Expert Systems with Applications 42 (2015) 1465–1478

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Review

State-of-the-art on mapping maintenance and challenges towards a fully automatic approach

Julio Cesar Dos Reis [a,b,*], Cédric Pruski [a], Chantal Reynaud-Delaître [b]

[a] Public Research Centre Henri Tudor, 6 avenue des Hauts-fourneaux, L-4362 Esch-sur-Alzette, Luxembourg
[b] LRI, University of Paris-Sud, Bât 650, 91405 Orsay Cedex, France

ARTICLE INFO

Article history:
Available online 16 September 2014

Keywords:
Knowledge management
Mapping evolution
Mapping adaptation
Mapping maintenance
Ontology alignment
Ontology evolution

ABSTRACT

In several domains, software applications have intensively used Knowledge Organization Systems (KOS) like database schemas, ontologies, taxonomies and thesauri and their associated semantic correspondences (*i.e.*, mappings). This underlines the relevance and capabilities of KOS and mappings to manage and integrate vast amounts of data. However, the dynamic nature of domain knowledge forces knowledge engineers to constantly modify KOS, to keep them up to date and useful. In this context, the maintenance of mappings affected by KOS evolution still remains an open research issue. Although this problem appears relevant for many different computer science fields, ranging from database to artificial intelligence, literature has so far only superficially addressed it to enable more flexible, automatic and precise solutions. This article presents, discusses and compares existing approaches for maintaining mappings and describes open research challenges.

1. Introduction

Semantic correspondences, or mappings, connect entities of different Knowledge Organization Systems (KOS) (Hodge, 2000), like ontologies, database schemas and thesauri to improve the capabilities of software applications in sharing, integrating and managing wide variety of data. Mappings allow computers to interpret data whose semantics are expressed using different KOS, through the definition of semantic relations that exist between their entities. For example, the biomedical concept named "*torso*" is an equivalent concept to "*trunk*" of another KOS. KOS and their associated mappings play a key role in a wide range of tasks including data annotation (Bodenreider & Stevens, 2006), information retrieval (Kitamura & Segawa, 2008), data integration (Lambrix, Strömbäch, & Tan, 2009) as well as knowledge representation (Bonacin, Pruski, & Da Silveira, 2013), and more generally, semantic interoperability.

Taking biomedical KOS as an example, mappings between two distinct KOS, named *Systematized Nomenclature of Medicine-Clinical Terms*[1] (SNOMED CT®) (SCT for short), and *International Classification of Disease – Clinical Modification*[2] (ICD-9-CM) (ICD for short), can easily support users accessing external data annotated with ICD from clinical records annotated with SCT. In this case, data referring to the same entity may be represented by different concepts' identifiers.

This context requires that mappings remain consistent (from the semantic point of view) over time, especially when KOS entities involved in mappings evolve. This problematic scenario occurs in many dynamic and knowledge-intensive domains where knowledge engineers might frequently modify interconnected KOS. In the Semantic Web for instance, up-to-date mappings could allow more trustable semantic searches and reasoning over integrated ontologies (Kitamura & Segawa, 2008). In biomedicine, mappings provide a semantic reference for understanding the meaning of data between different systems (Lambrix et al., 2009). Thus, updated mappings guarantee a consistent integration of models and effectively support software applications of different nature.

The huge size and dynamics of existing KOS, especially in the biomedical domain, forces knowledge engineers to periodically revise hundreds of thousands of mappings making it a time consuming, error prone and tedious task. The critical part played by mappings in decision support applications thus requires automatic tools to support engineers and domain experts in the maintenance process. However, the design and implementation of such tools must deal with several factors and can be tackled from various perspectives with their own drawbacks and advantages. These aspects represent serious obstacles towards a fully automatic solution, requiring a complete and exhaustive survey of existing approaches addressing the problem.

