different research works [11,32], where MedDRA terms are mapped to equivalent concepts (or at least assumed so) from several biomedical ontologies or terminologies, including SNOMED-CT. However, we believe this solution suffers from certain weaknesses:

(1) This strategy is limited because not every MedDRA term can be mapped to a unique SNOMED-CT concept having the same meaning (one-to-one mapping). For example, the MedDRA term *Joint dislocation postoperative* does not have a direct equivalent concept in SNOMED-CT and must be expressed through *several* Snomed-CT concepts, e.g. *Dislocation of joint* and *Postoperative complication*. Decomposing the meaning of a MedDRA term in that way is different than stating a mere conceptual equivalence: in the present example, the mapping relations correspond to subsumption (is-a) relations, or, to use SKOS (Simple Knowledge Organization System) mapping vocabulary, to skos:broadMatch relations. In other cases, domain relations are even necessary.

(2) This strategy does not permit to perform semantic reasoning based on the relations MedDRA terms have to each other, first of all hierarchical relations. The main added value of formalizing semantics using subsomptive relations: the principle of property inheritance is thus lost. The same applies to inferences based on formal properties of properties, e.g. on their transitive or inverse character. Reasoning can be made on the concepts and conceptual relations from the external resource used to define MedDRA, but not on MedDRA itself.

(3) Finally, an indirect mapping-based formalization of MedDRA semantics is simply unnecessary if the purpose is to provide a formal definition usable for reasoning. Directly defining the MedDRA terms avoids having to use a supplementary resource in the semantic reasoning procedure.

For these reasons, we have opted for a direct formalization of MedDRA terms. We have built an OWL-DL version of MedDRA named *OntoADR*, where MedDRA terms are defined by a set of

[2] As the ICH guide for MedDRA users explains, "unless users achieve consistency in how they assign terms to verbatim reports of symptoms, signs, diseases, etc., use of MedDRA cannot have the desired harmonizing effect in the exchange of coded data. [...] Consistent term selection promotes medical accuracy for sharing MedDRA-coded data and facilitates a common understanding of shared data among academic, commercial and regulatory entities." [14].

[3] The UMLS is a semantic network developed by the NLM (U.S. National Library of Medicine) to link terms from more than hundred controlled vocabularies and provide semantic definitions of terms [30]. Both SNOMED-CT and MedDRA are included.