MAISA - Maintenance of semantic annotations

Silvio Domingos Cardoso

universite
PARIS-SACLAY

NNT : 2018SACLS338

MAISA – Maintenance of Semantic Annotations

Thèse de doctorat de l'Université Paris-Saclay
préparée à l'Université Paris-Sud

École doctorale n°580 Sciences et technologies de l'information et de la communication (STIC)
Spécialité de doctorat : Informatique

Thèse présentée et soutenue à Orsay, le 07 Décembre 2018, par

M. Silvio Domingos Cardoso

Composition du Jury :

M. Pierre Zweigenbaum Directeur de recherche CNRS, LIMSI Université Paris-Sud	Président
M. Fabien Gandon Directeur de Recherche, INRIA, Université Côte d'Azur	Rapporteur
M. Jean Charlet Chargé de mission APHP, LIMICS - UMRS Paris 6	Rapporteur
M. Patrick Ruch Professeur, Geneva School of Business Administration - HES-SO	Examinateur
Mme Lina Soualmia Maître de Conférence, LIMICS - Normandie Universités	Examinateur
M. Cédric Pruski Chargé de recherche, Luxembourg Institute of Science and Technology - LIST	Examinateur
M. Marcos Da Silveira Chargé de recherche, Luxembourg Institute of Science and Technology - LIST	Examinateur
Mme Chantal Reynaud Professeure, LRI - Université Paris-Sud	Directeur de thèse

Thèse de doctorat

Acknowledgements

I would like to thank everyone who supported me during the development of this PhD. Especially my advisors, Chantal Reynaud-Delaître, Cédric Pruski and Marcos da Silveira for their support, patience and remarkable guidance to shape and develop this work.

I am also thankful to all Luxembourg Institute of Science and Technology (LIST) staff who directly or indirectly contributed to carry out this project, as well as, our patterns from University of Leipzig for the successful collaboration.

I would like to thank my family for their support in this stage of my life, my friends in Luxembourg for the good moments, in special to Da Vide, Stefano, Rafael, Gabriel, Vitor and Thayc. I am also thankful to all Brazilian community in Luxembourg and their immense support, in special to Clarissa Gomes, Marco Gomes, Alexandre Weimar and Clarissa Weimar.

Finally, I express my heartily gratitude to the Fonds National de la Recherche (FNR) for the financial support and the members of the jury who dedicated their efforts on reviewing this thesis.

Table of content

List of Figures

List of Tables

Chapter 1

Introduction

Contents

1.1 Motivation and research question

The use of Knowledge Organization Systems (KOS) [Hodge, 2000], such as classification schemes, thesauri or ontologies in the medical field, has proven to be of great value in tackling semantic interoperability issues. In many cases, KOS elements are used to annotate objects such as electronic health records (EHR), case report forms (CRF), genes or publications in order to make their semantics explicit for software applications.

This is the case in the biomedical domain, where the main interests in annotating documents are twofold for healthcare professionals: i) to transfer these documents to other institutions/people (e.g., to accelerate the reimbursement process, to request a second opinion, etc.), ii) to easily retrieve patient information. Secondary uses of these annotations are often seen in decision-support systems, public health analysis, patient recruitment for clinical trials, etc. For example, the well-known Gene Ontology (GO) is used to describe the molecular functions of genes and proteins, and scientific publications in MEDLINE are semantically annotated with concepts from the Medical Subject Headings (MeSH), facilitating the search for relevant medical information [Lowe and Barnett, 1994].

These diverse use cases for annotations show that they are widely utilized and can support various tasks in medical information systems, such as retrieving, sharing and exchanging information. However, the dynamic nature of medical KOS mean that annotations may be affected each time a new version of a KOS is released to include or review healthcare knowledge. New concepts can be added, obsolete ones removed and the definitions of existing concepts may be refined through the modification of their attribute values [Dos Reis et al., 2014]. The removal of a concept in a terminology engenders the removal of the semantics of the associated annotation, therefore making the annotated data incomprehensible for computers. More generally, changes in KOS may directly impact the annotations associated with changed concepts or changed data and new KOS versions can potentially invalidate previous annotations. As a result, many annotations can lose their relevance and value, thus hindering the intended use and exploitation of annotated data.

Consider, for instance the example in Figure 1.1. A subset of a document in PUBMED[1] is

[1] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1342315/

annotated with the term *Menstrual migraine*, an attribute from concept 625.4 of ICD-9-CM, version 2008AA. In version 2009AA, this attribute was removed and became a new concept with the ID 346.4. We consider this annotation as impacted, because the change in the KOS caused a mismatch between the annotation created with version 2008AA and the concept in the new KOS, version 2009AA. Furthermore, the *excludes* relationship in the ICD-9-CM guidelines states that: i) these concepts are considered mutually exclusive, and ii) the condition *Menstrual migraine* is not included in this particular code. In consequence of these changes, concepts 625.4 and 346.4 should not be assigned together and the terms from code 625.4 that were changed have to be encoded elsewhere.

Thus, software applications like search engines and data portals will be affected when retrieving information from documents. For example, doctors accessing EHRs from a hospital through a search engine will not be able to retrieve precise and complete information if the query specifies *Menstrual migraine* and no document was annotated with the right concept code at the query evaluation time. It illustrates the direct impact of KOS changes on semantic annotations and underlines the real need for advanced methods and tools able to keep semantic annotations up-to-date, avoiding the need for human intervention to this manually.

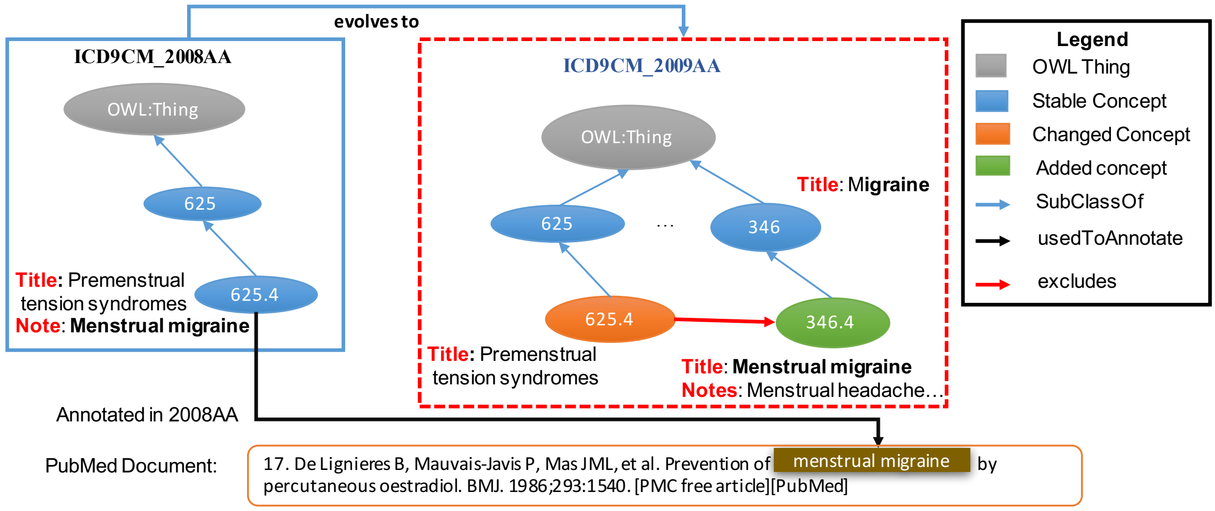

evolves to
ICD9CM_2008AA
OWL:Thing
625
625.4
Title: Premenstrual tension syndromes
Note: Menstrual migraine
ICD9CM_2009AA
OWL:Thing
625
...
346
Title: Migraine
625.4
346.4
Title: Premenstrual tension syndromes
Title: Menstrual migraine
Notes: Menstrual headache...
Legend
OWL Thing
Stable Concept
Changed Concept
Added concept
SubClassOf
usedToAnnotate
excludes
Annotated in 2008AA
PubMed Document:
17. De Lignieres B, Mauvais-Javis P, Mas JML, et al. Prevention of menstrual migraine by percutaneous oestradiol. BMJ. 1986;293:1540. [PMC free article][PubMed]

Figure 1.1: Annotation evolution case study.

Annotations cannot always be changed directly because: i) the content of the document cannot be accessed, and ii) the documents are encrypted and the meta-data (i.e., the annotations) can only be accessed, not modified. This is usually the case in EHR where data managers are not able to modify medical content without the intervention of health professionals. Nonetheless, users should still be able to use the latest KOS versions to access documents via annotations. An example of such a case is depicted in Figure 1.2, where the query *Tiotropium Bromide* is searched into an indexed database, which does not contain this entry. Thus, the query will not yield any results, i.e. the query will not return any documents associated with this term to the user, even though data containing the above-mentioned terms is present in the database.

The European eStandards project[2] is an example of where such cases could occur. In a scenario where multi-agent systems exchange health information from patient documents with national and cross-border institutions, it is important to consider the privacy of data while maintaining the high-level information that is exploitable by direct queries. Therefore, ad-hoc methods that adapt the annotations when they are accessed or queried by users are necessary.

In this global context, this thesis will design and implement methods and tools to support

[2] http://www.estandards-project.eu/

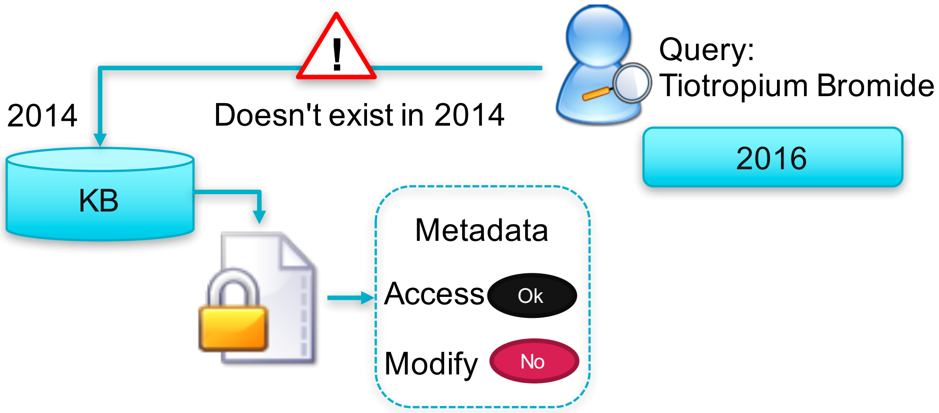


Figure 1.2: Annotations not modifiable leading to a query without results (problem 2)

the semi-automatic maintenance of the semantic annotations affected by KOS evolution in order to keep annotations exploitable over time. We aim to answer the following research question:

General Problem: **How can we automatically maintain the validity of biomedical annotations in the presence of changes in the underlying KOS over time?**

In particular, such a system has to be generic enough to deal with different terminologies having distinct structures. Furthermore, performance issues must be considered when trying to apply reasoning over thousands of concepts, e.g., Medical Subject Headings (MeSH) with 237,000 entries and Systematized Nomenclature of Medicine, Clinical Terms (SNOMED CT) with around 310,000 entries.

Thus, we formulated the following hypothesis to guide our investigation:

Hypothesis: The information from KOS, as well as information about KOS evolution, can be used to define a robust maintenance mechanism.

This brings us to the following specific challenges:

- **RQ1:** What is the impact of KOS changes on semantic annotations?
- **RQ2:** What is the most suitable model for addressing the annotation evolution problem?
- **RQ3:** How can we automatically maintain the validity of semantic annotations without re-annotating the content of all documents when KOS are updated?
- **RQ4:** Which methods can be used to keep the annotations searchable when the document and annotations cannot be changed directly?
- **RQ5:** Can we predict which KOS concept will change in the near future and what kind of changes will impact that concept?

Finally, the research problem investigated in this thesis is summarized in Table 1.1.

1.2 Methodology

This thesis aims to address the requirements for annotation maintenance comprehensively and to correct the shortcomings of previous approaches. To do this, we have adopted a methodology based on an iterative process, as depicted in Figure 1.3.

The methodology comprises three phases: i) Data-driven Analysis, ii) Design & Formalization, and iii) Implementation, which are performed in cycles in order to find answers for each of the research questions.

Table 1.1: Research Summary

Area	Knowledge Representation
Subject	Maintenance of Semantic Annotations
General Problem	Maintenance of semantic annotations after changes in the underlying KOS.
Hypothesis	Information from KOS, as well as information about KOS evolution, can be used to define a robust maintenance mechanism.
General Goal	Design and implementation of a semi-automatic method to maintain semantic annotations
Specific Research Questions	**RQ1:** What is the impact of KOS changes on semantic annotations? **RQ2:** What is a suitable model for addressing the annotation evolution problem? **RQ.3:** How can we automatically maintain the validity of semantic annotations without re-annotating the content of all documents when KOS are updated? **RQ4:** Which methods can be used to keep the annotations searchable when the document and annotations cannot be changed directly? **RQ5:** Can we predict which KOS concepts will change and impact the annotations in the near future?

- **Data-driven Analysis**: The main objective of this phase is to understand the typical annotation process and the evolution of annotations triggered by KOS changes. To do this, we used three major resources as primary data: i) automatic and manual annotations, ii) documents like EHRs and CRFs, iii) medical KOS. The outcomes provide insights and quantitative results, demonstrating the behaviour and correlation between the KOS changes and impacted annotations.

  For the scope of this work, the four KOS selected (MeSH[3], ICD-9-CM[4], SNOMED CT[5] and NCIt[6], described in OWL[7]) are used to evaluate our approach, which is in line with the Gruber definition [Gruber, 1993] of ontology. Thus, we will use the term ontology to refer to them. However, we highlight that the logical part of these models was not used for inferring new knowledge for reasoning issues [Schulz et al., 2007]. Furthermore, the models were used only to define concepts (and their attributes) and regions related to these concepts following the formality expressed in the UMLS metathesaurus[8] *((Concept A rdfs:subClassOf Concept B)* ; *(Concept A skos:prefLabel 'term 1'))*, etc.

- **Design & Formalization**: The main objective of this phase is to develop a model that describes semantic annotations and their evolution, as well as formal methods to keep them up-to-date automatically over time. We used the outputs from the previous analyses and the related works to build models and workflows that allow us to maintain the impacted annotations.

[3] https://meshb.nlm.nih.gov/search

[4] https://www.cdc.gov/nchs/icd/icd9cm.htm

[5] https://www.snomed.org/snomed-ct/

[6] https://ncit.nci.nih.gov/ncitbrowser/

[7] https://www.w3.org/TR/owl-ref/

[8] https://www.ncbi.nlm.nih.gov/books/NBK9685/

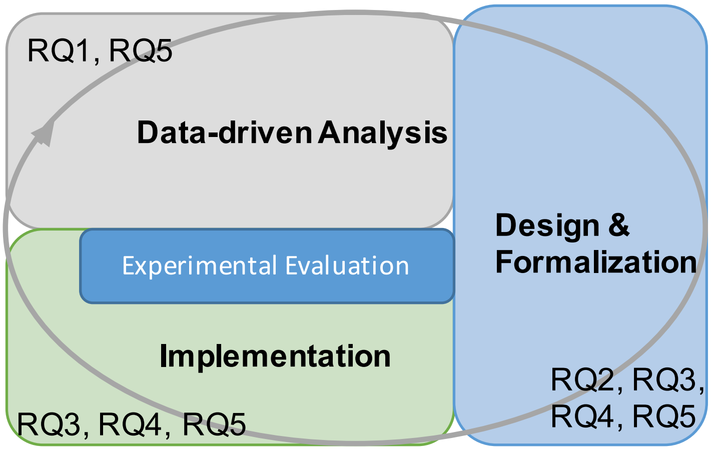


Figure 1.3: Schema of the methodology

- **Implementation**: Using Semantic Web technologies, we implemented the formalized models and workflows. The outcome is a prototype capable of: i) computing the evolution of annotations and ii) evaluating the performance of our algorithms.

1.3 Contributions

Each of the specific research questions allowed us to make significant contributions to the state-of-the art in the evolution of annotations. These contributions are briefly outlined below and are described in detailed in the upcoming chapters.

- Chapter 2: **Study of factors influencing the evolution of annotations and available models**. We addressed research questions RQ1 and RQ2.

  Contribution: Through quantitative results we highlighted the correlation between changes in the ontologies and changes in the annotations. Furthermore, we included features in the existing annotation formalism to support (semi-)automatic annotation maintenance mechanisms.

- Chapter 3: **Adapting semantic annotations**. Using the insights and results of our previous analysis on the evolution and adaptation of annotations over a period of time, we started to investigate RQ3.

  Contribution: We defined a method to automatically keep semantic annotations up-to-date. It is a four-level approach combining different methods to drive the annotation adaptation process.

- Chapter 4: **Semantic similarity measures to adapt semantic annotations**. It addresses a boundary question that emerged after several cycles in our methodology, and that is an important component in our architecture and in other methods that deal with semantic annotations.

  Contribution: We proposed an original approach combining lexical and ontology-based semantic similarity measures to improve the relatedness of clinical terms. We utilized this approach to improve our method for maintaining semantic annotations by the adding a new rule. Furthermore, we restructured our method to maintain the annotations through multiple versions of ontologies.

- Chapter 5: **Keeping the annotations searchable when the document and annotations cannot be changed directly**. This addresses RQ4, which focuses on the second variant case in the evolution of annotations, see Figure 1.2.

  Contribution: We built a knowledge base capable of: i) keeping the impacted annotations searchable without changing them, ii) using this knowledge graph as an alternative for background knowledge when dealing with the modifiable impacted annotations from RQ3.

- Chapter 6: **Predicting ontology changes**. We approached RQ5, which aims to provide support for the annotation maintenance process.

  Contribution: We built a machine learning model to predict i) whether a concept of an ontology will evolve in its next release, and ii) the type of change that affects it. For annotators, this work can support their decision concerning the term or the ontology that will be used to annotate specific types of documents (e.g., a more stable ontology would be preferable).

Chapter 2

Annotation Models and Factors Influencing the Evolution of Annotations

Contents

In order to answer RQ1 (*what is the impact of KOS changes on semantic annotations*) and RQ2 (*what is a suitable model to address the annotation evolution problem*) introduced in chapter 1, we conducted a study to understand how annotations are represented and whether the current models were capable of supporting their evolution. We also investigate how KOS evolution impact existing semantic annotations in a quantitative and qualitative way. Thus, in this chapter, we describe the main concepts related to semantic annotation and its associated representation models in section 2.1.1. In Section 2.1.2 we investigate existing approaches surveying the impact of KOS changes on semantic annotations. In sections 2.2 and 2.3 we describe the various experiments and the results we obtained regarding the evolution of annotations due to KOS changes. Finally, in section 2.4 we propose new features that should be taken into account in current models to cope with the evolution of annotations.

2.1 Study of existing annotation models and factors influencing annotation evolution

2.1.1 Existing annotation models

Semantic annotations are defined in the literature in many ways. According to Oren et al. [Oren et al., 2006], the term annotation can denote the process of annotating as well as the result of this process. Moreover, they distinguish three families of annotations. *Informal annotations* that are not machine-readable, (e.g. a handwritten margin annotation in a book). *Formal annotations* that are machine-understandable but are not defined using ontological terms, (e.g. highlights in an HTML document). Last, and the kind of annotation we are referring to in this thesis, *ontological annotations* are machine-understandable and are composed of elements from an ontology (see Figure 2.1).

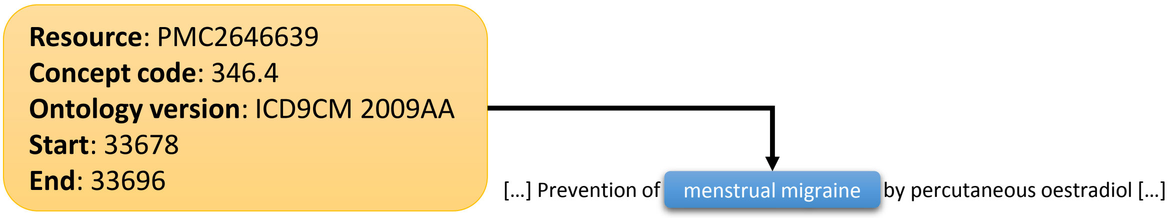

Resource: PMC2646639
Concept code: 346.4
Ontology version: ICD9CM 2009AA
Start: 33678
End: 33696
[...] Prevention of menstrual migraine by percutaneous oestradiol [...]

Figure 2.1: Example of annotation using the concept recognition process for a PubMed document. The term *menstrual migraine* is annotated with KOS concept 346.4, which belongs to ICD-9-CM, version 2009AA (UMLS)

To represent semantic annotations (SA) in the biomedical field, [Luong and Dieng-Kuntz, 2006] defined the following annotation model:

$$\text{SA} = (R_a, C_a, P_a, L, T_a) \tag{2.1}$$

Where:
R_a: set of resources, for instance, an RDF resource.
C_a: set of concept names defined in ontology ($C_a \subset R_a$)
P_a: set of properties, for instance, an rdf:type ($P_a \subset R_a$)
L: set of literal values, for example, “Fever”, “Malaria Fever”, etc.
T_a: set of triples (s,p,v) where $s \in R_a$, $p \in P_a$ and $v \in (R_a \cup L)$

In [Luong and Dieng-Kuntz, 2006], annotations are instances of an ontology represented in W3C Resource Description Framework (RDF) triples, s, p, o as following:

```
<ev:Person rdf:about="http://persinfo.com/John.Beeman">
        <ev:hasDisease rdf:resource='&ev;Malaria_Fever'>
    </ev:hasDisease>
</ev:Person>
```

[Groß et al., 2009] and [Hartung et al., 2008] defined an annotation mapping (AM), i.e. a set of annotations somehow related, considering versioning and quality information, aspects which were missing in [Luong and Dieng-Kuntz, 2006] model. In their work, an annotation mapping is defined as:

$$AM = (I_u, ON_v, Q, A) \tag{2.2}$$

Where:
$I_u = (I, t)$: is an instance source. It consists of a set of instances $I = \{i_j, ..., i_n\}$, e.g., molecular biological objects, such as genes or proteins, at timestamp t. Instances are described by an accession ID.
ON_v: is an ontology in version v that contains (C, R, t), it means a set of concepts $C = \{c_1, ..., c_n\}$ and relationships $R = \{r_1, ..., r_m\}$ released at time t.
Q: is a set of quality indicators (ratings) of annotations. The quality indicators may be numerical values or come from predefined quality taxonomies, e.g., the evidence codes for provenance

information or stability indicators.
A: is a set of annotations. A single annotation $a \in A$ is denoted by $a = (i, c, \{q\})$, i.e. an instance item $i \in I_u$ is annotated with an ontology concept $c \in ON_v$ and a set of quality indicators (ratings) $\{q\} \in Q$

Recently, the W3C[9] has published a new candidate recommendation for expressing and sharing annotations between systems, see Figure 2.2. It expresses the relationship between two or more resources by means of an RDF graph, using three major entities: i) Annotation, ii) Body and iii) Target.

The Body and Target, may be of any media type, e.g. Text, Audio, Video. Both are connected to the annotation through the *oa:hasBody* and *oa:hasTarget* relationships, which can be duplicated to create multiple body and/or target resources at the same time for an annotation. Furthermore, the relationship between these two entities may vary according to the intention of the annotation.

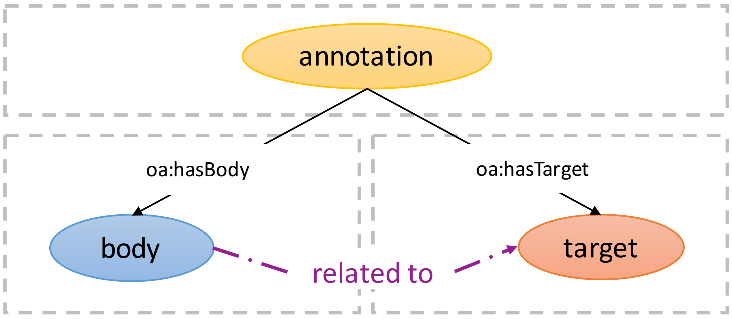


Figure 2.2: W3C web annotation data model

This model is the foundation of a more general framework for sharing and reusing annotated information across different hardware and software platforms, see Figure 2.3. Although, it specifies many features, we only describe those used to analyse the evolution of annotations in this thesis: *concept*, *target*, *motivation* and *agents*.

In Figure 2.3, the feature *target* allows the nature of the documents to be specified: text, video or image. In order to do this, it has to be associated with a source by using the relationship *oa:hasSource*, which informs where the desired media is located. In Figure 2.3, the resource used, *PMC2630914*, is a paper from PubMed. It is also necessary to define the corresponding part of media to be annotated. To do this, one uses *selectors*, whose main functionality is to describe the part of interest of a resource. In our example, the *selector1* defines the range of the annotated text segment in the document. Further, this first selector is refined by the inclusion of a *selector2*, that has a prefix and a suffix, allowing it to describe the information that comes before and after the annotation.

The second feature, *concept*, is not directly defined in the W3C annotation data model. Therefore, we used the *Body* resource, since this allows the resources used for annotation to be indicated, as well as supporting any kind of media to represent it. In Figure 2.3, the descriptor *D00558* from MeSH (mentioned in this thesis as concept), is attached to the annotation's body through the relationship *skos:related*. It creates an associative link between the body and the target indicating that the two are inherently "related", but that one is not in any way more general than the other. If the concept/ descriptor used is more general or specific than the target, we can use *skos:broader* or *skos:narrower*, respectively.

The *motivation* feature specifies the reasons for the annotation being created or for the inclusion of the body or target. In Figure 2.3, it is defined by the relationship *oa:motivatedBy*, where we used the motivation: *identifying*. This motivation assigns an identity to the target or

[9] `http://www.w3.org/TR/annotation-model/`

identifies what is being depicted or described in the target. Our example depicts the PubMed document *PMC2630914*.

Finally, the *agents* are indicated by the relationships *oa:annotatedBy* and *oa:serializedBy*, respectively. Both allow to define who produced the annotation, e.g., "A. Person" and the software utilized to serialize it, e.g., BioPortal. These features are useful for tracing the provenance of the annotation, e.g., the parameters used during the creation of the annotation. These parameters play a key role, because according to Funk et al. [Funk et al., 2014], entity recognition systems may vary from ontology to ontology and do not perform equally on natural language texts. Furthermore, we can define the date it was serialized through the relationship *oa:serializedAt* and when it was verified by the domain specialist by using *oa:annotatedAt*.

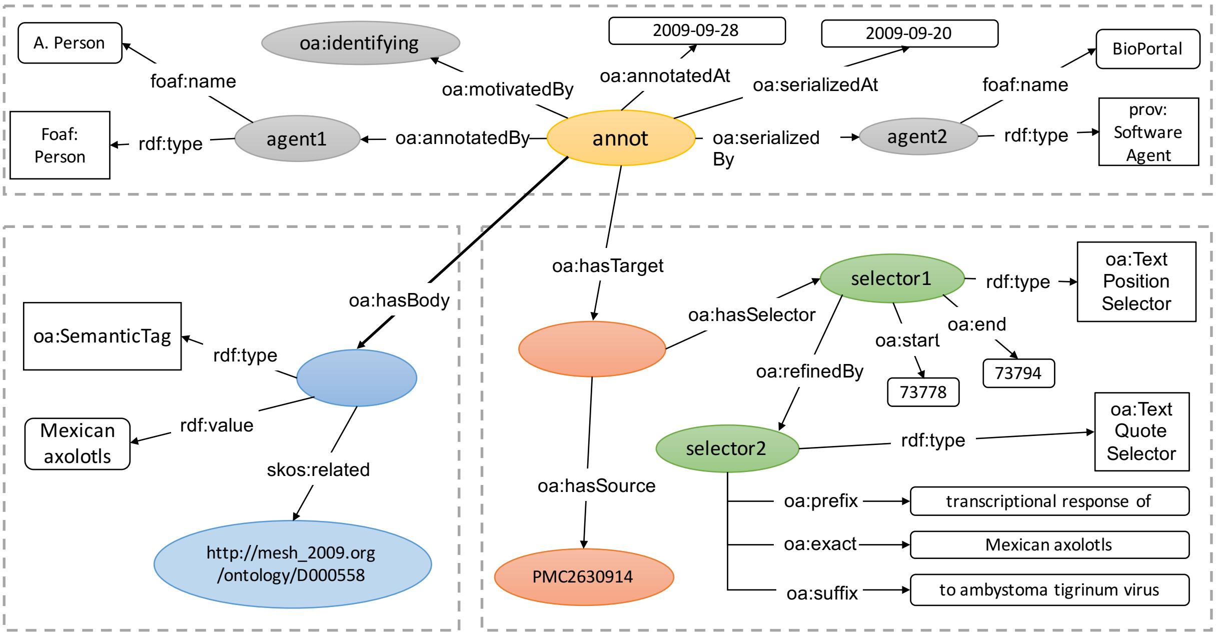


Figure 2.3: Instantiation of a W3C web annotation data model.

However, this model is still not sufficient to deal with evolution issues. The two main reasons are: i) the motivations available to describe annotations that have been changed. For example, *editing, moderating and replying*, are mostly related to manual editions and do not consider the impact of KOS changes; ii) there is no distinction between stable annotations and those that have evolved when we access *as:items*, i.e., a list of assorted annotations present in the target. Therefore, in section 2.3 we point out new features that must be considered when representing evolved annotations.

2.1.2 Related work in factors influencing the evolution of annotations

Recent analytical approaches to the evolution of the annotations focus on biological domain, in particular on GO-annotated documents. [Traverso-Ribon et al., 2015] developed the AnnEvol framework to compare two versions of a dataset (for instance, UnitProt-GOA and Swiss-Prot) and to verify those entities in the $dataset_{(i)}$ and $dataset_{(i+1)}$ (at two successive moments in time) that are similar and those that are different, using evolution criteria (e.g. obsolete, removed and added annotations).

[Groß et al., 2012] provided a method to evaluate to what degree changes of Gene Ontology (GO) and GO annotations (GOAs) may affect functional enrichment analyses, analysing two real-world experimental datasets, as well as 50 generated datasets. They proposed two types of stability measures to assess the impact of ontology and annotation changes. In contrast to AnnEvol, [Groß et al., 2012] deal with other types of change besides add and delete, such as, merge (merging of two or more categories into one category). They also verified strong structural changes, such as addR (insertion of a new relationship) and delR (deletion of an existing relationship). However, these changes do not significantly impact on GOAs, when compared to removal and merge changes that have a major affect in the annotations.

[Huntley et al., 2014] reported that GO annotations change for many reasons: i) the removal of *partOf* relationships, which are used to create inferred annotations; ii) taxon restriction, i.e., the redefinition of concepts and terms in order to remove ambiguity, such as the class GO:0051297 being replaced by GO:0007098. Furthermore, their work focuses on revision changes that were not considered in previous works, e.g. the detection of inappropriate terms used for manual annotation in past versions. These changes generally improve the accuracy of annotations and the underlying ontology. Finally, for those who are generating annotations, they recommend ensuring that they are using the most up-to-date version of GO and appropriate terms.

In summary, we concluded that the existing approaches evaluating the factors that influence the evolution of annotations only work with simple ontology changes (like addition and removal), and only study the evolution of GO ontology. The referenced works do not propose any methods to maintain the annotations. Therefore, it is necessary to further analyse the stability of KOS annotations based on different KOS like MeSH and verify possible features to be taken into account to properly maintain semantic annotations in biomedical and clinical use cases.

2.2 Experimental assessment of the impact of KOS evolution on semantic annotation

To bridge the gaps underlined in the previous section, we decided to conduct an empirical analysis regarding the evolution of KOS and annotations[10]. The lessons we learned through these experiments allowed us to come up with a new proposal to deal with semantic annotation evolution issues. The used material and the adopted assessment methodology are detailed in this section.

2.2.1 Material

As our objective is to analyse the evolution of semantic annotations, we worked on several versions of an annotated corpus of documents. Since there is no gold standard containing successive sets of annotated documents for multiple KOS, we had to build our own baseline (silver standard). To this end, we used two annotation tools (based on distinct annotation methods), two different medical standard KOS and their associated successive versions, an ontology *Diff* tool to be able to identify the evolution of the concepts used to produce the annotations and a collection of documents representative of the medical domain.

These documents were collected from the 2014 Clinical Decision Support Track (TREC 2014) campaign, which contains 733,138 biomedical articles about generic medical records. All documents from this database are open access documents from PubMed Central PMC. For our analyses we selected 5000 documents randomly.

The set of KOS is composed of several versions of medical KOS, represented in OWL format and used as a “reference ontology” for text annotation. In order to annotate the documents, we selected two KOS: International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM); and Medical Subject Headings (MeSH). We collected 13 official versions of each KOS released between 2004 and 2016 in UMLS (versions AA) and transformed them into OWL files[11].

[10] In this experiment annotations are similar to the lookups mentioned in [Lin et al., 2017]

[11] https://github.com/ncbo/umls2rdf

Regarding the annotation tools, the selection criteria were that they were open source, allowed the reference ontology to be selected, proposed APIs, had good documentation, and were extensively used for research and/or commercial purposes. We first selected General Architecture for Text Engineering (GATE) [Cunningham, 2002]. It provides support for Ontology-Aware NLP, allowing any ontology to be loaded as an RDF file, before using a gazetteer to obtain lookup annotations that have the class URIs and the text offset (offset is a pair {start, end} that indicates the distance, in terms of characters, from the beginning of the document. {start} indicates the position of the first character of the text while {end} indicates the position of the last character). The second tool selected is the NCBO Annotator. It is part of the NCBO Annotator framework and uses a dictionary built by extracting all concept labels and/or other associated attributes (e.g., synonyms) that syntactically identify concepts from the KOS [Whetzel et al., 2011]. Both annotators utilize different algorithms to produce annotations. GATE uses ontology-aware NLP and NCBO Annotator uses MGrep. Moreover, NCBO Annotator also allows other KOS to be used to annotate the terms, if a mapping exists between the concepts of both KOS. For instance, *melanoma* could also be annotated using UMLS Concept Unique Identifiers (CUIs) C0025202 (from NCI Thesaurus), or C0025202 (from SNOMED CT).

We used COnto-Diff [Hartung et al., 2013] in collaboration with our partners from the University of Leipzig, to determine an expressive and invertible *diff* evolution mapping between two versions of an ontology, although any other software capable of computing the diff between two ontologies can be used. COnto-Diff calculates basic change operations (insert/update/delete) from two KOS versions expressed in either OWL or OBO based on a predefined set of rules defining basic and complex transformations (e.g., concept merging, concept splitting, move of concept, etc).

2.2.2 Method

To identify and quantify the impact of changes affecting KOS concepts involved in annotations (as illustrated in Figure 1.1), we proposed the methodology depicted in Figure 2.4. The six steps of the methodology are the following:

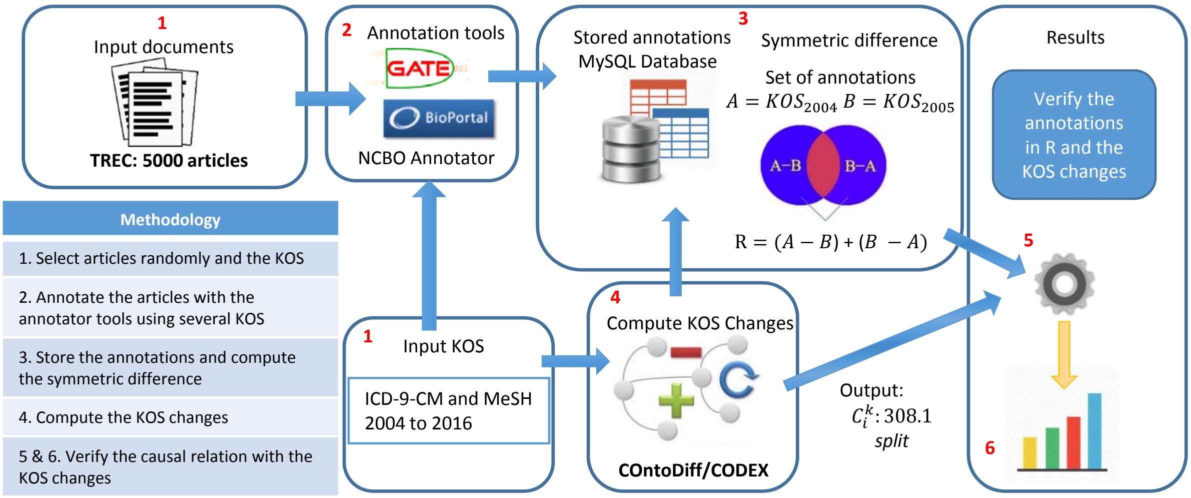

1
Input documents
TREC: 5000 articles
2
Annotation tools
GATE
BioPortal
NCBO Annotator
Stored annotations
MySQL Database
3
Symmetric difference
Set of annotations
$A = KOS_{2004}$ $B = KOS_{2005}$
A–B
B–A
$R = (A - B) + (B - A)$
Results
Verify the annotations in R and the KOS changes
5
6
Methodology
1. Select articles randomly and the KOS
2. Annotate the articles with the annotator tools using several KOS
3. Store the annotations and compute the symmetric difference
4. Compute the KOS changes
5 & 6. Verify the causal relation with the KOS changes
1
Input KOS
ICD-9-CM and MeSH 2004 to 2016
4
Compute KOS Changes
COntoDiff/CODEX
Output:
C_i^k: 308.1
split

Figure 2.4: The experimental protocol. The numbers in red correspond to the six steps explained in the text

1. We randomly selected 5000 documents from the TREC corpus and collected the 13 successive KOS versions of ICD-9-CM and MeSH (from 2004 to 2016).

2. We used GATE and NCBO Annotator to annotate these documents. We configured GATE and NCBO Annotator to use one specific KOS version and repeated the annotation process for each version. We filtered the annotations produced by both annotators according to [Doğan et al., 2014] (e.g., keep the longest match concept for an annotation).

3. We regrouped all annotations in one database. We then computed the symmetric difference $A_{m,n} \Delta A_{m,n+1}$ between the two annotation sets ($A_{m,n}$ and $A_{m,n+1}$) generated from a document R_m using two successive KOS versions (K_n and K_{n+1}) as the following:

$$\begin{aligned} &A_{m,n} \Delta A_{m,n+1} = \\ &\{a \mid a \in A_{m,n} \land a \notin A_{m,n+1}\} \cup \{a \mid a \in A_{m,n+1} \land a \notin A_{m,n}\} \end{aligned} \tag{2.3}$$

   a is an annotation that can be described as $\{i, m, n, \mathit{Offset}, c\}$ where i is an instance identifier, m is the document, n is the KOS used, *Offset* is the text position and c is the KOS concept. The symmetric difference allows annotations that have been removed, added and modified to be identified.

4. To identify KOS changes, each pair of two successive KOS versions was used as the input for COnto-Diff to compute the KOS difference. The difference was stored into another MySQL database and was reused to explain the changes.

5. We compared the 13 annotation sets of each document by pairs [2004-2005, 2005-2006, ...] to identify what had changed in the annotations and to find correlations with the KOS changes identified by COnto-Diff. An annotation a' is considered as an evolution of a if the *Offset* or/and the underlined concept c used in the annotation a are different from those of a' and there is an overlap of both *Offset*.

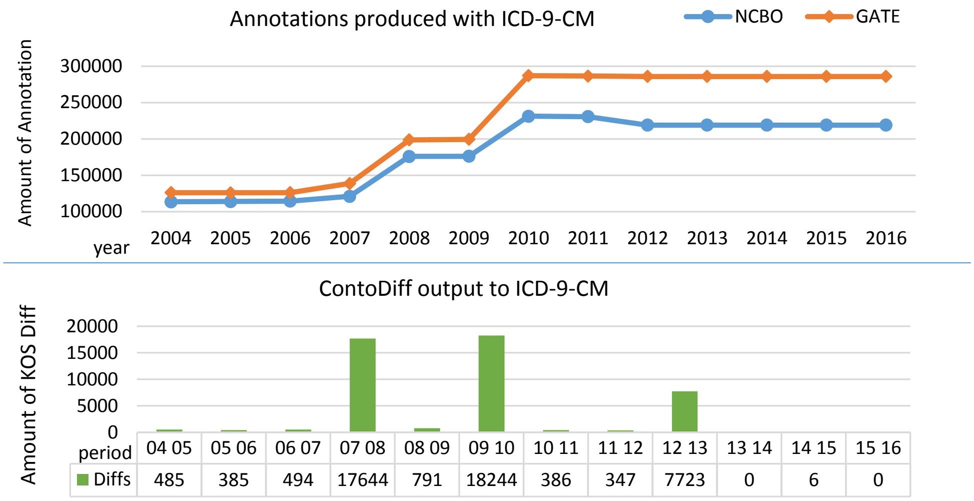
6. Finally, we analysed the generated subset of annotations/KOS changes in order to understand the impact of KOS changes on the annotations.

2.3 Results

The methodology described in the previous section has allowed us to produce more than 66 million annotations. The amount of annotations varies according to the annotation tools used (GATE or NCBO Annotator) as depicted in Figures 2.5 and 2.6. The difference between the two sets of annotations results from the method used to annotate the documents (they do not use only exact matches). A general observation can be made based on Figure 2.5 and 2.6.

We observe a huge increase in the amount of annotations produced in the periods 2007/2008 and 2009/2010 using ICD-9-CM (Figure 2.5). This increase is accompanied by the changes that occurred in the KOS during these periods according to COnto-Diff output. On the other hand, the amount of annotations in the period 2012-2013 did not increase even though there were many KOS changes. In the UMLS database, we observed an average of 8,746 words/label during this period and thus the annotators are not able to produce annotations for these changed labels. Hence, we can conclude that the change in the number of annotations does not necessarily correspond to the amount of KOS changes. In the chapter 3, we analysed what kinds of KOS changes trigger what types of annotation changes since not all kinds of changes in the KOS have the same impact on the annotations (e.g., some KOS changes do not change the annotations).

In order to verify whether a change in the annotations is triggered by the evolution of the KOS concepts or a gap in the annotator, we followed step 3 in Section 2.2.2. The first (quite evident) observation is that 100% of the annotation changes are caused by a KOS change even when the annotation methods do not only produce exact matches. This simple hypothesis had not



period	04 05	05 06	06 07	07 08	08 09	09 10	10 11	11 12	12 13	13 14	14 15	15 16
🟩 Diffs	485	385	494	17644	791	18244	386	347	7723	0	6	0

Figure 2.5: Amount of annotations and KOS changes (green) produced with 13 versions of ICD-9-CM. The annotations from NCBO Annotator are represented by blue circles and those from GATE by orange diamonds. The y-axis represents the amount of annotations/changes and the x-axis the KOS versions over time.

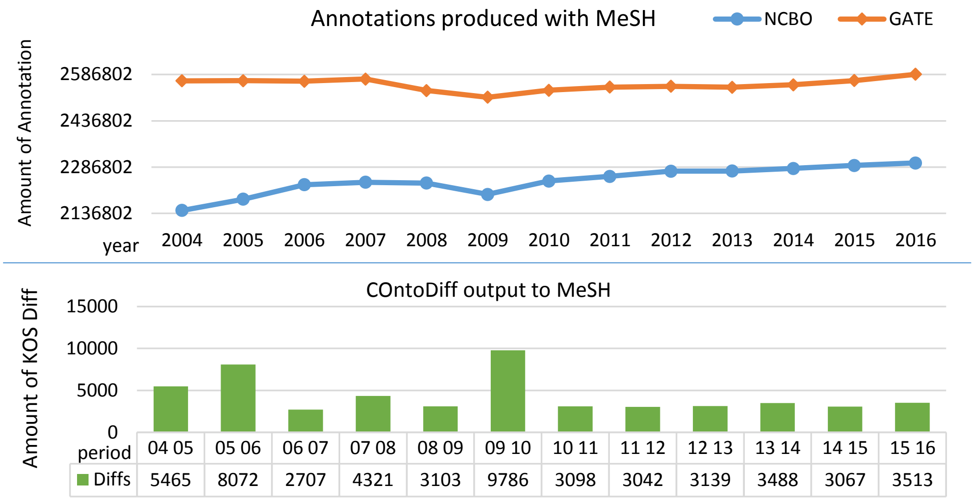
been demonstrated before in the literature. We continued our analyses regarding the evolution of annotations by refining the previous sets of symmetric difference (see step 5 in Section 2.2.2). If more than one concept candidate existed to annotate a text, we used the following selection criteria: (1) the most recent concept and the one with the largest offset, as proposed by [Doğan et al., 2014]. For instance, a text with the words *chronic kidney disease* can be annotated as *kidney disease* or *chronic kidney disease*; we selected only the latter concept. This decision can generate changes in the annotation from one KOS version to another (change operations). One of these changes is a shift of the offsets before and after the evolution while part of these offsets overlaps. For instance, in 2007, we had the annotation *"personality disorders"*. After a KOS change in 2008, the new annotation became *"schizoid personality"* (of which *"personality"* is overlapped with the previous offset). For such cases, we compute a (2) *chgOffset* operation. We formally define these conditions in Eq. (2.4):

$$Evolution(a_i, a_{i+1}) \rightarrow \begin{cases} recentCp(a_i, a_{i+1}) \land bigOffset(a_i, a_{i+1}), \text{ if } 1 \\ chgOffset(a_i, a_{i+1}) \qquad\qquad\qquad\qquad\quad , \text{ if } 2 \end{cases} \tag{2.4}$$

As a result, we observed that the new KOS versions do not necessarily produce more annotations despite the increasing size of the KOS over time [Da Silveira et al., 2015] (cf. Figures 2.7 and 2.8). Analysing the amount of annotations and the types of changes occurring in the KOS, we observed that some minor changes which do not affect the semantics of the concepts, still might impact the annotations. For instance, the concept 780.39 in ICD-9-CM, version 2007AA (Seizures) evolves to (Seizure) in ICD-9-CM, version 2008AA. However, neither annotator recognized that the concepts had the same meaning and therefore the associated annotations are different from one version to the next.

