Evolution of biomedical ontologies and mappings: Overview of recent approaches

Anika Groß [a,*], Cédric Pruski [b], Erhard Rahm [a]

[a] Institute of Computer Science, Universität Leipzig, P.O. Box 100920, 04009 Leipzig, Germany
[b] Luxembourg Institute of Science and Technology, 5 Avenue des Hauts-Fourneaux, L-4362 Esch-sur-Alzette, Luxembourg

ARTICLE INFO

ABSTRACT

Biomedical ontologies are heavily used to annotate data, and different ontologies are often interlinked by ontology mappings. These ontology-based mappings and annotations are used in many applications and analysis tasks. Since biomedical ontologies are continuously updated dependent artifacts can become outdated and need to undergo evolution as well. Hence there is a need for largely automated approaches to keep ontology-based mappings up-to-date in the presence of evolving ontologies. In this article, we survey current approaches and novel directions in the context of ontology and mapping evolution. We will discuss requirements for mapping adaptation and provide a comprehensive overview on existing approaches. We will further identify open challenges and outline ideas for future developments.

1. Introduction

Ontologies have gained much importance in the past two decades, especially in the biomedical domain [1,2]. Many different ontologies have been developed in various sub-disciplines. For instance, BioPortal [3] currently provides access to more than 500 different biomedical ontologies. Ontologies consist of defined concepts, that are typically structured within trees or acyclic graphs where the concept nodes are interconnected by *is-a*, *part-of* and other semantic relationships. One main application of ontologies is the semantic annotation of different kinds of data objects. For instance, the well-known Gene Ontology (GO) is used to describe molecular functions of genes and proteins [4] and to predict new gene functions [5]. Chemical entities can be described by the Chemical Entities of Biological Interest (ChEBI) ontology [6], and concepts of medical ontologies like SNOMED CT [7] are assigned to documents like electronic health records (EHRs) or case report forms (CRFs). However it is important to note, that in the biomedical domain the term "ontology" is often not used in the sense of formal, axiom-based ontologies but instead for a wide spectrum of simpler terminologies including a.o. thesauri, taxonomies and is-a-hierarchies. The well-known definition of Gruber "*An ontology is an explicit specification of a conceptualization.*" [8] leaves room for variation w.r.t. to the detail of specification [9]. There is a wide spectrum of ontologies of varying expressiveness ranging from simple controlled vocabularies and thesauri to informal and formal "is-a" structures, and, at the highest level of expressiveness, formal ontologies that specify disjoint classes, part-whole relationships and further kinds of logical constraints [9]. The W3C provides a definition for different kinds of non-formal ontologies and calls them knowledge organization systems (KOS). KOS denote a.o. thesauri, classification schemes, subject heading systems and taxonomies and can be expressed by the *Simple Knowledge Organization System* (SKOS) data model [10]. Throughout the paper we will use the term "ontology" for ontologies of varying expressiveness as done by most of the relevant work on biomedical ontology and ontology evolution.

Often there are several ontologies within one domain and they can contain overlapping information. Mappings between such related ontologies interrelate or link corresponding and semantically related concepts and are of high importance for data integration and ontology-based query and analysis tasks. For instance, these mappings support merging several related ontologies into one ontology (e.g. [11,12]). A prominent huge integrated data source is the Unified Medical Language System (UMLS) [13] built out of more than 100 biomedical ontologies. Moreover, ontology mappings can support a semantic search since ontology-based queries can be enhanced by involving additional ontologies that are interconnected via mappings. Typically, an ontology mapping covers a set of semantic correspondences (links) between the concepts of two different ontologies. The semi-automatic determination of ontology mappings (ontology matching) has been an active research area for more than a decade [14,15]. Similarly numerous approaches

* Corresponding author.
E-mail addresses: gross@informatik.uni-leipzig.de (A. Groß), cedric.pruski@list.lu (C. Pruski), rahm@informatik.uni-leipzig.de (E. Rahm).

have been proposed to determine biomedical annotation [16,17], i.e., to link biomedical objects or documents to describing ontology concepts. Such methods produce recommendations that support domain experts in finding correct and complete ontology mappings and annotations.

Usually, ontologies are not static but modified on a regular basis. This process is known as *ontology evolution*. For instance, ontologies need to be changed to incorporate new domain knowledge, remove design errors or to achieve changed requirements. Often ontology development is a collaborative process that is supported by tools such as Protégé [19] or OBO-Edit [20]. In the life sciences, many ontology consortia continuously release new ontology versions. For instance, GO releases a new version every day, while the National Cancer Institute Thesaurus (NCIT) [21] is published on a monthly basis. Fig. 1 exemplarily shows the history of changes between 2015–11 and 2016–04 in GO. Typically, new versions contain improved and extended knowledge such as new concepts (classes), relationships or attributes like synonyms. However, existing knowledge can also be revised or removed, e.g. concepts might be deleted or marked as obsolete. For instance in the shown time period for GO (see Fig. 1), new classes have been added continuously, in 03-2016 some concepts were set to obsolete, and some definitions and class labels have been deleted. To manage the evolution of ontologies it is essential to determine changes, e.g. by analyzing change logs or by computing the difference (Diff) between two given versions of an ontology. Such a Diff is useful to synchronize changes in collaborative ontology development and to adapt dependent applications.

