Mapping Adaptation between Biomedical Knowledge Organization Systems

Julio Cesar Dos dos Reis Reis

Université Paris-Sud

École doctorale d'informatique de Paris-Sud

Laboratoire de recherche en informatique

Discipline: Informatique

THÈSE DE DOCTORAT

Soutenue le 24 Octobre 2014

par

Julio Cesar DOS REIS

Mapping Adaptation between Biomedical Knowledge Organization Systems

Directrice de thèse :	Chantal Reynaud-Delaître	Professeur, Université de Paris-Sud, France
Co-encadrant de thèse :	Cédric Pruski	Chargé de recherche, CRP Henri Tudor, Luxembourg
Composition du jury :		
Rapporteurs :	Nathalie Aussenac-Gilles	Directrice de recherche, IRIT-CNRS, France
	Frank van Harmelen	Professeur, Université libre d'Amsterdam, Pays-Bas
Examinateurs :	Christine Froidevaux	Professeur, Université de Paris-Sud, France
	Erhard Rahm	Professeur, Université de Leipzig, Allemagne
	Stéfan J. Darmoni	Professeur, Université de Rouen, France

UNIVERSITY OF PARIS-SUD

COMPUTER SCIENCE DOCTORATE SCHOOL

COMPUTER SCIENCE LABORATORY

THESIS

defended on 24 of October 2014

for obtaining the degree of

DOCTOR OF PHILOSOPHY IN COMPUTER SCIENCE

by

JULIO CESAR DOS REIS

Mapping Adaptation between Biomedical Knowledge Organization Systems

Thesis supervisor:	Chantal Reynaud-Delaître	Professor, University of Paris-Sud, France
Thesis co-supervisor:	Cédric Pruski	Researcher, CRP Henri Tudor, Luxembourg
Jury composed by:		
Scientific reporters:	Nathalie Aussenac-Gilles	Research Director, IRIT-CNRS, France
	Frank van Harmelen	Professor, VU University Amsterdam, Netherlands
Examiners:	Christine Froidevaux	Professor, University of Paris-Sud, France
	Erhard Rahm	Professor, University of Leipzig, Germany
	Stéfan J. Darmoni	Professor, University of Rouen, France

Acknowledgments

I would like to express my heartily gratitude to several people for their immense support during the course of my PhD. First and foremost, I am especially grateful to my wife Laira, the love of my life, for her incommensurable patience at every single moment of this endeavor, and who waited so long for this moment. My deepest appreciation goes to my family for the unconditional encouragement, my parents Sebastião (dad, I wish you were here) and Zeni, as well as my lovely parents-in-law Cesar and Toninha.

I would like to thank my advisor, Chantal Reynaud-Delaître. Her guidance and comments enabled me to shape and complete this work. Our meetings in Orsay were very fruitful. I am also thankful to Cédric Pruski, for co-supervising the development of this thesis. I appreciate their confidence in my research work.

I am extremely grateful to my coworkers and colleagues at the CRP Henri Tudor in Luxembourg: Duy, Marcos, Asma, Veruska, Claude, François, Monika, Kaddour, Hamid and others. I must also thank Anika Gross and Michael Hartung from University of Leipzig for our successful collaboration. I would like to thank all friends in Luxembourg, especially Martina, Simone and Lisa for their help, friendship and the very great moments shared. I am also thankful to all colleagues in Brazil who I have kept in touch all these years.

Furthermore, I am deeply indebted to people who somehow taught me anything and let their contribution in this work. This includes the several anonymous reviewers who refereed this research in its publications, and the members of the jury who dedicated their efforts on reviewing this thesis.

Luxembourg, 04^{th} November 2014 Julio Cesar DOS REIS

To my sweet wife Laira,
thank you for being the heart of my mind.

Abstract

Modern biomedical information systems require exchanging and retrieving data between them, due to the overwhelming available data generated in this domain. Knowledge Organization Systems (KOSs) offer means to make the semantics of data explicit which, in turn, facilitates their exploitation and management. The evolution of semantic technologies has led to the development and publication of an ever increasing number of large KOSs for specific sub-domains like genomics, biology, anatomy, diseases, *etc.* The size of the biomedical field demands the combined use of several KOSs, but it is only possible through the definition of mappings. Mappings interconnect entities of domain-related KOSs via semantic relations. They play a key role as references to enable advanced interoperability tasks between systems, allowing software applications to interpret data annotated with different KOSs. However, to remain useful and reflect the most up-to-date knowledge of the domain, the KOSs evolve and new versions are periodically released. This potentially impacts established mappings demanding methods to ensure, as automatic as possible, their semantic consistency over time. Manual maintenance of mappings stands for an alternative only if a restricted number of mappings are available. Otherwise supporting methods are required for very large and highly dynamic KOSs.

To address such problem, this PhD thesis proposes an original approach to adapt mappings based on KOS changes detected in KOS evolution. The proposal consists in interpreting the established correspondences to identify the relevant KOS entities, on which the definition relies on, and based on the evolution of these entities to propose actions suited to modify mappings. Through this investigation, (i) we conduct in-depth experiments to understand the evolution of KOS mappings; we propose automatic methods (ii) to analyze mappings affected by KOS evolution, and (iii) to recognize the evolution of involved concepts in mappings via change patterns; finally (iv) we design techniques relying on heuristics explored by novel algorithms to adapt mappings. This research achieved a complete framework for mapping adaptation, named *DyKOSMap*, and an implementation of a software prototype. We thoroughly evaluated the proposed methods and the framework with real-world datasets containing several releases of mappings between biomedical KOSs. The obtained results from experimental validations demonstrated the overall effectiveness of the underlying principles in the proposed approach to adapt mappings. The scientific contributions of this thesis enable to largely automatically maintain mappings with a reasonable quality, which improves the support for mapping maintenance and consequently ensures a better interoperability over time.