We also observed that there are some periods in the KOS evolution history which are more stable and this stability is also reflected in the evolution of the annotations (e.g., the two periods 2010/2011 and 2013/2014 in ICD-9-CM Figures 2.5 and 2.7).

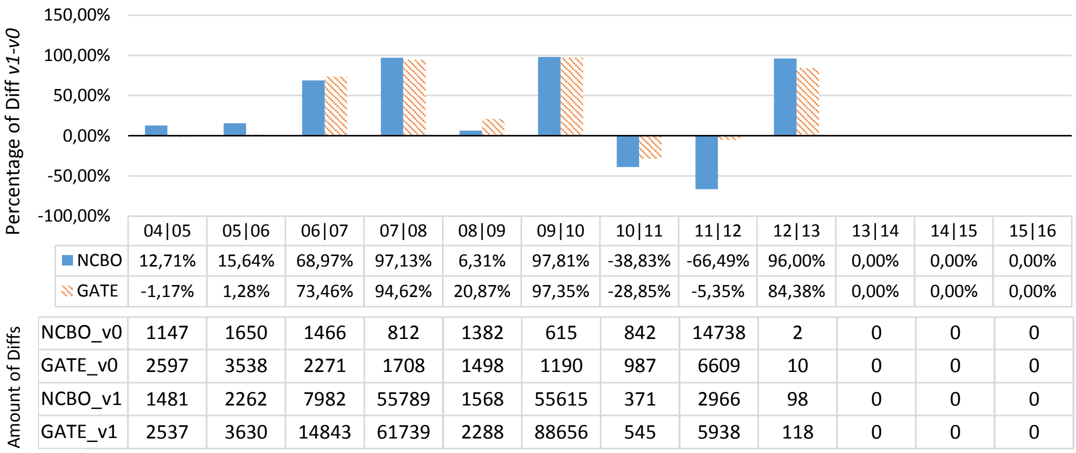
Changes in the KOS also have different impacts depending on the amount of annotations a concept is associated with. This is for instance the case for concept 084.4 of ICD-9-CM period 2007/2008 which is associated with 3143 annotations distributed in 162 documents in our corpus



period	04 05	05 06	06 07	07 08	08 09	09 10	10 11	11 12	12 13	13 14	14 15	15 16
🟩 Diffs	5465	8072	2707	4321	3103	9786	3098	3042	3139	3488	3067	3513

Figure 2.6: Amount of annotations and KOS changes (green) produced with 13 versions of MeSH. The annotations from NCBO Annotator are represented by blue circles and those from GATE by orange diamonds. The y-axis represents the amount of annotations/changes and the x-axis the KOS versions over time.

while concept V15.03 of ICD-9-CM period 2012/2013 is associated with only one annotation. If a single KOS change affects many annotations, the maintenance of the annotation may require a huge amount of time if carried out manually by domain experts. It will therefore be interesting, from an annotation perspective, to be able to identify concepts that will remain stable for annotation purposes in the next release since these annotations will not have to be maintained.

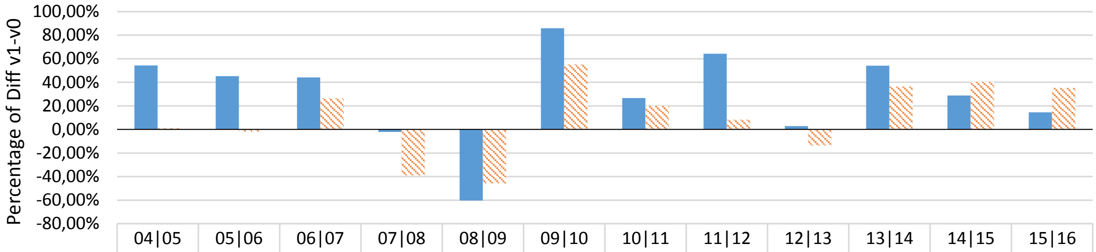


	04|05	05|06	06|07	07|08	08|09	09|10	10|11	11|12	12|13	13|14	14|15	15|16
■ NCBO	12,71%	15,64%	68,97%	97,13%	6,31%	97,81%	-38,83%	-66,49%	96,00%	0,00%	0,00%	0,00%
▨ GATE	-1,17%	1,28%	73,46%	94,62%	20,87%	97,35%	-28,85%	-5,35%	84,38%	0,00%	0,00%	0,00%

NCBO_v0	1147	1650	1466	812	1382	615	842	14738	2	0	0	0
GATE_v0	2597	3538	2271	1708	1498	1190	987	6609	10	0	0	0
NCBO_v1	1481	2262	7982	55789	1568	55615	371	2966	98	0	0	0
GATE_v1	2537	3630	14843	61739	2288	88656	545	5938	118	0	0	0

Figure 2.7: Differences in two successive annotation sets produced with ICD-9-CM. The blue (solid) colour represents the annotations that belong to NCBO Annotator, and the orange (hashed) colour to GATE.

We then analysed how these annotations evolve. In Table 2.1, we present five use cases showing how the annotations evolve over time and their relation with the evolution of KOS. The second column is related to the year in which each annotation was made, while the third column is associated with annotated text and the fourth column the KOS concept used. Finally, the fifth and sixth columns present the change behaviour of the annotation and the KOS concept, respectively.



	04|05	05|06	06|07	07|08	08|09	09|10	10|11	11|12	12|13	13|14	14|15	15|16
■ NCBO	54,29%	45,17%	44,26%	-2,13%	-60,30%	85,89%	26,57%	64,32%	2,76%	54,15%	28,91%	14,57%
▨ GATE	1,22%	-1,92%	26,43%	-38,79%	-45,86%	55,18%	19,86%	8,25%	-13,51%	36,30%	40,18%	35,41%

Amount of Diffs

NCBO_v0	15213	28464	5028	63378	48553	3566	20502	4671	7221	3519	11900	24290
GATE_v0	19831	49743	9780	66547	34012	9087	20957	14706	12927	7040	10198	18538
NCBO_v1	51350	75356	13014	60737	12026	46969	35341	21511	7631	11830	21580	32574
GATE_v1	20321	47872	16806	29348	12625	31461	31346	17350	9849	15065	23900	38864

Figure 2.8: Differences in two successive annotation sets produced with MeSH. The blue (solid) colour represents the annotations that belong to NCBO Annotator, and the orange (hashed) colour to GATE.

In the first use case (in 2008), the annotated text *hepatitis* is associated with concept 573.3, which did not change between 2008 and 2009 (i.e. a stable concept). In 2009, another concept (571.42) was also used to annotate the term *hepatitis*. Our selection criteria define that we will select the concept with the longest title (*autoimmune hepatitis*). We also observed that this concept (571.42) changed in 2009 (a split was detected).

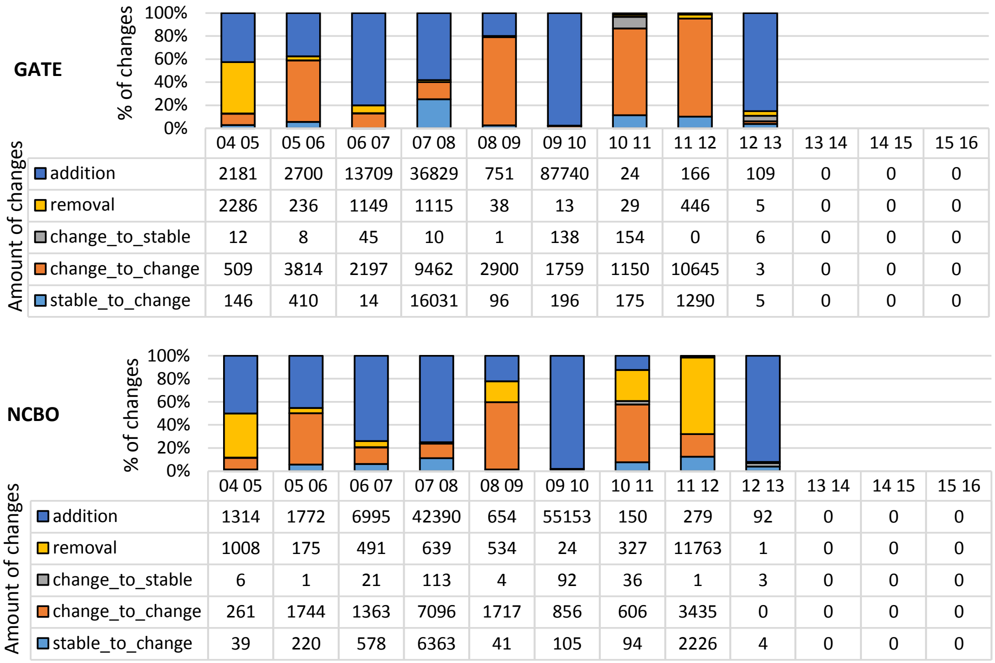
The second use case illustrates a situation where both concepts changed (i.e. 625.4 had an attribute deleted, and 346.4 is a new concept).

The third use case presents the inverse situation of use case 1, i.e., an annotation evolved from a change concept to a stable concept. In an in-depth analysis, this case was mainly observed when more general concepts were used to annotate the text. This behaviour occurs when the annotator is not able to determine whether a change in the concept has modified its meaning or not.

Use case	KOS version	Annotation		Annotation Behavior	KOS Behavior
		Annotated text	Concept		
1	2008 2009	hepatitis autoimmune hepatitis	573.3 571.42	change	stable concept split
2	2008 2009	menstrual migraine menstrual migraine	625.4 346.4	change	delAtt addC
3	2009 2010	acute renal failure renal failure	584.9 586	change	ChgAttValue stable concept
4	2008	abdominal tomography	88.02	addition	AddA
5	2004	bulimia	307.51	removal	ChgAttValue

Table 2.1: Use cases for annotation evolution. These different cases are referred in this section as: case 1: *stable_to_change*; case 2: *change_to_change*; case 3: *change_to_stable*; case 4: *addition*; case 5: *removal*.

The last two use cases describe the addition or removal of annotations. Regarding the removal of annotations, we also determined that there were some cases where the concept remained with the same meaning, but the annotator missed this knowledge and the annotation was removed from the document.



GATE

	04 05	05 06	06 07	07 08	08 09	09 10	10 11	11 12	12 13	13 14	14 15	15 16
🟦 addition	2181	2700	13709	36829	751	87740	24	166	109	0	0	0
🟨 removal	2286	236	1149	1115	38	13	29	446	5	0	0	0
⬛ change_to_stable	12	8	45	10	1	138	154	0	6	0	0	0
🟧 change_to_change	509	3814	2197	9462	2900	1759	1150	10645	3	0	0	0
🟦 stable_to_change	146	410	14	16031	96	196	175	1290	5	0	0	0

NCBO

	04 05	05 06	06 07	07 08	08 09	09 10	10 11	11 12	12 13	13 14	14 15	15 16
🟦 addition	1314	1772	6995	42390	654	55153	150	279	92	0	0	0
🟨 removal	1008	175	491	639	534	24	327	11763	1	0	0	0
⬛ change_to_stable	6	1	21	113	4	92	36	1	3	0	0	0
🟧 change_to_change	261	1744	1363	7096	1717	856	606	3435	0	0	0	0
🟦 stable_to_change	39	220	578	6363	41	105	94	2226	4	0	0	0

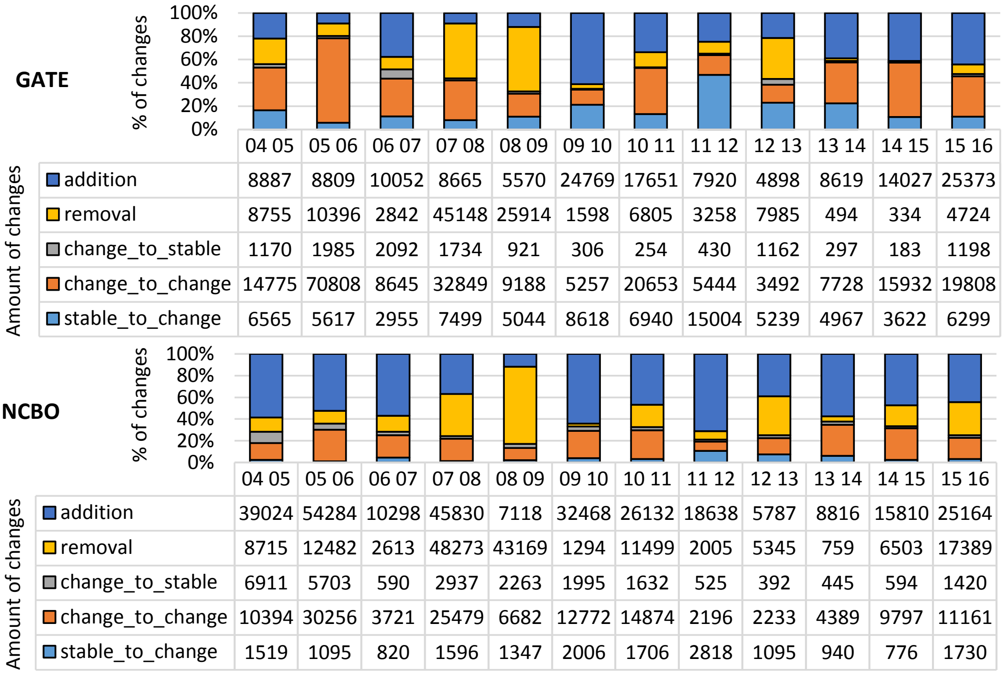
Figure 2.9: Distribution of changes of ICD-9-CM annotations. The y-axis represents the percentage of changes, the x-axis the KOS versions, and the amount of observed changes for each period is described below. The cases listed follow 2.1

Figures 2.9 and 2.10 show how often these use cases are observed in the corpus annotated with ICD-9-CM and MeSH using GATE and NCBO Annotator, respectively. In general, we observe that changes in ICD-9-CM have less impact on the annotations than those in MeSH. The low expressiveness of ICD-9-CM could be a reason for this as the annotators tend to apply exact match techniques for these kinds of KOS. Semantic-based techniques are used more for KOS with high expressiveness. These differences are better observed by comparing Figures 2.9 and 2.10 to see how the annotation technique influences the final annotation results regarding the expressiveness of the KOS. Use cases 2 and 5 (*change_to_change* and *removal*, respectively) are more frequent in the MeSH-based annotations. Thus, annotations based on ICD-9-CM evolve quite similarly for GATE and NCBO Annotator, while the annotations based on MeSH evolve differently, depending on the annotator used.

Taking into account the annotator techniques only, we observed that GATE also tended to preserve existing annotations while the rates of new annotations over deleted ones are quite similar for both annotators. More precisely, the rates of use cases 1 and 2 over the deleted ones (GATE has more than double that of NCBO) explain the results presented in Figure 2.5 (number of annotations increases faster for GATE).

2.4 A model supporting annotation evolution

The results presented in the previous section allow us to state that the evolution of the KOS has a direct impact on the definition of semantic annotations. However, we also showed that the modification of KOS concepts has different impacts depending on the technique that is implemented to generate the annotations. Furthermore, the evolution of KOS does not necessarily produce more information (see Figures 6 and 7). Actually, we have observed that KOS are becoming more and more precise over time, which indicates the addition of new specific concepts



GATE

	04 05	05 06	06 07	07 08	08 09	09 10	10 11	11 12	12 13	13 14	14 15	15 16
■ addition	8887	8809	10052	8665	5570	24769	17651	7920	4898	8619	14027	25373
■ removal	8755	10396	2842	45148	25914	1598	6805	3258	7985	494	334	4724
■ change_to_stable	1170	1985	2092	1734	921	306	254	430	1162	297	183	1198
■ change_to_change	14775	70808	8645	32849	9188	5257	20653	5444	3492	7728	15932	19808
■ stable_to_change	6565	5617	2955	7499	5044	8618	6940	15004	5239	4967	3622	6299

NCBO

	04 05	05 06	06 07	07 08	08 09	09 10	10 11	11 12	12 13	13 14	14 15	15 16
■ addition	39024	54284	10298	45830	7118	32468	26132	18638	5787	8816	15810	25164
■ removal	8715	12482	2613	48273	43169	1294	11499	2005	5345	759	6503	17389
■ change_to_stable	6911	5703	590	2937	2263	1995	1632	525	392	445	594	1420
■ change_to_change	10394	30256	3721	25479	6682	12772	14874	2196	2233	4389	9797	11161
■ stable_to_change	1519	1095	820	1596	1347	2006	1706	2818	1095	940	776	1730

Figure 2.10: istribution of changes of MeSH annotations. The y-axis represents the percentage of changes, the x-axis the KOS versions, and the amount of observed changes for each period is described below. The cases listed follow Table 2.1

whose labels are usually long (in terms of words) and therefore only very rarely appear in medical documents. Our study pointed out important features to take into account, at the semantic annotation model level, to facilitate the maintenance of annotation over time. These features can be used to extend the model proposed by [Groß et al., 2009] (see Sect. 2.1). In consequence, we define our evolution model as:

$$SAM = (I_u, ON_v, R_a, \mathit{Offset}, Q, H, A, SemRel, U_f)$$

Where :

- *Offset* is an element to describe the location of the element to be annotated in a given resource. From an evolution perspective, this is important for linking annotations from different versions and also for distinguishing annotations related to the same element but annotated differently.
- H is an element to describe which attribute of the concept (e.g., title, synonym, preferred terms, etc.) was used to produce the annotation. This element is extremely important since the annotation is usually defined based on the value of one concept attribute. If one of the attributes is changed in corresponding concept, but not the one used to annotate it, the annotation may not need to be modified.
- $SemRel$ is an element that describes the semantic relationship between the KOS concept and the annotated part of the resource. For instance, one sentence can be annotated as equivalent to a concept, more/less specific, partial match, etc. Thus, if a concept is removed, the annotated sentence can be linked to the super-class of the concept and the relation has to be changed to "less-specific".
- U_f is an element that points to the previous version of the annotation. This element is used to keep an evolution chain of annotations.

Our proposal, allowing annotation versions to be linked, can also be used to improve the W3C model by creating an additional motivation, known as "evolving", and a property, "evolved" that link the "annotation" to its past version, allowing a chain of annotations versions to be created, see Figure 2.11.

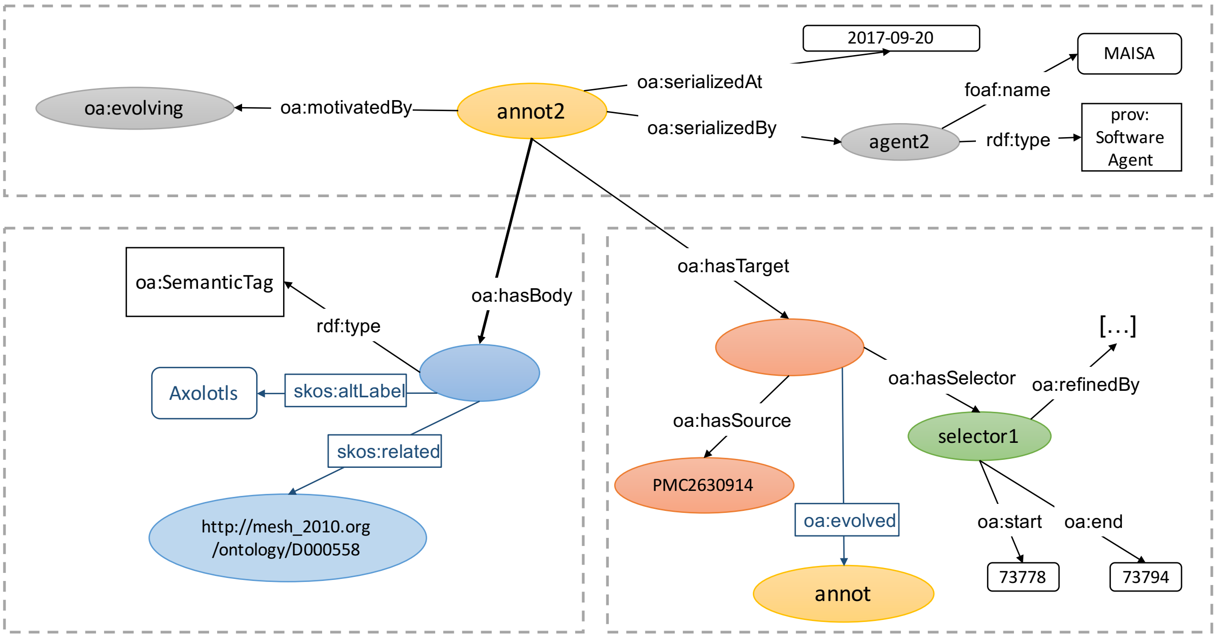


Figure 2.11: Proposed Evolution Model. The blue colours represent the modifications regarding the first model in Figure 2.3.

The improvements of our proposed model to the W3C model can be visualized in blue in Figure 2.11, when compared to Figure 2.3. The evolved annotation now contains the property "*oa:evolved*" in the *Target* that links it to the past version. Moreover, we modified the *Body* by adding the described feature *H*, represented by *skos:altLabel* indicating which concept attribute (e.g. title, synonym, preferred terms, etc.) was used to produce an annotation. This information can be used, for example, to determine whether the used attribute of the concept is impacted by the ontology evolution and whether this further triggers the corresponding annotation modification.

Finally, we included other *semantic type* relationships, *partial match*, and *other ontology region*, to indicate the relationship between a concept and a text segment. For instance, some changes in the above case *change_to_change* (see Figure 1.1), move a concept or its terms to another ontology region. Therefore, this relationship may help future evolution maintenance and curation by domain experts.

2.5 Conclusion

In this chapter, we approached RQ1 and RQ2 through an empirical analysis of the evolution of biomedical annotations and its relation to the KOS changes. For this, we used a set of documents annotated with GATE and NCBO Annotator using 13 different versions of two well-known biomedical KOS (ICD-9-CM and MeSH). We observed that there was a correlation between KOS and annotation changes. We then regrouped the annotation changes according to the type of information modified and the way it was modified. We obtained five different cases of changes (see Section 2.3) and verified how the annotations evolved during the KOS evolution. In a second step, we analysed different annotation models in order to verify whether they could represent (or whether we could infer from their elements) all the criteria required to classify the annotation changes. As a result of this step, we proposed an extended annotation model designed to support the evaluations and maintenance of annotations utilized in our maintenance method described in the next chapter.

Chapter 3

Direct maintenance of semantic annotations

Contents

In order to answer RQ3 (*How can we automatically maintain the validity of semantic annotations without re-annotating all document content when a KOS evolves?*) discussed in chapter 1, we designed an automatic approach for maintaining semantic annotations valid over time when the underlying KOS evolves. To do this, we designed a rule-based approach that considered the findings from chapter 2, which highlighted different aspects to take into account for the maintenance of semantic annotations.

Besides the set of rules, we used two other methods to improve the quality of our maintenance method, detailed in section 3.2. The first relies on the use of background knowledge (BK) [Pruski et al., 2016], while the second one exploits semantic change patterns (SCP) [Dos Reis et al., 2015a].

Thus, we divided this chapter up as follows. In section 3.1, we discuss the related work on semantic annotation maintenance processes and highlight possible improvements. Section 3.2 presents our method to overcome the limitations of existing approaches. In section 3.3 and 3.4, we describe the experiments and results, respectively. Finally, we discuss the results and highlight our contribution in section 3.5.

3.1 Existing approaches for maintaining semantic annotations valid over time

One possible solution to cope with the evolution of annotations is the re-annotation of documents [Tissaoui et al., 2011]. However, Funk et al. [Funk et al., 2014] point out that concept recognition systems vary from ontology to ontology and do not perform equally on natural language texts. Furthermore, the necessity of validating automatic generated annotations is a laborious and time-consuming task for domain experts [Doğan et al., 2014]. Therefore, it is vital to propose advanced methods and tools able to automatically maintain semantic annotations impacted by KOS evolution and/or changes in the annotated data or documents.

In the literature, three families of approaches dealing with annotation maintenance can be found. The first addresses the problem of automatic detection of inconsistent annotations [Eilbeck et al., 2009, Qin and Atluri, 2009, Köpke and Eder, 2011, Zavalina et al., 2015]. This is mainly done by the combined identification of concepts that have changed from one KOS version to the next and the set of annotations associated with them. However, mechanisms to support the correction of impacted annotations are not proposed.

The second family of approaches focuses on the automatic detection and manual correction of invalid annotations by using standalone applications [Maynard et al., 2007, Auer and Herre, 2007, Burger et al., 2010, Park et al., 2011]. These approaches only consider basic ontology changes, e.g., the deletion and addition of concepts in KOS. Nevertheless, more complex changes are also important and also need to be considered. Moreover, the requirement of human intervention to perform the maintenance is hardly applicable in the medical domain by virtue of the huge amount of annotations to be adapted.

Lastly, most advanced works implement an automatic detection and correction of the annotations [Luong and Dieng-Kuntz, 2006, Abgaz, 2013, Frost and Moore, 2014]. These approaches are based on different techniques, each of which is briefly described in this chapter.

[Luong and Dieng-Kuntz, 2006] developed the CoSWEM framework to investigate the evolution of annotations and to maintain them using a rule-based approach to detect and correct basic inconsistencies, such as deletion. This approach converts ontologies to RDF(S) files and detects annotations affected by the evolution of the ontologies, as well as potentially inconsistent annotations using CORESE. This work focuses on expressive and small-sized ontologies and can hardly be applied to large biomedical ones, because the implemented reasoning techniques require the power of description logics (not always used in biomedical controlled terminologies) to decide on the validity of the annotations.

[Abgaz, 2013] developed methods to facilitate the evolution of ontology-driven content management systems (OCMS). The evolution is done by analysing the impacts of change operations and selecting an optimal evolution strategy before the changes are permanently implemented. The proposed strategies, i) no-action, ii) cascade, iii) attach-to-parent, and iv) N-level cascading, were based on reasoning techniques and mostly deal with removal of concepts/meta-data as described below:

1. **No-action strategy**: This states that a given change operation, e.g. adding a concept, is implemented without adding consequential or corrective changes. For instance, even after the addition of a new class in the ontology, e.g. *avian influenza*, the documents referring to it and annotated with the class *influenza* will not be adapted.

2. **Cascade strategy**: This is the opposite of the no-action strategy. In this case, the changes will be propagated throughout the class and annotations. However, they only deal with cases of removal by propagating the deletion to all dependent entities.

3. **Attach-to-parent strategy**: This states that when a certain entity is deleted, its dependent entities are linked to the parent whenever it appears.

4. **N-level cascading**: This is a specific type of the cascade strategy. This strategy is applied to ontology classes that are found N distances from the target class. For example, if N is set to 2, the *N-level cascading* will apply the changes to up to two hierarchies.

[Frost and Moore, 2014] propose a novel algorithm for optimizing gene set annotations to best match the structure of specific empirical data sources. The proposed method uses entropy minimization over variable clusters (EMVC). It filters the annotations for each gene set to remove inconsistent annotations. The results show that EMVC can filter between 92% and 67% of the inconsistent annotations from MSigDB C4 v4.0 cancer modules using leukemia data and MSigDB C2 v1.0 using p53 data, respectively. This method is able to improve the annotations but does not produce good results for improving incomplete gene sets or identifying new gene sets. It is very sensitive to several algorithm parameters, specifically, the cluster method and it can be computationally expensive. Furthermore, the authors point out that EMVC only works in the gene set domain, meaning other domains cannot take advantage of this approach.

The literature review highlights that there is no annotation maintenance/adaptation framework able to cope with the specificity of the medical domain e.g., size of the KOS, number of annotations. Therefore, in this chapter we present a method to automatically maintain semantic annotations when the used KOS evolves. The method discussed in the upcoming sections is related to the *Direct maintenance of semantic annotations*, i.e., the first use case discussed in chapter 1. For this purpose, we proposed a set of rules based on the rigorous analysis of the evolution and adaptation of a set of annotations over a ten-year period of time, described in Chapter 2 [Cardoso et al., 2016].

3.2 Proposed approach for adapting semantic annotations

The proposed method aims to comprehensively address the requirements for annotation maintenance and to correct the shortcomings of previous approaches highlighted in section 3.1.

Figure 3.1 illustrates the maintenance process we propose. It is a multi-layer approach that we split according to inputs, processes and outputs. It allows the annotation maintenance to be optimized throughout the processes by considering more information at each step to maintain annotations that remain invalid after the correction that occurs in a previous step.

The different processes we have identified consist of: i) Automatically detecting inconsistent annotations caused by the evolution of the underlying KOS; ii) Using only the information gained from the evolution of the KOS to adapt impacted annotations; iii) Using information from the external KOS to maintain annotations that could not be maintained using local resources; iv) Using change patterns to finalize the maintenance and optimize the quality of the set of adapted annotations.

- **Identification of invalid annotations**: This consists of identifying invalid annotations by analysing the evolution of the associated KOS. To this end, it takes a set of annotations and two successive versions of the KOS used, namely K_n and K_{n+1}, as input. Concepts that have changed between K_n and K_{n+1} can be identified using an ontology *Diff* tool [Hartung et al., 2013]. Such tools also provide additional information specifying the type of changes that have affected these concepts. As is the case for ontology mapping adaptation [Groß et al., 2013], such information plays a key role in the maintenance task because it determines the type of correction to apply to the annotations at the next levels. For instance, the deletion of a concept attribute can lead to the deletion of annotations but the deletion of the same attribute value in the context of a concept split can lead to the migration of the annotation to the evolved version of the concept (i.e. the result of the split). The difficulty is therefore to consider not only basic ontological changes (i.e. addition/deletion of concepts) as is the case in existing approaches for annotation maintenance but complex changes (i.e. split/merge of concepts) to optimize both the maintenance process and the quality of the adapted annotations.

- **Annotation correction using ontology change rules**: This consists of using information derived from the set of annotations itself, as well as the data of the *Diff* between the two KOS versions K_n and K_{n+1}, to adapt the invalid annotations identified. At this level,

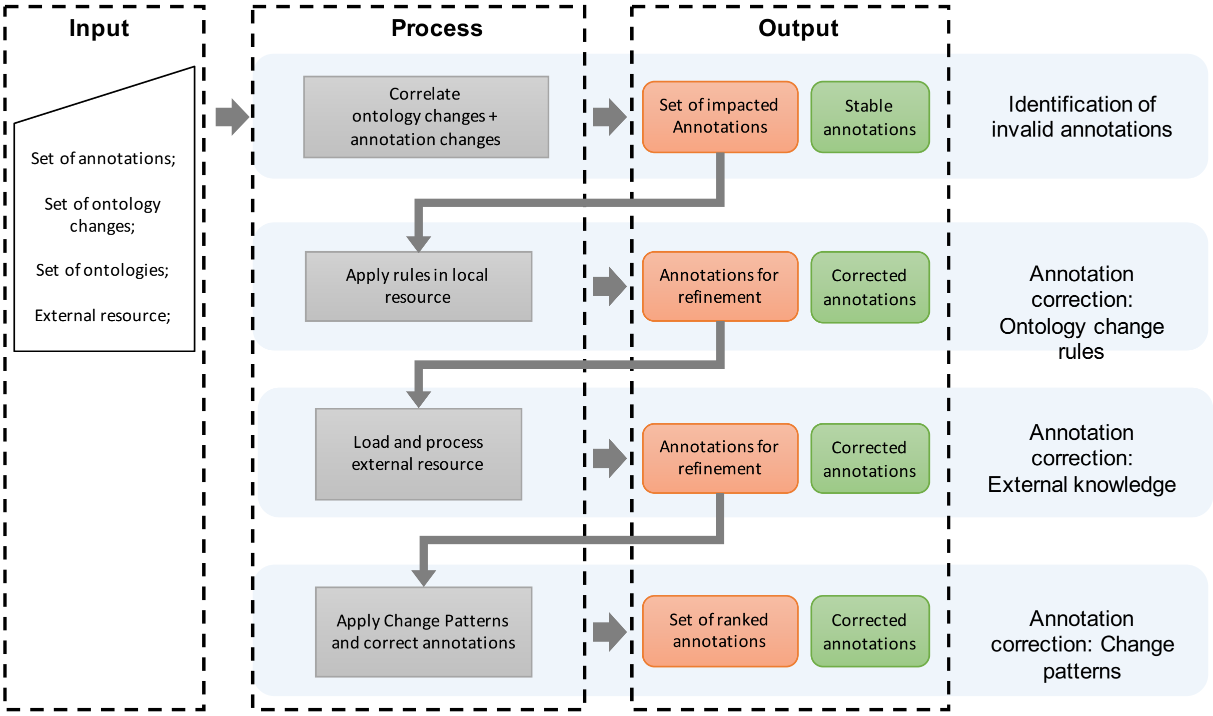

Input
Set of annotations;
Set of ontology changes;
Set of ontologies;
External resource;
Process
Correlate ontology changes + annotation changes
Apply rules in local resource
Load and process external resource
Apply Change Patterns and correct annotations
Output
Set of impacted Annotations
Stable annotations
Annotations for refinement
Corrected annotations
Annotations for refinement
Corrected annotations
Set of ranked annotations
Corrected annotations
Identification of invalid annotations
Annotation correction: Ontology change rules
Annotation correction: External knowledge
Annotation correction: Change patterns

Figure 3.1: The framework for supporting annotation maintenance. Source:[Cardoso et al., 2017b]

the correction of annotations can be specified in rules that combine the evolution context of the KOS and the status of the annotations, e.g., impacted or stable annotations. Under these conditions, the rules must specify the maintenance action to be performed.

Our maintenance process relies on the combined use of eight rules derived from our analysis of the evolution of annotations in chapter 2. Furthermore, we based our approach on the guidelines associated with semantic annotation [Doğan et al., 2014]. These guidelines aided us in defining the sequence and conditions for applying the rules.

The eight rules proposed are listed in Table 3.1. Each column represents one feature of the model described in chapter 2: the original concept code and the KOS version (CP_{v0}), the annotated text ($Annot_{v0}$), a prefix ($Prefix_{v0}$), and a suffix ($Suffix_{v0}$). We also added one column to show the changes observed ($Changed_{v1}$) and another to indicate the rule executed for the situation presented (one rule per line). The proposed rules are:

1. **MergeAnnot**: This rule will be applied when two parts of a document, annotated with different concepts, can be put together and annotated with only one (more specific) new concept. For instance, in 2004AA, the texts *"pregnancy"* and *"hypertension"* were annotated with the concept codes D011247 and D006973, respectively. In 2005AA, a new concept containing both terms was created and the annotation was evolved to concept D046110, see Table 3.1.
2. **IncreaseAnnot**: This rule increases the amount of information that can be annotated after the evolution of the underlying KOS. To do this, we compare the new label or attribute values of the candidate concept with the information surrounding the initial annotation (i.e., we take into account the prefix and suffix of the annotated text). Concretely, this action modifies the offset value in the annotation model and (if needed) the concept ID, e.g., D002403, *"cathepsin"* ↔ D056668, *"cathepsin l"*, see the second example in Table 3.1.

3. **ResurrectAnnot**: In some cases, one concept can be temporarily deleted from a KOS, leading to the deletion of the associated annotation. For instance, the annotation *chemiluminescence* in Table 3.1 was removed by a change in MeSH 2005AA. This rule allows the annotation to be re-activated when the concept is re-integrated to the KOS, e.g., concept D017083 in MeSH 2006AA. This rule was inspired by the findings of [Eilbeck et al., 2009], where they investigated the addition and deletion of gene annotations from release to release.

4. **PluralAnnot**: This rule verifies whether the change in the underlying concept or attribute value is due to a plural or singular variation ("agglutination" ↔ "agglutinations"). In this case, the change in the terminology does not imply a change in the impacted annotations since the semantics of the concept is not altered. Note that plurals are language-dependent rules and our evaluation only considers English KOS.

5. **ChangeConceptAnnot**: This rule changes the concept ID of the annotation due to the evolution of the concept. This situation arises when the label or the attribute value of the concept, used to create the annotation, is moved to another concept or used to create a new concept. For instance, concept D003704 changed to D057174 (referring to *"Semantic dementia"*) in MeSH 2009AA/2010AA.

6. **SplitAnnot**: This rule splits an existing annotation if the evolution of the underlying concept leads to the creation of two more precise annotations. For instance, the text *"diabetic foot ulcers"*, annotated in 2005AA with the MeSH code D017719, was split into two new annotations in 2006AA: D017719 ("*diabetic foot*") and D016523 ("*foot ulcers*"), see Table 3.1.

7. **PartialMatch**: This rule applies lexical and semantic algorithms to change the concept ID of an annotation. We further discuss how to implement this rule in Chapter 4, since it needs a deep investigation on semantic and lexical measures. Therefore, the experiments demonstrated in this chapter do not include this rule and we only listed it to facilitate future references regarding our framework.

8. **SuperClassAnnot**: This rule changes the concept ID to the *superClass* ID since no concept can be found to precisely maintain the annotation. It will also change the relation (i.e., "Equivalent" → "Is A") between the concept and the annotation. For instance, after checking whether any of the previous rules were executed with the annotation *"infective agents"*, the last example in Table 3.1 shows that, instead of deleting the annotation, we propose using the *superClass* to annotate the text. Thus, *"infective agents"* is a kind of *"other organism groupings"*. Note that this is only possible if the formalism used to annotate the text follows our proposed formalism in chapter 2.

The sequence in which the rules are executed is important to assure the quality of the annotations modified. Based on the propositions of the annotation guidelines [Doğan et al., 2014], we established the following sequence (without *PartialMatch*): *MergeAnnot, IncreaseAnnot, ResurrectAnnot, PluralAnnot, ChangeConceptAnnot, SplitAnnot, SuperClassAnnot.* First we ranked the rules that increase the information of an annotation (*i.e., MergeAnnot* and *IncreaseAnnot*), as suggested in the guideline *"Annotate the most specific concept that correctly describes the disease mention"*. The next rules (*ResurrectAnnot, PluralAnnot* and *ChangeConceptAnnot*) are mainly related to the structure of the KOS and text. We started with *ResurrectAnnot* because changing the concept ID (*ChangeConceptAnnot*) would increase the complexity in identifying the restoration of the concept. The *PluralAnnot* rule is an exception, because it does not affect the other rules and can be placed anywhere in the sequence. The *SplitAnnot* was placed close to the end of our process due to it only occurring in rare cases, as mentioned by [Doğan et al., 2014]. It respects the following recommendation: *"Annotate a disease mention using multiple concepts to logically describe the disease mention, using the "+" concatenator"*. Finally, the *SuperClassAnnot* was positioned at the end of our process as an alternative to the removal of the annotation.

The precision of the rules has some limitations. Thus, we decided to evaluate other potentially complementary methods in order to improve the quality of our outcomes. In this sense, we selected two other methods: background knowledge and semantic change patterns.

- **Annotation correction using external resource knowledge**: This consists of using information inferred from external knowledge sources to maintain the annotations that could not be corrected using local resources of the previous level. Actually, in many cases the drift of ontological concepts can be characterized only by considering the semantic relationships provided by other ontologies [Pruski et al., 2016]. Often, labels of concept are completely different from a syntactic point of view, before and after evolution. Therefore, considering local resources only does not allow their evolution to be characterized and, in turn, to be reused for annotation maintenance purposes. For example, the evolution of the label of concept D019684 in the MeSH from "Magnoliophyta" in 2009 to "Angiosperms" in 2010 requires an external knowledge source. Applying existing approaches to annotations associated with this concept would simply lead to the deletion of the annotations. But, the consideration of external resources (here mappings between SNOMED CT and MeSH, provided by Bioportal) show that these two terms are synonyms, therefore the annotation can be kept. Nevertheless, the nature of the external knowledge resources can vary. Depending on whether RDF datasets like BIO2RDF [Belleau et al., 2008] or expressive OWL ontologies contained in Bioportal [Noy et al., 2009] are considered, the inferred information can be of a different quality and can affect the quality of the maintenance process.

  The BK algorithm (see algorithm 1) presents an overview of the whole process. Figure 3.2 helps to illustrate how the algorithm works. The input of the algorithm is the concept ID (C_s), label (L_s), KOS target (KOS_t), and KOS source (KOS_s), e.g. (D019684, Magnoliophyta, MESH, {SNOMED CT, ICD-9-CM, NCIT}). After initializing the variables (lines 1 and 2), our method queries external sources using the impacted annotation label (line 3) and stores the resulting concepts (*Request*). For instance, concept 420928000, from SNOMED CT, is one candidate. Only concept candidates belonging to the source KOS KOS_s are kept (lines 4-5). Then, for each concept from *Request*, the mappings are collected (line 6). Only mappings to the target KOS (KOS_t) are kept (lines 7-10); these are the grey boxes in Figure 3.2.

  The next step of our algorithm (line 11) retrieves all stable ancestors of a source concept C_s within a specified period, e.g. (2009/2010). From all candidates that satisfy the previous conditions, we compute the similarity between the ancestors and the source concept (lines 12 to 13) using Jiang-Conrath (JC) [Jiang and Conrath, 1997]. Then, we take the most similar ancestor MSA (line 14). Finally we select the best candidate to maintain our annotation (lines 15-17). This is the mapping showing the highest similarity to the MSA.

- **Annotation correction using change patterns**: Information provided by the *Diff* and the use of external resources may not be sufficient to deal with invalid annotations. In this

Table 3.1: Examples of annotations computed by our rules. (CP_{v0}) concept code in specific year, ($Annot_{v0}$) annotated text, ($Prefix_v0$) prefix, ($Suffix_v0$) suffix, $Changed_{v1}$ result of applied rule.

CP_{v0}	$Annot_{v0}$	$Prefix_{v0}$	$Suffix_{v0}$	$Changed_{v1}$	Rule
D011247 in 2004AA	pregnancy	diabetes mellitus and	-induced hypertension	{D046110, pregnancy-induced hypertension}, 2005AA	MergeAnnot
D006973 in 2004AA	hypertension	pregnancy-induced	. Apgars were		
D002403 in 2009AA	cathepsin	responses [67]. a	l-like gene (ee049537) has	{D056668, cathepsin l}, 2010AA	IncreaseAnnot
D017083 in 2004AA	chemiluminescence	of western blot	were acquired	{D017083, chemiluminescence}, 2006AA	ResurrectAnnot
D000371 in 2009AA	agglutination	of antibodies. weakly reactive	required an adequate light	{D000371, agglutinations}, 2010AA	PluralAnnot
D003704 in 2009AA	Semantic dementia	frontotemporal dementia pnfa	?prion	{D057174, semantic dementia}, 2010AA	ChangeConceptAnnot
D017719 in 2005AA	diabetic foot ulcer	associated with	are recommended	{D017719, diabetic foot}{D016523, foot ulcers}, 2006AA	SplitAnnot
C50922 in 2009AA	infective agents	the most common	, the necrotic base of	{C14376, other organism groupings}, 2010AA	SuperClassAnnot

Algorithm 1: Similarity between mappings and ontology source in the Background Knowledge. MSA: the most similar ancestor.

```
Input: Concept source $C_s$; Label $L_s$ ; KOS source $KOS_s$; KOS Target $KOS_t$; KOS Changes $chgs$;
Output: Concept Target $C_t$
$MappingSet \leftarrow \emptyset$
$ValidMappings \leftarrow \emptyset$
$Request \leftarrow getConceptsFromBK(L_s)$
forall $cp \in Request$ do
    if $(cp \in KOS_s) == TRUE$ then
        $MappingSet \leftarrow getMappings(cp)$
forall $mapping \in MappingSet$ do
    $target \leftarrow getConceptTarget(mapping)$
    if $(target \in KOS_t) == TRUE$ then
        $ValidMappings \leftarrow target$
$Set\_Sup\_Classes \leftarrow getAllStableAncestor(C_s, KOS_{v0}, chgs)$
forall $obj \in Set\_Sup\_Classes$ do
    $calSemanticDistance(obj, C_s, KOS_{v0})$
$MSA \leftarrow getMostSimilarAncestor(Set\_Sup\_Classes)$
forall $obj \in ValidMapping$ do
    $calSemanticDistance(MSA, obj, KOS_{v1})$
$C_t \leftarrow getHighestSimilarity(ValidMappings)$
return $C_t$
```

case, the analysis of the morphosyntactic form of concept labels can reveal information to make decisions on annotation maintenance. This technique has already been explored in the context of ontology mapping adaptation [Dos Reis et al., 2015b, Dinh et al., 2014] but remains less relevant in terms of quality in the resulting maintenance decisions. Change Patterns are modifications observed in attribute values of a concept using linguistic-based features to identify the correlation between concepts over time. For instance, a *Partial Copy* between concepts is computed if and only if there is a partial overlap between words from an attribute present in the KOS version K_n and an attribute in the new KOS version K_{n+1} (i.e., the attribute a_0 becomes a_1).