* Corresponding author at: Public Research Centre Henri Tudor, 6 avenue des Hauts-fourneaux, L-4362 Esch-sur-Alzette, Luxembourg.
E-mail addresses: julio.dosreis@tudor.lu (J.C. Dos Reis), cedric.pruski@tudor.lu (C. Pruski), chantal.reynaud@lri.fr (C. Reynaud-Delaître).

[1] www.ihtsdo.org/snomed-ct.
[2] www.cdc.gov/nchs/icd/icd9cm.htm.

http://dx.doi.org/10.1016/j.eswa.2014.08.047
0957-4174/

This article provides a thorough survey on mapping maintenance. In summary, we make the following contributions:

- We formally define and illustrate the mapping maintenance problem, highlighting the complexity of the problem. We borrow examples from the biomedical domain, which are explored throughout this article.
- We systematically review the literature on the mapping maintenance problem, offering a complete state-of-the-art by presenting, comparing and discussing existing proposals in different suggested categories.
- We analyze lacks of existing approaches discussing general open issues that literature fails to address to reach fully automatic mapping maintenance. This allows us to underscore open research challenges.

We achieve the results through a careful and articulated analysis of the reviewed literature combined with our previously conducted empirical studies. The obtained results highlight various innovative and useful aspects that are essential for addressing mapping maintenance in a more complete and refined manner.

We structure the remainder of this article as follows: Section 2 describes the mapping maintenance problem and its specificities, providing necessary definitions. Section 3 reviews existing proposals organizing them by suggested categories. We provide an analytical comparison between them. Section 4 discusses our findings and challenges of different nature, representing open research issues. Finally, Section 5 wraps up with concluding remarks and outlines future work.

2. Mapping maintenance problem

A **KOS** K specifies a set of concepts interrelated by directed relationships. We define a set of concepts of a KOS K_x at time j, such that $j \in \mathbb{N}$, as $C\left(K_x{}^j\right) = \left\{c_i^j | i \in \mathbb{N}\right\}$. Each concept $c \in C\left(K_x^j\right)$ has a unique identifier and is associated with a set of attributes $A(c) = \{a_i | i \in [1 \ldots p]\}$ (*e.g.*, label, synonym, definition, *etc.*), where p is the number of attributes characterizing concept c. Furthermore, each attribute is defined for a particular objective, *e.g.*, "*label*" for denoting concept names or "*definition*" for giving the meaning in the context where the concept is used. A relationship *rel* between two concepts, $c_1 \in C\left(K_x{}^j\right)$ and $c_2 \in C\left(K_x{}^j\right)$ interrelates a particular concept and another one in the same KOS, *e.g.*, $rel(\text{"}is\text{—}a\text{"}, c_1, c_2)$, where the label of c_1 refers to "*brain cancer*", and "*cancer*" is the label of c_2, respectively.

Mappings denote the semantic correspondences between entities (most usually concepts) of different, but domain related KOS. More formally, given two KOS namely K_S and K_T, we define K_S as the source KOS and K_T the target KOS of mappings. A mapping m_{st}^j, established at time j, between two concepts $c_s^j \in C\left(K_S^j\right)$ (namely source concept) and $c_t^j \in C\left(K_T^j\right)$ (namely target concept) is given by:

$$m_{st}^j = \left(c_s^j, c_t^j, semType^j\right) \quad (1)$$

where $semType^j \in \{\bot, \equiv, \leqslant, \geqslant, \approx\}$ refers to the semantic relation between c_s^j and c_t^j. The $\bot$ stands for *unmappable*, $[\equiv]$ *equivalent*, $[\leqslant]$ *more specific than*, $[\geqslant]$ *less specific than* and $[\approx]$ *partially matched*, respectively. For instance, concepts can be equivalent (*e.g.*, "*torso*" $\equiv$ "*trunk*"), one concept can be less or more specific than the other (*e.g.*, "*lower limbys*" $\leqslant$"*limb segment*") or concepts can be somehow related ($\approx$). We define $\mathcal{M}_{ST}^j = \left\{\left(m_{st}^j\right)_i | i \in \mathbb{N}\right\}$ as the set of different mappings at time j between KOS K_S and K_T. Mappings are usually created either manually or using (semi-)automatic alignment methods (Euzenat & Shvaiko, 2007).