The evolution of ontologies has impact on ontology-based applications. For instance, ontology mappings and annotations can become invalid when the underlying ontologies are changed. This is especially critical in highly volatile domains such as the life sciences. Fig. 2 illustrates two ontologies ($O1$ and $O2$) and a mapping between them ($M_{O1,O2}$). In $O1$, one concept has been removed (red) while two concepts have been added to $O2$ (green). Another concept in $O2$ has been revised (blue) e.g., by changing the concept name. These ontology changes have impact on the set of correspondences (dashed lines) and might require changes in the mapping. In the example, one correspondence is associated to a deleted concept, and might therefore be removed. Moreover, the added and revised concept might lead to novel correspondences. Hence, ontology-based mappings can become out-dated as a consequence of ontology evolution. In order to keep mappings up-to-date they need to be migrated to currently valid ontology versions. On the one hand, a manual mapping maintenance can be very time consuming or even infeasible since ontologies and mappings can become very large. On the other hand, automated methods could be simply reapplied on the same data to obtain a valid mapping w.r.t.
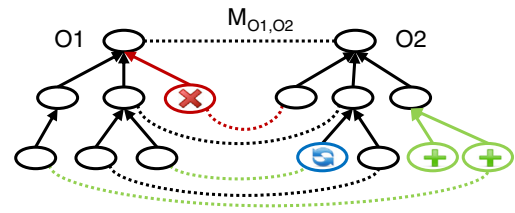

Fig. 2. Example ontology and mapping evolution.

the current ontology version. However, this can lead to a huge loss in quality since existing mappings might have been manually verified and corrected in the meantime. Just recomputing the results would discard this valuable knowledge. Moreover, usually a smaller part of an ontology is changed such that it seems likely to adapt only affected mapping parts. Therefore, it is useful to apply (semi-) automatic adaptation methods to migrate out-dated ontology-based mapping to currently valid ontology versions.

In this review, we will first introduce the problem of ontology and mapping evolution (Section 2) and then give an overview of recently proposed evolution methods for the biomedical domain and discuss open challenges:

- Methods for ontology evolution have been surveyed in several contexts before (e.g. [22–25]). Here we will focus on recent approaches that we see relevant for semi-automatic adaptation of ontology-based mappings and applications in the life sciences. This includes novel directions in ontology change detection and prediction and the visualization of ontology evolution. (Section 3)
- We will then discuss requirements for mapping evolution and provide a comparison and overview on existing (semi-) automatic adaptation strategies for ontology-based mappings. (Section 4)
- We will finally outline open challenges and future directions for the evolution of ontologies and ontology-based mappings and applications (Section 5).

2. Problem formulation

In this section, we will introduce the basic scenario of ontology and mapping evolution along with an illustrating example. An ontology $O = (C,A,R)$ consists of a set of concepts C (or classes) that are connected via a set of relationships R with different semantics such as *is-a* or *part-of*. Often ontologies form so-called Directed Acyclic Graphs.
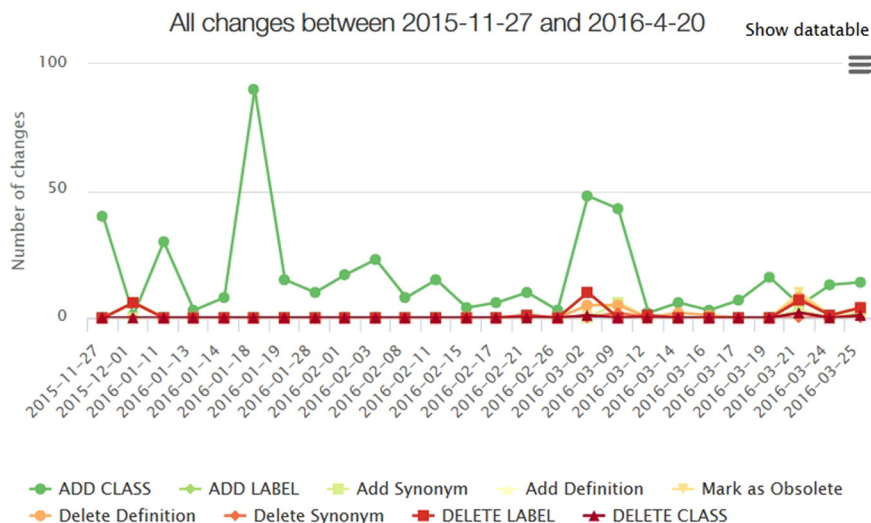

Fig. 1. History of changes in Gene Ontology (generated with ontology lookup service [18]).