For instance, the annotation "*Physiologic processes*", shown in Figure 3.3 and produced using MeSH in the 2008/2009 period was removed. This is due to a change in the attribute value in the definition of concept D010829 leading to "*Physiological Phenomena*". Assuming the following conditions, i) we do not have information inside the ontology to deal with this change, ii) the super class from concept D010829 is *Thing*, iii) external resources do not provide the necessary information to make decisions, the application of four change patterns (*total copy, total transfer, partial copy, partial transfer*) considering only the attributes in the same sub-ontology, e.g. the sub-classes from concept D010829, allows the concept associated with this annotation to be changed from D010829 to D055705.

- **Output**

  Our approach was designed to process the annotations according to different levels of granularity, but the outputs contain only two kinds of data.

  At the levels dealing with the correction of the annotations, the outputs are i) the corrected

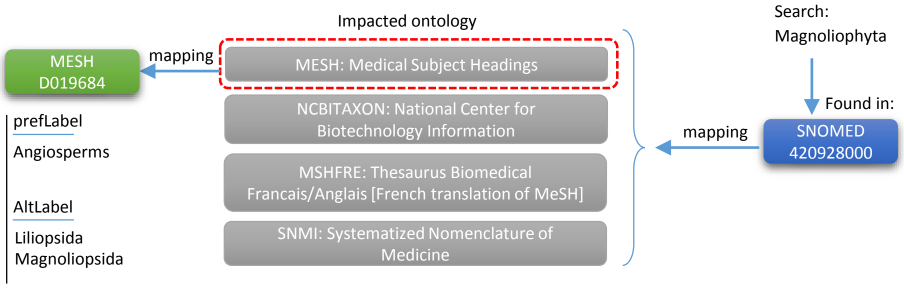

Impacted ontology
MESH D019684
mapping
MESH: Medical Subject Headings
NCBITAXON: National Center for Biotechnology Information
MSHFRE: Thesaurus Biomedical Francais/Anglais [French translation of MeSH]
SNMI: Systematized Nomenclature of Medicine
prefLabel
Angiosperms
AltLabel
Liliopsida
Magnoliopsida
Search: Magnoliophyta
Found in:
SNOMED 420928000
mapping

Figure 3.2: Use of BioPortal as Background Knowledge

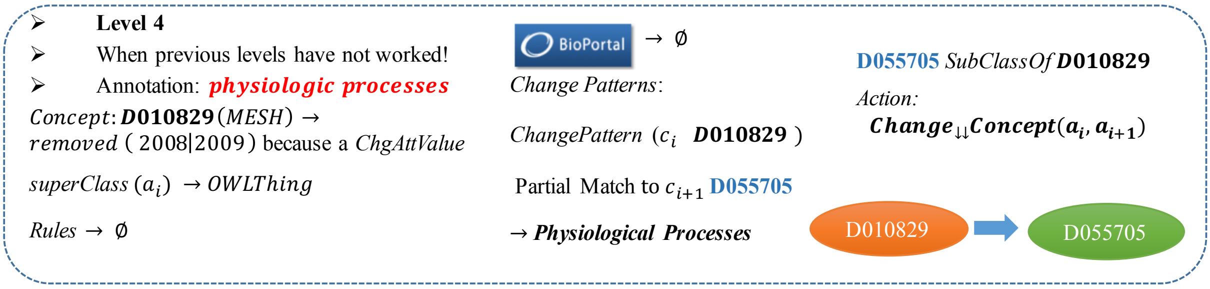

Level 4
When previous levels have not worked!
Annotation: physiologic processes
Concept: D010829 (MESH) → removed (2008|2009) because a ChgAttValue
superClass (a_i) → OWLThing
Rules → ∅
BioPortal → ∅
Change Patterns:
ChangePattern (c_i D010829)
Partial Match to c_{i+1} D055705
→ Physiological Processes
D055705 SubClassOf D010829
Action:
$\mathbf{Change}_{\downarrow\downarrow}\mathbf{Concept}(a_i, a_{i+1})$
D010829
D055705

Figure 3.3: Use of Change Pattern to maintain annotations

annotations, and ii) the set of annotations that need further investigation. Once corrected, the annotations are also enriched with evolution information making future modifications easier and enhancing their quality, see Figure 3.4.

In Figure 3.4, the *Body* was enriched with an element highlighting the type of evolution that occurred, i.e., *PluralAnnot*. This heuristic allows us to verify whether the ontology becomes more precise due to a split of the concept or due to a small review in the labels of a concept, as indicated by the *PluralAnnot* rule.

If invalid annotations remain, the definition of other levels exploiting different kinds of information for maintenance purposes need to be implemented. Our proposed approach flagged these annotations as unsolved, allowing future extensions that have to take into account the following aspects: i) the complexity of the evolution affecting KOS, ii) the nature of the annotation, and iii) the specificity of the kind of object, e.g., *images* or *videos*. The rules that are used at each level also need to be defined by considering the quality of the adapted annotations.

3.3 Experimental assessment

In this section, we introduce the method and material we used to evaluate our approach. The experiments we conducted consist of applying our approach to a set of annotated documents and comparing it to a corpus of reference representing an evolved version of the initial set of documents.

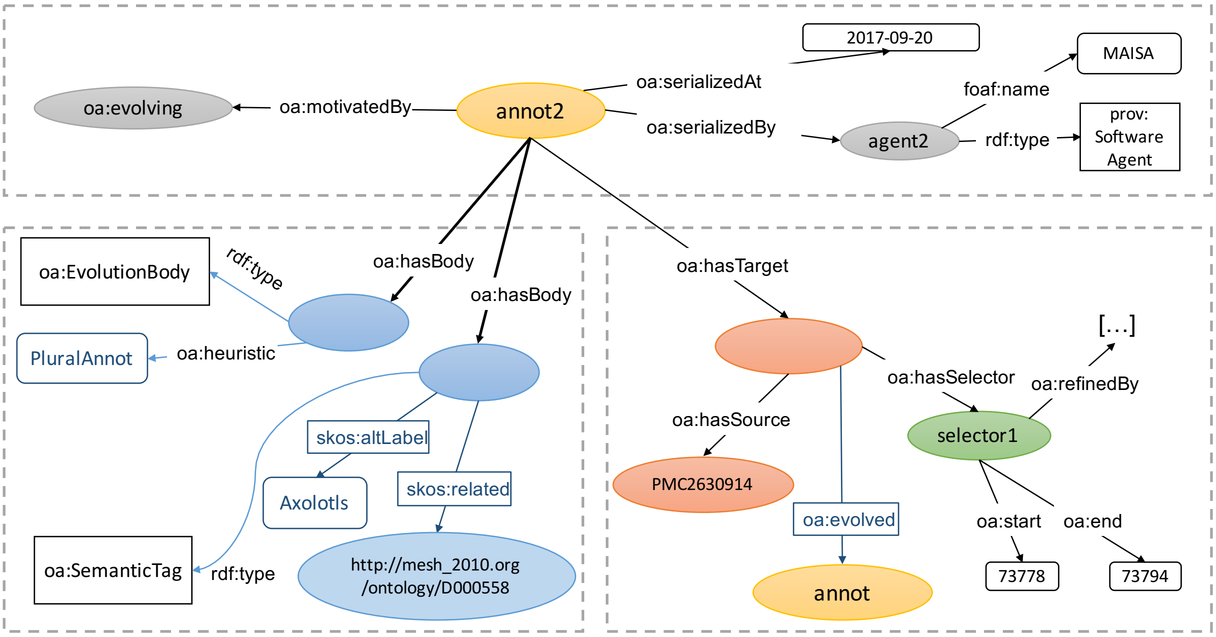


Figure 3.4: Extra information included in the body

Material

Our annotation maintenance process takes as its inputs:

- a set of outdated annotations,
- the old and new OWL versions of the KOS used to generate the annotations.

In our experiments, we used the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM), Medical Subject Headings (MeSH), National Cancer Institute Thesaurus (NCIt) and Systematized Nomenclature of Medicine - Clinical Terms (SNOMED CT). We used versions 2009AA and 2010AA downloaded from the UMLS and transformed into OWL files. We used COnto-Diff [Hartung et al., 2013] to identify the changes in the new version of the KOS since these changes are strongly correlated with the validity of annotations [Cardoso et al., 2016].

Silver Standard

Since there is no annotation baseline generated with sequential ontology versions exists, we had to build our own corpus of reference using the annotations produced in chapter 2 as a base resource. To do this, we randomly selected 500 annotations generated with the 2009AA version of the four KOS (around 125 annotations from each KOS) and asked three experts to manually validate/correct the evolution of the 500 selected annotations, according to the 2010AA version of the corresponding KOS. Each expert validated one-third of the annotations without discussing them with the others. The consolidated outcomes make up the silver standard, which can be downloaded from `https://git.list.lu/ELISA/AnnotationDataset`. We adopted the term “silver” to indicate that our reference is based on one viewpoint only.

We represent annotations following the W3C Web Annotation Data Model[12], discussed in chapter 2. To illustrate how an annotation is represented in our system, Table 3.2 contains the original annotation (from 2009) in the first line and the evolved annotation (from 2010) in the second line.

[12] https://github.com/anno4j/anno4j

KOS	Doc.	Concept	Annotated text	Start	End	Prefix	Suffix	KOS label
MeSH 2009AA	232	D019684	Magnolio phyta	4587	4600	during the evolution of	(angiosper ms) [5]. typical such	Magnolio phyta
MeSH 2010AA	232	D019684	Magnolio phyta	4587	4600	during the evolution of	(angiosper ms) [5]. typical such	Angios perm

Table 3.2: Example of an evolving annotation, extracted from our silver standard.

In our analysis, we used the following features: the name and version of the KOS, the reference of the resource document, the concept code, the annotated text followed by the start and end offset, i.e., the position in the document where this annotation is found, and the prefix and suffix, i.e., the information that comes before and after the annotation. The illustrative example in Table 3.2 shows one annotation produced with the MeSH 2009AA version using the PubMed document 232[13] and the concept D019684. The annotated text is "*Magnoliophyta*", and this can be found in the position [4587,4600]. We set up the system with four words as the prefix *"during the evolution of"* and a suffix *"(angiosperms) [5]. typical such"*. We can observe that the concept label used to annotate the text changed from 2009AA to 2010AA.

To measure the effectiveness of the proposed approach, we used classic well-known metrics from the literature, such as *precision*, *accuracy*, *recall*, *F1-score*, *ROC curve* to investigate/understand two characteristics of our method: i) the capacity of our framework to detect impacted annotations after changing a KOS concept; and ii) the ability to correctly adapt those impacted annotations into consistent ones. In this case, consistency means equivalency with the silver standard. We measured the efficiency of the `Rules` alone, the Background Knowledge (`BK`) alone, the Semantic Change Patterns (`SCP`), the Lexical Change Patterns (`LCP`) and the combinations of these techniques in order to determine whether they complement each other or not.

3.4 Results

When applying the four annotation maintenance techniques (`Rules`, `BK`, `SCP`, `LCP`) to our dataset, we can observe a significant difference in the results. For instance, as shown in Table 3.3, all four methods provide good precision, but there is a significant variation [0, 0.98] regarding the recall. In the first line of Table 3.3, we present the precision, recall and F1-Score resulting from applying the `Rules` method to four different subsets of our initial dataset (ICD-9-CM, MeSH, NCIt, and SNOMED CT). We also evaluate the consequence of combining the methods (2-by-2, and all together). For instance, the fifth line of the table presents the results of combining `Rules` and `BK` methods, while the tenth line shows the results of combining all four methods.

The goal of this first set of experiments was to evaluate whether the methods (or a combination of them) provide satisfactory quality (in terms of the F1-Score) to determine whether an impacted annotation will evolve or not. Note that we are not yet evaluating whether the annotation evolved correctly (this is part of the second evaluation step). A quick analysis shows that all methods can accurately identify some of the evolving annotations, but not all. From a practical point of view, if an error margin of 2% is acceptable for the domain, then KOS engineers can trust the method to identify the annotations that will change (i.e., the minimal observed precision was

[13] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2631504/

Method	ICD-9-CM			MeSH		
	P	R	F1	P	R	F1
`Rules`	1	0.982	0.991	0.991	0.941	0.966
`BK`	1	0.129	0.229	1	0.050	0.094
`SCP`	1	0.041	0.078	1	0.091	0.167
`LCP`	1	0.048	0.092	1	0.099	0.180
`Rules/BK`	1	0.982	0.991	0.991	0.941	0.966
`Rules/SCP`	1	0.982	0.991	0.991	0.941	0.966
`Rules/LCP`	1	0.982	0.991	0.991	0.941	0.966
`BK/SCP`	1	0.161	0.278	1	0.140	0.246
`BK/LCP`	1	0.161	0.278	1	0.149	0.259
CombineAll	1	0.982	0.991	0.991	0.941	0.966

Method	NCIT			SNOMED CT		
	P	R	F1	P	R	F1
`Rules`	0.980	0.942	0.961	0.929	0.812	0.867
`BK`	1	0.115	0.207	1	0.625	0.769
`SCP`	0	0	0	0	0	0
`LCP`	1	0.019	0.038	0	0	0
`Rules/BK`	0.980	0.942	0.961	0.929	0.812	0.867
`Rules/SCP`	0.980	0.942	0.961	0.929	0.812	0.867
`Rules/LCP`	0.980	0.942	0.961	0.929	0.812	0.867
`BK/SCP`	0.857	0.115	0.203	1	0.625	0.769
`BK/LCP`	1	0.135	0.237	1	0.625	0.769
CombineAll	0.979	0.940	0.959	0.929	0.812	0.867

Table 3.3: Precision (P), Recall (R) and F1-Score (F1) of impacted annotations computed using four different methods (`Rules`, `BK`, `SCP`, `LCP`) and the combination of them. The red and orange colours indicate low and medium recall, respectively.

98%). However, the recall can be significantly different according to the dataset and the method adopted. We detail the reasons for this in the next section. We would like to highlight that our rules have an excellent performance, obtaining in some cases an F1-Score of 99%.

The second evaluation process consists of applying methods to select which adaptation actions can make the annotation evolve correctly, and comparing the outcomes with the silver standard. The goal is to measure how precise our recommendations are. Table 3.4 describes the performance of each method regarding the four different datasets. Each experiment is represented by an Area Under the Curve (AUC) value, giving the probability that a randomly selected instance will correctly be adapted by our method [Hajian-Tilaki, 2013].

The AUC values of the analysed methods vary according to the dataset. A quick analysis of Table 3.4 shows that combining methods provides slightly better results than applying only one of them for MeSH and NCIt. The opposite is observed for ICD-9-CM and SNOMED CT where the `Rules` provide better results. Furthermore, SCP shows the lowest AUC values of all heuristics. We also verified that there are significant differences in the AUC between the KOS, like those between MeSH and NCIt. Detailed explanations on these observations are provided in the next section.

	ICD9CM		MeSH		NCIt		SNOMED CT
Method	AUC		AUC		AUC		AUC
Rules	0.862		0.882		0.731		0.844
BK	0.597		0.545		0.663		0.75
SCP	0.589		0.57		0.615		0.500
LCP	0.589		0.57		0.625		0.500
Rules/BK	0.83		0.87		0.731		0.844
Rules/SCP	0.821		0.878		0.731		0.833
Rules/LCP	0.839		0.878		0.74		0.844
BK/SCP	0.589		0.579		0.663		0.75
BK/LCP	0.589		0.579		0.673		0.75
CombineAll	0.821		0.887		0.73		0.815

Table 3.4: AUC values of developed heuristics used to maintain annotations. The red and blue colour highlight the lower and higher values for each dataset, respectively.

3.5 Discussion

The analysis of annotation evolution in the healthcare domain is an understudied topic. As explained in section 3.1, several works propose the automatic detection of inconsistent annotations, but few of them address the automatic correction of inconsistent annotations. The work presented in this thesis shows that some annotations can be preserved/adapted after the evolution of the KOS used to generate the annotations. Four methods were proposed: Domain Specific Rules, Background Knowledge, Semantic and Lexical Change Patterns. The outcomes presented in the results section demonstrated that we can obtain high AUC by applying these methods together in the automatic maintenance of annotations or to support domain experts in this activity.

When analysing each method, we observed that BK contributes to the precision of the annotation changes. The main characteristic of BK is that it depends on the richness of information in other sources (e.g., ontologies with overlapping concepts). Another aspect that can be deduced from the experiments is the dependency of the BK method on the expressivity and consistency of the KOS. For example, MESH D002544 has as synonym concepts that are siblings in other KOS (e.g. “Cerebral infarct left hemisphere” SNOMED CT 362323007 and “Cerebral infarct right hemisphere” SNOMED CT 362322002), leading to loose information when the system follows the KOS mappings that cross MeSH. Moreover, we observed that the BioPortal repository only contains the latest KOS version (i.e., from 2016), but our experiments use documents annotated with the 2009AA and 2010AA versions. Versioning is an aspect that has not yet been integrated into BioPortal, but it deserves to be considered in the future.

The analysis of SCP and LCP shows a good precision and low recall. The reason for this is that both methods consider only changes between concepts that are in the same neighbourhood (i.e., siblings, super- and sub-concepts). Thus, changes that move the concept to other branches of the KOS are not included, leading to an increased number of false negatives. For SNOMED CT, (see Table 3.3), we did not observe any cases with SCP and LCP in our dataset. Since, the used data was randomly selected, we consider that it was a coincidence. We did our analysis based on the results from the other KOS, see chapter 2. However, these heuristics are able to cover cases where the Rules or BK do not work. For instance, the annotation “*ubiquitin carboxy-terminal hydrolase*” NCIt C21490 correctly evolved to “*ubiquitin carboxyl-terminal hydrolase BAP1*”. This correct adaptation increased the AUC value to NCIt using the configuration: Rules/LCP, see Table 3.4. Thus, we concluded that this method is better than BK and its combinations (lines 2 and 5) for NCIt .

Domain-specific rules are defined to describe frequent patterns of changes, and they are expected to generate outcomes from them with good precision and recall. This was the case for the rules proposed in our experiments. However, our rules do not cover all annotation evolution cases perfectly. For example, in SNOMED CT the annotation *[PMC2633322; '31113003'; 'diverticulum'; 6412; 6424; 'association with the meckel'; ', the appendix,']* had to evolve to *"Meckel diverticulum"* ConceptID *37373007*, in order to be similar to our silver standard. In spite of this, *IncreaseAnnot* provided a different result. Applying the rules to adapt this annotation, we sometimes selected the ConceptID *37373007* and sometimes the ConceptID *127962001*. We identified two reasons for this behaviour:

1. In SNOMED CT the same label can be associated to two different ConceptIDs. For instance, the concept *37373007* (*Meckel diverticulum*) has two super-classes: *"Congenital anomaly of small intestine"* and *"Diverticulosis of small intestine"*, while the concept *127962001* (*Meckel diverticulum*) has the following super-classes *"Persistent embryonic structure"*, *"Structure of yolk stalk"*, *"Structure of distal portion of ileum"* and *"Diverticulum"*.

2. We used the Lucene search engine[14] to get the ConceptID of a given label. For this specific case, Lucene returns two ConceptIDs. However, the order of the results is not always the same. In our algorithm, we use the first result given by Lucene to execute the *IncreaseAnnot* rule. This explains the random outcome of our approach that occurs for SNOMED CT and NCIt when computing *"glycerol kinase gene"* codes (C75498, C75499).

Regarding the low value for BK method in MeSH and ICD-9-CM, we observed that some annotations did not correctly matched our silver standard. The example related to the annotation *"acute renal failure"*, concept 584.9 in ICD-9-CM version 2009 can illustrate the reason of the problem. In our silver standard the domain specialists indicated that the right adaptation is *"acute kidney failure" 584.9*, i.e., the same concept using the new term. Our method computed the right adaptation as the concept 584. After discussion with the domain specialists, we verified that our algorithm was correct. The problem comes from the mappings provided by BioPortal regarding MeSH↔ICD-9-CM, i.e., the MeSH concept *"acute renal failure"* has two mappings pointing to the concepts 584 and 584.9.

Another aspect to highlight is that depending on the context in which the maintenance methods were used (e.g., high expressive KOS), there are considerable differences in the sets of results. We also observed that a combination of methods can be used for a more complete set of evolution situations, as in the following:

- LCP, SCP and BK methods show low complementary results to identify whether the KOS evolution impacts the annotations, but an improvement was observed by combining the methods to identify the correct evolution of the annotation.

- On one hand, Rules increase the amount of corrected annotations of all SCP, LCP and BK analyzed cases.

3.6 Conclusion

In this chapter, we presented the possibility of using annotation maintenance tools that can keep track of KOS evolution. Moreover, it also demonstrates that the automatic correction/adaptation of annotations can reach a reasonable reliability rate. But, it is important to highlight that the role of human beings is still determinant in assuring the quality of the annotations in critical scenarios, as observed in the biomedical domain. Finally, our maintenance approach is done without a complete re-annotation of the document and ensures a high-quality annotation for the

[14] https://lucene.apache.org/core/

chosen concept. In next chapter, we discuss how to enhance the proposed method by including the *PartialMatch* rule that applies lexical and semantic algorithms to change the concept ID of an annotation.

Chapter 4

Semantic similarity measures to adapt semantic annotations

Contents

Measuring the similarity between concepts is a cornerstone of our approach to maintain annotations valid over time. As described in chapter 3, our architecture is able to adapt only annotations that fully match the associated label of the concept independently of the used technique used, i.e. `Rules` or `BK`. However, in some cases, there is a clear syntactic divergence between the annotation and the concept label. To overcome this limitation and help answering RQ3 and RQ4, we raise the following question: **How can we improve the existing similarity measures to enhance the relatedness between terms while taking the syntactic mismatch into account?** In order to approach this question, we assume that *the combination of semantic and lexical similarity measures (hybrid measures) can improve the relatedness between terms and be used to adapt the semantic annotations.*

We start this chapter by analysing related work in the field of semantic similarity measures. We then introduce our hybrid similarity measure that combines Lexical Similarity Measures (LSM) and ontology-based Semantic Similarity Measures (SSM) to improve the relatedness between biomedical terms. Using this hybrid measure we enrich our set of rules with a new one capable of maintaining annotations when the concept label and the annotated text are not the same. We assess the validity of our hybrid measure by first showing the correlation between the values obtained and the scores given by domain experts on a reference corpus using the

Spearman's rank correlation metric. We then use the Fisher's Z-Transformation to evaluate the stability of the utilized measures with respect to the evolution of KOS. Finally, we demonstrate that the new `PartialMatch` rule we derive from this metric is able to outperform other techniques and statistically improve the number of corrected annotations by using the Sign Test statistical method.

4.1 Existing Similarity Measures

The literature in this domain reveals several families of similarity measures. For instance, [Gomaa and Fahmy, 2013] described string-based, corpus-based and knowledge-based metrics. The former two groups are related to LSM and rely on syntactic or lexical aspects of the terms to compare strings, such as "*Failure of the kidney*" with "*Kidney failure*" [Oliva et al., 2011]. The latter group is related to the semantic aspects of the terms (SSM) such as the equivalence between "*Myocardium*" and "*Cardiac Muscle*".

In the biomedical domain, various LSMs have been used and evaluated in order to improve the retrieval of biomedical documents [Soualmia et al., 2012], support the mapping adaptation process [Dos Reis et al., 2015a], improve the semantic relatedness between terms in named entity recognition process, e.g., "ammonium" ↔ "ammonium ion" [Rudniy et al., 2014], etc. The results show that the LSMs are capable of improving the relatedness between terms. However, different thresholds must be considered since the metrics perform differently according to the domain of application. Notice that this kind of similarity measure does not take semantic aspects of the terms into account. Consider for instance terms like "*Cancer*" and "*Malignancy*". While they are completely disjointed from a lexical point of view, they are semantically equivalent.

To overcome this barrier, SSM measures were introduced. They exploit the meaning of the terms in a corpus and evaluate their similarity according to, for instance, the distribution of the words or the co-occurrence of terms [Mihalcea et al., 2006]. Semantic similarity measures can also rely on knowledge representation models such as thesauri or ontologies where structural properties of the model (e.g. hierarchy of concepts in an ontology) are used to compute the similarity between concepts [Lord et al., 2003]. It is used in a wide range of applications: automatic annotation validate in Gene Ontology [Couto et al., 2006], information retrieval algorithms [Sy et al., 2012], Linked Data paradigms [Meymandpour and Davis, 2016], etc.

Different categories of SSM exist [Couto et al., 2006, Harispe et al., 2014, Meymandpour and Davis, 2016, Resnik, 1995a] to evaluate similarity between concepts:

1. *Edge-based* measures estimate the similarity of two concepts as a function of the distance separating two concepts in the ontology.

2. *Feature-based* measures relies on the taxonomic interpretation of the feature model proposed in Tversky [Tversky, 1977]; generally, the representation of a concept corresponds to a set of neighbouring concepts or instances. Feature-based strategies root semantic similarity in the context of classical binary or distance measures (e.g. set-based measures, vector-based measures).

3. *IC-based* measures assess the similarity of concepts as a function of the Information Content (IC) from their Most Informative Common Ancestor (MICA), e.g. the deepest concept that subsumes two verified concepts. [Resnik, 1995a, Harispe et al., 2014].

4. *Hybrid* measures combine the approaches described above.

These measures have been extensively evaluated across multiple benchmarks [Pesquita et al., 2009, Garla and Brandt, 2012, Harispe et al., 2014, Costa and Leal, 2016]. As a result, the IC-based measures generally outperform the edge-based ones. One of the main drawbacks of

Feature-based measures is that they consider dimensions as mutually orthogonal and do not exploit relationships that link concepts. Finally, the hybrid-measures require specifics parameters, thus making a generalization of multiple KOS difficult.

Similarity measures have been used independently. Their combination, especially the couple LSM/ontology-based SSM, remains under-explored. [Pereira Nunes et al., 2013] proposed an approach combining co-occurrence-based and semantic-based measures for entity linking. It produces a semantic connectivity score capable of measuring the relatedness of web resources like *Charlotte Bobcats* and *Carmelo Anthony* on Dbpedia.

[Sánchez et al., 2012] rely on the combination of the WordNet and MeSH terminologies during the process to measure the relatedness of terms. This is done by: i) an assessment of the semantic overlapping between subsumer pairs and ii) an evaluation of their structural similarities analyzing the ontologies to which they belong, instead of relying solely on the terminological matching between subsumer labels. As a result, the accuracy obtained in the multi-ontology scenario almost reaches (but rarely surpasses) that obtained in an ideal mono-ontology setting.

[Peng et al., 2018] proposed NETSIM2, a network-based method capable of calculating the similarity between two Gene Ontology terms by combining the information from co-function network and GO global structure through a random walk with restart-based method. The experimental results using Enzyme Commission and a biological process show that NETSIM2 performs best among all standard measures on Yeast[15] and Arabidopsis[16] datasets. Furthermore, NETSIM2 can significantly improve the performance of semantic similarity measurements, especially when dealing with incomplete species description.

The literature review highlights a lack of hybrid measures combining LSMs and ontology-based SSMs. Furthermore, none of the existing approaches confirms whether the SSMs are stable when applying them to dynamic KOS, i.e. whether the similarity values remain acceptable after changes in the KOS. In this chapter, we present our approach, which combines both methods in order to enhance the similarity between concepts, and an experiment to verify the stability of these measures while using consecutive KOS versions.

4.2 Background

In this section, we introduce the statistical methods used to evaluate the experiments we carried out in this chapter.

Spearman's Rank Correlation

Spearman's Rank Correlation (cf. equation 4.1) is a statistical method that measures the coefficient strength of a linear relationship between paired data [Press et al., 1988]. In other words, it verifies whether the values produced by automatic similarity measures, e.g. values within the interval [0, 1], and scores given by domain specialists, e.g. values in intervals of [0, 1500], are correlated.

To compute the r_s, i.e. the correlation value, we have to follow four steps:

1. Rank the automatic similarity value and the domain specialist judgment in columns three and four of Table 4.1. This process will create two ordered sets as detailed in columns five and six.

2. Using the previous rank (columns five and six), we subtract them to obtain the difference of rankings d_i in column seven.

[15] http://www.yeastgenome.org/

[16] http://ftp.plantcyc.org/Pathways

3. We compute the square of d_i to obtain d_i^2 in column seven.
4. We apply the equation 4.1 on d_i^2, to obtain the correlation value r_s. The variable n in this equation is related to the number of cases observed in the dataset, such that $n \geq 10$.

As result of this process, we will have a correlation value r_s such that $-1 \leq r_s \leq 1$, where -1 indicates total disconnection and 1, a strong relationship. In this example the computed r_s is 0.393 which indicates that there is not a strong correlation between the automatic values and the domain specialist judgments.

$$r_s = 1 - \frac{6\sum_i d_i^2}{n\left(n^2-1\right)} \tag{4.1}$$

Concept 1	Concept 2	Automatic similarity value	Specialist judgment	rank Automatic	rank specialist	d_i	d_i^2
Enalapril	Lisinopril	0.54	1280.0	1.0	8.0	-7.0	49.0
Mycosis	Histoplasmosis	0.4	1282.5	2.0	4.0	-2.0	4.0
Dizziness	Vertigo	0.25	1287.0	3.0	1.0	2.0	4.0
Emaciation	Cachexia	0.44	1290.25	4.0	6.0	-2.0	4.0
Convulsion	Epilepsy	0.32	1302.75	5.0	2.0	3.0	9.0
Thalassemia	Hemoglobinopathy	0.46	1307.0	6.0	7.0	-1.0	1.0
Cefazolin	Keflex	0.43	1323.0	7.0	5.0	2.0	4.0
Lipitor	Zocor	0.33	1330.75	8.0	3.0	5.0	25.0
Medrol	Prednisolone	0.55	1387.5	9.0	9.0	0.0	0.0
Sinemet	Sinemet	1.0	1533.5	10.0	10.0	0.0	0.0

Table 4.1: Example using Spearman's Rank Correlation

Fisher's Z-Transformation

Fisher's Z-Transformation is a statistical method that allows the verification of whether two nonzero's r_s, i.e. Spearman's Rank Correlation are statistically different [Press et al., 1988]. Thus, we can verify whether the r_s from an automatic similarity method is better than other r_s' obtained from another method.

The comparison between the above correlations is done in three steps. First, the conversion of r_s and r_s' into z_1 and z_2 by applying equation 4.2, i.e. converting two sampling distributions into normal distributions.

Second, we obtain the probability value p such that $0 \leq p \leq 1$ through equation 4.3, where $N1$ and $N2$ are the number of data in the dataset.

$$z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) \tag{4.2}$$

$$p = erfc\left(\frac{|z_1 - z_2|}{\sqrt{2}\sqrt{\frac{1}{N1-3}+\frac{1}{N2-3}}}\right) \tag{4.3}$$

Finally, to ensure differences, we test our null hypotheses $H_0 : r_s = r_s'$ case $p > 0.05$ and vice versa. Nevertheless, this can only be done if $N1$ and $N2$, are moderately large ($N \geq 10$).

Sign Test

The Sign Test is a non-parametric test used to verify whether or not two groups are equally sized, i.e., the amount of success cases remains the same, before and after an applied procedure [Dixon and Mood, 1946]. This test ignores the actual values of data and uses $+$ or $-$ signs during the calculations, see Table 4.2.

In this example, in the first column we have the values of procedure X, followed by the values of procedure Y in column two. The difference of both methods $X - Y$ is shown in column three. In the fourth column, the values were translated to signs: $(+)$ indicating that procedure X provides better results, $(-)$ indicating that procedure Y is better, and (NA) there is no difference between the values.

The null hypothesis of the Sign Test is: H_0 : *Population median difference = 0*, i.e., the amount of $+$ sings (r^+) and $-$ sings (r^-) does not differ significantly from equality. To calculate this test we compute a binomial distribution[17] using as input: i) the amount of success cases given by $max(r^-, r^+) = 9$, for the example of Table 4.2; ii) the number of trials $n = 12$, i.e., the sum of r^- and r^+ excluding NA. In this example, we keep the null hypothesis because the result (*p-value*) is 0.146, indicating that there is no evidence for a difference between the two procedures when using $\alpha = 0.05$, i.e., H_0 : *Procedure X = Procedure Y* case $p > 0.05$ and vice versa.

Procedure X	**Procedure Y**	**X-Y**	**Sign of x-y**
443	57	386	+
443	88	355	+
370	370	0	NA
436	587	-151	-
422	463	-41	-
423	463	-40	-
424	463	-39	-
243	88	155	+
1000	1000	0	NA
236	310	-74	-
222	321	-99	-
223	333	-110	-
224	587	-363	-
224	632	-408	-

Table 4.2: Example using Sign Test statistical method.

4.3 A Hybrid LSM/Ontology-based SSM

In order to combine LSM and ontology-based SSM measures, we utilized a weighted arithmetic mean, see equation 4.4. It calculates the similarity between two concepts c_i, c_j by applying classic similarity measures over two respective concepts, e.g., *C0035078:Renal failure* ↔ *C0035078:Kidney failure* and assigning weights to each similarity.

In summary, the values LSM_{score} and SSM_{score} represent the similarity scores given by metrics like *Levenshtein* and *Resnik 1995 GraSM* in the interval of [0,1]. The variables α and τ are the weights in the interval of [0, 1] increasing in 0.1, excluding both α and τ equal to zero.

[17] `https://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.stats.binom_test.html`

This allows us to change the contribution of each measure to calculate the final similarity, e.g., in $\alpha = 0.8$ and $\tau = 0.3$ the semantic measure has more importance in the final score, however it is smoothly adjusted by the lexical measure.

$$simi_{(c_i,c_j)} = \frac{(SSM_{score} * \alpha) + (LSM_{score} * \tau)}{\alpha + \tau} \tag{4.4}$$

4.4 Partial Match Rule to Adapt Annotations

The major limitation of our method to adapt semantic annotations demonstrated in Chapter 3, was related to the limited scope of the change patterns to correct the annotations that are still invalid after the application of the `Rules` and `BK` methods. As change patterns consider the local evolution of a concept, i.e., *siblings*, *super-* and *sub-classes*, we added a new rule, called `PartialMatch`, able to deal with the evolution of a concept outside the neighbourhood.

This rule changes the term and/or the concept ID of an annotation considering lexical and semantic similarity measures, e.g. SNOMED CT 398624005:*Ethanol* → 420140004:*Allergy to ethanol*. The algorithm 2 gives an overview of the whole process. The input of the algorithm is the annotation before its evolution a_0, the KOS before and after the evolution KOS_{v0} and KOS_{v1}; the KOS changes *chgs*; and the weights τ and α for the hybrid measure.

Our method starts by gathering the information of the annotation in lines (1-2). We retrieve the used concept c_0 and the attribute att_0 following our model presented in Chapter 2. Then, our algorithm retrieves all stable ancestors $Set\_Sup\_Classes$ of a source concept c_0 within a specified period, e.g., 2009/2010 (line 3). Using these ancestors, the most similar ancestor (MSA) is kept (line 4) according to its similarity with c_0. This similarity respects the approach we proposed and utilizes the *preferred label* and *concept ID* to find the LSM_{score} and SSM_{score}, respectively.

At line 5, our algorithm computes a lexical view of attribute att_0 in KOS_{v1}. i.e., it retrieves all the related labels from stable and added concepts, which have the occurrence of at least, one common word, e.g., when querying *167696007:feces examination* in SNOMED CT 2010 we have 779 concepts as lexical view: [*162089003: Feces normal*; *167635008:Feces examination: growth; etc*].

The next steps (lines 6-11) compute the similarity between the MSA and the concepts from the lexical view. Once this process is complete, the best concept is chosen to maintain the annotation (line 12). Finally, at line 13 we adapt the annotation by changing the *concept ID* and/or the attribute value (e.g., by including additional information about the occurred evolution) as illustrated in Figure 3.4.

Using this configuration, our new `PartialMatch` rule is capable of maintaining semantic annotations by changing the term/concept even if it is in a different ontology region. For example, *167696007:feces examination* in SNOMED CT 2009 will evolve to *167592004:examination of feces* in SNOMED CT 2010.

4.5 Experimental assessment

In this section, we introduce the material and measures we used to evaluate our approach. We start by describing the assessment for the hybrid measure in Section 4.5.1 and then, in Section 4.5.2, we assess the `PartialMatch` rule.

4.5.1 Assessment for the hybrid LSM/Ontology-based SSM

The experiments we have conducted consisted of instantiating our hybrid LSM/Ontology-based SSM with a set of measures listed in Table 4.3, and comparing the results to a corpus of reference

Algorithm 2: PartialMatchAnnot: Partial matches between an attribute value and an annotation

```
Input: An impacted annotation at version 0 $a_0$ ; $KOS_{v0}$; $KOS_{v1}$; KOS Changes $chgs$; $\tau$
       and $\alpha$ values in the range of [0.1, 1]
Output: An evolved annotation $a_1$
$c_0 \leftarrow getConceptFromAnnot(a_0)$
$att_0 \leftarrow getAttributeFromAnnot(a_0)$
$Set\_Sup\_Classes \leftarrow getAllStableAncestor(c_0, KOS_{v0}, chgs)$
$MSA \leftarrow getMostSimilarAncestor(Set\_Sup\_Classes, c_0, KOS_{v0})$
$actions \leftarrow lexicalView(att_0, chgs, KOS_{v1})$
if $(MSA \neq \emptyset \wedge actions \neq \emptyset) == TRUE$ then
    forall $actions$ do
        $SSM_{score} \leftarrow SemanticSimi(MSA, actions_i, Ont_{v1})$
        $LSM_{score} \leftarrow LexicalSimi(att_0, actions_i)$
        $simi \leftarrow hybridMeasure(SSM_{score}, LSM_{score}, \tau, \alpha)$
        $actions_i \leftarrow updateSimilarity(actions_i, simi)$
    $bestCp \leftarrow getHighestSimilarity(actions)$
    $a_1 \leftarrow buildEvolvedAnnot(bestCp, a_0)$
    return $a_1$
return $a_0$
```

built by domain specialists that contain several pairs of concepts and their similarities.

Gold Standard

We used the three reference datasets suggested by [McInnes and Pedersen, 2013] to evaluate our approach. We first used *MayoSRS* [Pakhomov et al., 2011]. It contains 101 pairs of concept labels together with a score assigned to each pair denoting their relatedness. The value of the score, ranging from 0 to 10, is determined by domain experts. 0 represents a low correlation while 10 denotes a strong one.

The second dataset we used is a subset of *MayoSRS* [Pakhomov et al., 2011] made up of 30 pairs of concept labels. For this dataset, a distinction is made between the two categories of experts: coders and physicians and the values of the relatedness score is ranging from 1 (unrelated) to 4 (almost synonymous).

The third dataset is the UMNSRS described in [Pakhomov et al., 2010]. Bigger than the two previous ones, it is composed of 725 concept label pairs whose similarity was evaluated by four medical experts. The similarity score of each pair was given based on a continuous scale ranging from 0 to 1500. These values were assigned experimentally by users.

Knowledge Organization Systems

In our experiments we used: Medical Subject Headings (MeSH) and Systematized Nomenclature of Medicine - Clinical Terms (SNOMED CT). We used the versions 2009AA to 2014AA (excluding the AB versions), downloaded from the UMLS and transformed into OWL files. Furthermore, we considered the evolution of concepts using the following configuration $SUPRESS =' Y'$ in Table $MRCONSO$ from the UMLS thesaurus. We chose these terminologies because the terms present in the gold standard, e.g. *Lymphoid hyperplasia* were not present in other KOS, e.g. NCIt.

Measures

We used a total of 12 Lexical measures, 11 Semantic measures and 9 Information Content techniques (Table 4.3), i.e., methods included in the SSMs to extract the features of concepts when calculating the similarity [Resnik, 1995b].

We reused/adapted the implementation of the Lexical measures that we collected from open-source projects, published in Github [18], and related works from the biomedical domain [Lin et al., 2017, Rudniy et al., 2014]. The semantic measures are those implemented in the framework proposed by [Harispe et al., 2014]. The Jiang Conrath (equation 4.8) is an example. It calculates the similarity between concepts c_1, c_2 by computing the Most Informative Common Ancestor (MICA) and the Information Content (IC) which avoids the dependency of a corpus, e.g. documents, to calculate the concept usage.

The IC can be calculated as in the example of equation 4.5. In this equation, [Seco et al., 2004] computes the IC according to the *log* of the number of descendants from a concept c including itself, divided by the *log* of the total number of concepts in the ontology. The MICA is calculated according to equation 4.7. In this equation, the biggest IC from a set of common ancestors from concepts c_1, c_2 is selected according to the MICA score. To find the common ancestors we utilize the equation 4.6, which simply makes an intersection between all parents from c_1 and c_2.

$$IC_{Seco}(c) = 1 - \left(\frac{log(numberInclusiveDescendants)}{log(numberConceptsOnto)} \right) \tag{4.5}$$

$$setAncestors = intersection(parents(c_1), parents(c_2)) \tag{4.6}$$

$$MICA = \max(\{IC(x) : setAncestors_x\}) \tag{4.7}$$

$$sim_{JC}(c_1, c_2) = 1 - \frac{IC(c_1) + IC(c_2) - 2 \cdot MICA(c_1, c_2)}{2} \tag{4.8}$$

In total, 1188 hybrid measures and 118800 combinations were evaluated considering the weights described in section 4.3. Therefore, and for better readability, we list only the top-10 measurements in our experiments.

Lexical Measure (LSM)	**Semantic Measure (SSM)**	**Information Content (IC)**
Levenshtein	Jiang Conrath 1997 Norm	Resnik Unpropagated 1995
Smith Waterman	Feature Tversky Ratio Model	Sanchez 2011
Jaccard	Tversky IC Ratio Model	Sanchez 2011 b adapted
Cosine	Lin 1998	Seco 2004
Block Distance	Lin 1998 GraSM	Zhou 2008
Euclidean Distance	Mazandy 2012	Harispe 2012
Longest Common Substring	Jaccard IC	Depth Max Non linear
Jaro Winkler	Resnik 1995 GraSM	Depth Max Linear
LACP	Resnik 1995	Ancestors Norm
Tf-idf	Jaccard 3W IC	
AnnoMap	Sim IC 2010	
Bigram		

Table 4.3: Lexical and Semantic Measures utilized.

[18] https://github.com/tdebatty/java-string-similarity

4.5.2 Assessment for the Partial Match Rule

The experiments we conducted to assess the `PartialMatch` rule consisted of including it in the method discussed in Chapter 3 and verifying whether this rule improved the amount of adapted annotations.

Terminologies

As described in chapter 3, our maintenance method utilized four well-known KOS: ICD-9-CM, MeSH, NCIt and SNOMED CT. However, we worked with two versions to adapt the annotations in Chapter 3, i.e., from version 2009 to version 2010. In this chapter, we included multiple versions, e.g., 2009 to 2016, in the experiments to verify whether the `PartialMatch` rule also performed well over a longer time.

Silver Standard

For these experiments, we adapted the silver standard described in Chapter 3 by including the reference for the annotations in 2016. Table 4.4 shows an illustrative example related to our silver standard. It shows one annotation produced with MeSH:2009AA using the PubMed document 232[19] and the concept D009133. The annotated text is "muscular atrophy", and can be found at position [5561,5577] of the document. We customized our system to have a maximum of four words as prefix *"(HD), spinal and bulbar"* and as suffixes *", drpla and various"*. It can be observed that the concept label and ID used to annotate the text increased and changed, respectively, from 2009AA to 2010AA. Furthermore, in 2016AA the concept ID changed again, from D055534 to D020966. Therefore, we have an annotation impacted multiple times by the evolution of MeSH.