This definition of mapping differs from the one largely accepted by the database community where a schema mapping specifies how data instances of one schema correspond to data instances of another (Velegrakis, Miller, & Popa, 2004a). Mapping maintenance has been historically studied in database schemas. In this context, they represent mappings in a declarative way as queries or view definitions, having a very general form of mapping. They describe a mapping q from a schema S (called the source schema) to schema T (called the target schema), as an assertion of the form: $Q_S \longrightarrow Q_T$, where Q_S consists in a query over S and Q_T refers to a query over T (Velegrakis et al., 2004a). When handling mappings between *XML* models they use a similar definition of mapping. Observe that the definition of KOS mappings in terms of correspondences between concepts differs from a schema view or query. A unique query may involve a set of dependent equivalent schema correspondences, which changes the abstraction level of the mapping definition.

The evolution of a KOS (Klein & Noy, 2003) in terms of atomic or complex changes affecting its entities may invalidate previously determined mappings (Dos Reis, Pruski, Silveira, & Reynaud, 2012). In other words, given a mapping m_{st}^j, due to the modifications affecting the concept c_s^j or c_t^j, the type of semantic relation $semType_{st}^j$ no longer represents the correct semantic link between c_s^j and c_t^j.

Fig. 1 presents the investigated scenario of the mapping maintenance problem. Given two versions of the same source KOS, namely K_S^0 at time j and K_S^1 at time $j+1$, we always have at least one target KOS K_T and an initial set of valid mappings M_{ST}^0 between K_S^0 and K_T^0 at time j. Since we consider KOS evolution, we examine different versions of each KOS.

If K_S or K_T evolves (*cf.* Fig. 1), represented by a set of KOS change operations (*diff*), we need to determine the set of updated mappings M_{ST}^1, since the evolution probably impacts mappings in M_{ST}^0. This problem consists in determining how to perform changes in existing mappings, which for instance includes redefining the semantic relation between the new version of c_s and c_t. The semantic validity stands for the logical consistency of the mappings. For instance, we might handle mappings established with removed concepts.

We can have a simplified view of the investigated scenario considering the evolution of only one KOS per time. The results of the mapping maintenance task must consist in a set of up-to-date and most complete possible mappings in M_{ST}^1. Complete means that the highest coverage between both KOSs is obtained. More generally, we define mapping maintenance as follows:

Mapping maintenance refers to the task aiming to keep existing mappings in an updated and valid state, reflecting changes affecting

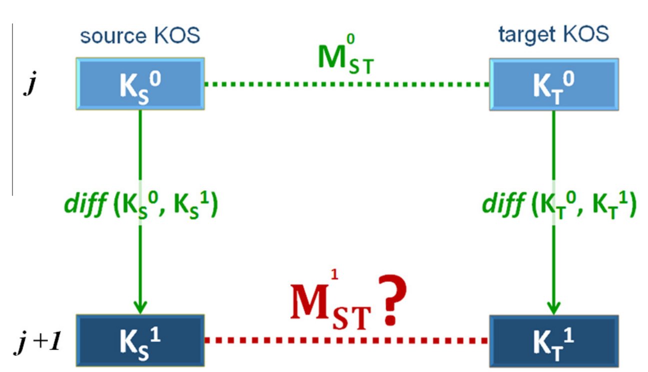

Fig. 1. The mapping maintenance problem. This figure shows the mapping maintenance problem. A source and a target KOS interrelated via a sets of mappings M_{ST}^0. New versions of these KOSs may trigger KOS changes that can affect existing mappings.