KOS	Doc.	Concept	Annotation	Start	End	Prefix	Suffix
2009AA	232	D009133	muscular atrophy	5561	5577	(HD), spinal and bulbar	, drpla and various forms
2010AA	232	D055534	spinal and bulbar muscular atrophy	5543	5577	(HD),	, drpla and various forms
2016AA	232	D020966	spinal and bulbar muscular atrophy	5543	5577	(HD),	, drpla and various forms

Table 4.4: Example of an evolving annotation from our silver standard. The red color indicates the changes that occurred in the annotation at KOS evolution time. Source: [Cardoso et al., 2017a]

4.6 Experimental Setup

This section describes the methods utilized to evaluate both approaches, i.e., the hybrid measure and the `PartialMatch` rule. We start by describing the set-ups for the hybrid LSM/SSM in Section 4.6.1 and then in Section 4.6.2 we detail the set-up for the `PartialMatch` rule.

[19] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2638829/

4.6.1 Experimental set-up for the hybrid LSM/SSM

In order to evaluate the capacity of our proposed metric to improve the similarity between pairs of concepts, as well as the stability of SSMs over time, we defined the three following different configurations:

- **Set-up 1** aims to verify the stability of semantic measures over time, i.e., whether the evolution of the KOS impacts the relatedness between automatic similarity values and domain specialist responses. To determine this, we followed three steps: i) we established the gold standard and semantic measure, e.g., dataset: *MiniMayoSRS*, SSM: *Jiang Conrath 1997 Norm* and IC: *Sanchez 2011*; ii) we computed the similarity results using consecutive versions of an ontology, e.g., MeSH 2009, 2010 and 2011; and iii) we used Fisher's Z-Transformation to verify whether the results obtained were statistically different. If the p-value obtained from Fisher's Z-Transformation was under 0.05, we assumed that the KOS changes impacted on the similarity values and vice-versa.

- **Set-up 2** verified the amount of hybrid measures (Lexical×Semantic) that outperformed the approaches that used one single method. In this configuration, we fixed the ontology, then grouped the results from all datasets to verify how many results from the combined methods outperformed the results from the single method. This setup ($dataset \times ontology\_version \times measures$) generates 25920 possibilities. We only present the overall results and the top 10 cases in the following sections.

- The objective of **set-up 3** is to point out the best measure over all datasets. To do this, we tested two possibilities: i) ranking with respect to the ontology. We fixed the ontology and then analysed the performance from all combined measures across the datasets. This step computed the hybrid measures rank according to *dense pandas.Series.rank* [20]; ii) overall ranking regardless of the ontology and dataset. In this step, we created a new ranking according to the lowest standard deviation, summation and average of the previous ranks.

4.6.2 Experimental Setup for the Partial Match Rule

In order to evaluate the capacity of our method to adapt impacted annotations into consistent ones, we utilized the two different configurations described below:

- **Set-up 4** utilized only single methods, i.e., `Rules`, `PartialMatch`, `SCP` and `LCP` without their combinations. However, we analyzed only the new `PartialMatch` approach proposed in Section 4.4. We intend to verify whether this new rule outperforms the other techniques, `SCP` and `LCP`, respectively.

- The **Set-up 5** determined the position of the `PartialMatch` rule in our framework to adapt semantic annotations. We tested two possibilities: i) Before the *SuperClassAnnot*; ii) After the *SuperClassAnnot*. After the determination of the `PartialMatch` positioning, we evaluated the framework with all proposed extensions by comparing it to the configuration utilized in Chapter 3.

For both Setups, we tested the effectiveness of the methods as follows: i) the capacity of our framework to detect impacted annotations after changing a KOS concept; and ii) the ability to correctly evolve the impacted annotations into consistent ones. In this case, consistency means being consistent with the silver standard.

[20] https://pandas.pydata.org/pandas-docs/version/0.21/generated/pandas.Series.rank.html

Metrics

To evaluate whether our predictions were similar to the silver standard, we used classic well-known metrics, such as, *Precision*, *Recall*, *F1-score*, *Area Under the Curve* (AUC) and *Accuracy* [Powers, 2011]. The comparison between the results from Chapter 3 and those obtained after the extension of our maintenance method is done by using the *Sign Test* method [Dixon and Mood, 1946]. Finally all percentages concerns to the use of *Percentage Change*, i.e., $\frac{(V_2 - V_1)}{|V_1|} * 100$

4.7 Results and Discussion

This section describes the results obtained when assessing the hybrid measure and the `PartialMatch` rule. We start by describing the results for the hybrid LSM/SSM in Section 4.7.1 and then, in Section 4.7.2, we detail the results for the `PartialMatch` rule.

4.7.1 Hybrid LSM/Ontology-based SSM

The results for **setup 1**, i.e., the stability of SSMs over time, can be observed in Table 4.5. In the first column we have the semantic measures described in Section 4.5.1, and in the second column the versions of the ontology used. The comparison between the versions, column three, was done through a Cartesian product between the versions 2009 to 2014, where we removed the redundant combination of years, i.e., we removed for instance 2014×2013 and be kept 2013×2014.

For readability reasons, we listed only a sub set of the results in Table 4.5, but all values can be found on `https://git.list.lu/ELISA/SemanticSimi`. This subset is related to the UMNSRS dataset which has the lower and higher Z-Fisher among all results.

The Z-Fisher in Table 4.5 shows that none of the SSMs had significant differences over time. For the purpose of considering statistical differences between the SSMs, the Z-Fisher must be lower than 0.05. As verified in our result, the lowest Z-Fisher value was 0.277, for the SSM configuration: *Resnik Unpropagated 1995* using IC: *Seco 2004* in the period *2010* & *2014*, red colour in Table 4.5.

We thus verified that KOS changes do not impact the SSMs over the time. However, the balance of the dataset, i.e. the size and the amount of impacted concepts used to calculate the SSMs, may impact the final results. We observed that the percentage of impacted concepts in these datasets was 2.8%, while the percentage of impacted concepts used to calculate the SSMs in an ontology region, i.e., *subClass*, *superClass* and *Siblings* was 5.53%. Furthermore, the top-k hybrid measures in our overall rank of **setup 3** (Table 4.7), utilized the measures with the lowest Z-Fisher in Table 4.5. This result highlights that the evolution of the ontologies played a key role during the process of calculating the SSMs similarity. Thus, future work on SSMs must include other pairs of impacted concepts to verify whether the stability of these measures and the obtained rank continue to be the same.

Regarding the **set-up 2**, i.e., the percentage of hybrid measures that outperformed the single SSMs, we observed that some hybrid measures, e.g. *AnnoMap* × *Zhou 2008* × *Resnik 1995 GraSM* had a better Spearman's correlation score than the single SSMs in 91.6% of cases, see Table 4.6. In this Table, the first column is related to the LSM measures, and the second column to the Information Content (IC) and semantic measures. Finally, column three shows the percentage of hybrid measures that outperformed the single SSMs.

In Table 4.6, AnnoMap is present in 76% of the LSMs that contributed to increasing the Spearman's correlation of the hybrid measures and, in consequence, outperformed the single SSMs. The similarity computed by AnnoMap [Lin et al., 2017], see equation 4.9, is also based on the combined similarity score from different string similarity functions, in particular TF/IDF,

IC & SSM Measure	Year	Z-Fisher
Seco 2004 & Jiang Conrath 1997 Norm	2009 & 2010	0.519871
Seco 2004 & Jiang Conrath 1997 Norm	2010 & 2011	0.880821
Seco 2004 & Jiang Conrath 1997 Norm	2010 & 2014	0.277042
Seco 2004 & Jiang Conrath 1997 Norm	2011 & 2012	0.991348
Seco 2004 & Jiang Conrath 1997 Norm	2012 & 2013	0.991341
Seco 2004 & Jiang Conrath 1997 Norm	2013 & 2014	0.356598
Ancestors Norm & Resnik 1995 GraSM	2009 & 2010	0.69417
Ancestors Norm & Resnik 1995 GraSM	2010 & 2011	0.832429
Ancestors Norm & Resnik 1995 GraSM	2011 & 2012	1.0
Ancestors Norm & Resnik 1995 GraSM	2012 & 2013	1.0
Ancestors Norm & Resnik 1995 GraSM	2013 & 2014	0.793019

Table 4.5: Stability of SSMs over time using UMNSRS dataset. We are considering the p-value of 0.05 as threshold of statistical significance. The red color indicates the lowest Z-Fisher obtained in our experiments and the orange indicates the higher

Trigram and LCS (longest common substring).

$$sim_{AnnoMap}(t_1, t_2) = MAX(\mathit{TfIdf}, TriGram, LCS) \tag{4.9}$$

When analysing the results of the best hybrid measures in **set-up 2**, we verified that the similarity score obtained with the hybrid measures was considerably improved. In practice, when calculated by single SSMs methods, *Pain↔Morphine* CUIs: C0030193 and C0026549, obtains a similarity score of 0.27. Using the hybrid measure (equation 4.3), the similarity score increased to 0.56 and better matched the score given by domain specialists in the UMNSRS dataset, i.e., 996.75 in a scale of [0, 1500].

Nevertheless, we also verified that some hybrid measures never improved the single SSMs, e.g. LSM: *Block distance*, SSM *Resnik Unpropagated 1995* with IC *Sim IC 2010.* This occurred mainly for LSMs as Block distance, Jaccard and TF/IDF, which considered strings as orthogonal spaces. In parallel, Information Content (IC) focused only on the positioning of concepts in an ontology without exploring additional resources like: Ancestors Norm and Max Linear, which also demonstrated no improvement.

Regarding the LSMs *tf-idf*, *Jaccard*, and *Block distance*, we verified that these methods lose the information contained in the prefix [Gusfield, 1997], e.g.,"Renal failure" ↔ "Kidney failure". When we verified the scores given by the domain specialists in MiniMayoSRS dataset (4.0), these terms were classified as strongly related. The similarity measures computed using Cosine or Jaccard were 0.5 and 0.33, respectively (medium to poorly related terms). On the other hand, methods like LACP provided a similarity of 0.77, which better matches the scores from the domain specialists, increasing the Spearman's correlation value.

The SSMs also reported a similar behaviour, e.g. *Ancestors Norm*, which computes the IC scores according to the number of ancestors from a concept divided by the total of number of concepts in an ontology, i.e., $ic = \frac{nbAncestors(v)}{nbConceptInOnto}$. This IC failed to describe the region of a concept in the KOS, because concepts with the same number of ancestors, but in different ontology regions, will have the same IC. The main drawback of this approach is that is does not consider information such as *siblings* to calculate the IC. As discussed in [Meng et al., 2012] and also verified during our experiments, the IC of a concept is dependent on its topology in the taxonomy. Thus, exploring information as *siblings* will enhance the representation of the concept topology, i.e. the spatial region it has in the ontology. It is widely utilized in other domains, e.g., ontology prediction, mapping alignment [Pesquita and Couto, 2012, Da Silveira et al., 2015].

Table 4.6: Percentage of hybrid measures that outperforms the single SSMs

Lexical	**IC** & **Semantic**	%
AnnoMap	Zhou 2008 & Resnik 1995 GraSM	91.67
	Resnik Unpropagated 1995 & Tversky IC Contrast Model	91.67
	Seco 2004 & Tversky IC Contrast Model	91.67
	Resnik Unpropagated 1995 & Resnik 1995 GraSM	87.5
	Sanchez 2011 b adapted & Resnik 1995	87.5
	Seco 2004 & Resnik 1995	87.5
	Harispe 2012 & Jiang Conrath 1997 Norm	87.5
	Zhou 2008 & Resnik 1995	87.5
	Seco 2004 & Resnik 1995 GraSM	87.5
	Sanchez 2011 b adapted & Resnik 1995 GraSM	87.5
Longest Common Substring	Sanchez 2011 b adapted & Tversky IC Contrast Model	87.5
AnnoMap	Resnik Unpropagated 1995 & Resnik 1995	87.5
Longest Common Substring	Harispe 2012 & Jiang Conrath 1997 Norm	83.33
LACP	Sanchez 2011 & Jian Conrath 1997 Norm	83.33

Regarding the first result of **set-up 3**, i.e., the overall rank of the measures for each ontology, we verified that hybrid measures performed better than the single SSMs for both KOS, MeSH and SNOMED CT. We are considering top measures to be those that have the minimum average rank, summation and standard deviation.

In our experiments, we verified that the most highly performing hybrid measure for MeSH was: Lexical: *AnnoMap*, IC & SSM: *Seco 2004* & *Jiang Conrath 1997 Norm*, (weights: $\alpha = 0.8$ and $\tau = 0.4$ or $\alpha = 1.0$ and $\tau = 0.5$). We also observed that this hybrid measure was ranked in the top 3 for MeSH, but with different weights, see Appendix 2 for more details.

In SNOMED CT, another hybrid measure was ranked as the most highly performing. The combination: Lexical: *AnnoMap*, IC & SSM: *Sanchez 2011b adapted* & *Jiang Conrath 1997 Norm*, $\alpha = 1$ and $\tau = 0.9$ was first in the rank. Moreover, MeSH and SNOMED CT had different measures in their top rank, (see Appendix 3 for more details).

Regarding this change in the ranks of both ontologies, we verified that the dataset utilized (MayoSRS, Coders/Physicians, UMNSRS) did not contain many repeated concepts, and the measures that had less performance in the UMNSRS dataset were those with higher Spearman's score in Coders/Physicians and MayoSRS. The complexity associated with UMNSRS can be partially be justified by: i) the amount of cases to match with the domain specialists scores, around 175 in UMNSRS and 30 in the others datasets; ii) as discussed in [Pakhomov et al., 2010] and also verified in our experiments, the relation *similarity* ↔ *relatedness* is unidirectional, i.e., the terms that have high similarity are often related but the opposite is not true. For instance,the semantic similarity of *Sinemet*↔*Sinemet* CUIs: C0023570 and C0006982 is 0.93, while *Pain*↔*Morphine* CUIs: C0030193 and C0026549 is 0.27. In UMNSRS dataset, the case *Pain*↔*Morphine* is more frequent and in consequence reduces the quantity of appropriated matches between domain specialists and automatic semantic measures.

When we recalculated the ranks to produce the overall rank, the second experiment from **Set-up 3**, we can better observe the good performance of our approach, see Table 4.7. The two first columns of Table 4.7 indicate the configuration for lexical and semantic measures. The

Table 4.7: Overall rank combining all ontologies and dataset. The values inside the columns five to eight indicate the obtained ranks.

Lexical	**IC & SSM**	**Alpha and Tau**	**Ontology**	**Coders**	**Physicians**	**Mayo**	**UMNSRS**	$\sum$	$\bar{X}$	σ
AnnoMap	Seco 2004 & Jiang Conrath 1997 Norm	0.8 and 0.5	SNOMED CT	357	373	189	461	1958	242.5	151.7
			MeSH	138	137	7	296			
AnnoMap	Resnik Unpropagated 1995 & Jiang Conrath 1997 Norm	0.8 and 0.5	SNOMED CT	357	373	189	461	1959	243	151.74
			MeSH	138	137	7	297			
AnnoMap	Sanchez 2011 b Adapted & Jiang Conrath 1997 Norm	0.9 and 0.6 1.0 and 0.7 0.3 and 0.2 0.6 and 0.4	SNOMED CT	356	340	103	511	1999	240	161.38
			MeSH	113	76	140	360			
AnnoMap	Seco 2004 & Jiang Conrath 1997 Norm	0.5 and 0.3	SNOMED CT	391	412	195	433	2012	242	152.95
			MeSH	138	137	17	289			
AnnoMap	Resnik Unpropagated 1995 & Jiang Conrath 1997 Norm	0.5 and 0.3 1.0 and 0.6	SNOMED CT	391	412	195	433	2013	242.5	152.98
			MeSH	138	137	17	290			

weights α and τ used to tune the contribution of each measure, are listed in the third column. Column four corresponds to the ontology used, followed by the ranks obtained with each dataset in columns five to eight. The final rank remained the same, when considering the minimum average, summation and standard deviation, presented in columns nine to eleven, as ranking criteria. In our results, we verified that the best performance corresponded to the hybrid measure: Lexical: *AnnoMap*, IC & SSM: *Seco 2004* & *Jiang Conrath 1997 Norm*, $\alpha = 0.8$ and $\tau = 0.5$. It was ranked in the Top 8 in MeSH.

Figure 4.1 shows the behaviour of this hybrid measure using SNOMED CT as the reference ontology. In each sub-figure from Figure 4.1, we have one dataset using SNOMED CT version 2012 and the weights for the SSM and LSM represented in the x-axis and y-axis, respectively. The weights increase by a factor of 0.1 and demonstrate the importance for each measure, e.g., a x-axis equal to 0.0 and y-axis equal to 1.0 show that semantic measures have more importance, while the main diagonals of the tables show that the measures have the same importance. We decided to present the results of this hybrid measure for SNOMED CT, because even if its final rank is far from the top rank for SNOMED CT, it was ranked as the best hybrid measure in the overall rank.

As a primary observation, we verified that the best τ and α, indicated by the yellow colour in Figure 4.1, remained in the main diagonal for the MiniMayoSRS dataset (Coders). In this dataset, we can reduce the complexity to find an appropriate τ and α, because they have the same importance. In the MiniMayoSRS dataset (Physicians), the best values of τ and α remain in the upper triangular part.

The main difference we noticed for the best combination, *AnnoMap*, IC & SSM: *Seco 2004* & *Jiang Conrath 1997 Norm*, $\alpha = 0.8$ and $\tau = 0.5$, was the similarity results in UMNSRS dataset. The obtained results were not better than the single SSMs. The reason for this, was that *AnnoMap* does not provide good similarity for the terms in the UMNSRS dataset. The Spearman's correlation value obtained from this lexical measure was -0.113. On the other hand, Lexical measures such as *LACP* have a better Spearman's correlation value, 0.113, and are a better choice to combine with the SSMs. For instance, using the configuration LSM: Lexical IC & SSM: *Ancestors Norm* & *Lin 1998 GraSM*, $\alpha = 0.8$ and $\tau = 0.1$, we have the best Spearman's correlation value for UMNSRS dataset 0.462, which outperforms the single SSMs 0.456.

4.7.2 Partial Match Rule

The results regarding the analyses of **setup 4** are demonstrated in Figures 4.2 and 4.3. The results in Figure 4.2 concern to the ability of these methods to detect impacted annotations, while Figure 4.3 showed the ability to propose correct adaptations for the impacted annotations. We utilized the references (2009/2010) and (2009/2016) of our silver standard to evaluate `PartialMatch`.

As observed in Figure 4.2, the precision of all methods to detect impacted annotations is high for both periods (2009/2010) or (2009/2016). However, the recall varies according to the terminology used and the year. SNOMED CT and NCIt in (2009/2010) showed the highest recall for the `PartialMatch`, while `SCP` and `LCP` show null values or close to 0.

In the period (2009/2016), SNOMED CT showed an improvement of 18% ($V_1 = 0.75, V_2 = 0.885$) when compared to the years (2009/2010), while the other terminologies had a smooth variation. In short, the proposed rule outperformed the `SCP` and `LCP` in all KOS and in all years, to detect the impacted annotations. This result was very clear when we observed SNOMED CT in Figure 4.2.

The `PartialMatch` rule also demonstrated good results for adapting the annotations in all KOS versions (2009/2010) and (2009/2016), see Figure 4.3. The AUC value for NCIt in 2009/2010 was 13.9% higher than `LCP` and `SCP` ($V_1 = 0.625, V_2 = 0.712$), followed by an improvement of 14% in 2009/2016 ($V_1 = 0.614, V_2 = 0.702$). This difference increased when

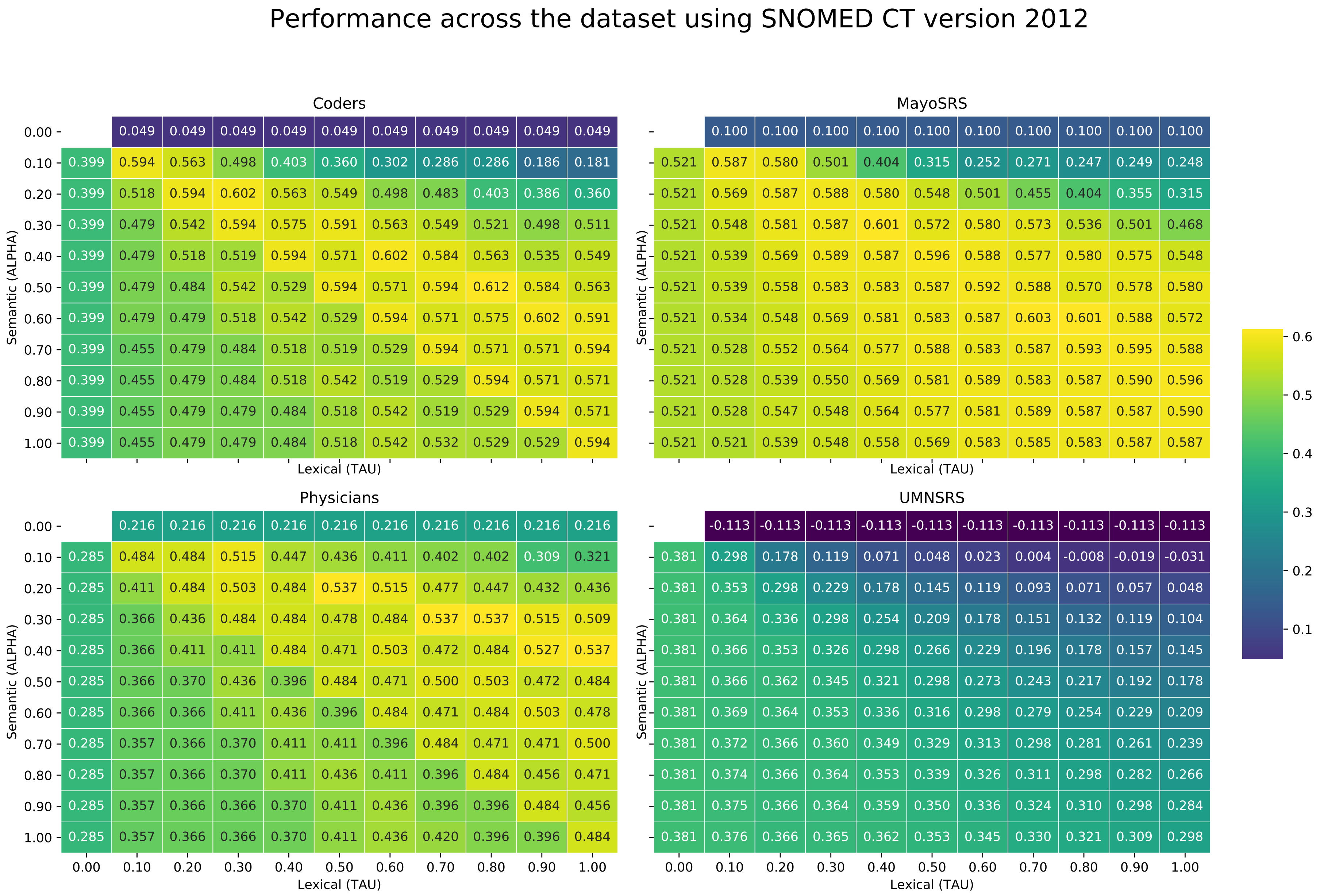
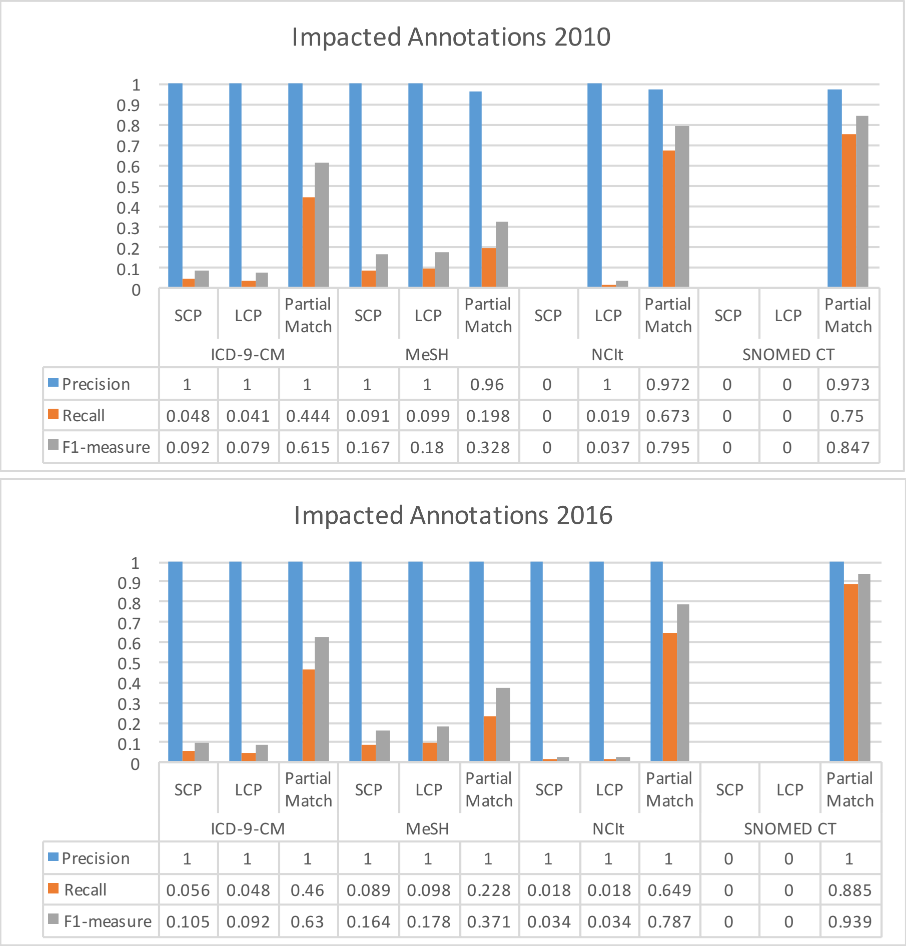


Figure 4.1: Performance of our best hybrid measure across the dataset in SNOMED CT version 2012. The Yellow colour refers to the best values, while the blue values the less performing.

we observed the F1-Score, reaching 55% for LCP ($V_1 = 0.371, V_2 = 0.575$) and 65% for SCP in 2009/2016 ($V_1 = 0.348, V_2 = 0.575$).

The other terminologies, ICD-9-CM and MeSH, also showed better results for the PartialMatch rule in 2009/2010 and 2009/2016. The difference observed in ICD-9-CM was 20% for the F1-Score in PartialMatch ($V_1 = 0.303; V_2 = 0.379$) and 4.75% for AUC ($V_1 = 0.589, V_2 = 0.617$), while in MeSH this difference was less expressive, reaching 0.7% for AUC ($V_1 = 0.57, V_2 = 0.574$) and 5% for F1-Score ($V_1 = 0.246, V_2 = 0.259$).

When analysing the results provided by each method (PartialMatch, SCP and LCP), we verified that the inclusion of PartialMatch that uses SSMs to find candidate concepts in other ontology regions produced a better adaptation for the impacted annotations. The main reason was that PartialMatch covered more situations than only the neighbourhood utilized in Change Patterns. Therefore, it could also be extended to future LCP and SCP versions to increase the definition of the context of the concept.

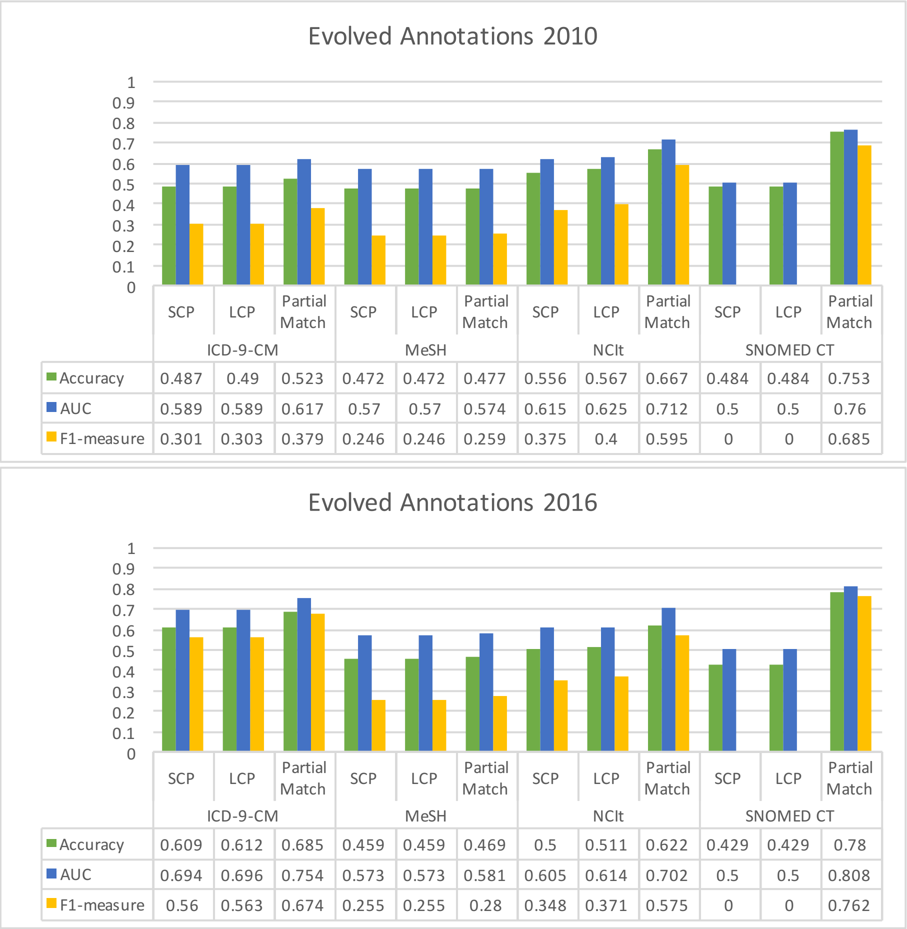


	ICD-9-CM			MeSH			NCIt			SNOMED CT		
	SCP	LCP	Partial Match	SCP	LCP	Partial Match	SCP	LCP	Partial Match	SCP	LCP	Partial Match
🟦 Precision	1	1	1	1	1	0.96	0	1	0.972	0	0	0.973
🟧 Recall	0.048	0.041	0.444	0.091	0.099	0.198	0	0.019	0.673	0	0	0.75
■ F1-measure	0.092	0.079	0.615	0.167	0.18	0.328	0	0.037	0.795	0	0	0.847

	ICD-9-CM			MeSH			NCIt			SNOMED CT		
	SCP	LCP	Partial Match	SCP	LCP	Partial Match	SCP	LCP	Partial Match	SCP	LCP	Partial Match
🟦 Precision	1	1	1	1	1	1	1	1	1	0	0	1
🟧 Recall	0.056	0.048	0.46	0.089	0.098	0.228	0.018	0.018	0.649	0	0	0.885
■ F1-measure	0.105	0.092	0.63	0.164	0.178	0.371	0.034	0.034	0.787	0	0	0.939

Figure 4.2: Performance of methods in *Setup 4* to detect impacted annotations.

Regarding the position of PartialMatch in **Setup 5**, we verified that only ICD-9-CM 2009/2010 and SNOMED-CT 2010/2016 exhibited a better performance when the PartialMatch was placed after the *SuperClassAnnot*. However, this is a minor improvement. The huge impact is observed in NCIt when the PartialMatch rule is placed before the *SuperClassAnnot*. The F1-Score shows a positive difference of 8.5% in 2009/2010 ($V_1 = 0.735, V_2 = 0.798$) and 6.6% in 2009/2016 ($V_1 = 0.750, V_2 = 0.800$) (see Appendix 1 for more details). Therefore, our next results concerning the adaptation of semantic annotations, only used the PartialMatch **before**



	SCP	LCP	Partial Match	SCP	LCP	Partial Match	SCP	LCP	Partial Match	SCP	LCP	Partial Match
	ICD-9-CM			MeSH			NCIt			SNOMED CT		
🟩 Accuracy	0.487	0.49	0.523	0.472	0.472	0.477	0.556	0.567	0.667	0.484	0.484	0.753
🟦 AUC	0.589	0.589	0.617	0.57	0.57	0.574	0.615	0.625	0.712	0.5	0.5	0.76
🟨 F1-measure	0.301	0.303	0.379	0.246	0.246	0.259	0.375	0.4	0.595	0	0	0.685

	SCP	LCP	Partial Match	SCP	LCP	Partial Match	SCP	LCP	Partial Match	SCP	LCP	Partial Match
	ICD-9-CM			MeSH			NCIt			SNOMED CT		
🟩 Accuracy	0.609	0.612	0.685	0.459	0.459	0.469	0.5	0.511	0.622	0.429	0.429	0.78
🟦 AUC	0.694	0.696	0.754	0.573	0.573	0.581	0.605	0.614	0.702	0.5	0.5	0.808
🟨 F1-measure	0.56	0.563	0.674	0.255	0.255	0.28	0.348	0.371	0.575	0	0	0.762

Figure 4.3: Performance of methods in *Setup 4* to adapt impacted annotations.

the *SuperClassAnnot*.

In Table 4.8, we presented the results obtained when adapting the annotation in the period 2009/2010. In the first column, we stated the KOS used, i.e., ICD, MeSH, NCIt and SNOMED CT. The configuration described in Chapters 3 and 4 is mentioned in the second column and the third column shows the methods followed by the metrics.

We observed that the eight rules used in our framework (`Rules+`), were capable of reaching the same results as the combination with the `BK` method. Furthermore, the use of `BK` for annotations generated with SNOMED CT leaded to a smooth decrease in the values. `Rules+` was also capable of outperforming (or achieving the same results) the best methods of Chapter 3 (`Rules`) with a 7.8% of improvement in AUC for NCIt ($V_1 = 0.74, V_2 = 0.798$) and 15% for the F1-Score ($V_1 = 0.649, V_2 = 0.747$), see Table 4.8.

Analysing the previous results, we observed that no annotation was adapted by the `BK` method in ICD-9-CM, MeSH, NCIt and SNOMED CT. It occurred in both versions used, 2009/2010 and 2009/2016. The variation in the AUC refers only to the results of *IncreaseAnnot* rule discussed in Chapter 3, i.e., with a term that increases the information of the impacted annotation occurring in different concepts. In Fact, the `BK` technique did not compute any annotations because all of them were adapted at a previous layer (`Rules+`).

In general, the new `PartialMatch` rule is capable of providing adaptations that are not feasible with only `BK`. This is for instance the case of the annotation *"postoperative myocardial infarction"*. In fact, no mappings contained in BioPortal exist with the terminologies used. Even if `PartialMatch` rule shows a good performance, some improvements are still possible. For

KOS	Setup	Method	ACC	AUC	F1
ICD9CM	Chapter 4	BK & `Rules+`	0.856	0.879	0.863
		`Rules+`	0.834	0.862	0.839
	Chapter 3	`Rules`	0.834	0.862	0.839
MeSH	Chapter 4	BK & `Rules+`	0.867	0.891	0.877
		`Rules+`	0.867	0.891	0.877
	Chapter 3	CombineAll	0.862	0.887	0.872
NCIT	Chapter 4	BK & `Rules+`	0.767	0.798	0.747
		`Rules+`	0.767	0.798	0.747
	Chapter 3	LCP & `Rules`	0.7	0.74	0.649
SNOMED CT	Chapter 4	BK & `Rules+`	0.828	0.833	0.8
		`Rules+`	0.849	0.854	0.829
	Chapter 3	BK & `Rules`	0.839	0.844	0.815

Table 4.8: Results regarding the adaptation of annotations of *Setup 5* during the period 2009/2010. The blue values indicate the best pipelines and `Rules+` indicates the eight used rules in our framework, while `Rules` refers to methods applied in chapter 3.

instance, the adaptation of *C11197:"folfox"* to *C11197:"folfox regimen"*, is not aligned to our silver standard. It should evolve to *C63590:"FOLFOX-4 Regimen"*, which also considers the suffix of this annotation. We will need to verify whether the inclusion of thresholds in future versions of `PartialMatch` overcomes this limitation.

The good performance of `Rules+` is also confirmed in the period 2009/2016. The results in Table 4.9 show that, except for SNOMED CT, it is capable of outperforming all the other techniques. Furthermore, it shows significant differences when compared with the results from Chapter 3. In NCIt the F1-Score is 13.34% higher than Chapter 3 ($V_1 = 0.667, V_2 = 0.756$) and for ICD-9-CM `Rules+` shows a positive difference of 17.26% ($V_1 = 0.643, V_2 = 0.754$).

The analyses of these adaptations demonstrated that the way these terminologies have been changed as well as their internal structure, have a remarkable influence on the adaptations. For example, in MeSH the reuse of CODEs and synonyms aids the adaptation method, mainly seen in the adaptations of *ChangeConceptAnnot*. In SNOMED CT the generation of new IDs which move the entire concept to another region or add new ones also have a positive impact. This was mainly verified for the adaptations using *ChangeConceptAnnot* and `PartialMatch` rule.

On the other hand, in ICD-9-CM which has a basic structure of a tree with a maximum depth of three without many synonyms, rules such as *ChangeConceptAnnot* or *IncreaseAnnot* did not have a good performance. The drawback is that the application of semantic techniques or string similarity methods do not aid in the maintenance task, as verified with other KOS. For example, the annotation *brain injury*:854.00 was extended in 2010 to *traumatic brain injury*:V80.01 which is located in a different region. Then, in 2011, it evolved again, because the concept V80.01, became more specific *"screening for traumatic brain injury"*. In our silver standard, the domain specialists decided to reduce the expressivity of this annotation by returning to the first concept, V80.01, and reducing the annotated text.

We verified that when applying the current `Rules` of our framework, we were unable to provide a good adaptation for this annotation. `PartialMatch` was not able to find a reasonable result since it produced weak results for semantic similarity between concepts V80.01 and 854.00. Furthermore, the string similarity value of "screening for traumatic brain injury" is higher than *"brain injury"* when compared to *"traumatic brain injury"*. Therefore, future versions also have to deal with the reduction of expressivity in annotations through multiple versions.

Finally, we compared our method Rules+ against the best method from Chapter 3, using the Sign Test [Dixon and Mood, 1946], see Table 4.10. In the first column of this table, we stated the KOS used, followed by the two compared methods in columns two and three. In columns four to six we have the signs obtained during the process of computing the Sign Test. Finally, in column seven we have the p-value which indicates whether one method differs from another. We are considering $p < 0.05$ as a significant difference.

The results of Table 4.10 show that NCIt had significant differences between the methods. Its *p-value* was lower than 0.05 leading us to refuse the null hypothesis (*Population median difference = 0*). It means that the Rules+ are able to maintain more annotations than LCP & Rules (method from Chapter 3).

We observed that few annotations remain invalid and marked as unsolved by our method. In such cases, the extension of the Rules+ to adapt these annotations is very complex. Basically, the concepts are in different ontology regions and use terms that are very different to those used in the annotations. More in-depth studies on string matching combined to semantic similarity measures are required to address the missing cases.

KOS	Setup	Method	ACC	AUC	F1
ICD9CM	Chapter 4	BK & Rules+	0.73	0.781	0.719
		Rules+	0.757	0.803	0.754
	Chapter 3	SCP & Rules	0.676	0.737	0.643
MeSH	Chapter 4	BK & Rules+	0.875	0.901	0.89
		Rules+	0.88	0.905	0.895
	Chapter 3	SCP & Rules	0.859	0.888	0.874
		CombineAll	0.859	0.888	0.874
NCIT	Chapter 4	BK & Rules+	0.75	0.8	0.75
		Rules+	0.753	0.804	0.756
	Chapter 3	BK & Rules	0.685	0.75	0.667
		SCP & Rules	0.685	0.75	0.667
		LCP & Rules	0.685	0.75	0.667
		CombineAll	0.685	0.75	0.667
SNOMED CT	Chapter 4	BK & Rules+	0.901	0.913	0.905
		Rules+	0.912	0.923	0.917
	Chapter 3	Rules	0.923	0.933	0.928

Table 4.9: Results regarding the adaptation of annotations for *Setup 3* during the period 2009/2016. The blue values indicate the best cases and Rules+ indicates the eight used rules in our framework, while Rules refers to methods applied in chapter 3..

4.8 Conclusion

In this chapter, we presented an approach that combined Lexical and Semantic measures to enhance the concept similarity. Our experimental analysis demonstrated that together they could outperform methods based on only one similarity measure (i.e., SSM or LSM). We observed that the use of lexical similarity approaches was important to improve the quality of the results

KOS	Method 1	Method 2	Signs			p-value
			+	NA	-	
ICD-9-CM	Rules Chapter 3	BK & Rules+ Chapter 4	17	149	21	0.6271
MeSH	CombineAll Chapter 3	Rules+ Chapter 4	2	190	3	1.0
NCIt	LCP/Rules Chapter 3	Rules+ Chapter 4	0	84	6	0.03125
SNOMED CT	Rules Chapter 3	Rules+ Chapter 4	0	92	1	1.0

Table 4.10: Results regarding the Sign Test. The values in blue refers to p-value (probability) $p < 0.05$ which indicates we rejected the null hypothesis (H_0 : *Population median difference = 0*) and red values $p \geq 0.05$, indicates we supported the null hypothesis.

provided by the SSMs and vice-versa.

Using an approach that combined LSM and SSM, we presented the `PartialMatch` rule for maintaining semantic annotations affected by the evolution of KOS. Our experimental analysis demonstrated that `PartialMatch` was capable of achieving good results to adapt annotations using one or multiple successive KOS versions. We observed that the use of semantic similarity approaches was important for determining the relatedness during the evolution process.

In the next chapter, we discuss how to keep annotations searchable without applying the direct maintenance approach and how to integrate this new method with the one described in Chapter 3 in a general architecture to maintain semantic annotations.

Chapter 5

Ad-hoc maintenance of semantic annotations

Contents

In this chapter we present the indirect maintenance approach, which addresses the problem of searching for annotations when the KOS used to annotate documents have changed, but the annotations cannot be updated. It answers RQ4 (*Which methods can be used to keep the annotations searchable when the document and annotations cannot be changed directly?*) and differs from the direct maintenance approach discussed in Chapter 3 in the sense that the indirect maintenance does not change the annotation.

This method applies to cases where: i) the annotations cannot be modified, ii) The documents are confidential and cannot be accessed (only the metadata can be accessed), and iii) The metadata are read-only. However, annotations impacted by KOS evolution must not be lost, methods to search for annotated documents are needed to execute tasks in areas such as public health, research, patient history, etc.

To address this problem we proposed a searchable knowledge base (KB) containing the KOS evolution history. It allows us to navigate through complex relationships related to the KOS evolution, e.g., *addConcepts*, *move*, *split*, and then search for documents by matching the metadata with concepts from past and present KOS versions.

The originality of the proposed approach relies on the fact that we exploit the evolution and structure of the KOS to construct a knowledge graph that can be used to enrich queries when searching for medical documents. For instance, when querying biomedical data sources, the queries are extended with information from the knowledge graph that we created with information about the KOS evolution. Existing approaches [Esch et al., 2015, Butt et al., 2015, Rashid and

Nisar, 2016, Lee et al., 2016, El-Dsouky et al., 2016, Roberts et al., 2016] only use the current version of the KOS to extract concepts and create queries. The evolution aspect is neglected and those dealing with historical data will probably get an incomplete set of results.

The chapter is structured in the following way: In Section 5.1, we introduce relevant notions to understand the problem of indirect annotation maintenance. In Section 5.2, we discuss the related work in the field of graph generator and information retrieval, highlighting the drawbacks and possible improvements. In Section 5.3, we detail our approach to managing the indirect maintenance of semantic annotations. In Section 5.5, we discuss the integration of both the direct and indirect maintenance methods. The Section 5.6 describes the methodology utilized to evaluate our approach, as well as the dataset utilized. Finally, in Section 5.6.1 and 5.6.2, we present the results and discuss them, respectively.

5.1 Problem Statement

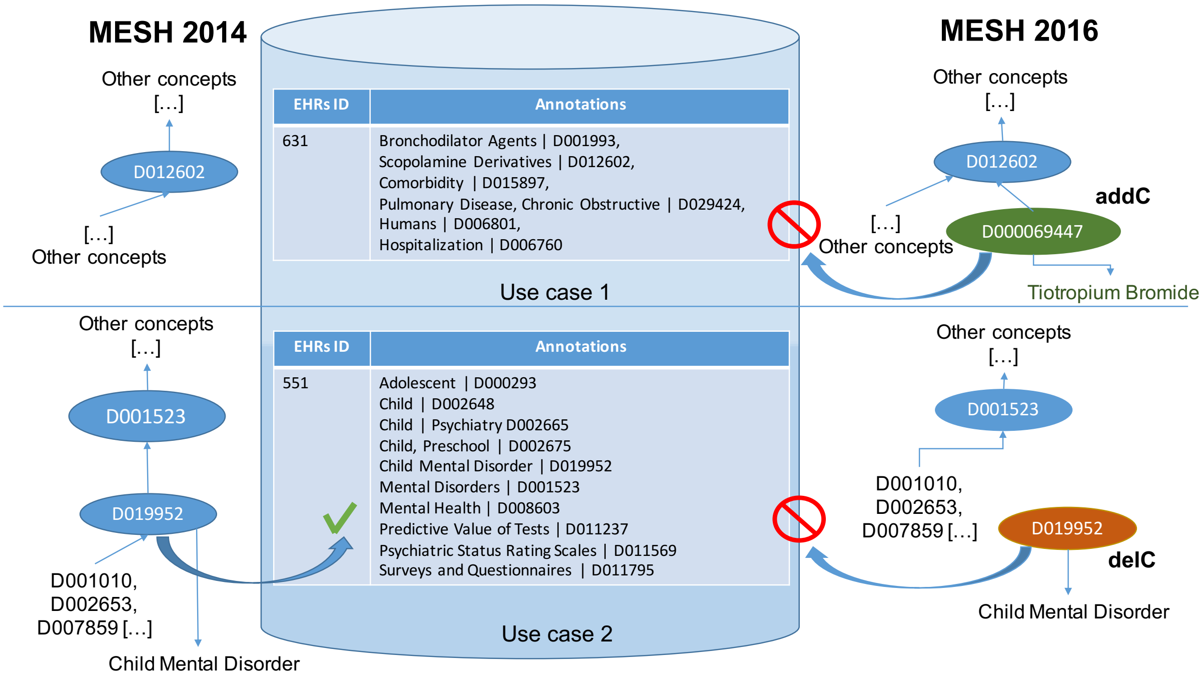
In our work, we split the problem of keeping semantic annotations searchable through our indirect maintenance method into two sub-problems: i) how to organize and store the information related to the KOS evolution; ii) how to use the stored information about KOS evolution to improve the quality of results when searching for documents over a long period of time. Notice that both problems are related to information retrieval tasks.

For the specific context of our experiments, we consider that our source of information contains only Electronic Health Records (EHRs). Thus, for readability purposes, we use the acronym EHR in this chapter to make reference to our data source. However, an indirect maintenance method can be applied to any data source having the following properties:

1. Composed of documents with metadata that are annotations;
2. Annotations linking a document (or part of a document) with concepts from an ontology;
3. Ontology evolving over time, so that different documents can potentially be annotated with different versions of the ontology;
4. Annotations using only use the terminological part of the ontology (e.g., concept, labels and synonyms), as highlighted in Section 1.2 (Data-driven Analysis).

The structure of the data source was also simplified for readability purposes. We assume that the information available is similar to the use cases presented in Figure 5.1. In this example, our knowledge base (KB) contains two annotated EHRs. The first column contains the *EHRs ID*, i.e., the identification of the patient report, with *Annotations*, i.e., the concepts describing the patient report, in the second column. The content of the patient report is not available.

Figure 5.1 illustrates two use-cases where the evolution of the ontology (in this case, MeSH) can impact access to the information. On the left-hand side of the Figure 5.1, we have the version 2014 of MeSH and on the right-hand side, the version 2016 of MeSH. The upper part of the figure shows the consequence of including a new concept in MeSH. The concept D000069447:*Tiotropium Bromide* (green colour) was added. As a consequence, from 2016 onwards, all EHR somehow related to the description of the drug *Tiotropium Bromide* will have this new concept in the annotation set. However, the EHRs metadata created before 2016 will not have this concept. If a user queries the system using this new concept, no EHRs created before 2016 (e.g. EHR 631) will be retrieved. Using our approach, the system will be able to return EHR 631, as well as other EHRs containing this annotation. The procedure for that will be detailed later on in this chapter, but in short, we created a knowledge graph capable of creating a link between the concept *Tiotropium Bromide* and the concept *Chronic Obstructive Pulmonary Disease (COPD)*. Thus, the query will be automatically enriched to include the latter concept for retrieving EHRs



EHRs ID	Annotations
631	Bronchodilator Agents | D001993, Scopolamine Derivatives | D012602, Comorbidity | D015897, Pulmonary Disease, Chronic Obstructive | D029424, Humans | D006801, Hospitalization | D006760

EHRs ID	Annotations
551	Adolescent | D000293 Child | D002648 Child | Psychiatry D002665 Child, Preschool | D002675 Child Mental Disorder | D019952 Mental Disorders | D001523 Mental Health | D008603 Predictive Value of Tests | D011237 Psychiatric Status Rating Scales | D011569 Surveys and Questionnaires | D011795

Figure 5.1: Knowledge base containing the immutable documents/annotations.

created before 2016 (this new query extension will not be applied to documents created after 2016), and EHR 631 will be part of the results of the search process. This procedure increases the recall of a search, but the precision can be penalized.

Use case 2, illustrated at bottom of Figure 5.1, is a more complex case. When a concept is deleted from the MeSH, the ontology-based search system can no longer create queries using this concept. Thus, in 2016, end-users will not be able to search for EHRs using the term *Child Mental Disorder*. To keep documents containing this term searchable and the previous version of concept D019952 reusable, we propose to use our knowledge graph to go back to the last version of MeSH where the concept D019952:*Child Mental Disorder* exists (i.e., 2014), select a set of related concepts that can replace D019952 in queries made in 2016, then use this set of concepts to enrich the query and retrieve recent EHRs related to EHR 551.

Thus, when using outdated ontologies to query EHRs annotated with the last version of the ontology, part of the information become unsearchable, as illustrated in Figure 5.1. We need a system that can address this problem. For instance, an end-user that types a query with concepts from MeSH version 2014 should be able to find relevant documents annotated with MeSH 2016 and vice-versa.

The current approaches supporting the maintenance of semantic annotations only work with direct maintenance, as discussed in chapter 3. In this chapter, we propose a new `Ad-hoc` approach that combines methods from other domains (e.g., Evolving Graph Generator (EGG), and Information Retrieval (IR)) to tackle the problem of indirect maintenance.

Our approach deals with the refereed problem in a two-step process. First, we built a knowledge graph through EGG techniques that describe the KOS evolution and its complex/temporal relationships (detailed in Section 5.3). Then we develop a search algorithm, based on the state-of-the-art of IR domain, to exploit the knowledge graph and enrich queries in order to retrieve the documents over a long period of time. The major gap when using these approaches alone is the temporal/evolutionary aspect, further discussed in the related work section below.

5.2 Related Work

Several domains, such as medicine, social sciences, finance, etc. represent data represented as graphs. We can also consider the part of the KOS that we are using in this work as a graph (we refer to it as the KOS graph). Graphs provide flexibility to represent the domain knowledge with its complex relationships. In this work, we used graphs to represent the evolution of annotations. For instance, in the approach presented earlier to maintain annotations (chapter 3 and 4), we used the adapted W3C annotation model (see Figure 2.11) to represent the evolution of an annotation *annot2* from its old version *annot*. The newly created relationship *evolved* was used as following: `:annot2 oa:evolved :annot`. This representation format allows us to process a wider range of queries to exploit the graph, for instance, to calculate the shortest path between two annotations that have evolved over a period of time (e.g. 2009 to 2014).

The work presented in this chapter was inspired by the Evolving Graph Sequence (EGS) domain. EGS utilizes and analyses many graph snapshots from various periods of time [Kosmatopoulos et al., 2016] and connects them to represent their evolution. Recent works in EGS mostly deal with i) the modelling of the graph in order to reduce the storage space of multiple snapshots [Caro et al., 2015, Moffitt and Stoyanovich, 2017]; ii) historical reachability queries, which compare whether two nodes of a graph are connected over time [Akiba et al., 2014, Semertzidis and Pitoura, 2016]; iii) efficient snapshot retrieval, i.e., index management of large historical evolving graphs, in order to speed up the retrieval process [Kirsten et al., 2009, Khurana and Deshpande, 2013, Labouseur et al., 2015].

A common aspect of EGS approaches is that the information from these graphs, e.g. nodes, relationships, validity periods are already defined and no further inferred data is needed. However, in some particular cases more information must be included, e.g. the complex relationships related to the KOS evolution (see Section 5.1) [Alami et al., 2017]. These evolution aspects lead to the development of more sophisticated techniques than those proposed in Evolving Graph Generator (EGG) domain.

In EGG, [Bagan et al., 2017] proposed *gMark*, a framework to generate synthetic graphs and query workloads. Using a *graph configuration* file, which contains i) the number of predicates and nodes, as well as their properties, *gMark* builds a synthetic graph for the informed configuration. This setup removes the technical constraints that are commonly hardcoded and makes the customization of such systems by domain specialists more difficult when applied in other domains.

[Alami et al., 2017] proposed an extension of *gMark* that includes temporal constraints given by the user. For instance, when creating a named graph in TriG format[21] `:G1 { :hotel2 a ex:Building ; ex:availableRooms "12"^^<rdfs:double> .}`. The number of available rooms computed for `<hotel2>`, can assume values in the interval of [1,100] from one snapshot to another.

The main drawback regarding these two approaches [Bagan et al., 2017, Alami et al., 2017] is the difficulty of generalizing them for the evolution of biomedical KOS graphs. Since the KOS graph is already there and we do not want to recreate the concepts or instances, we need to focus only on the information relating to its evolution to enrich the KG, e.g., `:G1 { :D000069447 a ex:KOSConcept ; ex:ChangeType addConcept; ex:date "2003-10-02"^^xsd:date ; ex:knows _: D019952 . }`. This example includes a new connection between the concept source D000069477 and the concept target D019952.

K-NN graph-building techniques update existing graphs by attaching new nodes/relationships to their K nearest neighbours. The problem with these approaches is the lack of consideration of temporal aspects as well as their performance when processing multiple versions [Debatty et al., 2016]. Another aspect of K-NN that makes this technique hardly applicable to our indirect

[21] `https://www.w3.org/TR/trig/`

maintenance approach, (see Section 5.1) is the way information retrieval algorithms are currently implemented in existing approaches that do not deal with temporal aspects.

In EGS, existing approaches [Akiba et al., 2014, Akiba et al., 2015, Semertzidis and Pitoura, 2016, Semertzidis and Pitoura, 2018] only work with historical reachability queries, i.e., given two concepts X and Y, they check whether both concepts are connected over a period of time. In our case, we need to deal with only one concept as input (we assume that end-users do not need to know the equivalent of concept X in previous KOS versions).

Most generic approaches in the biomedical field, which implements the exploratory search [Esch et al., 2015, Butt et al., 2015, Rashid and Nisar, 2016, Lee et al., 2016, El-Dsouky et al., 2016, Roberts et al., 2016] only work with the current KOS version, neglecting the historical/evolutionary aspect of the KOS.

In the literature review, we did not find works dealing with the indirect maintenance of semantic annotations. Furthermore, none of the related approaches, using graphs, are completely adequate for the aspects involved in the evolution of KOS and/or exploratory search problems (as explained before). In this chapter, we present our approach to overcoming these problems. We built an `Ad-hoc` searchable knowledge base considering the evolution of KOS to overcome the problem of indirect maintenance.This approach is based on K-NN graph building algorithms [Dong et al., 2011, Debatty et al., 2016] as shown in the next section.

5.3 Knowledge graph to represent evolving ontology

Figure 5.2 gives a simplified overview of our knowledge graph (KG) representing evolving ontology. The turquoise and dark blue boxes represent stable and changed concepts, respectively. As illustrated in Figure 5.2, the initial version of our KG, i.e., MeSH 2009, contains concept D009133:*Muscular Atrophy*. In 2010, a new related concept was added to specialize the existing one (D055534:*Spinal and Bulbar Muscular Atrophy*). To represent this evolution, we created a *highLvlChg* relationship between both concepts. Thus, documents associated with D009133 and/or D055534 can be retrieved if we find one of these two concepts in the query (e.g., *Muscular Atrophy*).

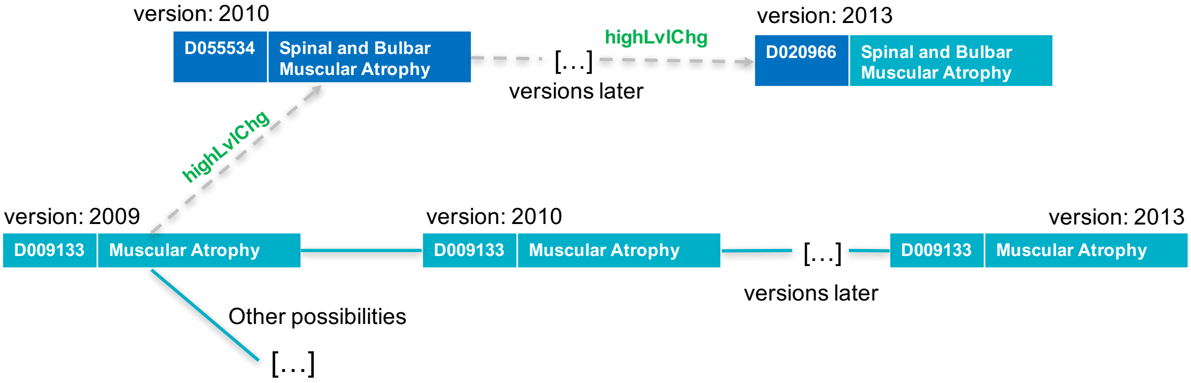


Figure 5.2: The proposed `Ad-hoc` history of concepts. The turquoise and dark blue boxes indicate the changed and stable concepts, respectively.

We propose a KG that deals with the multi-versioning and/or evolution of KOS. Figure 5.2 shows the evolution of concept D055534 to D020966 during the period 2010 to 2013 (this concept remains stable until 2012 and moved to another region of the KOS in 2013). To represent this evolution, we created a *highLvlChg* relationship between both concepts allowing us to navigate through past and current versions of concepts. Thus, even if the concept D055534 was moved to another region of the KOS, we are still able to use the KG to retrieve the previous versions of the concept (*D055534* and *D009133*). To perform this navigation, we added the following features to the KG, where the features described in [Debatty et al., 2016] were extended by additional ones to cope with the evolutionary aspect of the KOS:

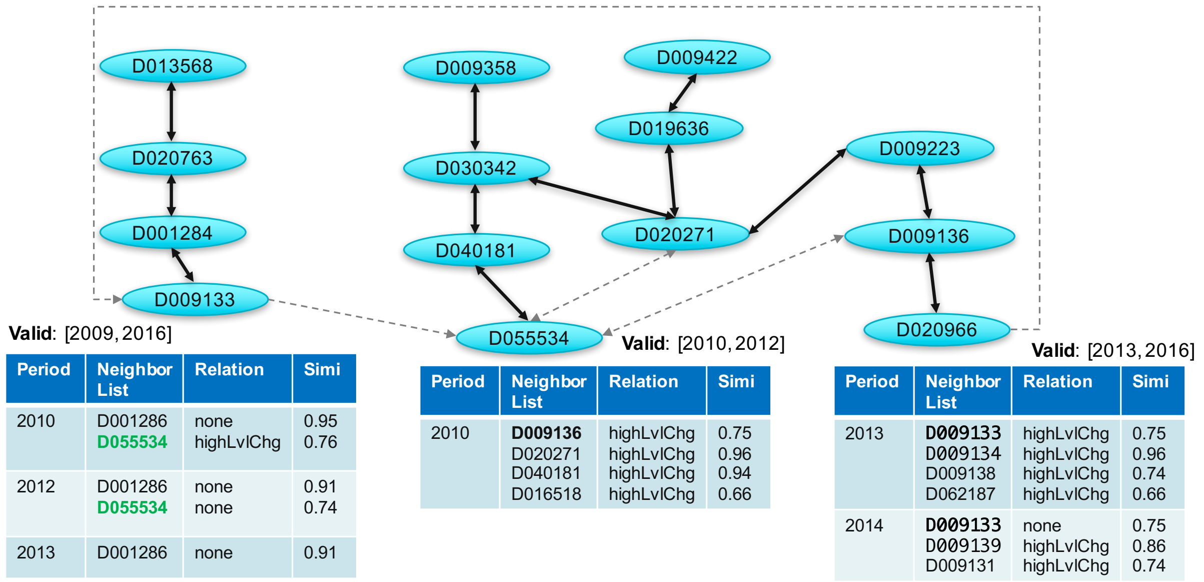
- **Edge directions**: Considering concepts as vertices of a graph, Edges materialize the relationship between concepts. For example, in Figure 5.3, concept D009133:*Muscular Atrophy* and D001284:*Atrophy* are related in our graph. Since our KG is a digraph [Debatty et al., 2016], one can distinguish between the subject and the object of the relationship. In this work, we consider the **structural relationships** that are *superClass*, *subClass*, *siblings* and *none* as uniquely labelled edges, depicted as black arrows in Figure 5.3. Regarding the concepts that emerged from the evolution of the KOS and their connection within the KG, we utilized the same principle of digraphs, but the connections may not follow the ontology structure. For instance, concept *D055534* in Figure 5.3 is connected to its *superClass* and shares some similarities with two other concepts *D009136, D016518*, from other regions of the ontology. These connections are illustrated in dashed grey arrows in Figure 5.3 but are associated with vertices in our KG.

- **Similarity value**: It indicates the degree of similarity between two concepts (or two versions of the same concept). We used the hybrid measure described in Chapter 4 to compute it. When a new KOS version is added to our KG, for each pair of connected vertices, the value of the similarity is either calculated or update as depicted in Figure 5.3.

- **Validity periods**: Versioning and storage capacity is an important feature present in our KG. To reduce the required storage capacity, we used methods like those described in [Caro et al., 2015, Moffitt and Stoyanovich, 2017]. These methods labeled the validity period of concepts and their relationships on the graph nodes and edges, see Figure 5.3. Applying this method avoid duplicating the whole KOS into the KG for every new version.

- **Relationships**: In order to include more semantics in our KG, we created two types of semantic relationship: **evolutionary relationship** associated with vertices; and **structural relationship** associated with edges. Evolutionary relationships are *highLvlChg*, *lowLvlChg* and *none. highLvlChg* includes *delC*, *addC*, *split*, *move* and *chgAttValue*; *lowLvlChg* includes *delA* and *addA*; *none* means that the connected vertices had a KOS change at some point in its evolution. Structural relationship are *superClass*, *subClass* and *siblings.* Figure 5.2 shows only the evolutionary relationships while Figure 5.3 shows both. For instance, *highLvlChg* is an evolutionary relationship indicating that, from one version to another, a major change in the KOS was observed (in this case, a concept was added). Figure 5.3 uses *Super* to indicate that concept D001284 subsumes concept D009133. The importance of having these two types of relationships becomes evident when the system needs to define strategies to enrich queries. For instance, a query using the term *Spinal and Bulbar Muscular Atrophy*, from MeSH 2013, will not find documents before 2010. But, knowing the history of this concept, the query can be enriched to return all documents, created before 2010, that also contain the term *Muscular Atrophy.*

  When the query contains outdated concepts, the KG can be used to include additional terms in the query. For instance, consider the situation where one system uses MeSH 2009 to request documents containing the concept D009133 to another system that uses MeSH 2013 to annotate its documents. Thus, using the KG, the query can be enriched with the concepts D055534, D009136 and D020966, connected through the path (D009133, D055534, D009136, D020966) in the period [2009, 2013], to retrieve documents associated with *Spinal and Bulbar Muscular Atrophy* and *Muscular Atrophy.* In such cases, the relation and similarity values become very helpful for selecting additional terms (see Figure 5.3).



Valid: [2009, 2016]

Period	Neighbor List	Relation	Simi
2010	D001286 **D055534**	none highLvlChg	0.95 0.76
2012	D001286 **D055534**	none none	0.91 0.74
2013	D001286	none	0.91

Valid: [2010, 2012]

Period	Neighbor List	Relation	Simi
2010	**D009136** D020271 D040181 D016518	highLvlChg highLvlChg highLvlChg highLvlChg	0.75 0.96 0.94 0.66

Valid: [2013, 2016]

Period	Neighbor List	Relation	Simi
2013	**D009133** D009134 D009138 D062187	highLvlChg highLvlChg highLvlChg highLvlChg	0.75 0.96 0.74 0.66
2014	**D009133** D009139 D009131	none highLvlChg highLvlChg	0.75 0.86 0.74

Figure 5.3: KG proposed for the indirect maintenance of semantic annotations. The black arrows indicate the connection following the ontology structure, while the dashed grey lines indicate the connections created by our approach. Each node contains the validity period and its neighbours. For each triple (node, period, neighbour), one relation and similarity describe their link.

We formalize our KG as a direct graph $G = (V, E)$, where V is the set of vertices and E the set of edges. The set of vertices is denoted by:

$$\begin{aligned} V = \{(c, p, \mathrm{NL}) | c \in O, p \in \mathbb{N}, \\ \mathrm{NL} = \{(c_i, RE, simi_V) | c_i \in O, RE \in \{highLvl, lowLvl, none\}, simi_V \in \mathbb{R}\}\} \end{aligned}$$

where, c is a concept from an ontology O in a period p, e.g. 2009, containing a neighbour list NL inferred by the KNN graph approach of Debatty. Each NL contains a concept target c_i; the relation emerged from the evolution RE, whose values $highLvl$ and $lowLvl$ denote some KOS change and $none$ indicates that they are connected because they have a high similarity value $simi_V$. The set of edges E is denoted by:

$$E = \{(u, v, p, SR, simi_E) | u, v \in V, p \in \mathbb{N}, SR = \{super, sub, sib\}, simi_E \in \mathbb{R}\}$$

Where, u and v are vertices belonging to V and are connected during a period p, e.g. 2009. Since edges respect the ontology structure, each connection has a semantic relationship SR, which is one of the values $superClass, subClass$ and $sibling$. Finally, $simi_E$ represents the similarity between u and v.

5.4 Indirect maintenance of semantic annotations

This Section describes the method for the indirect maintenance of semantic annotations. We start describing the workflow of Figure 5.4 with the method to construct the KG, called *Offline phase*, then, we detail a use-case implementing this KG for the indirect maintenance of semantic annotations, called *Online phase*.

5.4.1 Offline phase

It starts by taking as its input the multiple KOS versions provided by the user in OWL. In the *Build enhanced KOS* process, we use the first KOS version as a bootstrap and enrich it with

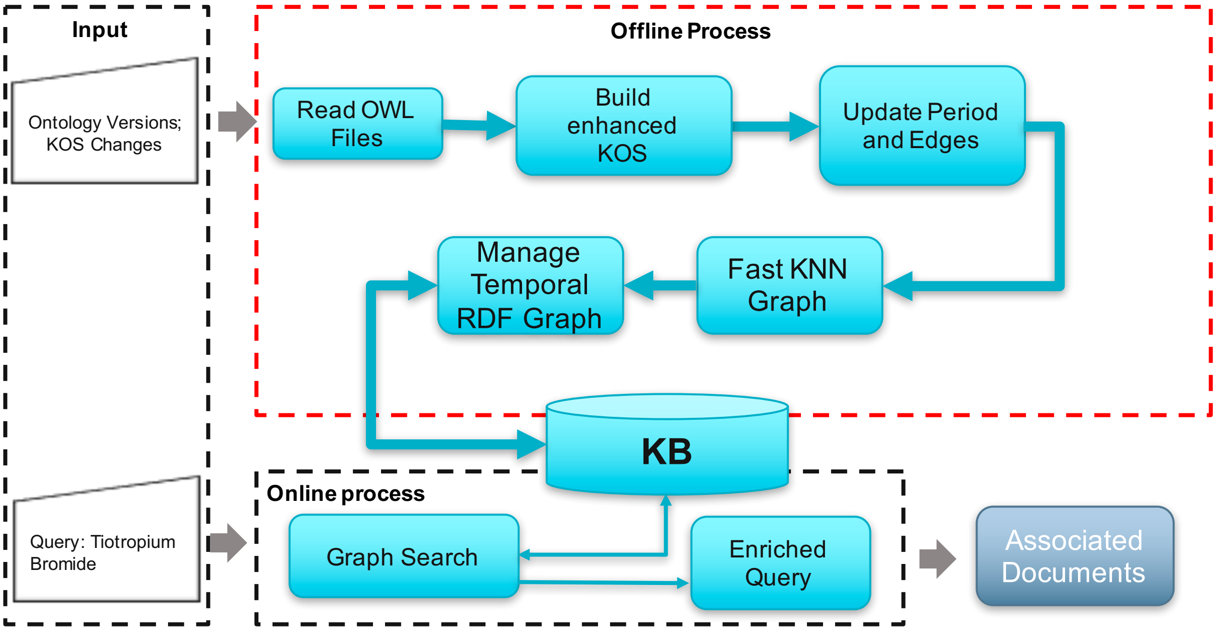


Figure 5.4: Workflow to compute the indirect maintenance of annotations.

the features discussed in Section 5.3, i.e. *edge directions*, *similarity values*, *valid periods* and *relationships*. The resulting graph serves as an initial KG that will iteratively be enriched with information acquired from the next versions of the KOS considered. To update the validity periods and edges of each concept, we first compute the ontology changes using Conto-Diff [Hartung et al., 2013] in consecutive versions. Then, we update the concepts and relationships in two steps:

1. Update Concept Period: In this case, if a concept is not present in the KOS changes, we assume it belongs to both versions. Therefore, we update the validity period accordingly, e.g., D009133:*Muscular atrophy* is not present in the KOS changes of 2010 then its corresponding validity period is [2009, 2010]. In contrast, the validity period of concept D016821:*Phytomastigophorea*, deleted in 2010, is [2009].

2. Update Period and Edges: After computing all the concepts' validity periods, we start to walk through the KG to update the similarity, relationship and period of edges, considering two cases.

   (a) The source and target concepts belong to the same hierarchy and have the same validity period. We then update the structural relationship associated with the considered edge's information and similarity.

   (b) The concept source has changed or all of its neighbours have changed. We then add it to a temporary list to be updated during the *Fast Knn Graph* process.

At that stage, we have the KG containing all up-to-date structural information and a list of concepts to be computed in the *Fast Knn Graph* process. In our example implementing MeSH 2009/2010, the *Fast KNN Graph* process will update the Neighbour List of all changed concepts containing one of the following KOS changes: *split*, *move*, *addC*, *addA*, *addLeaf* and *addInner*.

The *Fast KNN Graph* process follows Debatty's approach [Debatty et al., 2016] to update the NL list from a given impacted concept. This two-step process is as follows:

1. Search for the k-nearest neighbours of a given node in the graph. This is done using the improved Graph Nearest Neighbour Search (iGNNS) algorithm [Debatty et al., 2016], which iteratively explores the neighbours of neighbours down to a fixed depth.

2. *Update the selected neighbours*: This phase adjusts the Neighbour List of a given source and target concepts using the k-nearest neighbours found by the iGNNS.

The computational cost of updating a node in Debatty's approach is $O\left(\frac{n}{speedup} + k^{DEPTH}\right)$. In this equation n is the size of the graph and *speedup* is a parameter to tune the algorithm, in the sense that it reduces the number of elements to compare when finding the k-nearest neighbours. Finally, k^{DEPTH} refers to the maximum depth that the algorithm will walk through to find the most similar vertices of the graph. Depending on the size of the graph and the *speedup* parameter, this approach can be time consuming. Therefore, we enhanced this approach by including the following modifications:

- *Lexical View*: We reduced the complexity of computing a new node of Debatty's algorithm by implementing the lexical view discussed in Section 4.4. In [Debatty et al., 2016], to find the k-nearest neighbours, $\frac{n}{speedup}$ comparisons are necessary. The inclusion of the lexical view will reduce the number of comparisons from $\frac{n}{speedup}$ to $\frac{lexicalView(node_i)}{speedup}$, where $lexicalView(node_i)$ is the amount of concepts returned by the Lucene search engine[22]. In the worst case scenario, i.e., the concepts do not have *lexicalView*, the number of comparisons remains $\frac{n}{speedup}$.

- *Temporal Aspects*: [Debatty et al., 2016] utilizes a tuple {node, NeighborList} to access the graph. We included a third feature *Period* that indicates the validity of the edge and node, respectively. Thus, our KG represent the nodes and their connections via a triple {node, Period, NeighborList}. The nodes/connections that are not valid in a given period are excluded from the list of candidates.

The last process from the offline phase deals with the *Management of Temporal RDF Graph*. It stores and manages the generated KG. From a practical point of view, the KG is represented in a TriG format [23], which allows a RDF Dataset to be written in a compact textual format. Moreover, the number of nodes in this KG is calculated by the sum of the impacted concepts over time plus the number of nodes from the last KOS version: $(\sum_i^n impactedConcepts_i) + lastVersion$.

5.4.2 Online Phase

This consists of querying our KG using a set of terms from any KOS version as input to retrieve the history of those terms. To access the KG we utilized the iGNNS algorithm proposed by Debatty, however any search algorithm can be used to interact with the knowledge graph. To perform the search, we included the following modifications in Debatty's algorithm:

- *Filters*: When querying the KG, users can specify the type of relationship to be used.

- *Periods*: Our algorithm explores the evolution of the KOS in two ways:

    1. We look for the related concepts in preceding versions, e.g. from 2009 until 2013. In this case the starting point of our KG is t and we incrementally search the neighbours of $t+1$, $t+2$ until the last version available in our KG.

[22] https://lucene.apache.org/core/

[23] https://www.w3.org/TR/trig/

2. The search is performed in anterior versions, e.g., 2012, 2011, 2010. Similarly, we define a starting point t to match the query with our KG and incrementally search the neighbours having $t-1$, $t-2$ until the first available version.

During this search, we consider only the top-k most similar neighbours to go deeper into the graph. For example, using a Neighbour List of 15 and the top-5 most similar neighbours, we consider only the subset of neighbours most similar to the triggered concept, i.e., the one used to reach these neighbours to go deeper into the graph.

- *Top-k*: Many concepts can be retrieved by the search. Therefore, we restricted the algorithm to return the top-k most similar concepts regarding the query (first matched concepts). This top-k configuration is informed by the user, but we also restricted the algorithm when a similarity equal to 1.0 is found in the results. In this case, we return the exact amount of concepts having similarity equal to 1.0 instead of the top-k.

After retrieving the most similar concepts given a query specified by the user, our workflow returns the possible concepts as output to enrich the query. These concepts will later be used to retrieve documents in health facilities. At this stage domain experts can access the EHRs stored in a database (see Figure 5.1) following their own implementation.

5.5 Maintenance of Semantic Annotations, the MAISA framework

In this Section, we describe MAISA, our final framework to maintain semantic annotations valid over time when the underlying KOS change. MAISA (Maintenance of Semantic annotations) encompasses the methods discussed in Chapter 3 as well as the indirect maintenance method described in Section 5.4.

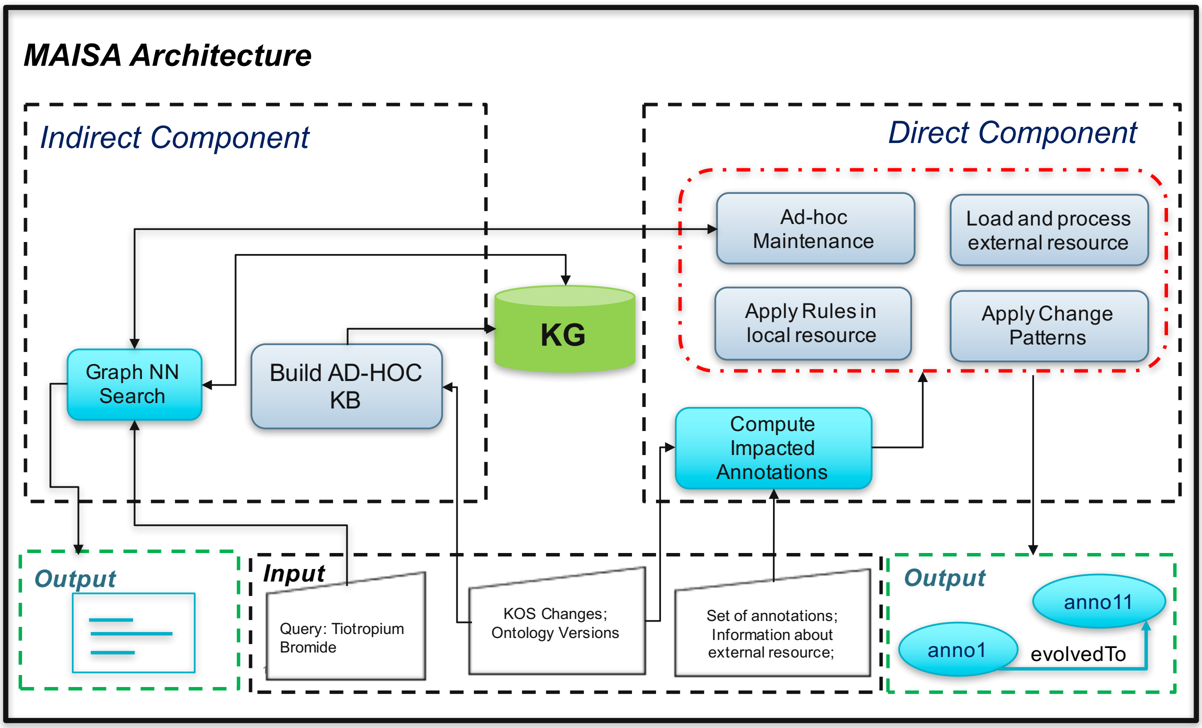

MAISA Architecture
Indirect Component
Direct Component
Ad-hoc Maintenance
Load and process external resource
Apply Rules in local resource
Apply Change Patterns
KG
Graph NN Search
Build AD-HOC KB
Compute Impacted Annotations
Output
Input
Query: Tiotropium Bromide
KOS Changes; Ontology Versions
Set of annotations; Information about external resource;
Output
anno11
anno1
evolvedTo

Figure 5.5: The MAISA Framework.

Figure 5.5 illustrates the architecture of the MAISA framework. In the upper left part, we have the indirect component discussed in this chapter. The grey and blue boxes refer respectively to the *offline* and *online* phase discussed in Section 5.4. The green box in the middle refers to our `Ad-hoc` knowledge graph that contains the KOS history.

As previously discussed, our indirect maintenance method takes as its input a term, representing user's queries, to enrich with information dealing with the history of the concepts associated with this term. We implemented a Restfull service to provide these concepts and leaft the implementation of accessing the EHRs to health facilities. These split processes provide more security when utilizing our method to query, because no patient data is needed and no access to sensitive database is required.

Regarding the direct component illustrated in the upper right part, the grey boxes refer to the maintenance methods from Chapter 3, while the blue box concerns the process that identifies the invalid annotations by analysing the evolution of the associated KOS (see Section 3.2).

The difference in this component regarding the workflow described in Chapter 3, is the inclusion of the *Ad-hoc Maintenance* method to adapt the annotations. This adaptation utilizes the result from *Graph NN Search* from the indirect component, but here we search for the most similar concept. This new method was adopted as a second option for the use of `BK` method due to the lack of available mappings between ontologies.

Finally, inputs and outputs illustrated at the bottom of Figure 5.5 remain the same from those discussed in Chapter 3 and Section 5.4, i.e., we continue using a *Set of annotations*; *Set of KOS changes*; *Set of ontologies*; *Information about external resources* for the direct maintenance, while in the indirect maintenance we utilize the *Ontology Versions*; *KOS Changes* and *keyword query*.

5.6 Material and Methods to evaluate the direct and indirect approaches

This section describes the resources and methods used to evaluate our direct and indirect approaches. The main resources are: i) the terminologies; ii) the data set utilized to evaluate our approach; iii) the method used in the indirect maintenance and iv) the metrics for evaluation.

Terminologies

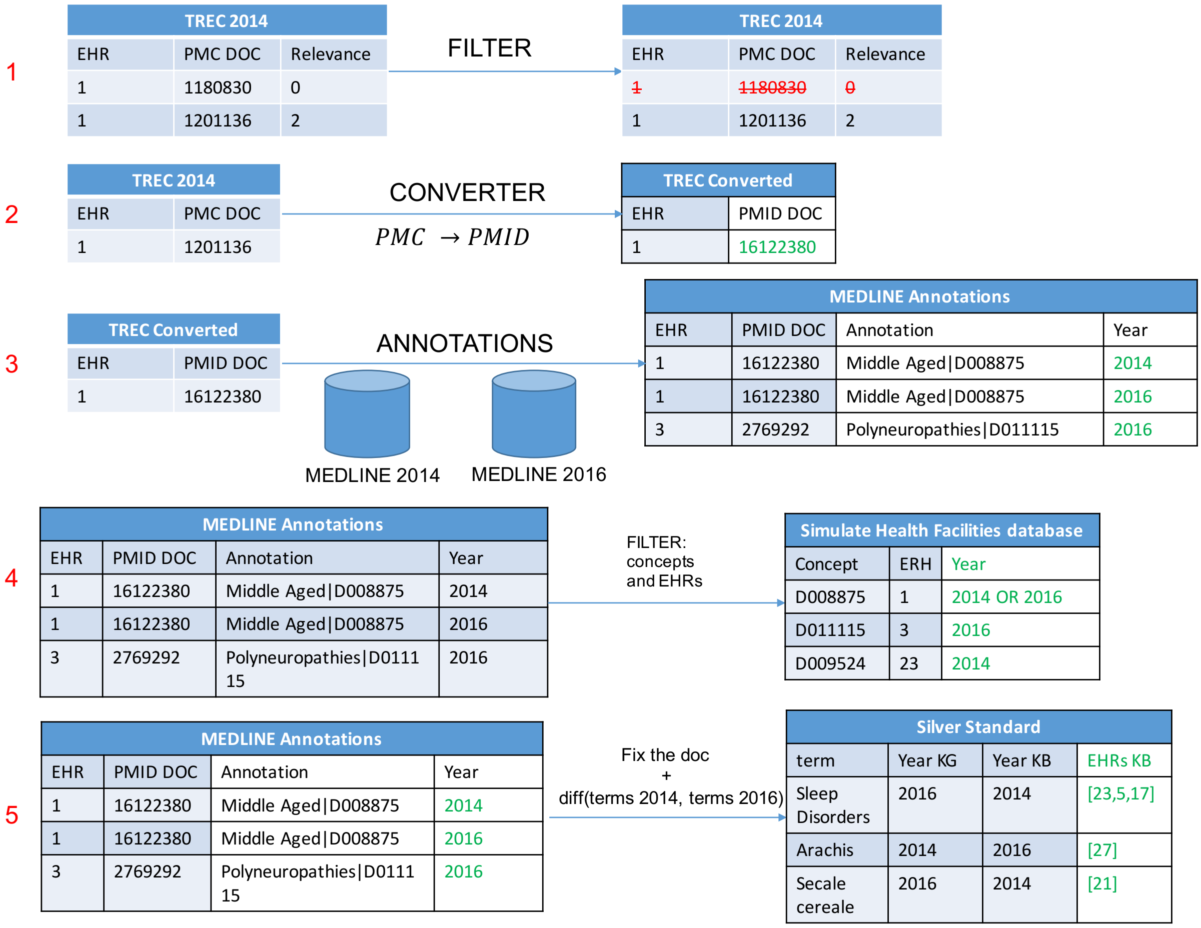
As described in Section 5.3, our method implements consecutive KOS versions in order to build an `Ad-hoc` knowledge graph that is used in the indirect maintenance approach. In our experiments, we used: Medical Subject Headings (MeSH) versions 2014AA to 2016AA (excluding the AB versions), downloaded from UMLS and transformed by ourselves into OWL files. To compute the difference between the terminologies, we use COnto-Diff [Hartung et al., 2013].

Silver Standard

Since no annotation baseline representing annotated EHRs in different periods exists, we had to build our own corpus of reference. Figure 5.6 illustrates the following methodology:

1. We selected the EHRs from the TREC Clinical Decision Support Track version 2014[24]. In this corpus each row has one domain-specialist opinion indicating whether a document is related to an EHR. The value 0 indicates no relation, while 2 means strong relation. In our silver standard, we utilized all rows that have a value equal to two in order to avoid narrow or non-related documents.

[24] `http://www.trec-cds.org/2014.html`



TREC 2014		
EHR	PMC DOC	Relevance
1	1180830	0
1	1201136	2

TREC 2014		
EHR	PMC DOC	Relevance
~~1~~	~~1180830~~	~~0~~
1	1201136	2

TREC 2014	
EHR	PMC DOC
1	1201136

TREC Converted	
EHR	PMID DOC
1	16122380

TREC Converted	
EHR	PMID DOC
1	16122380

MEDLINE Annotations			
EHR	PMID DOC	Annotation	Year
1	16122380	Middle Aged|D008875	2014
1	16122380	Middle Aged|D008875	2016
3	2769292	Polyneuropathies|D011115	2016

MEDLINE Annotations			
EHR	PMID DOC	Annotation	Year
1	16122380	Middle Aged|D008875	2014
1	16122380	Middle Aged|D008875	2016
3	2769292	Polyneuropathies|D011115	2016

Simulate Health Facilities database		
Concept	ERH	Year
D008875	1	2014 OR 2016
D011115	3	2016
D009524	23	2014

MEDLINE Annotations			
EHR	PMID DOC	Annotation	Year
1	16122380	Middle Aged|D008875	2014
1	16122380	Middle Aged|D008875	2016
3	2769292	Polyneuropathies|D011115	2016

Silver Standard			
term	Year KG	Year KB	EHRs KB
Sleep Disorders	2016	2014	[23,5,17]
Arachis	2014	2016	[27]
Secale cereale	2016	2014	[21]

Figure 5.6: Methodology to build the silver standard of annotated EHRs.

2. The document identifier present in the TREC corpus is associated with the documents available on the PubMed Central (PMC) repository, i.e., using this identifier we can retrieve a document from the PMC website, e.g., the document 1180830[25]. These documents represent the subjects (drug descriptions, disease definitions, etc.) related to the EHR, but no metadata are associated with them. Thus, we had to use a converter[26] which translates the PMC identifier to a PMID identifier. This PMID identifier allows us to retrieve the annotations from MEDLINE/PubMed Baseline.

3. The MEDLINE/PubMed Baseline[27] is a dataset of annotations related to PubMed papers. These annotations are generated each year and represent a static view of the data each time a baseline is released. Thus, the MeSH vocabulary updates do not change annotations from previous versions. In our silver standard, we are using the annotations from versions 2014 and 2016, gathered by associating the PMID identifiers from previous phase with the MEDLINE baseline.

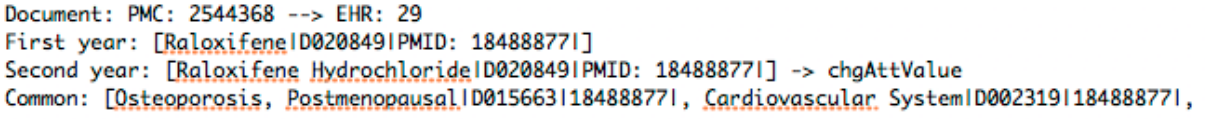
4. After gathering all annotations, we built a knowledge base (KB). Notice that this KB is different from our KG; the former contains the set of metadata that we are using to evaluate our approach while the latter contains a graph describing the evolution of an ontology. The KB is composed of the following data [Concept, EHRs, Years], which indicate

[25] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1180830/
[26] https://www.ncbi.nlm.nih.gov/pmc/pmctopmid/
[27] https://mbr.nlm.nih.gov/

which concept is related to which EHR in a specific year. We used it to simulate the environment described in Section 5.1 where the query is different from the annotated EHRs. Nevertheless, health facilities will have their own storage format for these associations and our format will be only used for evaluation purposes.

5. Using the annotations collected, we built a set of reference queries to evaluate our method. We first grouped a set of documents from MEDLINE annotations and performed a diff between their annotations. To illustrate what this diff looks like, in Figure 5.7 we included an extract of our dataset. It contains the annotation [*Raloxifene*:D020849] created in 2014 and referring to document 2544368 (EHR 29). The diff shows that in the second year (2016), there was a change in the attribute value and the annotation became [*Raloxifene Hydrochloride*:D020849]. The common annotations emerging from this diff corresponded to the data present in both years. Since we search for differences to be in line with the example of Section 5.1, the common annotations were discarded from our queries. To generate our queries, we used the term (from the annotation diff), the year to interact with in our KG, the year to retrieve the documents in the KB from the previous phase and the EHR ids. For instance, at bottom of Figure 5.6, we elaborated a query with the term *Sleep Disorders*, which will start interacting with the KG in version 2016 and retrieve the associated EHRs [5, 17, 23] from the simulated KB in version 2014.

The goal of our `Ad-hoc` method is to verify whether the concepts returned from the *Fast search* process, when interacting with the `Ad-hoc` KG described in Section 5.3, allow a complete and precise retrieval of documents. The dataset is available at `https://git.list.lu/ELISA/AnnotationDataset`. It contains a total of 23 queries to evaluate the scenario, query 2014 and KB 2016, while for the opposite, query 2014 and KB 2016 have 27 queries.



```
Document: PMC: 2544368 --> EHR: 29
First year: [Raloxifene|D020849|PMID: 18488877|]
Second year: [Raloxifene Hydrochloride|D020849|PMID: 18488877|] -> chgAttValue
Common: [Osteoporosis, Postmenopausal|D015663|18488877|, Cardiovascular System|D002319|18488877|,
```

Figure 5.7: Example of the diff between the terms from MEDLINE annotations; the first year refers to 2014 and the second year to 2016.

Experimental Setup

In order to evaluate the indirect maintenance, we implemented two different setup described below.

Set-up 1: We search for EHR documents. To do this, we used: i) the `Ad-hoc` knowledge graph described in Section 5.3; ii) the queries from our silver standard, and iii) the simulated KB that represents the database from health facilities.

We are running the queries from our silver standard considering two opposite cases: i) terms taken from KOS version 2014 verified on EHRs annotated in 2016 and ii) terms extracted from KOS version 2016 evaluated on EHRs annotated in 2014. Both cases implement the *Fast Search* process, which interacts with the `Ad-hoc` KG built with the following number of nearest neighbours in {2, 4, 6, 10}. We used two similarity measures: i) the best hybrid approach described in chapter 4 and ii) only a single semantic measure.

In order to increase the precision of the search, we applied the filters described in Section 5.4, by building the power set of the filters, i.e., the set of all subsets of a set. Moreover, we only selected the top-k most similar concepts considering $k \in \{5, 10, 15, 20\}$.

The experimental configuration $Conf$ listed in the results (cf Section 5.6.1) can be expressed as:

$$Conf = (method\_K\_x\_d\_y\_f\_name\_top\_k\})$$

where:

- **method** = {combined/semantic}: denotes the hybrid or single semantic measure used to build the `Ad-hoc` KG.
- **K_x**: is the number of **x** nearest neighbours used to build the `Ad-hoc` KG.
- **d_y**: is the **y** maximum depth specified for the search process.
- **f_name**: is the **name** of the filter utilized when querying the `Ad-hoc` KG. In this field each filter is described by its three first letters, e.g., *hig* means we are considering only concepts that have *highLvl* relationship.
- **top_k**: is the **k** most similar concepts returned.

Set-up 2: We applied the indirect maintenance process, over the silver standard described in Chapter 3 using version 2009/2010. To do this, we selected the best configuration obtained for **set-up 1** and computed the evolution of the annotations. With this configuration we evaluate: i) the capacity of our framework to detect impacted annotations after a change in the KOS, and ii) the ability to correctly turn the impacted annotations into consistent ones. The results of the `Ad-hoc` method are compared to the `BK` method of the direct method, since we proposed it as a second option for KOS lacking available mappings on the web.

Metrics

To evaluate the effectiveness of our indirect maintenance method for retrieving EHRs, we used the average of Precision, Recall, F1-score and fall-out, i.e., the percentage of non-relevant documents retrieved in a query [Ishioka, 2003]. Regarding **Setup 2**, we used the classic Precision, Recall, F1-score and Area Under the Curve (AUC)[Powers, 2011].

5.6.1 Experimental results

The results regarding the analyses of **setup 1** are depicted in Figures 5.8 and 5.10. In both figures, the y-axis represents the average of Recall from all queries and the x-axis the average Precision. The curves inside these plots refer to the different F1-scores values (f). The legend refers to the configurations described in Section 5.6, e.g., f_hig means we use the filter *highLvl* during the query process.

In the first case, the initial query contains a term from 2014 and the EHRs documents were annotated in 2016 (Figure 5.8). The preliminary results demonstrate that our method had significant variation regarding the impact of the number of neighbours used to build the KG. The lowest F1-Score (0.40) is associated with the configuration *(combined_k_2_d_1_f_hig_non_top_10)*, while the best score was obtained with the configuration *(semantic_k_10_d_1_f_hig_non_top_5)*, reaching the F1-score of 0.82. The fall-out value for the best configuration shows good results, as illustrated in Figure 5.9. The average ratio of non-relevant documents returned in the queries was 1.51%.

In the second case, the query contains a concept existing in 2016 and the EHRs was annotated in 2014 (Figure 5.10). We observed that our method demonstrated an average performance. In this case, the maximum F1-score reached was 0.55 for the configuration *(semantic_k_6_d_1_f_hig_non_top_5)*, and the minimum was 0.26 for the configuration *(combined_k_2_d_1_f_hig_non_top_10)*. We observed a maximum fall-out in this scenario of 7.23% for

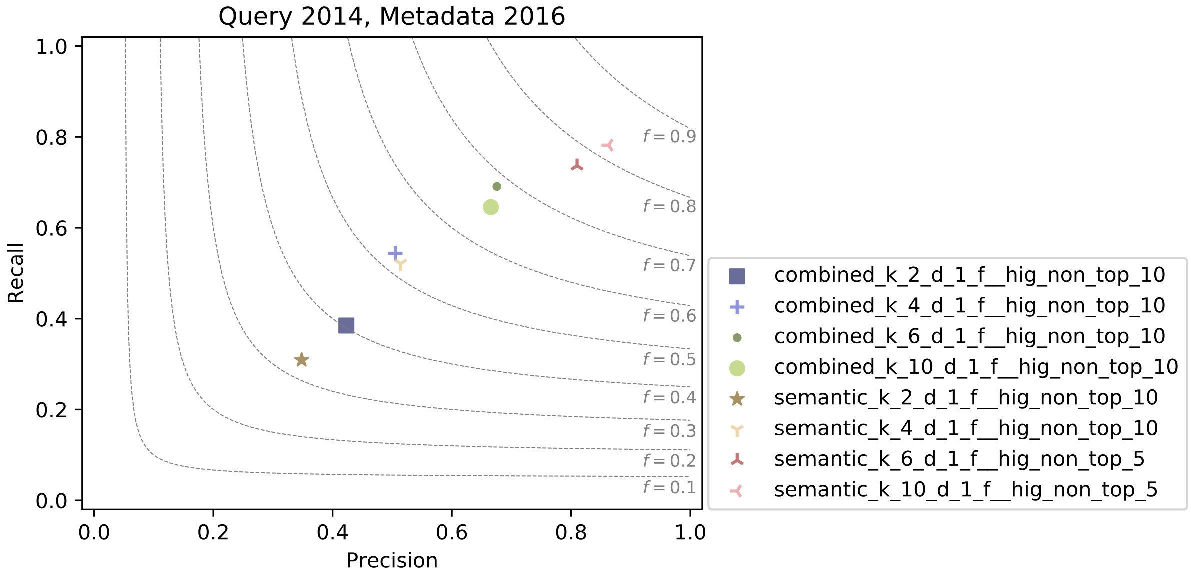


Figure 5.8: Maximum average Precision, Recall and F1-score regarding the best configuration for each KG. The query used is a term that exists only in 2014; the EHRs were annotated in 2016.

Method	ICD-9-CM				MeSH		
	P	R	F1		P	R	F1
BK	1	0.129	0.229		1	0.050	0.094
KG	1	0.500	0.667		0.974	0.306	0.465

Method	NCIt				SNOMED CT		
	P	R	F1		P	R	F1
BK	1	0.115	0.207		1	0.625	0.769
KG	0.975	0.750	0.848		0.917	0.668	0.786

Table 5.1: Precision (P), Recall (R) and F1-Score (F1) of impacted annotations computed using BK and Ad-hoc method. The red and orange colours indicate low and medium recall, respectively.

the configuration *(combined_k_2_d_1_f_hig_non_top_10)*, and a minimum of 1.41% for *(semantic_k_10_d_1_f_hig_non_top_5)*, see Figure 5.11. Detailed explanations of these observations are provided in Section 5.6.2.

The results for **set-up 2**, i.e., the use of Ad-hoc method for the direct maintenance of semantic annotations, are shown in Tables 5.1 and 5.2. In Table 5.1, we demonstrate the capacity of our method to detect impacted annotations. We utilized the values of precision, recall, and F1-Score to compare Ad-hoc and BK methods when applying them to the silver standard of chapter 3.

As a primary result, we observed that our Ad-hoc method is able to detect more impacted annotations than BK for all KOS, i.e. ICD-9-CM, MeSH, NCIt and SNOMED CT. The Recall and F1-Score of Ad-hoc in Table 5.1 is higher than BK in all these KOS. Regarding the precision, Ad-hoc method shows a small variation in all the KOS used.

The use of Ad-hoc method for the direct maintenance, Table 5.2, achieved similar results as BK in all KOS used. The AUC in MeSH is the same for both methods, while in other KOS the observed AUC shows small variations. In the next Section we discuss these results in detail and explain why both methods have close AUC in Table 5.2 and a distinct F1-Score.

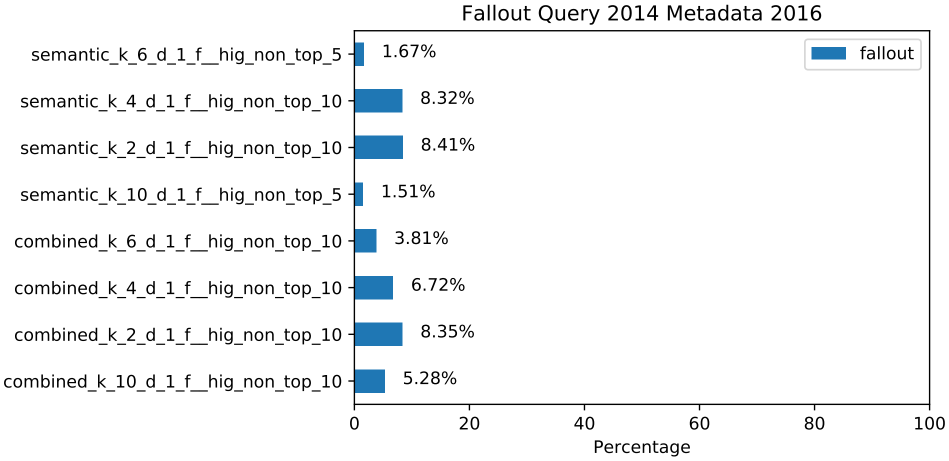


Figure 5.9: Fall-out for best configurations observed in setup 1, when the query was a term from 2014 and the EHRs were annotated in 2016.

Method	ICD9CM	MeSH	NCIt	SNOMED CT
	AUC	AUC	AUC	AUC
BK	0.597	0.545	0.663	0.75
KG	0.593	0.545	0.615	0.74

Table 5.2: AUC values of BK and Ad-hoc method when used to maintain annotations.

5.6.2 Discussion

The outcomes presented in Section 5.6.1 demonstrated that we were able to obtain high F1-Score and AUC for both direct and indirect maintenance processes.

When analysing the results obtained for **Setup 1**, we verified that the Ad-hoc method returns precise results for 2014 queries. For instance, when querying using *Sulfamethoxazole-trimethoprim combination*, concept D015662, was affected by a (*chgAttValue*) in 2015. Our algorithm was able to find concepts D013420:*Sulfamethoxazole*, D015662:*Trimethoprim, Sulfamethoxazole Drug Combination* and D014295:*Trimethoprim*. The returned terms were added to the initial query as a conjunction of terms and the system returned the right set of EHRs when evaluating the enriched query. When creating a KG, we need to consider the number of neighbours and the semantic similarity method. In our experiments with different KGs, we observed that increasing the number of neighbours improves the quality of the results (Figure 5.10). The semantic similarity measures also play a key role in this process. For instance, the best F1-Score in the second scenario (Figure 5.10) was obtained with the configuration that utilized the single semantic similarity with 6 neighbors.

The differences verified in the results between Figures 5.8 and 5.10, pointed out that after a concept becomes more specific, tracking back its previous version is a complex task. For instance, according to BioPortal[28], to retrieve documents associated with the term *Serine-Arginine Splicing Factors*:D000068103 added in 2016, we had to utilize the concept *nuclear proteins*:D009687 (2007-2015). With our Ad-hoc KG, we were able to reach the concept D009687 starting from D000068103 in 2016. The shortest path calculated between D000068103 and D009687 using Dijkstra's algorithm is (D000068103, D001120, D005903, D000067816, D011506, D009687). However, the complexity involved to reach the final node D009687, resulted in our approach not retrieving the EHRs associated with *Serine-Arginine Splicing Factors*. In our

[28] https://bioportal.bioontology.org/

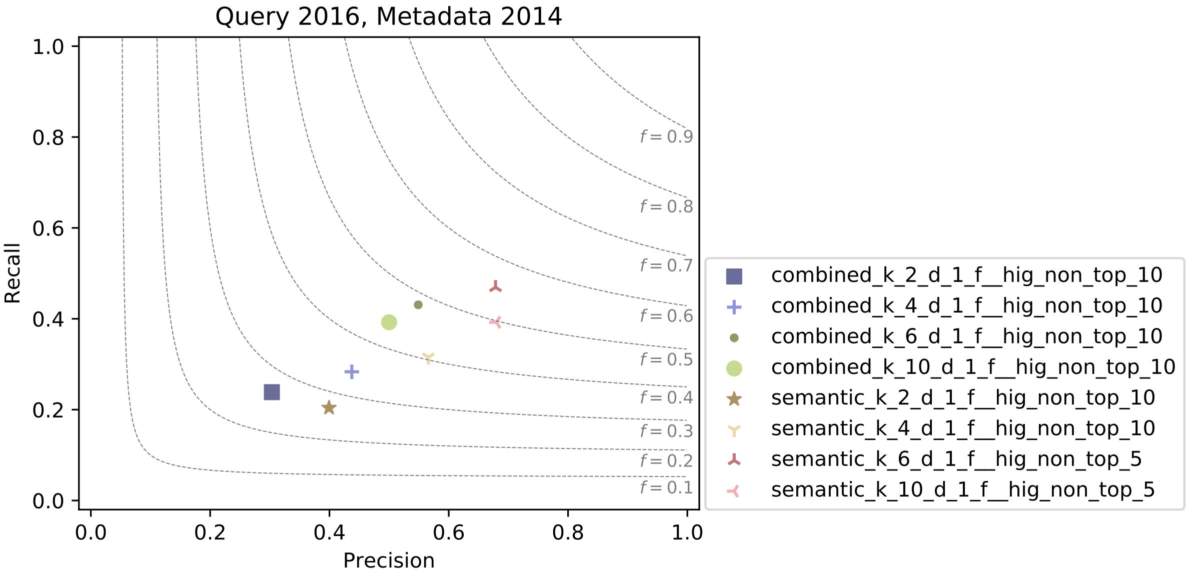


Figure 5.10: Maximum average Precision, Recall and F1-score regarding the best configuration for each KG. The query used is a term that existed in 2016 only and the EHRs were annotated in 2014.

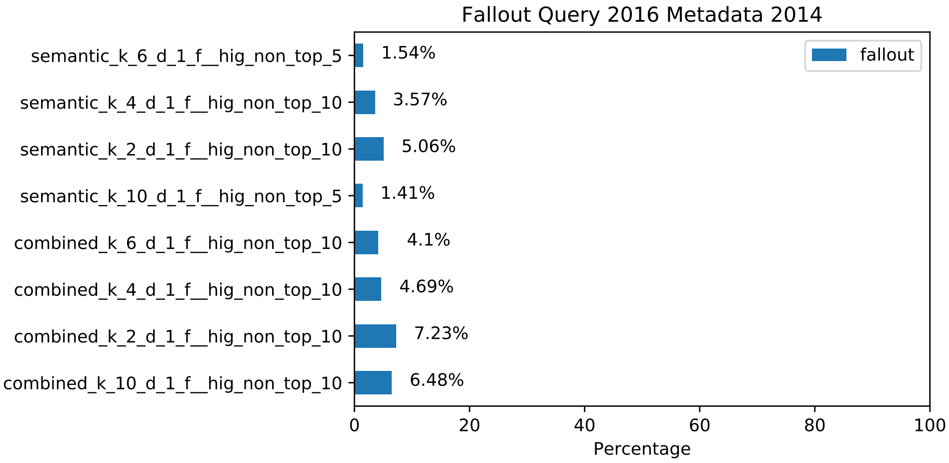


Figure 5.11: Fall-out for best configurations observed in setup 1, when the query is a term that exists only in 2016 and the EHRs were annotated in 2014.

algorithm, the path between both concepts had to pass through Neighbour Lists (NL) with low similarities, e.g. the NL between concepts D000068103 and D001120, has similarity equal to 0.354, difficultly encountered by our algorithm. Regarding the fall-out observed in both scenarios, we noticed that our method prioritized precision over recall when limiting the results by the top-k. It is also observed in Figures 5.8 and 5.10 that the F1-score is always under the main diagonal.

Moreover, we verified that when using a lower value of k to build the KG, the search algorithm does not provided good results. It keeps the F1-score low as well as the fall-out. On the other hand, when $k \geq 6$, almost all queries gave relevant results.

Finally, we verified that KOS changes and semantic similarity play a key role in building the `Ad-hoc` KG, resulting in a good quality of the enriched query. The best configurations in Figures 5.8 and 5.10 always utilized the filters *highLvl* and *none*. The first one is related to the evolution of the KOS, while the second informs that the concepts are somehow related by their similarity (i.e., no need to consider hierarchical relation between the concepts).

The analyses of **Set-up 2**, i.e. the use of `Ad-hoc` method to directly maintain the annotations, also demonstrated good results. We observed that the `Ad-hoc` method provided more adaptations than `BK`; this was mainly observed in the F1-Score of Table 5.1.

Analysing the results, we observed that `BK` flagged many annotations as *Unsolved*, i.e. no adaption was computed and we utilized the same concept after the evolution to evolve the annotation. This occasionally matched with the silver standard from chapter 3 and increased the AUC in Table 5.2.

Another aspect to highlight is that our `Ad-hoc` method provided adaptations that were not feasible with only `BK` in **Set-up 2**. As for chapter 4 with the `PartialMatch` rule, the annotation *"postoperative myocardial infarction"*, was correctly adapted in SNOMED CT by the `Ad-hoc` method. Actually, BioPortal has no mappings for this term, leading to use of the previous term that does not match with the silver standard. We thus, recommend the use of `Ad-hoc` method as an alternative for the usage of `BK` when there are few or no mappings to external resources available.

5.7 Conclusion

In this chapter, we addressed question RQ4 through the development of a method able to keep impacted annotations searchable without direct changes. It is one of the aspects of MAISA, a framework to maintain semantic annotations either by direct or indirect maintenance approaches.

The experimental analyses demonstrated that this method is capable of achieving good results when querying for EHRs or directly adapting the annotations. We observed that using a temporal `Ad-hoc` KG provides good representation for multiple ontology versions and their evolution.

This approach contributes to the state of the art of annotation maintenance by including a new method to adapt them, and to the Evolving Graph Generator domain by improving existing techniques. We also build a new resource for the Semantic Web: a knowledge graph that contains ontologies and their history. In the next chapter we discuss how to anticipate the evolution of concepts associated with annotations.

Chapter 6

Predicting ontology changes

Contents

In previous chapters, we discussed how to keep the validity of the semantic annotations when the underline KOS evolves over time directly or in an `Ad-hoc` manner. Depending on the amount of annotations that are affected, domain experts are solicited for a laborious validation task. Tools that support them to identify the impact of KOS changes in the annotations or even to foresee potential KOS evolution and alert them will potentially facilitate their work and improve the quality of their work. The work presented in this chapter contributes to address this need. We propose a stochastic model, using machine learning (ML) techniques, for identifying whether a concept of an ontology will evolve or not in the next release of the ontology and specify the type of change. In our global approach, this work was developed to answer RQ5: *Can we predict which KOS concept will change in the near future and the type of change that affect that concept?* The main objective of this analysis is to support domain experts during the annotation phase (or maintenance phase) by alerting them about risks of choosing concept and/or an ontology to annotate biomedical documents due to the high probability of evolution these concepts have.

First, we aim at identifying concepts whose definition requires to be revised, but to be consistent with our approach of maintenance of annotations, we went further in our analysis and we also develop a model to identify the type of non-logical changes that will affect each concept [Klein et al., 2002]. We considered the following types of changes: i) the extension of the ontology *i.e.* the addition of new concepts; ii) the modification of the description of a concept *e.g.* modification of the label or attribute value; iii) the removal of a concept; and iv) whether a concept will move to another part of the ontology. We named these four types of changes as *Extension*, *Change Description*, *Removal*, and *Move*, respectively (see Section 6.1).

We base our proposal on state-of-the-art approaches of the field [Chandrashekar and Sahin, 2014, Pesquita and Couto, 2012, Tsatsaronis et al., 2013] and extend them in several ways by adding new features that were identified as playing a key role in the evolution, by evaluating different techniques to deal with unbalanced datasets, and by analysing the impact of different machine-learning methods on different types (in terms of expressivity, size and dynamics) of ontologies. In addition to classical features selection, mainly based on structural information (see Section 6.1) derived directly from the ontology, we used Web information obtained by querying relevant scientific publications in the domain, the subset of information accessible through UMLS (Unified Medical Language System [29]), and also temporal information like the past evolution of the considered concept, as well as ontology region stability. Moreover, unlike existing work that clearly focuses on one dedicated ontology *i.e.* Gene Ontology for Pesquita and Couto [Pesquita and Couto, 2012] and MeSH for Tsatsaronis et al. [Tsatsaronis et al., 2013] and on the extension of the ontology, our method has been designed to cope with any existing ontology. We therefore propose an experimental validation of our model on four OWL versions of standard biomedical ontologies having different sizes, levels of expressivity and evolution frequencies: ICD-9-CM, MeSH, NCI thesaurus and SNOMED CT. Furthermore, we also compare our model to existing models when possible.

The remainder of the chapter is structured as follows: Section 6.1 introduces relevant notions and presents related work from the field "predicting ontology evolution". Section 6.2 presents the material and methods we used to design our approach. Section 6.3 shows the experimental results we obtained for the evolution of biomedical ontologies and Section 6.4 discusses them. Finally, Section 6.5 concludes the chapter.

6.1 Background

6.1.1 Problem statement

The main problem addressed in this chapter is the identification of needs for the evolution of the non-logical part of an ontology. We divided this problem into:

1. The identification of the set of concepts that need to be revised (associated with the function $Evolv_K$, defined below),

2. The recommendation of the types of revision that need to be implemented to update the concept considered (associated with the function $IdentTypeOfChange_K$, defined below).

In this chapter, $O^t = (C^t, R^t, A^t)$ represents version t of an ontology where C^t denotes the set of concepts, R^t the set of relationships between the concepts and A^t the set of axioms. Following the definition provided by Wang et al. [Wang et al., 2011], we define the meaning $M(c^t)$ of a concept $c^t \in C^t$ as a triple

$$M(c^t) = (label(c^t), int(c^t), ext(c^t))$$

In this definition, $label(c^t)$ represents the label of c^t, $int(c^t)$ is a set of properties *e.g.* object and datatype properties in OWL, or more generally speaking concept attributes, and $ext(c^t)$ is the extension of c^t (the set of individuals).

By

$$K = Struct(c^t) \cup Temp(c^t) \cup Rel(c^t)$$

we denote the context for our work. $Struct(c^t)$ represents the structural characteristics of c^t. It includes the intrinsic characteristic of a concept *e.g.*, the number of attributes defining a concept,

[29] https://www.nlm.nih.gov/research/umls/

or the number of siblings, superconcepts and subconcepts. $Temp(c^t)$ denotes the temporal characteristics of c^t, which includes aspects dealing with the history of a concept. In this work, we considered i) the stability of c^t obtained by measuring the elapsed time between t and the version l, with $0 < l < t$ and $M(c^t) \neq M(c^l)$ and ii) the stability of the neighbourhood of c^t (see Table 6.2). $Rel(c^t)$ considers the relational aspect of c^t acquired from external sources of information from the Web (see Section 6.2.3). Given one concept $c^t \in C^t$, our goal was to identify whether the meaning of c^t was still up-to-date at time $t+1$ in a given context K. Therefore, regarding this problem, the function $Evolv_K$ is defined as follows:

$$\begin{aligned} Evolve_K : C^t &\longrightarrow \{0,1\} \\ c^t &\longrightarrow \begin{cases} 0 & if \quad M(c^t) = M(c^{t+1}) \\ 1 & otherwise \end{cases} \end{aligned}$$

The first challenge of this work was to find an alternative to correctly execute this function when $M(c^{t+1})$ is unknown. In a detailed analysis on the evolution process, we observed that a concept could evolve in different ways. Complementary to the previous problem, knowing that a concept will evolve, we aimed to detect the type of revision required to update c^t and obtain c^{t+1}. We assumed that four types of revisions were possible

$$RevType = \{Extension, Removal, ChgDescription, Move\}$$

where *Extension* refers to new concepts to be added as subconcepts of c^t at time $t+1$. This type of revision was shown as relevant in [Pesquita and Couto, 2012]. *Removal* refers to the complete removal of c^t at time $t+1$. *ChgDescription* denotes the modification in the label as well as in the attributes structure and attribute values of c^t at time $t+1$. *Move* refers to changes in at least one superconcept of c^t at time $t+1$ (*i.e.* the set of superconcepts of c^t is different from the set of superconcepts of c^{t+1}, implying a move of c^t to another part of O^t). These revision categories regroup the ontological modifications identified by the literature from the field ontology evolution [Malone and Stevens, 2013, Klein et al., 2002, Stojanovic et al., 2002, Hartung et al., 2013, Hartung et al., 2009]. We focused on the non-logical part of the ontologies. To cover the logical part, we invite you to read [Gonçalves et al., 2011, Konev et al., 2012]. To detect the revisions we were interested in, we used the COnto-Diff tool [Hartung et al., 2013], but other *diff* tools such as PROMPT-Diff [Noy et al., 2002] may also be used. The inputs into the tool were the two versions of the ontology, and the output is the set of concepts and the revision actions associated with them.

Knowing that a concept had evolved, without having any other information about c^{t+1}, the second challenge of our work was to determine what type of revision was applied to the concept. In other words, in the perspective of the "identification of revision needs for a concept", we wanted to provide complementary information about what type of revision (from *RevType*) would be appropriated to keep the concept up-to-date. We associated this problem with the following function:

$$IdentTypeOfChange_K : C^t \longrightarrow RevType$$

This function takes the concept $c^t \in C^t$ and its structural, temporal and relational characteristics (see the various features in Table 6.2) as the input and returns the type of revision recommended for c^t. We will explain this function throughout the remainder of this chapter.

6.1.2 Related work

Identifying concepts that need revision can be seen, to a certain extent, as a concept-drift (or evolution) prediction problem [Groß et al., 2016]. In this context, works such as [Pesquita and Couto, 2012, Tsatsaronis et al., 2013] have proposed techniques to predict the modification of a concept, as well as the extension of biomedical ontologies using machine-learning techniques (ML). Using such approaches assumes that information encoded in the ontology or its resources, such as annotations, mappings and instances, govern its evolution and can be further exploited to predict how the ontology content will change in a future version [Pesquita and Couto, 2012].

Pesquita & Couto [Pesquita and Couto, 2012] have proposed the use of supervised learning classifiers like SVM, Naive Bayes, Bayesian Networks, Multilayer Perceptron and Decision Table to predict the extension of Gene Ontology. They obtained encouraging results with an average F-measure of 0.79 using the Bayesian Network. The features used to achieve these results were based on previous handcrafted rules and a series of guidelines for capturing changes [Stojanovic, 2004] and dealt with temporal information derived from previous versions of the ontologies considered, as well as information obtained from the usage of the ontology, such as citations and annotations. A relevant pattern observed in their results was that the GO concepts that have many children or many annotations and/or citations tend to be extended in future versions.

In the same vein, Tsatsaronis *et al.* [Tsatsaronis et al., 2013] complemented the work of Pesquita & Couto [Pesquita and Couto, 2012] and introduced temporal features in the learning classifiers. Their method takes one period of time as the parameter to train the classifier and another period of time representing the prediction time window, i.e. the period of time in which the change should take place. The temporal features and other features were used to predict the evolution of Medical Subject Headings (MeSH) over a period ranging from 1999 to 2012. As result, the authors showed that the Random Forest classifier can reach an F1-score of 66.4% to predict the extension of MeSH. Furthermore, they pointed that the use of temporal features as *temporal all children*, *temporal direct children*, aids significantly in the prediction. They based their approach on the following features:

1. **Structural features:** give information about the definition of a concept and the structure of its surrounding neighbourhood in the ontology.

2. **Annotation features:** based on the number of annotations corresponding to a concept label.

3. **Citation features:** based on the number of citations of a concept label in an external corpus of documents, e.g., number of scientific articles in PubMed mentioning a given concept label.

4. **Hybrid features:** combination of some of the previous features.

Another initiative to predict the extensional drift of concepts is the work of Meroño-Peñuela *et al.* [Meroño-Peñuela et al., 2013]. This work uses resources given by the Dutch historical censuses data set (CEDAR) [Meroño-Peñuela et al., 2017] to estimate whether a concept will experience a change in its extensional definition using statistical Linked Data. In this work, the authors do not consider the intentional component of the meaning of a concept but only the extensional one (i.e. the individuals).

Recently, Cano-Basave *et al.* [Cano-Basave et al., 2016] addressed the problem of ontology forecasting in the scientific research domain. They introduced a novel framework to be added in the ontology for the identification of emerging semantic concepts in the research domain. This is based on the integration of lexical novelty and adoption priors acquired from historical data. However, this approach only considers information external to the ontology and not the ontology itself or temporal information.

Despite the recent advances in the prediction of concept-drift or ontology forecasting, we observed the following deficiencies. First, existing approaches do not consider the intrinsic structural features, e.g., the total number of distinct attribute values of a concept. Second, they are corpus-dependent and require a significant number of annotations to predict the evolution of a concept. Third, the concept drift detection does not exploit existing semantically rich information. To overcome these shortfalls, our approach promotes the use of background knowledge and includes new temporal and structural features such as region stability over time.

6.2 Material and Methods

The identification of concepts to be revised ($Evolve_K$) and the recommendation of types of revision ($IdentTypeOfChange_K$) need to take into account intrinsic characteristics of the ontologies, as well as knowledge from external sources (e.g. the web). In this work, we analysed several types of information, selecting the information that improved the quality of the identification and recommendation functions. These functions are based on machine learning (ML) techniques. We evaluated several ML techniques to select the most precise one, according to the features and data used. In this section, we describe the various datasets used in our study, as well as the methods used to achieve the results presented in Section 6.3.

6.2.1 Material

As a starting point, we used the following datasets[30]:

- 10 successive versions of the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM), of the Medical Subject Headings (MeSH), of the NCI Thesaurus (NCIt) and of the Systematized Nomenclature of Medicine - Clinical Terms (SNOMED CT) covering the period ranging from 2004 to 2013. For each terminology, we used only the version published (or in-use) in the Unified Medical Language System (UMLS)[31] at the end of January of each year.

- The Diff between the mentioned OWL ontology versions was obtained using COnto-Diff [Hartung et al., 2013]. The Diff consists of the identification of basic (insert/update/delete) and complex transformations e.g., concept merging, concept splitting, move of concept, etc. that lead one ontology version to its successor. The ontological changes that could be identified by COnto-Diff were then regrouped to form the *RevType* set (see Section 6.1).

- The content of PubMed (about 17 million scientific articles) as a source of external information to find relevant publications from the medical domain for the periods 2011 to 2012 and 2012 to 2013.

- External termino-ontological resources (denoted as background knowledge in the remainder of the chapter). In our experiments, these resources were provided by the UMLS [Bodenreider, 2004]. The UMLS contains 131 resources and their associated versions, which are important for identification purposes, in order to compare concepts from corresponding versions. Since we evaluated concept descriptions from 2004 to 2013, we needed to compare the version of the concept to the description of another concept in UMLS at the corresponding moment in time.

6.2.2 Methods

In this section, we present the methodology adopted to solve the two problems defined in the $Evolve_K$ and $IdentTypeOfChange_K$ functions. This methodology has three objectives:

[30] https://git.list.lu/ELISA/predictingOntologyChange.git

[31] https://www.nlm.nih.gov/research/umls/licensedcontent/umlsarchives04.html

1. Identify the best ML technique and the most relevant features that play a significant role in the identification problem,
2. Select the best ML technique for recommending the revision types for ontology concepts, as well as the relevant features,
3. Evaluate the quality of the selected techniques and features.

To achieve these objectives, we set up the following methodology. The first step consisted of the generation of a dataset from the material described in Section 6.2.1, characterized with a set of pre-selected features derived from state-of-the-art field approaches, as well as experiments detailed in [Cardoso et al., 2016] (see Table 6.2). This dataset was then used to evaluate the performance of five two-class classifier methods (Boosted Trees, Random Forest, Decision Tree, Logistic Regression and SVM). These were pre-selected according to the specificity of our problem, the characteristics of the dataset (fewer than 20 features and fewer than 1000 data points by class), the required capacity to provide explainable class boundaries, the fact that there were dependencies between features, and the literature review [Hearst et al., 1998], which highlighted the good generalization performance of SVM. Neural Networks and Deep Learning were not included in our experiments because we were looking for comprehensive class boundaries and/or methods that perform well with a small quantity of training data. A summary of the decision criteria used in this work is presented in Table 6.1, where crosses indicate under-performance with the associated criteria.

	small #Features	#raw data < 1000	feature dependency	explainable model	unbalanced dataset
Boosted Trees	√	√	√	√	√
Random Forest	√	√	√	√	√
Decision Tree	√	√	√	√	√
Logistic	√	√	×	√	×
SVM	√	√	×	×	×
Neural Network	√	×	√	×	√
Deep Learn	√	×	√	×	√

Table 6.1: Criteria used to select the classification methods used in this work

Knowing that the ontologies considered had a ratio between changing and stable concepts of close to 10% (per evolution step) [Dos Reis et al., 2014], we decided to apply a balancing method. Two main methods are used in the literature [Kotsiantis et al., 2006]: Undersampling and oversampling. The limitation of the former method is that relevant information can be lost when reducing the size of the bigger group, which could impact the accuracy of our prediction. The latter method can transform rare cases into frequent ones. This can also be an obstacle to improving the quality of the $Evolve_K$ and IdentTypeOfChange$_K$ functions.

Six variations of these two methods were evaluated in this work:

- Undersampling 100%, which consists of randomly selecting the same quantity of elements from the bigger group as from the smaller group;

- Oversampling 100%, which implies randomly duplicating elements from the smaller group until it reaches the size of the bigger group;

- Undersampling 50%, which applies the same technique of undersampling 100% while keeping the size of the bigger group 150% bigger than the smaller group;

- Oversampling 50%, which applies the same technique of oversampling 100% while keeping the size of the bigger group 150% bigger than the smaller group;

- Undersampling 75%, follows the same principle of undersampling 50%, while keeping the size of the bigger group 125% bigger than the smaller group;

- Oversampling 75%, follows the same principle of oversampling 50%, while keeping the size of the bigger group 125% bigger than the smaller group.

The second sub-objective of our experimental study consisted of evaluating and selecting features that significantly impact the quality of the identification/recommendation functions. We started this set of experiments by proposing 17 features categorized into five different aspects, indicated by the symbols I, S, T, SW, and E in Table 6.2:Intrinsic characteristics of a concept, indicated by the symbol (I); Structural characteristics of the concept in the ontology, indicated by the symbol (S); Temporal aspect, indicated by the symbol (T); Semantic aspect of the concept, indicated by the symbol (SW); External related information, indicated by the symbol (E).

We used the Boosted Tree classifier to establish a rank of features. As input, we provided the "classifier method", as well as the 17 features with data collected up to 2012 (this data was split into training and test data). We evaluated different configurations of the classification method to verify the impact of the configurations on the ranking process. To establish the rank, we used the importance measurement provided by the GraphLab library[32]. The importance of feature X is determined by the sum of the occurrence of X as a branching node in all trees. We added the importance value calculated for each ontology and ranked the features based on this value. The rank was used for the selection of features; we added features one by one following the rank order and calculated the accuracy of the model. We repeated this operation 10 times and used the average and variance of the accuracy to select the best features.

We trained our model with the ontology versions from 2004 to 2012, i.e., classifiers and features were evaluated with data from this period, and only data from 2013 was used (our gold standard) to evaluate the quality (and generality) of our model. The experiments were implemented in a two-step process:

- We applied the trained models to 2000 concepts (1000 from each group), randomly extracted from the ontology version 2012, in order to identify the concepts that are candidates to be reviewed (outcome of $Evolv_K$). We compared these candidate concepts with those that evolved significantly in 2013 and calculated the accuracy and F1-score. For the sake of readability, we present some data with their accuracy values and other with their F1-score. Regarding the number of cases in our approach, we referred to the work of Krejcie & Morgan [Krejcie and Morgan, 1970] in order to determine a significant number of cases to train and evaluate our classifier.

- We used only the cases that evolved (from 2004 to 2012) to train our model to recommend the type of revision needed (outcome of $IdentTypeOfChange_K$). Notice that we did not balance the number of cases for each revision type. We evaluated the quality of the model with data from 2013.

[32] `https://turi.com/products/create/docs/index.html`

- We compared our approach with the state-of-the-art. In this, we used only part of the data (only for MeSH) to train the model and compared the results obtained with those presented in [Tsatsaronis et al., 2013].

6.2.3 Feature Engineering

Similar to the existing approaches for ontology evolution prediction (see Section 6.1), our model takes into account intrinsic and structural information extracted from the concept and its neighbourhood (indicated by "I" and "S" in Table 6.2 respectively) and Web information obtained after querying data portals such as PubMed ("E" in Table 6.2). However, our approach also deals with temporal information obtained by analysing the history of the considered ontology ("T" in Table 6.2), as well as semantic information obtained from the UMLS ("SW" in Table 6.2). There are 17 features defined in Table 6.2 that can be grouped as follows:

- Temporal features, included in the evaluation after observing their impact on semantic annotations [Cardoso et al., 2016]. We noticed that, if a concept is part of an unstable region, i.e. directly surrounded by concepts that change frequently over time, this concept is also more likely to change.We therefore want to verify whether this feature also plays a role in the identification of concepts requiring revision. According to our formalism, given a concept c^t, the two features dealing with temporal aspects form the $Temp(c^t)$ context.
- Background knowledge (BK), materialized by external ontologies. This information is potentially relevant for ontology maintenance tasks [Sabou et al., 2008]. In this sense, we used background knowledge to generate new features, evaluating their relevance for our identification model. For instance, we evaluated whether high similarity between the analysed concept and the siblings of matched concepts from other ontologies would indicate a trend for evolution. In this work, similarity was obtained by measuring the cosine similarity between the attribute values of c_t and those of the corresponding concepts in the background knowledge. The set of nine features dealing with background knowledge ("E" and "SW" in Table 6.2) made up the $Rel(c^t)$ context.
- Structural information, represented by the $Struct(c^t)$ context, denoted characteristics linked to the description of c^t, like the number of attributes of c^t, as well as semantic information about the super, subconcepts and siblings of c^t. These features are labelled with "S" and "I" in Table 6.2.

Feature	Description
(I) Num_att(c^t)	The total number of distinct attributes of c^t.
(I) Att_length(c^t)	Sum of the length of each attribute value of c^t.
(S) dir_children(c^t)	Number of direct sub-concepts of c^t.
(S) all_children(c^t)	Number of all subsumed concepts (direct and inferred) of c^t.
(S) siblings(c^t)	Number of concepts that share at least one super concept with c^t.
(S) isLeaf(c^t)	Gives an indication if c^t has no subconcept.
(T) Region_stability(c^t)	Coefficient measuring the stability of the neighbourhood of c^t. The neighbourhood includes the superconcepts, subconcepts, and siblings of c^t. The coefficient is obtained by dividing the number of concepts in the neighbourhood of c^t that have evolved

Continued on next page

Table 6.2 – continued from previous page

Feature	Description
	in the last version by the total number of concepts of the neighborhood.
(T) Last_evolution(c^t)	Indicates how many releases have been published since the last evolution of c^t.
(SW) Similarity_max(c^t)	Max cosine similarity between attribute values with its equivalents in BK. It is obtained by computing the Cartesian product between the set of attribute values of c_t and the set of attribute values of equivalent concepts of c_t defined in the UMLS.
(SW) Similarity_average(c^t)	Average of cosine similarity between attribute values of equivalent concepts of c^t in BK.
(SW) Max_simSup(c^t)	Max cosine similarity between attribute values with the superconcept of c^t in BK.
(SW) Max_simSib(c^t)	Max cosine similarity between attribute values with the sibling concepts of c^t in BK.
(SW) Max_simSub(c^t)	Max cosine similarity between attribute values with the subconcepts of c^t in BK.
(E) PubArtT(c^t)	Number of PubMed articles citing $label(c^t)$ in the previous release.
(E) PubArtT1(c^t)	Number of PubMed articles citing $label(c^t)$ in the current release.
(E) DiffArt(c^t)	The absolute difference in number of PubMed articles citing $label(c^t)$ between both release.
(E) DiffArtRatio(c^t)	The difference in the ratio of the number of PubMed articles.

Table 6.2: List of pre-selected features

The dataset mentioned in Section 6.2.2 was generated according to these features. It means that for each concept c^t considered, we computed the value of the corresponding features (see Table 6.2). An illustrative example of how the value of each feature was assigned is shown with the concept number 171.0 (*Malignant neoplasm of connective and other soft tissue of head, face, and neck*), from the ICD-9-CM version 2012. This concept has five attributes, including one "Abbreviation in any source vocabulary" (Mal neo soft tissue head), one "Metathesaurus preferred term" (Malignant neoplasm of connective and other soft tissue of the head, face, and neck), and three "Metathesaurus entry terms" (Malignant neoplasm of cartilage of the ear; Malignant neoplasm of cartilage of the eyelid; and Malignant neoplasm of connective and other soft tissue of the head, face and neck). The total length of these five attributes is equal to 258 characters (Att_length=258). The last time that this concept evolved was in 2008, i.e. four years earlier (Last_evolution=4). We also observed that none of the neighbours (siblings, super, subconcepts) evolved in 2012 (Region_stability=0). When we compared the concept label with that of the neighbourhood (for instance, with the superconcept "Malignant neoplasm of connective and other soft tissue"), using the cosine similarity measure, we obtained a value close to 0 (Similarity_max=0). Another observation is that this term did not appear in any publications between 2011 and 2012 (PubArtT=0).

6.3 Results

In this section, we present the experimental results that we obtained for the selection of the classifier as well as for the relevance of the features in the identification of concepts that need revision. We further discuss our results with respect to the recommendation of the type of the revision and compare our work with [Tsatsaronis et al., 2013].

6.3.1 Classifier selection

Our problem, identifying the concepts needing revision, can be seen as a classification problem, where we have two classes (i.e., those needing revision and those not needing revision) that we aim to populate with the concepts from our dataset. Several classification methods could have potentially been applied to our problem (Boosted Trees, Random Forest, Decision Tree, Logistic Regression and SVM), but we needed to narrow it down to one. The first set of experiments intended to support the classifier selection decision. First, we calculated the accuracy of each method (with different configurations, i.e., a combination of ML techniques and balance sampling methods) using the dataset made up of concepts belonging to ontologies released in 2012 and before. We randomly selected 2000 concepts (1000 from each class) and used 80% for training and 20% for validation (the selection of concepts was repeated for each experiment in order to assure the generality of the approach). We repeated the experiment 10 times to determine the average accuracy and the variance of the prediction. We also tested repeating the experiment 5, 15 and 20 times, but observed that over 10 times, the average accuracy does not change significantly. The decision criterion is based on the best average accuracy with the minimum variance. As we considered two distinct problems, i.e. the identification of concepts needing revision and recommendation of the type of revision, we evaluated the classification methods according to these two problems.

Figure 6.1 illustrates the outcomes of this set of experiments using the dataset containing MeSH concepts for $Evolv_K$. We used the GraphLab [33] libraries with the standard configuration and analysed their performance based on accuracy and F1-measures as follows: The best accuracies were selected and ranked; where two or more methods had the same (or very close) accuracy, we used the variance to improve the rank. Where the variance was not sufficient, we also used the F1-measure to support the decision. The outcome of this analysis was the selection of the Boosted Trees Classifier using the Oversampling 100% method. Table 6.3 shows that this classifier demonstrated the best performance for the four ontologies (on average this classifier had an F1-measure of 86%, while Random Forest had an F1-measure of 83%). The depth of the trees in the Boosted Trees Classifier is also a parameter that can interfere with the quality of the prediction. We evaluated seven different configurations (depth = 2, 5, 10, 50, 100, 200, 500) and measured the classification error 30 times. We used the average of these 30 experiments and the outcomes of these experiments can be seen in Figure 6.2. We observed that the training error falls to zero when the depth is higher than 50, thus we used this configuration to evaluate the relevance of the features.

Our problem for recommending the type of revision can be seen as a classification problem where we have multiple classes (i.e., Move, Extension, Removal, ChgDescription) and look to populate these classes with the concepts from our dataset. We evaluated the same classification methods as for $Evolv_K$ and chose one.

For this, we followed the same methodology as for $Evolv_K$ to select the best classification method, but used only the set of modified concepts as our input. The results are shown in Table 6.4. They show the number of cases we used to train and evaluate the classifiers. For this problem, the best classifier is Random Forest with Oversampling 100%. Even if we observed

[33] https://turi.com/products/create/docs/graphlab.toolkits.classifier.html

	ICD-9-CM		NCIt		MeSH		SNOMED CT	
	acu	var	acu	var	acu	var	acu	var
BoostedTrees	0.83	0.003	0.77	0.01	0.77	0.02	0.69	0.01
RandomForest	0.82	0.004	0.76	0.02	0.75	0.02	0.68	0.01
DecisionTree	0.81	0.004	0.75	0.02	0.72	0.03	0.66	0.02
SVM	0.68	0.005	0.61	0.01	0.71	0.01	0.65	0.01
Logistic	0.78	0.003	0.64	0.04	0.73	0.02	0.65	0.01

Table 6.3: Average accuracy and variance of the prediction for $Evolv_K$ calculated for all configurations with four different datasets: ICD-9-CM, NCIt, MeSH, SNOMED CT

that Random Forest and Boosted Trees have the same accuracy average, when we consider the variance and the F1-measure, Boosted Trees prove to be less precise than Random Forest. We excluded the SVM method from the evaluation because for this problem, the dataset is highly unbalanced (which contributed to reducing the accuracy of SVM). Furthermore, SVM showed the lowest performance among the classifiers for the $Evolv_K$. Table 6.5 shows the average accuracy for the recommendation of the type of revision, applying the Random Forest model to each ontology.

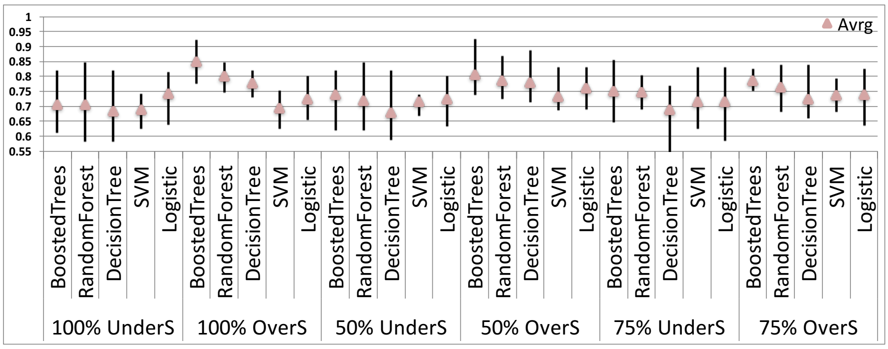


Figure 6.1: Accuracy and variance of five two-class classifiers according to four different balancing methods, using the MeSH dataset

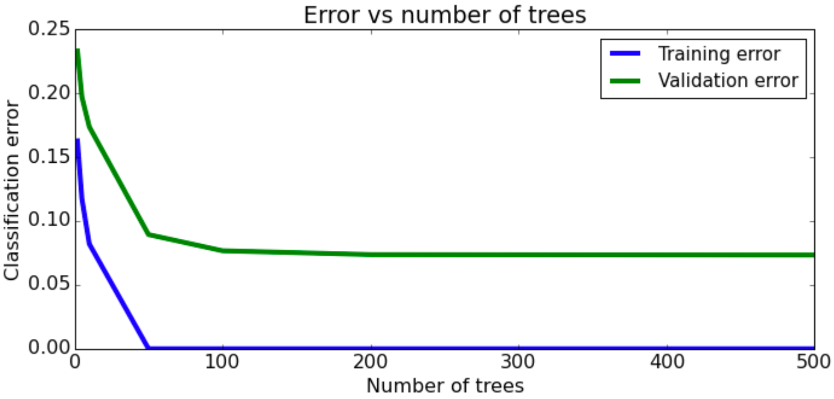


Figure 6.2: Classification Error according to the number of trees for the Boosted Trees Classifier

	ICD-9-CM	NCIt	MeSH	SNOMED CT	Total
BoostedTrees	0.87	0.76	0.51	0.71	0.71
RandomForest	0.84	0.77	0.54	0.70	0.71
DecisionTree	0.83	0.75	0.53	0.67	0.69
Logistic	0.82	0.76	0.5	0.52	0.65

Table 6.4: Average accuracy of the prediction for $IdentTypeOfChange_K$ calculated for all configurations with the four different datasets

		ICD-9-CM	NCIt	MeSH	SNOMED CT	Total
Move	P	0	0.5	0.33	0.91	0.435
	R	0	0.11	0.09	0.62	0.205
	F	0	0.31	0.21	0.76	0.32
Extension	P	0.837	0	0.58	0	0.354
	R	0.99	0	0.71	0	0.425
	F	0.91	0	0.65	0	0.39
Removal	P	0	0.64	0.65	0.72	0.502
	R	0	0.32	0.44	0.91	0.417
	F	0	0.48	0.55	0.81	0.46
ChgDesc.	P	0.826	0.81	0.63	0.65	0.73
	R	0.13	0.98	0.63	0.51	0.562
	F	0.48	0.9	0.63	0.58	0.647

Table 6.5: Precision (P), Recall (R) and F-score (F) for the prediction of $IdentTypeOfChange_K$ using the Random Forest classifier

6.3.2 Feature selection

Once one classifier had been selected for each problem, the next phase consisted of ranking the features described in Section 6.2.3 with respect to their relevance for both problems.

Features for the identification of concepts needing revision

In our experimental study, we used tree-based estimators (based on the Boosted Tree classifier) to determine the importance of each feature for problems 1 and 2. The idea was to select features that could perform well for the identification and recommendation problems independently of the ontology structure. Our assumption was based on the fact that the use of these four ontologies could potentially increase the generality of our model, since these ontologies have different structures and expressivity levels. Figure 6.3 shows the sum of the importance weights computed by the tree-based estimator for each ontology and for each feature. The outcome of this experiment allowed us to produce the following ranking (in order of importance): PubArtT, dir_children, Num_att, Similarity_max, PubArtT1, Max_simSib, Region_stability, Att_length, Last_evolution, Max_simSup, DiffArt, all_children, siblings, Max_simSub, DiffArtRatio, Similarity_average, isLeaf.

Then, we measured the accuracy of the model with different subsets of features. To do this, we used the test set. We generated a model with only one feature (always selected in the same order as in the ranking), trained it with the training set from 2004 to 2012, and evaluated it with the test set. We repeated the process, adding one new feature with each new iteration. We repeated the whole feature selection process 10 times in order to obtain the average value, as well as the standard deviation of the identification problem. In Figure 6.4 , we observe the average accuracy when accumulating the features. For instance, the third value of the MeSH dataset underlines the accuracy of the model when using the three features PubArtT, dir_children, and

	ICD-9-CM		NCIt		MeSH		SCT	
	Train.	Eval.	Train.	Eval.	Train.	Eval.	Train.	Eval.
Move	0	0	91	24	21	5	110	25
Extension	916	227	47	10	113	28	21	3
Removal	0	0	53	14	24	6	311	76
ChgDesc.	186	50	621	163	108	27	168	45

Table 6.6: Number of cases for training and evaluating the classifiers

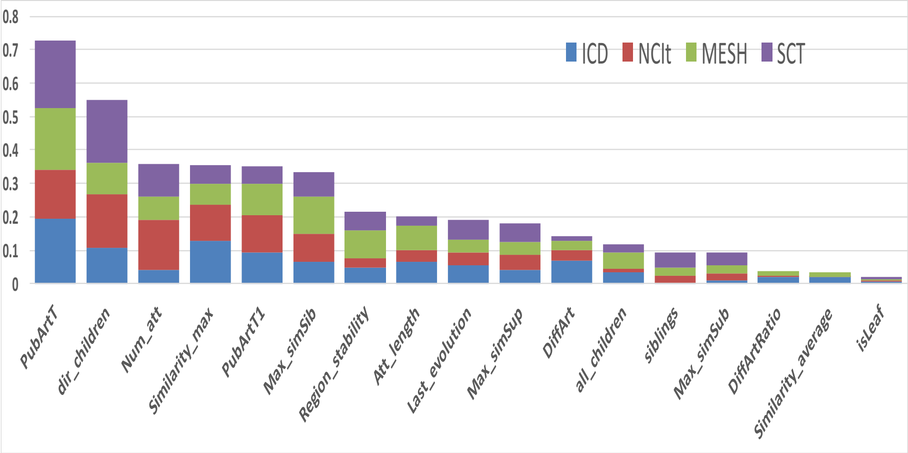


Figure 6.3: Feature relevance for the four ontologies, according to the tree-based estimator for $Evolv_K$

Num_att. The vertical line over each point indicates the standard deviation calculated based on the outcomes after a series of 10 experiments. Notice that, when considered alone, a feature can have a negative impact on the prediction of one dataset, but a positive impact on the others. Another important aspect extracted from the experiments was that the accuracy increases until (+/-) the addition of the 11^{th} feature. After that, it remains quite stable. This is true for almost all datasets. For example, ICD-9-CM and NCIt reach the upper value with 9 features, MeSH with 11 features, and SNOMED CT with 13 features. However, after adding the 12^{th} feature, the accuracy of the model decreased for MeSH and NCIt, indicating that these features add substantial noise to the system.

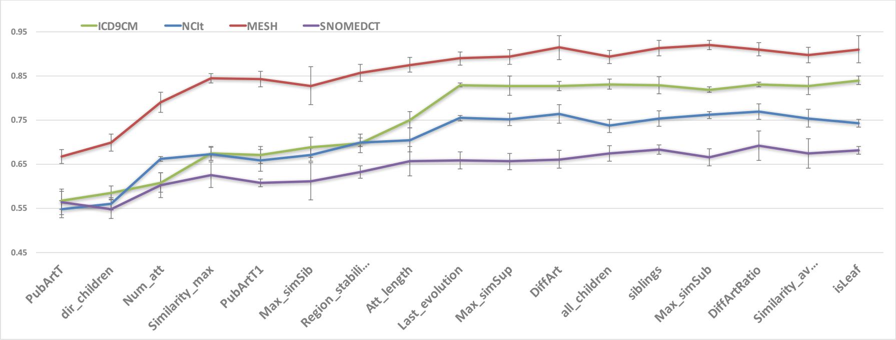


Figure 6.4: Accuracy when increasing the number of features for $Evolv_K$

Features for recommending the type of revision for a concept

Regarding $IdentTypeOfChange_K$, we followed the same feature engineering process to identify the most relevant features for recommending the type of revision. However, for this problem, the selected classifier was Random Forest and the number of cases for training and evaluating listed in Table 6.6.

The results, depicted in Figure 6.5, show that the same set of features is relevant for indicating the way the concept will evolve. However, unlike for $Evolv_K$, the importance of each feature is different. There are some minor permutations, for instance, *dir_children* is the most relevant feature for $IdentTypeOfChange_K$ while it is $PubArtT$ for $Evolv_K$. In general, the results show that the consideration of external sources of information for the prediction is extremely relevant since features related to PubMed and to the UMLS are among the most relevant. Moreover, Figure 6.6 illustrates the added value of temporal features since the region of stability drastically increases the precision, especially for $IdentTypeOfChange_K$.

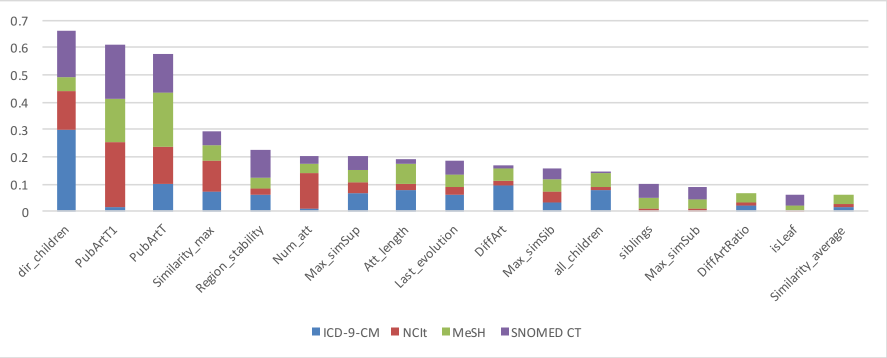


Figure 6.5: Feature relevance for recommending the type of revision

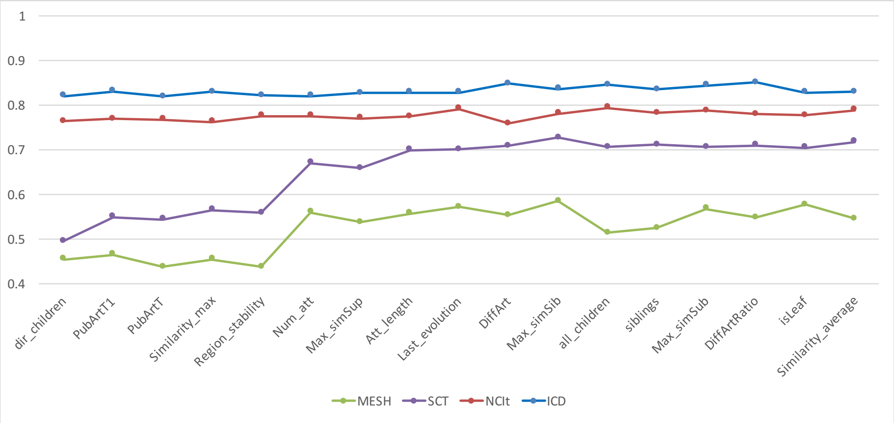


Figure 6.6: Accuracy when increasing the number of features for $IdentTypeOfChange_K$

Comparison with related work

The experiments we conducted allowed us to draw a comparison with the approach described in [Tsatsaronis et al., 2013]. The two approaches could not use exactly the same set of data since the data used in [Tsatsaronis et al., 2013] was not made available. Thus, this comparison considered (our hypothesis) that the selected data used in our experiments were quite similar to those used by Tsatsaronis et al. In their work, Tsatsaronis et al. measured the ability of their model to predict the extension of MeSH. However, they considered several sub-parts of MeSH independently (i.e. Organisms, Diseases, Chemicals and Drugs). In order to compare the approaches, we carried out our experiments using only data from MeSH, without making distinctions between the sub-parts (we have a more general approach). Table 6.7 shows the results of Precision, Recall and the F1-score we obtained compared to those of [Tsatsaronis et al., 2013]. Another modification we made in our approach to make the results more comparable was the selection of the Random Forest classifier and we used the same number of MeSH versions, i.e., 10. Moreover, as their method considers a modular time frame for the prediction i.e., the considered concept would evolve over the next y years, we reduced this period to one year to be comparable. We noticed that the structure of MeSH did not change significantly over these 10 years [Dos Reis et al., 2014].

	Our approach	Tsatsaronis *et al.*		
		Orga. Tree	Dis. Tree	Chem. and drugs Tree
Precision	0.58	0.43	0.47	0.35
Recall	0.71	0.14	0.09	0.11
F1-score	0.65	0.21	0.1	0.16

Table 6.7: Comparison for the prediction of the extension of MeSH

The results in Table 6.7 reveal that even with a more general approach, our method performed better in terms of precision, recall and F1-score than the one described in [Tsatsaronis et al., 2013]. The selected set of features, in particular the consideration of background knowledge, significantly improved the precision and recall of the identification of concepts that need revision (in the MeSH). However, this comparison had some limitations that can partially justify the differences showed in Table 6.7:

- The data used to train and evaluate the classifier. We randomly selected a balanced set of elements of MeSH between those that evolve and those that remain stable while Tsatsaronis *et al.* [Tsatsaronis et al., 2013] used the MetaCost method [Domingos, 1999] to balance their data. In order to be comparable, we assumed two hypotheses: (1) that the data sources are quite similar (MeSH, same period, balanced data), (2) that the classifiers use the same parameters. If these two hypotheses are correct, than we can imagine that the main difference in the results came from the features that we selected. Thus, our features were more relevant for solving the first problem. However, any information about classifier parameters was found in their publication [Tsatsaronis et al., 2013].

- Some differences in the definition of comparable features, especially temporal ones. In our work, we distinguish between the evolution of a concept and the evolution of its neighbourhood while the distinction made by Tsatsaronis et al. concerns all features, e.g. evolution of siblings, and of superconcepts or subconcepts.

6.4 Discussion

When analysing each feature and its impact on the identification/recommendation model, we formulated the hypothesis below, which can potentially explain the model behaviour.

Features dealing with **background knowledge** are the most important for the two functions that we developed ($Evolv_K$ and $dentTypeOfChange_K$). This can be justified by the fact that ontologies aim to reflect the real world, with the web containing relevant information from the real world that can be taken into account. Therefore, if specific knowledge within this domain evolves, then the part of the ontology related to this evolving knowledge must also evolve. $PubArtT(c^t)$ extracts the number of articles citing $label(c^t)$ from scientific publications (published in the year preceding the prediction period, i.e., in 2012). High values indicate that this knowledge is of utmost importance for the domain. We highlighted that PubMed uses MeSH to index documents. Thus, this feature has a higher impact on the quality of the identification of MeSH concepts than on the other ontologies. On the other hand, we observed that $PubArtT1(c^t)$ has a negative impact on $Evolv_k$ and a positive impact on $IdentTypeOfChange_K$. We are convinced that the date of publication of the ontologies considered plays a key role in this behaviour. For instance, MeSH concept D009369 was changed in January 2013, and with all papers published from January onwards using the new/changed concept (the impact of $PubArtT1(c^t)$ will probably be positive in this case). However, the date of new releases of ontologies are different (some are in October or even monthly). Our approach used the 31 January version, which probably created a bias for $PubArtT1(c^t)$ for some ontologies. This can potentially explain the negative impact of this feature in some ontologies. Additional experiments are needed to identify the impact of PubMed citations over a longer period of time. Moreover, background knowledge gives a helpful support to improve the quality of our method. When the maximum similarity with mapped concepts (from other ontologies) is low, it can potentially indicate that the current definition of the concept needs to be reviewed. This is why $Similarity\_max(c^t)$ is important for the prediction (for both problems). However, the similarity with existing superconcepts ($Max\_simSup(c^t)$) in the BK is less important than the similarity with subconcepts ($Max\_simSub(c^t)$) and siblings ($Max\_simSib(c^t)$). It shows that ontologies evolve in order to become more precise rather than more abstract. These features have limited impact on flat ontologies like ICD-9-CM where subconcepts can not easily be added.

Our experiments have shown that features derived from the **structural characteristics** of the ontologies are also important for the prediction. The relevance of the feature $dir\_children(c^t)$ has been observed in several works on ontology evolution prediction, according to the papers we have analysed (see Section 6.1). In combination with $PubArtT(c^t)$, this feature improves knowledge on the necessity of creating a more specialized (or not) concept. Few direct subconcepts and many citations constitute a good situation for concept evolution. Another good situation for concept evolution can be observed when we take into account the number of attributes ($Num\_att(c^t)$). A high number of attributes can be a good indicator that the definition of the concept is unclear or ambiguous. Thus, the probability that these attributes are partially transferred to another (new/modified) concept is higher. Another explanation could be that in our work, the change in a concept includes a change in any of its attributes, therefore it is more probable that concepts with many attributes evolve. Additional experiments are needed to verify which explanation occurs most frequently. There is an obvious correlation between $Att\_length(c^t)$ and $Num\_att(c^t)$, however, $Att\_length(c^t)$ also includes cases where a concept has few properties but is described with long strings. These cases are rarer than those with a high number of properties (which explains the lower importance weight). Nevertheless, this distinction allows the context of the concept to be refined and the quality of our predictions to be improved. It should be mentioned that the impact of this feature is more important for ICD-9-CM because of the existing limitation on the depth of the hierarchy. This limitation requires more precise properties (implying long value strings) to describe the concepts. We highlight that two features $all\_children(c^t)$ and $siblings(c^t)$ were pointed out by other authors as important for identifying concepts that need revision. In our experiments, we observed that their importance is low or even negative for $Evolv_k$. However, these structural properties may have (depending on the ontology) a positive impact, especially when predicting the extension.

Lastly, features dealing with temporal aspects have a strong positive impact on the two problems (identification and recommendation). When a concept belongs to a part of the ontology that evolves frequently, then the probability that this concept will evolve is high. This is what highlights the behaviour of the $Region\_stability(c^t)$ feature. Furthermore, $Last\_evolution(c^t)$ shows that stable concepts, i.e. concepts that have not changed over the past 10 years, have a lower chance of evolving in the future. This feature adds this notion to the model, which improves the quality of the prediction for all datasets.

During the implementation of this work, we held discussions with some potential end-users working in biomedical terminology management organizations in order to understand how the ontology maintenance tasks are planned and executed. Our understanding indicates that candidate concepts are selected based on ontology users' reports/requests and on the outcomes of specialized workgroups/task forces (we call this group of candidates manually generated candidates). Our approach can be integrated into this maintenance task to promote proactive maintenance action. We believe that the contributions of our approach are twofold: (1) It provides an extra source of information (that can potentially complement the manually generated one) with candidates and with suggestions of the type of revision that each concept needs; (2) It supports the creation of ranks (indicating priorities to execute changes in the ontologies) of concepts to be changed based on the combination of manually and automatically generated sets of candidate concepts. Our idea of selecting classifiers that have explainable models fits into this ambition of integrating manual and automatic candidate generation. Instead of trusting black boxes, the user can, for instance, extract the boosted tree decision rule and use it to create a "ranking rule".

A good integration of our approach in the ontology maintenance tasks can help improve the prediction model. The feedback of the users can be used to train the model in order to customize it (e.g., trained to apply to one specific ontology) and/or improve its precision.

Finally, we would like to reiterate that although our approach was developed to help reviewers in their daily tasks, our goal is not to substitute them. However, if the biomedical terminology management organizations decide that the prediction accuracy obtained with our approach is acceptable for some of their objectives, the role of our approach can change. For instance, if an organization decides that having a daily updated version of the ontology with some errors (e.g., beta version) is more relevant than keeping the same version of the ontology for a long period (e.g., one year), then little work would be required to implement an automatic ontology updates tool.

6.5 Conclusion

In this chapter, we analysed methods empirically to i) identify the concepts needing revision and ii) recommending the type of revision necessary answering RQ5 of this thesis. We selected the biomedical domain for our case study since it is highly dynamic and covered by more than 500 ontologies with distinct levels of expressivity. Thus, we collected information on 4 standard ontologies from this domain over a period of 10 years (a total of 40 ontologies were analysed). We also analysed extra information collected from publicly available sources like PubMed and UMLS. We grouped this information into temporal, structural and relational categories. Compared with the state-of-the-art, we proposed several new sub-groups for each category (features of our analysis process) and we evaluated their impact on the quality of our analysis. Finally, our analysis allows the most efficient supervised learning classifier technique and the most important set of features to be selected. We observed that the Boosted Trees Classifier is the best one for identifying concepts that need revision while Random Forest is more efficient for recommending the type of revision. In this work, we have observed that the consideration of background knowledge plays a key role in the identification of and recommendations for problems when sources like UMLS and BioPortal are available. We obtained good accuracy that fluctuates between 68% and 91% according to the dataset analysed; in an overall analysis (all datasets together), we achieved more than 71% accuracy. Next chapter will focus on the conclusion of our work, highlighting the main contributions of this thesis.

Chapter 7

Conclusions and perspectives

Contents

In this work, we have addressed the problem of the adaptation of semantic annotations impacted by the evolution of KOS. To tackle this main problem, two scenarios were identified based on real cases. In the first, annotations as well as the associated documents were available and modifiable while in the second, annotations were only accessible but not modifiable and no access to the associated documents was provided, for instance, for patient data inside Hospital Information Systems. We approached the problem with an empirical analysis of the evolution of the KOS and its impact on a set of millions of annotations. The result of this analysis showed a strong correlation between these two phenomena and has served as the basis of our solution. This has lead us to, on one hand, to the definition of a rule-based system to modify the annotations and, on the other hand, to the design of a knowledge graph for representing the evolution of concepts contained in KOS allowing us to retrieve the evolution of a concept and, consequently, the evolution of associated annotations.

7.1 Summary of the contributions

This PhD thesis has made contributions in the fields of Annotation Maintenance, KOS evolution, Semantic Similarity measures and dynamic Knowledge Graphs. We also believe that the research carried out in this thesis contributes to the development of the Semantic Web and its correlated fields, e.g. information retrieval and semantic interoperability. We summarize the main scientific contributions reflecting the research questions below:

RQ1: What is the impact of KOS changes on semantic annotations?

Through empirical analysis on factors influencing the evolution of annotations in chapter 2, we observed that the changes on annotations are strongly correlated to changes in the KOS. For this, we used a set of documents annotated with GATE and NCBO Annotator using 13 different versions of two well-known biomedical KOS (ICD-9-CM and MeSH). We also verified that the same behaviour occurred in the MEDLINE annotations utilized in Chapter 6. In this chapter, we utilized the impacted annotations as queries for our experiments.

RQ2: What is a suitable model for addressing the annotation evolution problem?

In chapter 2, we analysed different annotation models in order to verify whether they could represent (or whether we could infer from their elements) all the criteria required to classify the annotation changes. As a result, we proposed extensions for the W3C Open Annotation Data Model, e.g., the inclusion of an *evolved* relationship, which links an evolved annotation to its past version. These new features allow a better covering of evolutionary aspects, e.g. tracing back all the possible evolutions of an annotation, in order to maintain future versions through rules as *ResurrectAnnot* (see Chapter 3).

RQ.3: How can we automatically maintain the validity of semantic annotations without re-annotating the content of all documents when KOS are updated?

The novel MAISA framework presented the possibility of using methods such as Domain Specif Rules, Background Knowledge and Change Patterns to automatically keep semantic annotations up-to-date. Through empirical experiments, we demonstrated that the correction/adaptation of annotations can reach a reasonable reliability rate. However, it is important to highlight that the role of domain experts is still determinant in assuring the quality of the annotations in critical scenarios, as observed in the biomedical domain. Finally, our maintenance approach is done without a complete re-annotation of the document, since we reutilize the information present in the annotations to evolve the impacted ones.

RQ4: Which methods can be used to keep the annotations searchable when the document and annotations cannot be changed directly?

In chapter 5, we proposed a method able to keep impacted annotations searchable without directly changing them. The experimental analyses demonstrated that our proposed method was capable of achieving good results when querying EHRs and could be used as an alternative method to `BK`, utilized in chapter 3 to direct adapt the semantic annotations. We observed that using a dynamic *Knowledge Graph*, as we proposed, provides a good representation for multiple ontology versions and the evolutionary link between them. This approach contributes to the state-of-art of annotation maintenance by including a new method to cope with the evolution of annotations, based on techniques from the Evolving Graph Generator domain.

RQ5: Can we predict which KOS concepts will change and impact the annotations in the near future?

In chapter 6, we discussed how to enhance the MAISA framework by including a new method able to anticipate the evolution of concepts associated with annotations. This method supports domain experts during the annotation/maintenance phase by alerting them to the risks of choosing a concept and/or an ontology to annotate biomedical documents. Our approach showed more than 71% of accuracy when identifying concepts that needed revision. Furthermore, we contributed to the state-of-the-art by including temporal features, e.g. information about the last evolution date, which helped to increase the accuracy of the utilized methods (see Chapter 6).

In summary, all the main research questions investigated in this thesis were answered satisfactorily, allowing us to keep our hypothesis: *The use of information from KOS, as well as information about KOS evolution, can be used to define a robust maintenance mechanism.* We also approached more challenges that had emerged from our methodology in Section 5.6, which leaded to a complementary question.

Complementary Question: How can we improve the existing similarity measures to enhance the relatedness between terms while taking the syntactic mismatch into account?

In chapter 4, we presented an approach that combined Lexical and Semantic measures to enhance the concept similarity. Our experimental analysis demonstrated that together, they can outperform methods based on only one similarity measure (i.e., semantic or lexical). Using this hybrid measure, we introduced the `PartialMatch` rule for maintaining semantic annotations affected by the evolution of KOS. Our experimental analysis demonstrated that PartialMatch was capable of achieving good results to adapt annotations using one or multiple successive KOS versions. We observed that the use of semantic similarity approaches was important to determine the relatedness during the evolution process.

7.2 Directions for future work

Even though MAISA showed good results for keeping semantic annotations up-to-date when the utilized KOS had evolved, we identified some limitations offering relevant perspectives for further improvements and an extension of this work. The improvements listed in this section are related to the research field of semantic annotations and Semantic Web. They were not implemented due to the limited time to develop the work and the enlargement of scope.

We recommend the following enhancements to better evaluate our approach:

1. Conduct additional experiments using annotations and KOS from other domains outside of the biomedical domain. For instance, we can utilize the Eurovoc thesaurus[34], which was released after 2014, to verify whether the changes that emerged after a new version impacted annotations from LinkedEP dataset[35], i.e., the plenary debates of the European Parliament [Van Aggelen et al., 2014].

2. Qualitative evaluation involving domain experts. As highlighted during this thesis, domain experts play an important role in the final process. Therefore, we suggest to evaluating MAISA in a real environment, e.g. a hospital, to verify the user opinions.

3. The silver standard used to evaluate our experiments contains around 125 annotations per KOS. Therefore, we advise to enhancing the silver standard by including new cases of annotation maintenance. These new annotations can, for example, be evaluated using the Inter-annotator Agreement, i.e., a metric to quantify the agreement between the domain experts regarding the evolution of the annotations.

Regarding the maintenance process, we suggest the following topics to enhance the upcoming approaches.

1. Implement new Rules. As discussed in Chapter 4, domain specialists sometimes opted to reduce the expressiveness of the annotations. In further versions of MAISA or in approaches dealing with the evolution of annotations, rules describing how to cover such cases must be implemented. Moreover, undiscovered cases in this thesis can be explored through the utilization of techniques as Association Rule Mining [Hipp et al., 2000]. These techniques allow the discovering of new patterns from to be discovered in a large amount of data.

2. Improve the search algorithm utilized in the Ad-hoc phase by using more sophisticated techniques, such as [Marie et al., 2013]. Nevertheless, the temporal aspect also must be considered, since it is not present in the approaches from the state-of-the-art.

3. Improve the method to enrich queries, paying particular attention to the query language [Pruski et al., 2011]. In this work, we simply construct queries as a conjunction of terms favouring recall on precision. Additional work can propose a query language that offers other operators leading to complex queries giving better precision and recall when searching for documents.

[34] https://publications.europa.eu/en/web/eu-vocabularies/

[35] http://linkedpolitics.ops.few.vu.nl/web/html/home.html

4. [Pakhomov et al., 2010] utilized a touch screen to verify the relatedness of UMLS concepts and domain expert judgments. Inspired in the approach, advanced Human Interface techniques can also be used to understand how domain experts evolve their annotations and create new rules.

Finally, we can explore several research topics representing the extension of this thesis:

1. In chapter 5, we presented an `Ad-hoc` method to enhance the ontology information with its evolution. Our results demonstrated that the ontology history contains meaningful information to describe the concepts and their relatedness over time. The utilization of ontology changes and the `Ad-hoc` method which verifies the relatedness of concepts can also be explored in the context of Concept Learning/ Ontology Learning techniques, more specifically in those focusing on the T-box part of the ontology, i.e. the concepts [Zhang et al., 2016]. The current approaches are applied in general on small ontologies due the use of reasoning techniques during the learning phase. Different approaches in the same domain, e.g., [Bamunusinghe and Alahakoon, 2007] implementing neural networks, do not cover temporal aspects relying on a single ontology version and working as a black box, because the adjustments made by the neural network techniques are complex to explain.

2. The evolution of ontologies plays a key role when adapting semantic annotations. Therefore, models and/or languages that better describe how the concepts must evolve in new ontology versions can be an asset to mechanisms that adapt semantic annotations.

3. Extend the maintenance of semantic annotations to the Linked Open Data (LOD). In this scenario, methods that automatically adapt the existing links and annotations on the LOD by creating new ones and linking them to past versions could be explored.

4. Implement a measure that combines LSM and SSMs and also considers the KOS evolution. As observed in chapter 4, the use of lexical similarity approaches is important to improve the quality of the results provided by the SSMs and vice-versa. Nonetheless, these measures do not contains evolutionary aspects, therefore we propose the investigation of similarity measures that also consider the KOS changes.

In summary, this thesis contributed with new approaches to direct or indirect maintain the semantic annotations impacted by the evolution of KOS. The shortcomings observed in the state-of-the-art were considered and utilized to build the MAISA architecture. This architecture contains an original method to maintain the annotations indirectly that can reach reasonable results. Overall, the results obtained during the thesis were encouraging and allowed new improvements in the domain. All work described in this thesis has been the subject of several papers published at international conferences and in international journals. A list of these publications is given on the next page.

Publications

International Journal

- **Cardoso, S. D.**, Pruski, C. and Da Silveira, M. Supporting biomedical ontology evolution by identifying outdated concepts and the required type of change. Journal of Biomedical Informatics, Volume 87, 2018, Pages 1-11, ISSN 1532-0464, `https://doi.org/10.1016/j.jbi.2018.08.013`.

- **Silvio Cardoso**, Chantal Reynaud-Delaître, Marcos Da Silveira, Ying-Chi Lin, Anika Groß, Erhard Rahm, Cédric Pruski. Evolving semantic annotations through multiple versions of controlled medical terminologies. Health and Technology, November 2018, Volume 8, Issue 5, pp 361-376. `https://doi:10.1007/s12553-018-0261-3`

International Conference

- **Silvio Cardoso**, Marcos Da Silveira, Ying-Chi Lin, Victor Christen, Erhard Rahm, Chantal Reynaud and Cédric Pruski. Combining semantic and lexical measures to evaluate medical terms similarity. Data Integration in the Life Sciences. Hannover, Germany, 2018.

- Victor Christen, Ying-Chi Lin, Anika Groß, **Silvio Domingos Cardoso**, Cédric Pruski, Marcos Da Silveira and Erhard Rahm. A learning-based approach to combine medical annotation results. Data Integration in the Life Sciences. Hannover, Germany, 2018.

- **Cardoso, S. D.**, Reynaud-Delaître, C. Da Silveira, M. and Pruski, C. Combining rules, background knowledge and change patterns to maintain semantic annotations. AMIA 2017 Annual Symposium. Nov 4-8 Washington, D.C.

- Ying-Chi, L., Christen, V., **Cardoso, S.D.**, Da Silveira, M., Pruski, C., Rahm, E.: Evaluating and improving annotation tools for medical forms. In: 12th International Conference on Data Integration in Life Sciences (2017)

- Chaabane M., **Cardoso S. D.**, Pruski C. and Da Silveira M. DyKOSMap : from a prototype to a web application. In: 12th International Conference on Data Integration in Life Sciences (2017)

- **Cardoso, S. D.**, Pruski, C., Da Silveira, M., Lin, Y., Groß, A., Rahm, E., and Reynaud-Delaître, C. (2016). Towards a Multi-level Approach for the Maintenance of Semantic Annotations. In 10th International Conference on Health Informatics, HEALTHINF 2017, Porto, Portugal, February 21-23, 2017, pages 401-406.

- **Cardoso, S. D.**, Automatic maintenance of semantic annotations. In Doctoral Consortium Knowledge Engineering and Knowledge Management - 20th International Conference, EKAW 2016, Bologna, Italy, November 19-23, 2016.

- **Cardoso, S. D.**, Pruski, C., Da Silveira, M., Lin, Y., Groß, A., Rahm, E., and Reynaud-Delaître, C. (2016). Leveraging the impact of ontology evolution on semantic annotations. In Knowledge Engineering and Knowledge Management - 20th International Conference, EKAW 2016, Bologna, Italy, November 19-23, 2016, Proceedings, pages 68-82.

Appendix

Table 1: Experiments to determine the place of `PartialMatch`. The blue values are related to the best performance.

KOS	Year	Rule positioning	Accuracy	AUC	F1-Score
ICD-9-CM	2009/2010	Before	0.834	0.862	0.839
		After	0.845	0.871	0.851
	2009/2016	Before	0.757	0.803	0.754
		After	0.741	0.789	0.733
Mesh	2009/2010	Before	0.867	0.891	0.877
		After	0.851	0.878	0.861
	2009/2016	Before	0.88	0.905	0.895
		After	0.859	0.888	0.874
NCIt	2009/2010	Before	0.767	0.798	0.747
		After	0.697	0.735	0.64
	2009/2016	Before	0.753	0.804	0.756
		After	0.685	0.75	0.667
SNOMED-CT	2009/2010	Before	0.849	0.854	0.829
		After	0.839	0.844	0.815
	2009/2016	Before	0.912	0.923	0.917
		After	0.923	0.933	0.928

Table 2: Overall Rank in MeSH. The values inside the columns five to eight indicates the obtained ranks.

Lexical	**IC**	**Semantic**	**Alpha and Tau**	**Coders**	**Physicians**	**Mayo**	**UMNSRS**	$\bar{X}$	$\sum$	σ
AnnoMap	Seco 2004	Jiang Conrath 1997 Norm	0.8 and 0.4 1.0 and 0.5	138.0	137	36.0	256.0	137.5	567.0	89.95
AnnoMap	Resnik Unpropagated 1995	Jiang Conrath 1997 Norm	0.8 and 0.4	138.0	137	36.0	256.0	137.5	567.0	89.95
AnnoMap	Seco 2004	Jiang Conrath 1997 Norm	0.4 and 0.2 0.2 and 0.1 0.6 and 0.3	138.0	137	36.0	257.0	137.5	568.0	90.37
AnnoMap	Resnik Unpropagated 1995	Jiang Conrath 1997 Norm	0.4 and 0.2 1.0 and 0.5 0.6 and 0.3 0.2 and 0.1 0.9 and 0.5	138.0	137	36.0	257.0	137.5	568.0	90.37
AnnoMap	Seco 2004	Jiang Conrath 1997 Norm	0.9 and 0.5	138.0	137	23.0	273.0	137.5	571.0	102.24
SmithWaterman	Resnik Unpropagated 1995	Jiang Conrath 1997 Norm	0.4 and 0.1 0.8 and 0.2	174.0	174	88.0	140.0	157.0	576.0	40.63
SmithWaterman	Seco 2004	Jiang Conrath 1997 Norm	0.4 and 0.1 0.8 and 0.2	174.0	174	88.0	140.0	157.0	576.0	40.63
AnnoMap	Seco 2004	Jiang Conrath 1997 Norm	0.8 and 0.5	138.0	137	7.0	296.0	137.5	578.0	118.26
AnnoMap	Resnik Unpropagated 1995	Jiang Conrath 1997 Norm	0.8 and 0.5	138.0	137	7.0	297.0	137.5	579.0	118.69

Table 3: Overall Rank in SNOMED CT. The values inside the columns five to eight indicates the obtained ranks.

Lexical	**IC**	**Semantic**	**Alpha and Tau**	**Coders**	**Physicians**	**Mayo**	**UMNSRS**	$\sum$	$\bar{X}$	σ
AnnoMap	Sanchez 2011 b adapted	Jiang Conrath 1997 Norm	1.0 and 0.9	183	195	69	700	1147	189	281.3
AnnoMap	Resnik Unpropagated 1995	Sim IC 2010	0.3 and 0.1 0.6 and 0.2 0.9 and 0.3	42	201	45	886	1174	123	401.92
AnnoMap	Seco 2004	Sim IC 2010	0.6 and 0.2 0.9 and 0.3 0.3 and 0.1	42	201	47	886	1176	124	401.5
AnnoMap	Resnik Unpropagated 1995	Sim IC 2010	1.0 and 0.3	39	184	132	828	1183	158	359.87
AnnoMap	Seco 2004	Sim IC 2010	1.0 and 0.3	39	184	132	828	1183	158	359.87
AnnoMap	Sanchez 2011 b adapted	Jiang Conrath 1997 Norm	0.7 and 0.6	232	261	64	650	1207	246.5	247.87
AnnoMap	Seco 2004	Sim IC 2010	0.9 and 0.3 0.3 and 0.1 0.6 and 0.2	42	252	74	859	1227	163	379.58
Anno	Sanchez 2011 b adapted	Jiang Conrath 1997 Norm	0.8 and 0.7	232	261	65	671	1229	246.5	257.42
AnnoMap	Sanchez 2011 b adapted	Jiang Conrath 1997 Norm	0.1 and 0.1 0.2 and 0.2 0.3 and 0.3 0.4 and 0.4 0.5 and 0.5 0.6 and 0.6 0.7 and 0.7 0.8 and 0.8 0.9 and 0.9 1.0 and 1.0	190	209	46	791	1236	199.5	329.47

References

[Abgaz, 2013] Abgaz, Y. M. (2013). *Change impact analysis for evolving ontology-based content management*. PhD thesis, Dublin City University. Citation on page 22.

[Akiba et al., 2015] Akiba, T., Hayashi, T., Nori, N., Iwata, Y., and Yoshida, Y. (2015). Efficient top-k shortest-path distance queries on large networks by pruned landmark labeling. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, pages 2–8. AAAI Press. Citation on page 63.

[Akiba et al., 2014] Akiba, T., Iwata, Y., and Yoshida, Y. (2014). Dynamic and historical shortest-path distance queries on large evolving networks by pruned landmark labeling. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, pages 237–248. ACM. Citations on pages 62 and 63.

[Alami et al., 2017] Alami, K., Ciucanu, R., and Mephu Nguifo, E. (2017). EGG: A framework for generating evolving RDF graphs. In *ISWC Posters & Demonstrations*. Citation on page 62.

[Auer and Herre, 2007] Auer, S. and Herre, H. (2007). A versioning and evolution framework for rdf knowledge bases. *Perspectives of Systems Informatics: 6th International Andrei Ershov Memorial Conference, PSI 2006, Novosibirsk, Russia, June 27-30, 2006. Revised Papers*, pages 55–69. Citation on page 22.

[Bagan et al., 2017] Bagan, G., Bonifati, A., Ciucanu, R., Fletcher, G. H. L., Lemay, A., and Advokaat, N. (2017). gmark: Schema-driven generation of graphs and queries. *IEEE Transactions on Knowledge and Data Engineering*, 29(4):856–869. Citation on page 62.

[Bamunusinghe and Alahakoon, 2007] Bamunusinghe, J. and Alahakoon, D. (2007). Concept learning from visual experiences using unsupervised neural networks. In *2007 Third International Conference on Information and Automation for Sustainability*, pages 46–51. Citation on page 98.

[Belleau et al., 2008] Belleau, F., Nolin, M.-A., Tourigny, N., Rigault, P., and Morissette, J. (2008). Bio2rdf: towards a mashup to build bioinformatics knowledge systems. *Journal of biomedical informatics*, 41(5):706–716. Citation on page 26.

[Bodenreider, 2004] Bodenreider, O. (2004). The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270. Citation on page 81.

[Burger et al., 2010] Burger, T., Morozova, O., Zaihrayeu, I., Andrews, P., and Pane, J. (2010). Report on methods and algorithms for linking user-generated semantic annotations to semantic web and supporting their evolution in time. Technical report, University of Trento. Citation on page 22.

[Butt et al., 2015] Butt, A. S., Haller, A., and Xie, L. (2015). A taxonomy of semantic web data retrieval techniques. In *Proceedings of the 8th International Conference on Knowledge Capture*, K-CAP 2015, pages 9:1–9:9, New York, NY, USA. ACM. Citations on pages 60 and 63.

[Cano-Basave et al., 2016] Cano-Basave, A. E., Osborne, F., and Salatino, A. A. (2016). Ontology forecasting in scientific literature: Semantic concepts prediction based on innovation-adoption priors. In *Knowledge Engineering and Knowledge Management: 20th International Conference, EKAW 2016, Bologna, Italy, November 19-23, 2016, Proceedings 20*, pages 51–67. Springer. Citation on page 80.

[Cardoso et al., 2016] Cardoso, S. D., Pruski, C., Da Silveira, M., Lin, Y.-C., Groß, A., Rahm, E., and Reynaud-Delaître, C. (2016). Leveraging the impact of ontology evolution on semantic annotations. In Blomqvist, E., Ciancarini, P., Poggi, F., and Vitali, F., editors, *Knowledge Engineering and Knowledge Management*, pages 68–82, Cham. Springer International Publishing. Citations on pages 23, 30, 82, and 84.

[Cardoso et al., 2017a] Cardoso, S. D., Reynaud-Delaître, C., Da Silveira, M., and Pruski, C. (2017a). Combining rules, background knowledge and change patterns to maintain semantic annotations. *AMIA Annu Symp Proc*, 2017:505–514. Citations on pages xi and 45.

[Cardoso et al., 2017b] Cardoso, S. D., Reynaud-Delaître, C., Silveira, M. D., Lin, Y.-C., Groß, A., Rahm, E., and Pruski, C. (2017b). Towards a multi-level approach for the maintenance of semantic annotations. In *Proceedings of the 10th International Joint Conference on Biomedical Engineering Systems and Technologies - Volume 5: HEALTHINF, (BIOSTEC 2017)*, pages 401–406. INSTICC, SciTePress. Citations on pages ix and 24.

[Caro et al., 2015] Caro, D., Rodríguez, M. A., and Brisaboa, N. R. (2015). Data structures for temporal graphs based on compact sequence representations. *Inf. Syst.*, 51(C):1–26. Citations on pages 62 and 64.

[Chandrashekar and Sahin, 2014] Chandrashekar, G. and Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering*, 40(1):16 – 28. 40th-year commemorative issue. Citation on page 78.

[Costa and Leal, 2016] Costa, T. and Leal, J. P. (2016). *Semantic Measures: How Similar? How Related?*, pages 431–438. Springer International Publishing, Cham. Citation on page 38.

[Couto et al., 2006] Couto, F. M., Silva, M. J., Lee, V., Dimmer, E., Camon, E., Apweiler, R., Kirsch, H., and Rebholz-Schuhmann, D. (2006). Goannotator: linking protein go annotations to evidence text. *Journal of Biomedical Discovery and Collaboration*, 1(1):19. Citation on page 38.

[Cunningham, 2002] Cunningham, H. (2002). Gate, a general architecture for text engineering. *Computers and the Humanities*, 36(2):223–254. Citation on page 12.

[Da Silveira et al., 2015] Da Silveira, M., Dos Reis, J. C., and Pruski, C. (2015). Management of dynamic biomedical terminologies: Current status and future challenges. *Yearbook of Medical informatics*, 10(1):125–133. Citations on pages 14 and 49.

[Debatty et al., 2016] Debatty, T., Michiardi, P., and Mees, W. (2016). Fast online k-nn graph building. *CoRR*, abs/1602.06819. Citations on pages 62, 63, 64, 66, and 67.

[Dinh et al., 2014] Dinh, D., Dos Reis, J. C., Pruski, C., Da Silveira, M., and Reynaud-Delaître, C. (2014). Identifying change patterns of concept attributes in ontology evolution. In *European Semantic Web Conference*, pages 768–783. Springer. Citation on page 28.

[Dixon and Mood, 1946] Dixon, W. J. and Mood, A. M. (1946). The statistical sign test. *Journal of the American Statistical Association*, 41(236):557–566. Citations on pages 41, 47, and 56.

[Doğan et al., 2014] Doğan, R. I., Leaman, R., and Lu, Z. (2014). NCBI disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47:1–10. Citations on pages 13, 14, 21, 24, 25, and 26.

[Domingos, 1999] Domingos, P. (1999). Metacost: A general method for making classifiers cost-sensitive. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '99, pages 155–164, New York, NY, USA. ACM. Citation on page 91.

[Dong et al., 2011] Dong, W., Moses, C., and Li, K. (2011). Efficient k-nearest neighbor graph construction for generic similarity measures. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pages 577–586, New York, NY, USA. ACM. Citation on page 63.

[Dos Reis et al., 2015a] Dos Reis, J. C., Dinh, D., Da Silveira, M., Pruski, C., and Reynaud-Delaître, C. (2015a). Recognizing lexical and semantic change patterns in evolving life science ontologies to inform mapping adaptation. *Artificial intelligence in medicine*, 63(3):153–170. Citations on pages 21 and 38.

[Dos Reis et al., 2014] Dos Reis, J. C., Pruski, C., Da Silveira, M., and Reynaud-Delaître, C. (2014). Understanding semantic mapping evolution by observing changes in biomedical ontologies. *Journal of biomedical informatics*, 47:71–82. Citations on pages 1, 82, and 91.

[Dos Reis et al., 2015b] Dos Reis, J. C., Pruski, C., Da Silveira, M., and Reynaud-Delaître, C. (2015b). DyKOSMap: A framework for mapping adaptation between biomedical knowledge organization systems. *Journal of biomedical informatics*, 55:153–173. Citation on page 28.

[Eilbeck et al., 2009] Eilbeck, K., Moore, B., Holt, C., and Yandell, M. (2009). Quantitative measures for the management and comparison of annotated genomes. *BMC Bioinformatics*, 10(1):67. Citations on pages 22 and 25.

[El-Dsouky et al., 2016] El-Dsouky, A. I., Ali, H. A., and Rashed, R. S. (2016). Ranking documents based on the semantic relations using analytical hierarchy process. *International Journal of Advanced Computer Science and Applications*, 7(2). Citations on pages 60 and 63.

[Esch et al., 2015] Esch, M., Chen, J., Colmsee, C., Klapperstück, M., Grafahrend-Belau, E., Scholz, U., and Lange, M. (2015). Lailaps: The plant science search engine. *Plant and Cell Physiology*, 56(1):e8. Citations on pages 60 and 63.

[Frost and Moore, 2014] Frost, H. R. and Moore, J. H. (2014). Optimization of gene set annotations via entropy minimization over variable clusters (emvc). *Bioinformatics (Oxford, England)*, 30(12):1698–1706. Citation on page 22.

[Funk et al., 2014] Funk, C., Baumgartner, W., Garcia, B., Roeder, C., Bada, M., Cohen, K. B., Hunter, L. E., and Verspoor, K. (2014). Large-scale biomedical concept recognition: an evaluation of current automatic annotators and their parameters. *BMC Bioinformatics*, 15(1):1–29. Citations on pages 10 and 21.

[Garla and Brandt, 2012] Garla, V. N. and Brandt, C. (2012). Semantic similarity in the biomedical domain: an evaluation across knowledge sources. *BMC Bioinformatics*, 13(1):261. Citation on page 38.

[Gomaa and Fahmy, 2013] Gomaa, W. H. and Fahmy, A. A. (2013). A survey of text similarity approaches. *International Journal of Computer Applications*, 68(13). Citation on page 38.

[Gonçalves et al., 2011] Gonçalves, R. S., Parsia, B., and Sattler, U. (2011). Categorising logical differences between owl ontologies. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 1541–1546. ACM. Citation on page 79.

[Groß et al., 2013] Groß, A., Dos Reis, J. C., Hartung, M., Pruski, C., and Rahm, E. (2013). Semi-automatic adaptation of mappings between life science ontologies. In *International Conference on Data Integration in the Life Sciences*, pages 90–104. Springer. Citation on page 23.

[Groß et al., 2009] Groß, A., Hartung, M., Kirsten, T., and Rahm, E. (2009). Estimating the quality of ontology-based annotations by considering evolutionary changes. In *DILS*, pages 71–87. Citations on pages 8 and 18.

[Groß et al., 2012] Groß, A., Hartung, M., Prüfer, K., Kelso, J., and Rahm, E. (2012). Impact of ontology evolution on functional analyses. *Bioinformatics*, 28(20):2671–2677. Citations on pages 10 and 11.

[Groß et al., 2016] Groß, A., Pruski, C., and Rahm, E. (2016). Evolution of biomedical ontologies and mappings: Overview of recent approaches. *Computational and structural biotechnology journal*, 14:333–340. Citation on page 79.

[Gruber, 1993] Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge acquisition*, 5(2):199–220. Citation on page 4.

[Gusfield, 1997] Gusfield, D. (1997). *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology.* Cambridge University Press, New York, NY, USA. Citation on page 48.

[Hajian-Tilaki, 2013] Hajian-Tilaki, K. (2013). Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. *Caspian Journal of Internal Medicine*, 4(2):627–635. Citation on page 32.

[Harispe et al., 2014] Harispe, S., Sánchez, D., Ranwez, S., Janaqi, S., and Montmain, J. (2014). A framework for unifying ontology-based semantic similarity measures: A study in the biomedical domain. *Journal of Biomedical Informatics*, 48:38 – 53. Citations on pages 38 and 44.

[Hartung et al., 2013] Hartung, M., Groß, A., and Rahm, E. (2013). Conto–diff: generation of complex evolution mappings for life science ontologies. *Journal of biomedical informatics*, 46(1):15–32. Citations on pages 12, 23, 30, 66, 69, 79, and 81.

[Hartung et al., 2009] Hartung, M., Kirsten, T., Gross, A., and Rahm, E. (2009). Onex: Exploring changes in life science ontologies. *BMC bioinformatics*, 10(1):1. Citation on page 79.

[Hartung et al., 2008] Hartung, M., Kirsten, T., and Rahm, E. (2008). Analyzing the evolution of life science ontologies and mappings. In *DILS*, pages 11–27. Citation on page 8.

[Hearst et al., 1998] Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., and Scholkopf, B. (1998). Support vector machines. *IEEE Intelligent Systems and their Applications*, 13(4):18–28. Citation on page 82.

[Hipp et al., 2000] Hipp, J., Güntzer, U., and Nakhaeizadeh, G. (2000). Algorithms for association rule mining &mdash; a general survey and comparison. *SIGKDD Explor. Newsl.*, 2(1):58–64. Citation on page 97.

[Hodge, 2000] Hodge, G. (2000). Systems of knowledge organization for digital libraries: Beyond traditional authority files. Reports - Descriptive. Citation on page 1.

[Huntley et al., 2014] Huntley, R. P., Sawford, T., Martin, M. J., and O'Donovan, C. (2014). Understanding how and why the gene ontology and its annotations evolve: the go within uniprot. *GigaScience*, 3(4). Citation on page 11.

[Ishioka, 2003] Ishioka, T. (2003). Evaluation of criteria for information retrieval. In *Proceedings IEEE/WIC International Conference on Web Intelligence (WI 2003)*, pages 425–431. Citation on page 72.

[Jiang and Conrath, 1997] Jiang, J. J. and Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. *CoRR*, cmp-lg/9709008. Citation on page 26.

[Khurana and Deshpande, 2013] Khurana, U. and Deshpande, A. (2013). Efficient snapshot retrieval over historical graph data. In *2013 IEEE 29th International Conference on Data Engineering (ICDE)*, pages 997–1008. Citation on page 62.

[Kirsten et al., 2009] Kirsten, T., Hartung, M., Groß, A., and Rahm, E. (2009). Efficient management of biomedical ontology versions. In Meersman, R., Herrero, P., and Dillon, T., editors, *On the Move to Meaningful Internet Systems: OTM 2009 Workshops*, pages 574–583, Berlin, Heidelberg. Springer Berlin Heidelberg. Citation on page 62.

[Klein et al., 2002] Klein, M., Fensel, D., Kiryakov, A., and Ognyanov, D. (2002). Ontology versioning and change detection on the web. *Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web*, pages 247–259. Citations on pages 77 and 79.

[Konev et al., 2012] Konev, B., Ludwig, M., and Wolter, F. (2012). Logical difference computation with cex2. 5. *Automated Reasoning*, pages 371–377. Citation on page 79.

[Köpke and Eder, 2011] Köpke, J. and Eder, J. (2011). Semantic invalidation of annotations due to ontology evolution. In Meersman, R., Dillon, T., Herrero, P., Kumar, A., Reichert, M., Qing, L., Ooi, B.-C., Damiani, E., Schmidt, D., White, J., Hauswirth, M., Hitzler, P., and Mohania, M., editors, *On the Move to Meaningful Internet Systems: OTM 2011*, volume 7045 of *Lecture Notes in Computer Science*, pages 763–780. Springer Berlin Heidelberg. Citation on page 22.

[Kosmatopoulos et al., 2016] Kosmatopoulos, A., Giannakopoulou, K., Papadopoulos, A. N., and Tsichlas, K. (2016). An overview of methods for handling evolving graph sequences. In *Revised Selected Papers of the First International Workshop on Algorithmic Aspects of Cloud Computing - Volume 9511*, ALGOCLOUD 2015, pages 181–192, Berlin, Heidelberg. Springer-Verlag. Citation on page 62.

[Kotsiantis et al., 2006] Kotsiantis, S., Kanellopoulos, D., Pintelas, P., et al. (2006). Handling imbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering*, 30(1):25–36. Citation on page 82.

[Krejcie and Morgan, 1970] Krejcie, R. V. and Morgan, D. W. (1970). Determining sample size for research activities. *Educational and Psychological Measurement*, 30(3):607–610. Citation on page 83.

[Labouseur et al., 2015] Labouseur, A. G., Birnbaum, J., Olsen, Jr., P. W., Spillane, S. R., Vijayan, J., Hwang, J.-H., and Han, W.-S. (2015). The G* graph database: Efficiently managing large distributed dynamic graphs. *Distrib. Parallel Databases*, 33(4):479–514. Citation on page 62.

[Lee et al., 2016] Lee, S., Kim, D., Lee, K., Choi, J., Kim, S., Jeon, M., Lim, S., Choi, D., Kim, S., Tan, A.-C., and Kang, J. (2016). Best: Next-generation biomedical entity search tool for knowledge discovery from biomedical literature. *PLOS ONE*, 11(10):1–16. Citations on pages 60 and 63.

[Lin et al., 2017] Lin, Y.-C., Christen, V., Groß, A., Cardoso, S. D., Pruski, C., Da Silveira, M., and Rahm, E. (2017). Evaluating and improving annotation tools for medical forms. In Da Silveira, M., Pruski, C., and Schneider, R., editors, *Data Integration in the Life Sciences*, pages 1–16, Cham. Springer International Publishing. Citations on pages 11, 44, and 47.

[Lord et al., 2003] Lord, P. W., Stevens, R. D., Brass, A., and Goble, C. A. (2003). Investigating semantic similarity measures across the gene ontology: the relationship between sequence and annotation. *Bioinformatics*, 19(10):1275–1283. Citation on page 38.

[Lowe and Barnett, 1994] Lowe, H. J. and Barnett, G. O. (1994). Understanding and using the medical subject headings (mesh) vocabulary to perform literature searches. *Jama*, 271(14):1103–1108. Citation on page 1.

[Luong and Dieng-Kuntz, 2006] Luong, P.-H. and Dieng-Kuntz, R. (2006). A rule-based approach for semantic annotation evolution in the coswem system. In Kone, M. and Lemire, D., editors, *Canadian Semantic Web*, volume 2 of *Semantic Web and Beyond*, pages 103–120. Springer US. Citations on pages 8 and 22.

[Malone and Stevens, 2013] Malone, J. and Stevens, R. (2013). Measuring the level of activity in community built bio-ontologies. *Journal of biomedical informatics*, 46(1):5–14. Citation on page 79.

[Marie et al., 2013] Marie, N., Gandon, F., Myriam, R., and Rodio, F. (2013). Discovery Hub: on-the-fly linked data exploratory search. In *I-Semantics 2013*, Graz, Austria. Citation on page 97.

[Maynard et al., 2007] Maynard, D., Peters, W., and Sabou, M. (2007). Change management for metadata evolution. Citation on page 22.

[McInnes and Pedersen, 2013] McInnes, B. T. and Pedersen, T. (2013). Evaluating measures of semantic similarity and relatedness to disambiguate terms in biomedical text. *Journal of Biomedical Informatics*, 46(6):1116 – 1124. Special Section: Social Media Environments. Citation on page 43.

[Meng et al., 2012] Meng, L., Gu, J., and Zhou, Z. (2012). A new model of information content based on concept's topology for measuring semantic similarity in wordnet 1. *International Journal of Grid and Distributed Computing*, 5(3):81–94. Citation on page 48.

[Meroño-Peñuela et al., 2017] Meroño-Peñuela, A., Ashkpour, A., and Guéret, C. (2017). CEDAR: The dutch historical censuses as linked open data. *Semantic Web*, 8(2):297–310. Citation on page 80.

[Meroño-Peñuela et al., 2013] Meroño-Peñuela, A., Guéret, C., Hoekstra, R., Schlobach, S., et al. (2013). Detecting and reporting extensional concept drift in statistical linked data. *1st*

International Workshop on Semantic Statistics (SemStats 2013), ISWC. CEUR. Citation on page 80.

[Meymandpour and Davis, 2016] Meymandpour, R. and Davis, J. G. (2016). A semantic similarity measure for linked data: An information content-based approach. *Knowledge-Based Systems*, 109:276 – 293. Citation on page 38.

[Mihalcea et al., 2006] Mihalcea, R., Corley, C., Strapparava, C., et al. (2006). Corpus-based and knowledge-based measures of text semantic similarity. In *AAAI*, volume 6, pages 775–780. Citation on page 38.

[Moffitt and Stoyanovich, 2017] Moffitt, V. Z. and Stoyanovich, J. (2017). Towards sequenced semantics for evolving graphs. In *EDBT*. Citations on pages 62 and 64.

[Noy et al., 2002] Noy, N. F., Musen, M. A., et al. (2002). Promptdiff: A fixed-point algorithm for comparing ontology versions. *AAAI/IAAI*, 2002:744–750. Citation on page 79.

[Noy et al., 2009] Noy, N. F., Shah, N. H., Whetzel, P. L., Dai, B., Dorf, M., Griffith, N., Jonquet, C., Rubin, D. L., Storey, M.-A., Chute, C. G., et al. (2009). Bioportal: ontologies and integrated data resources at the click of a mouse. *Nucleic acids research*, page gkp440. Citation on page 26.

[Oliva et al., 2011] Oliva, J., Serrano, J. I., del Castillo, M. D., and Iglesias, Á. (2011). Symss: A syntax-based measure for short-text semantic similarity. *Data & Knowledge Engineering*, 70(4):390–405. Citation on page 38.

[Oren et al., 2006] Oren, E., Möller, K., Scerri, S., Handschuh, S., and Sintek, M. (2006). What are semantic annotations? *Technical Report. DERI Galway*, 9:62. Citation on page 7.

[Pakhomov et al., 2010] Pakhomov, S., McInnes, B., Adam, T., Liu, Y., Pedersen, T., and Melton, G. B. (2010). Semantic similarity and relatedness between clinical terms: An experimental study. *AMIA. Annual Symposium proceedings. AMIA Symposium*, 2010:572–6. Citations on pages 43, 49, and 98.

[Pakhomov et al., 2011] Pakhomov, S. V., Pedersen, T., McInnes, B., Melton, G. B., Ruggieri, A., and Chute, C. G. (2011). Towards a framework for developing semantic relatedness reference standards. *Journal of Biomedical Informatics*, 44(2):251 – 265. Citation on page 43.

[Park et al., 2011] Park, Y. R., Kim, J. J. H., Lee, H. W., and Yoon, Y. J. (2011). GOChase-II: correcting semantic inconsistencies from Gene Ontology-based annotations for gene products. *BMC Bioinformatics*, 12 Suppl 1(Suppl 1):S40. Citation on page 22.

[Peng et al., 2018] Peng, J., Zhang, X., Hui, W., Lu, J., Li, Q., Liu, S., and Shang, X. (2018). Improving the measurement of semantic similarity by combining gene ontology and co-functional network: a random walk based approach. *BMC Systems Biology*, 12(2):18. Citation on page 39.

[Pereira Nunes et al., 2013] Pereira Nunes, B., Dietze, S., Casanova, M. A., Kawase, R., Fetahu, B., and Nejdl, W. (2013). Combining a co-occurrence-based and a semantic measure for entity linking. In Cimiano, P., Corcho, O., Presutti, V., Hollink, L., and Rudolph, S., editors, *The Semantic Web: Semantics and Big Data*, pages 548–562, Berlin, Heidelberg. Springer Berlin Heidelberg. Citation on page 39.

[Pesquita and Couto, 2012] Pesquita, C. and Couto, F. M. (2012). Predicting the extension of biomedical ontologies. *PLoS Comput Biol*, 8(9):e1002630. Citations on pages 49, 78, 79, and 80.

[Pesquita et al., 2009] Pesquita, C., Faria, D., Falcão, A. O., Lord, P., and Couto, F. M. (2009). Semantic similarity in biomedical ontologies. *PLOS Computational Biology*, 5(7):1–12. Citation on page 38.

[Powers, 2011] Powers, D. M. (2011). Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. *Journal of Machine Learning Technologies*. Citations on pages 47 and 72.

[Press et al., 1988] Press, W. H., Flannery, B. P., Teukolsky, S. A., and Vetterling, W. T. (1988). *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, New York, NY, USA. Citations on pages 39 and 40.

[Pruski et al., 2016] Pruski, C., Dos Reis, J. C., and Da Silveira, M. (2016). Capturing the relationship between evolving biomedical concepts via background knowledge. In *the 9th Semantic Web Applications and Tools for Life Sciences International Conference*. Citations on pages 21 and 26.

[Pruski et al., 2011] Pruski, C., Guelfi, N., and Reynaud, C. (2011). Adaptive ontology-based web information retrieval: The target framework. *International Journal of Web Portals (IJWP)*, 3(3):41–58. Citation on page 97.

[Qin and Atluri, 2009] Qin, L. and Atluri, V. (2009). Evaluating the validity of data instances against ontology evolution over the semantic web. *Information and Software Technology*, 51(1):83 – 97. Citation on page 22.

[Rashid and Nisar, 2016] Rashid, J. and Nisar, M. W. (2016). A study on semantic searching, semantic search engines and technologies used for semantic search engines. *Information Technology and Computer Science(IJITCS)*, 8(10):82–89. Citations on pages 60 and 63.

[Resnik, 1995a] Resnik, P. (1995a). Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1*, IJCAI'95, pages 448–453, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc. Citation on page 38.

[Resnik, 1995b] Resnik, P. (1995b). Using information content to evaluate semantic similarity in a taxonomy. *CoRR*, abs/cmp-lg/9511007. Citation on page 44.

[Roberts et al., 2016] Roberts, K., Simpson, M., Demner-Fushman, D., Voorhees, E., and Hersh, W. (2016). State-of-the-art in biomedical literature retrieval for clinical cases: a survey of the trec 2014 cds track. *Information Retrieval Journal*, 19(1):113–148. Citations on pages 60 and 63.

[Rudniy et al., 2014] Rudniy, A., Song, M., and Geller, J. (2014). Mapping biological entities using the longest approximately common prefix method. *BMC Bioinformatics*, 15(1):187. Citations on pages 38 and 44.

[Sabou et al., 2008] Sabou, M., d'Aquin, M., and Motta, E. (2008). Exploring the semantic web as background knowledge for ontology matching. *Journal on data semantics XI*, pages 156–190. Citation on page 84.

[Sánchez et al., 2012] Sánchez, D., Solé-Ribalta, A., Batet, M., and Serratosa, F. (2012). Enabling semantic similarity estimation across multiple ontologies: An evaluation in the biomedical domain. *Journal of Biomedical Informatics*, 45(1):141 – 155. Citation on page 39.

[Schulz et al., 2007] Schulz, S., Suntisrivaraporn, B., Baader, F., et al. (2007). SNOMED CT's problem list: ontologists' and logicians' therapy suggestions. In *Medinfo 2007: Proceedings of the 12th World Congress on Health (Medical) Informatics; Building Sustainable Health Systems*, page 802. IOS Press. Citation on page 4.

[Seco et al., 2004] Seco, N., Veale, T., and Hayes, J. (2004). An intrinsic information content metric for semantic similarity in wordnet. In *Proceedings of the 16th European Conference on Artificial Intelligence*, ECAI'04, pages 1089–1090, Amsterdam, The Netherlands, The Netherlands. IOS Press. Citation on page 44.

[Semertzidis and Pitoura, 2016] Semertzidis, K. and Pitoura, E. (2016). Durable graph pattern queries on historical graphs. In *2016 IEEE 32nd International Conference on Data Engineering (ICDE)*, pages 541–552. Citations on pages 62 and 63.

[Semertzidis and Pitoura, 2018] Semertzidis, K. and Pitoura, E. (2018). Top-k durable graph pattern queries on temporal graphs. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–1. Citation on page 63.

[Soualmia et al., 2012] Soualmia, L. F., Prieur-Gaston, E., Moalla, Z., Lecroq, T., and Darmoni, S. J. (2012). Matching health information seekers' queries to medical terms. *BMC Bioinformatics*, 13(14):S11. Citation on page 38.

[Stojanovic, 2004] Stojanovic, L. (2004). *Methods and tools for ontology evolution.* PhD thesis, Karlsruhe Institute of Technology, Germany. Citation on page 80.

[Stojanovic et al., 2002] Stojanovic, L., Maedche, A., Motik, B., and Stojanovic, N. (2002). User-driven ontology evolution management. *Knowledge engineering and knowledge management: ontologies and the semantic web*, pages 133–140. Citation on page 79.

[Sy et al., 2012] Sy, M.-F., Ranwez, S., Montmain, J., Regnault, A., Crampes, M., and Ranwez, V. (2012). User centered and ontology based information retrieval system for life sciences. *BMC Bioinformatics*, 13(1):S4. Citation on page 38.

[Tissaoui et al., 2011] Tissaoui, A., Aussenac-Gilles, N., Hernandez, N., and Laublet, P. (2011). EVONTO - Joint evolution of ontologies and semantic annotations. (short paper). In Dietz, J., editor, *International Conference on Knowledge Engineering and Ontology Development (KEOD), Paris, 26/10/2011-29/10/2011*, pages 226–231. Citation on page 21.

[Traverso-Ribon et al., 2015] Traverso-Ribon, I., Vidal, M.-E., and Palma, G. (2015). Annevol: An evolutionary framework to description ontology-based annotations. In Ashish, N. and Ambite, J.-L., editors, *Data Integration in the Life Sciences*, volume 9162 of *Lecture Notes in Computer Science*, pages 87–103. Springer International Publishing. Citation on page 10.

[Tsatsaronis et al., 2013] Tsatsaronis, G., Varlamis, I., Kanhabua, N., and Nørvåg, K. (2013). Temporal classifiers for predicting the expansion of medical subject headings. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 98–113. Springer. Citations on pages 78, 79, 80, 84, 86, 90, and 91.

[Tversky, 1977] Tversky, A. (1977). Features of similarity. *Psychological Review*, 84(4):327–352. Citation on page 38.

[Van Aggelen et al., 2014] Van Aggelen, A., Hollink, L., Kemman, M., Kleppe, M., and Beunders, H. (2014). The debates of the european parliament as linked open data. *Semantic Web*, 8. Citation on page 97.

[Wang et al., 2011] Wang, S., Schlobach, S., and Klein, M. (2011). Concept drift and how to identify it. *Web Semantics: Science, Services and Agents on the World Wide Web*, 9(3):247–265. Citation on page 78.

[Whetzel et al., 2011] Whetzel, P. L., Noy, N. F., Shah, N. H., Alexander, P. R., Nyulas, C., Tudorache, T., and Musen, M. A. (2011). Bioportal: enhanced functionality via new web services from the national center for biomedical ontology to access and use ontologies in software applications. *Nucleic acids research*, 39(suppl 2):W541–W545. Citation on page 12.

[Zavalina et al., 2015] Zavalina, O. L., Kizhakkethil, P., Alemneh, D. G., Phillips, M. E., and Tarver, H. (2015). Building a framework of metadata change to support knowledge management. *Journal of Information &amp; Knowledge Management*, 14(01):1550005. Citation on page 22.

[Zhang et al., 2016] Zhang, D., Yang, Z., Wang, N., Wang, B., and Zhao, H. (2016). Ontology extension based on axiomatic cognitive model for ontology learning. In *2016 2nd IEEE International Conference on Computer and Communications (ICCC)*, pages 825–829. Citation on page 98.

universite PARIS-SACLAY

ÉCOLE DOCTORALE
Sciences et technologies
de l'information
et de la communication (STIC)

Synthèse : Les annotations sémantiques sont utilisées dans de nombreux domaines comme celui de la santé et servent à différentes tâches, notamment la recherche et le partage d'information, ou encore l'aide à la décision. Les annotations sont produites en associant à des documents digitaux des labels de concepts provenant des systèmes d'organisation de la connaissance (Knowledge Organization Systems, ou KOS, en anglais) comme les ontologies. Elles permettent alors aux ordinateurs d'interpréter, connecter et utiliser de manière automatique de grandes quantités de données. Cependant, la nature dynamique de la connaissance engendre régulièrement de profondes modifications au niveau du contenu des KOS provoquant ainsi un décalage entre la définition des concepts et les annotations.

Considérez l'exemple de Figure 1. Une partie d'un document de PUBMED[1] est annoté avec le terme *Migraine menstruelle*, un attribut du concept 625.4 du CIM-9 (ICD-9-CM en anglais), version 2008AA[2]. Dans la version 2009AA, cet attribut a été supprimé et est devenu le titre d'un nouveau concept, avec l'ID 346.4. Nous considérons donc que cette annotation a été impactée, car la modification du KOS a provoqué une non-concordance entre l'annotation créée avec la version 2008AA et le concept modifié du KOS dans sa version 2009AA. De plus, la relation *exclusion* dans les directives de l'ICD-9-CM indique que ces attributs n'appartiennent plus au même concept. La conséquence de ces modifications est que l'annotation a perdu sa signification et nécessite une correction.

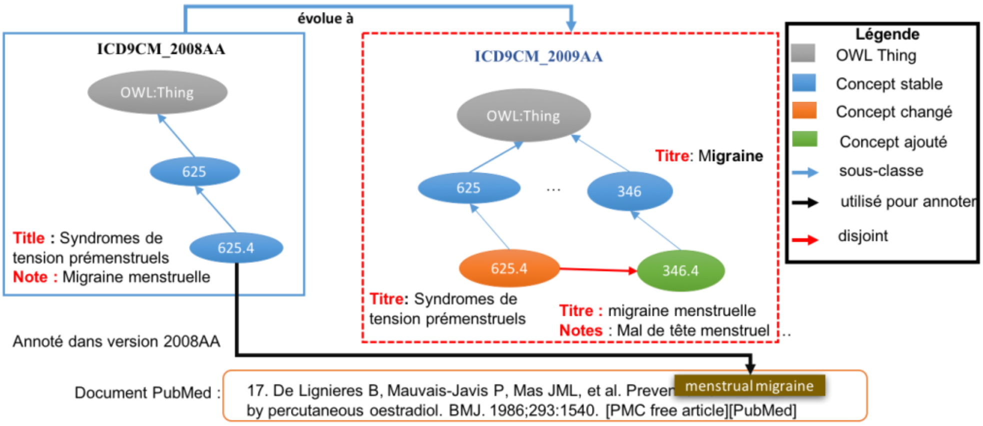

évolue à
ICD9CM_2008AA
OWL:Thing
625
625.4
Title : Syndromes de tension prémenstruels
Note : Migraine menstruelle
Annoté dans version 2008AA
ICD9CM_2009AA
OWL:Thing
625
…
346
625.4
346.4
Titre: Migraine
Titre: Syndromes de tension prémenstruels
Titre : migraine menstruelle
Notes : Mal de tête menstruel …
Légende
OWL Thing
Concept stable
Concept changé
Concept ajouté
sous-classe
utilisé pour annoter
disjoint
Document PubMed :
17. De Lignieres B, Mauvais-Javis P, Mas JML, et al. Preven
menstrual migraine
by percutaneous oestradiol. BMJ. 1986;293:1540. [PMC free article][PubMed]

Figure - Cas d'étude d'annotation

L'impact du changement de KOS se propage aussi à d'autres systèmes tels que les moteurs de recherche et les portails de données qui perdent en précision lors de la récupération d'informations à partir de documents annotés. Un médecin qui souhaite, par exemple, accéder au dossier d'un patient par mots clés, via un moteur de recherche, ne pourra pas récupérer des informations précises et complètes si la requête contient *migraine menstruelle*. Ce cas illustre

[1] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1342315/
[2] https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/release/notes.html

l'impact direct des changements de KOS sur les annotations sémantiques et souligne le besoin réel de méthodes et d'outils avancés capables de mettre à jour les annotations sémantiques tout en évitant une intervention humaine laborieuse et coûteuse.

Nous nous intéressons aussi au cas où les annotations ne peuvent pas être modifiées car, pour des raisons légales ou de confidentialité, seules les métadonnées peuvent être lues. Nous proposons une méthode alternative permettant l'accessibilité aux données sans les modifier. Nous avons créé un graphe de connaissances (GC) qui garde l'historique des évolutions de KOS. Ce graphe permet de naviguer dans des relations complexes liées aux aspects structurels et évolutifs de KOS. Par exemple, il permet de décrire qu'un concept a été *ajouté*, *déplacé*, *divisé*, … Ces informations supplémentaires constituent la base de notre méthode alternative de recherche d'information, permettant la recherche des documents correspondant aux métadonnées des concepts du KOS de différentes versions.

Dans ce contexte incluant ces deux types de scénarios, cette thèse propose et implémente des méthodes et outils pour promouvoir la maintenance semi-automatique des annotations sémantiques affectées par l'évolution de KOS, afin de garder ces annotations exploitables dans le temps. Nous visons à répondre à la question de recherche suivante :

Comment maintenir automatiquement la validité des annotations biomédicales en présence de modifications dans le KOS utilisé pour annoter ?

Le système proposé est suffisamment générique pour traiter des terminologies ayant des structures distinctes. Les problèmes de performance sont pris en compte lors de l'application de notre méthode sur des milliers de concepts, par exemple en provenance du MeSH (237000 concepts) ou SNOMED CT (310000 concepts).

Nous avons formulé l'hypothèse suivante pour guider nos études : Les informations du KOS, ainsi que des informations sur l'évolution du KOS, peuvent être utilisées pour définir un mécanisme robuste de maintenance.

Cela nous a amenés à traiter plusieurs sous-problèmes :

- **RQ1:** Quel est l'impact des changements du KOS sur les annotations sémantiques ?
- **RQ2:** Quel est le modèle le plus approprié pour résoudre le problème d'évolution des annotations ?
- **RQ3:** Suite à la publication d'une nouvelle version du KOS, comment maintenir automatiquement la validité des annotations sémantiques sans annoter à nouveau le contenu de tous les documents ?
- **RQ4:** Quelles méthodes peuvent être utilisées pour permettre la recherche des annotations lorsque le document et les annotations ne peuvent pas être modifiés ?
- **RQ5:** Pouvons-nous prédire quel concept du KOS changera dans un futur proche et quel type de changement aura un impact sur ce concept ?

Pour chaque sous-problème, rédigé sous forme de question de recherche (RQ), nous avons apporté des contributions significatives à l'état de l'art. Ces contributions sont brièvement décrites ci-dessous.

Les travaux effectués pour répondre la **RQ1**(Quel est l'impact des modifications de KOS sur les annotations sémantiques ?) ont montré, par une analyse empirique des facteurs influençant l'évolution des annotations (décrite dans le chapitre 1), que les modifications apportées aux annotations sont fortement corrélées aux modifications du KOS. Pour réaliser cette analyse, nous avons utilisé un ensemble de documents annotés avec GATE et NCBO Annotator et 13 versions différentes de deux KOS biomédicaux bien connu, ICD-9-CM et MeSH.

La question **RQ2** (Quel est le modèle le plus approprié pour résoudre le problème d'évolution des annotations ?) a été traitée dans le chapitre 2. Nous avons analysé différents modèles d'annotation afin de vérifier s'ils étaient appropriés pour représenter (ou si nous pouvions déduire de leurs éléments) tous les critères requis pour identifier les changements apportés aux annotations. Nos études ont permis de proposer des extensions au modèle de données *W3C Open Annotation* afin de résoudre ce problème, par exemple l'inclusion de la relation *evolved*, qui lie une annotation évoluée à sa version antérieure. Ces nouvelles fonctionnalités permettent une meilleure couverture des aspects liés à l'évolution. Par exemple, il est maintenant possible de retracer toutes les évolutions d'une annotation et de les réutiliser pour la recherche d'information (voir le chapitre 3).

Les travaux nécessaires pour répondre à la question **RQ3** (Suite à la publication d'une nouvelle version du KOS, comment maintenir automatiquement la validité des annotations sémantiques sans annoter à nouveau le contenu de tous les documents ?) nous a amenés à proposer une nouvelle architecture, appelée MAISA, à base de règles, qui combine des informations provenant de l'évolution des KOS et des connaissances extraites du Web, pour tenir à jour automatiquement les annotations sémantiques. Nous avons démontré par des expériences empiriques que la correction / adaptation des annotations peut atteindre un taux de fiabilité satisfaisant. Cependant, il est important de souligner que le rôle des experts du domaine est toujours déterminant pour garantir la qualité des annotations dans des scénarios critiques, comme cela a été observé dans le domaine biomédical. Notre approche de maintenance se fait sans ré-annotation complète du document, puisque nous réutilisons les informations présentes dans les annotations pour faire évoluer celles qui ont été impactées.

Afin de répondre à la question **RQ4** (Quelles méthodes peuvent être utilisées pour permettre la recherche des annotations lorsque le document et les annotations ne peuvent pas être modifiés ?), nous avons proposé, dans le chapitre 5, une méthode permettant de rechercher les annotations impactées sans les modifier. Les analyses expérimentales ont démontré que notre méthode est capable d'atteindre de bons résultats lorsqu'on interroge les GC et qu'elle peut être utilisée comme une méthode alternative à la méthode présentée dans le chapitre 3, pour adapter/modifier les annotations sémantiques. Nous avons observé que l'utilisation d'un graphe de connaissances permet d'avoir une bonne représentation des différentes versions de l'ontologie et du lien évolutif entre elles. Cette approche contribue à l'état actuel de la maintenance des annotations en intégrant une nouvelle méthode pour les maintenir.

Le problème de l'anticipation pointé par la question **RQ5** (Pouvons-nous prédire quel concept du KOS changera dans un futur proche et quel type de changement aura un impact sur ce concept ?) a été traité dans le chapitre 6. Nous avons analysé comment améliorer notre approche en incluant une nouvelle méthode capable d'anticiper l'évolution des concepts associés aux annotations. Cette méthode peut être utile aux experts du domaine pendant les phases d'annotation et de maintenance. Elle peut, par exemple, alerter sur les risques liés au choix du concept et / ou d'une ontologie pour annoter les documents biomédicaux. Notre approche a montré plus de 71% de précision lors de l'identification des concepts nécessitant une révision. En plus, nous avons contribué à l'état de la technique en incluant des fonctionnalités temporelles, par exemple, des informations sur la dernière date d'évolution d'une annotation, ce qui contribue à l'amélioration de la précision des méthodes utilisées.

En résumé, toutes les questions de recherche identifiées dans cette thèse (RQ1-5) ont reçu une réponse satisfaisante, ce qui permet de valider notre hypothèse : *Les informations du KOS, ainsi que des informations sur l'évolution de KOS, peuvent être utilisées pour définir un mécanisme robuste de maintenance*. Nous avons également dû traiter un problème secondaire, mais tout à fait intéressant, lié aux choix de la mesure de similarité la plus appropriée dans MAISA. Ce problème a été présenté dans le chapitre 4 et montre qu'une approche combinant des mesures lexicales et sémantiques peut améliorer la mesure de similarité entre les concepts. Nos analyses expérimentales ont démontré qu'une méthode exploitant ces mesures peut surpasser les méthodes basées sur une seule mesure de similarité (sémantique ou lexicale). En utilisant cette mesure hybride, nous avons introduit la règle *Correspondance_Partielle* pour maintenir les annotations sémantiques affectées par l'évolution du KOS. Nos analyses expérimentales ont démontré que la règle *Correspondance_Partielle* permet d'obtenir de bons résultats pour adapter les annotations en utilisant une ou plusieurs versions successives des KOSs.

En résumé, cette thèse a contribué à la maintenance des annotations sémantiques affectées par l'évolution de KOS. Nous nous sommes basés sur l'état de l'art dans le domaine pour construire l'architecture MAISA. Cette architecture contient une méthode originale pour gérer les annotations. Les résultats obtenus au cours de la thèse sont encourageants et ont permis d'apporter des améliorations pour le domaine. Tous les travaux décrits dans cette thèse ont fait l'objet de plusieurs articles publiés dans des conférences et des revues internationales.

université PARIS-SACLAY

ÉCOLE DOCTORALE
Sciences et technologies
de l'information
et de la communication (STIC)

Titre : MAISA - La maintenance des annotations sémantiques

Mots clés : Annotations sémantiques, maintenance des annotations, évolution des annotations, évolution d'ontologie, ontologies de la santé

Résumé : Les annotations sémantiques sont utilisées dans de nombreux domaines comme celui de la santé et servent à différentes tâches notamment la recherche et le partage d'information ou encore l'aide à la décision. Les annotations sont produites en associant à des documents digitaux des labels de concepts provenant des systèmes d'organisation de la connaissance (Knowledge Organization Systems, ou KOS, en anglais) comme les ontologies. Elles permettent alors aux ordinateurs d'interpréter, connecter et d'utiliser de manière automatique de grandes quantités de données. Cependant, la nature dynamique de la connaissance engendre régulièrement de profondes modifications au niveau du contenu des KOS provoquant ainsi un décalage entre la définition des concepts et les annotations. Une adaptation des annotations à ces changements est nécessaire pour garantir une bonne utilisation par les applications informatiques. De plus, la quantité importante d'annotations affectées rend impossible une adaptation manuelle. Dans ce mémoire de thèse, nous proposons une approche originale appelée MAISA pour résoudre le problème de l'adaptation des annotations sémantiques engendrée par l'évolution des KOS et pour lequel nous distinguons deux cas. Dans le premier cas, nous considérons que les annotations sont directement modifiables. Pour traiter ce problème nous avons défini une approche à base de règles combinant des informations provenant de l'évolution des KOS et des connaissances extraites du Web. Dans le deuxième cas, nous considérons que les annotations ne sont pas modifiables comme c'est bien souvent le cas des annotations associées aux données des patients. L'objectif ici étant de pouvoir retrouver les documents annotés avec une version du KOS donnée lorsque l'utilisateur interroge le système stockant ces documents avec le vocabulaire du même KOS mais d'une version différente. Pour gérer ce décalage de versions, nous avons proposé un graphe de connaissance représentant un KOS et son historique et un mécanisme d'enrichissement de requêtes permettant d'extraire de ce graphe l'historique d'un concept pour l'ajouter à la requête initiale. Nous proposons une évaluation expérimentale de notre approche pour la maintenance des annotations à partir de cas réels construits sur quatre KOS du domaine de la santé : ICD-9-CM, MeSH, NCIt et SNOMED CT. Nous montrons à travers l'utilisation des métriques classiques que l'approche proposée permet, dans les deux cas considérés, d'améliorer la maintenance des annotations sémantiques.

Title: MAISA - Maintenance of Semantic Annotations

Keywords: Semantic annotation, annotation maintenance, evolution of annotations, ontology evolution, biomedical controlled terminologies

Abstract: Semantic annotations are often used in a wide range of applications ranging from information retrieval to decision support. Annotations are produced through the association of concept labels from Knowledge Organization System (KOS), i.e. ontology, thesaurus, dictionaries, with pieces of digital information, e.g. images or texts. Annotations enable machines to interpret, link, and use a vast amount of data. However, the dynamic nature of KOS may affect annotations each time a new version of a KOS is released. New concepts can be added, obsolete ones removed and the definition of existing concepts may be refined through the modification of their labels/properties. As a result, many annotations can lose their relevance, thus hindering the intended use and exploitation of annotated data. To solve this problem, methods to maintain the annotations up-to-date are required. In this thesis we propose a framework called MAISA to tackle the problem of adapting outdated annotations when the KOS utilized to create them change. We distinguish two different cases. In the first one we consider that annotations are directly modifiable. In this case, we proposed a rule-based approach implementing information derived from the evolution of KOS as well as external knowledge from the Web. In the second case, we consider that the annotations are not modifiable. The goal is then to keep the annotated documents searchable even if the annotation are produced with a given KOS version but the user used another version to query them. In this case, we designed a knowledge graph that represents a KOS and its successive evolution and propose a method to extract the history of a concept and add the gained label to the initial query allowing to deal with annotation evolution. We experimentally evaluated MAISA on realistic cases-studies built from four well-known biomedical KOS: ICD-9-CM, MeSH, NCIt and SNOMED CT. We show that the proposed maintenance method allow to maintain semantic annotations using standard metrics.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France