Keywords: KOS; Biomedical KOS; Biomedical Ontologies; Ontology Evolution; KOS Evolution; Ontology Changes; Change Patterns; Ontology Mappings; KOS Mappings; Mapping Adaptation; Mapping Maintenance; Mapping Evolution.

Résumé

Les systèmes d'information biomédicaux actuels reposent sur l'exploitation de données provenant de sources multiples. Les Systèmes d'Organisation de la Connaissance (*Knowledge Organization Systems*, ou KOS, en anglais) permettent d'expliciter la sémantique de ces données, ce qui facilite leur gestion et leur exploitation. Bénéficiant de l'évolution des technologies du Web sémantique, un nombre toujours croissant de KOSs a été élaboré et publié dans des domaines spécifiques tels que la génomique, la biologie, l'anatomie, les pathologies, *etc.* Leur utilisation combinée, nécessaire pour couvrir tout le domaine biomédical, repose sur la définition de mises en correspondance entre leurs éléments ou mappings. Les mappings connectent les entités des KOSs liées au même domaine via des relations sémantiques. Ils jouent un rôle majeur pour l'interopérabilité entre systèmes, en permettant aux applications d'interpréter les données annotées avec différents KOSs. Cependant, les KOSs évoluent et de nouvelles versions sont régulièrement publiées de façon à correspondre à des vues du domaine les plus à jour possible. Les validité des mappings ayant été préalablement établis peuvent alors être remis en cause. Des méthodes sont nécessaires pour assurer leur cohérence sémantique au fil du temps. La maintenance manuelle des mappings est une possibilité lorsque le nombre de mappings est restreint. En présence de KOSs volumineux et évoluant très rapidement, des méthodes les plus automatiques possibles sont indispensables.

Cette thèse de doctorat propose une approche originale pour adapter les mappings basés sur les changements détectés dans l'évolution de KOSs du domaine biomédical. Notre proposition consiste à comprendre précisément les mappings entre KOSs, à exploiter les types de changements intervenant lorsque les KOSs évoluent, puis à proposer des actions de modification des mappings appropriées. Nos contributions sont multiples : (i) nous avons réalisé un travail expérimental approfondi pour comprendre l'évolution des mappings entre KOSs; nous proposons des méthodes automatiques (ii) pour analyser les mappings affectés par l'évolution de KOSs, et (iii) pour reconnaître l'évolution des concepts impliqués dans les mappings via des patrons de changement; enfin (iv) nous proposons des techniques d'adaptation des mappings à base d'heuristiques. Nous proposons un cadre complet pour l'adaptation des mappings, appelé *DyKOSMap*, et un prototype logiciel. Nous avons évalué les méthodes proposées et le cadre formel avec des jeux de données réelles contenant plusieurs versions de mappings entre KOSs du domaine biomédical. Les résultats des expérimentations ont démontré l'efficacité des principes sous-jacents à l'approche proposée. La maintenance des mappings, en grande partie automatique, est de bonne qualité.

Mots-clés : Système d'Organisation de la Connaissance, Domaine biomédical, Ontologies biomédicales, Evolution d'ontologies; Evolution de KOSs dans le domaine biomédical, Changements dans les ontologies, Mappings, Maintenance des mappings, Adaptation des mappings, Evolution des mappings.

Contents

List of Figures **5**

List of Tables **7**

1 Introduction **9**

1.1 Problem and research question 11

1.2 Objective and contributions 12

1.3 Thesis outline 14

Part I Analysis of bibliography and empirical data **15**

2 State-of-the-art **19**

Introduction 19

2.1 Definitions 20

2.1.1 KOS model 20

2.1.2 KOS mappings 21

2.2 Mapping maintenance 22

2.2.1 The mapping maintenance problem 22

2.2.2 Mapping revision 24

2.2.3 Mapping calculation 27

2.2.4 Mapping adaptation 29

2.3 Mapping interpretation 32

2.4 KOS evolution characterization 34

2.4.1 Types of KOS change operations 34

2.4.2 KOS change patterns 36

Summary 39

3 Understanding mappings evolution **41**

Introduction 41

3.1 Experimental scenario . . . 43
3.1.1 Materials . . . 43
3.1.2 Mapping evolution characterization . . . 45
3.1.3 Experiments organization . . . 46
3.2 Effects of KOS evolution on mapping changes . . . 46
3.2.1 Experimental procedure . . . 46
3.2.2 General analyses . . . 47
3.2.3 Specific analyses . . . 50
3.3 Interdependencies between mapping change operations . . . 56
3.3.1 Experimental procedure . . . 56
3.3.2 Results of the quantitative analysis . . . 58
3.3.3 Results of the qualitative analysis . . . 60
3.4 Mapping evolution under complex KOS changes . . . 62
3.4.1 Experimental procedure . . . 62
3.4.2 ICD-9-CM cases . . . 65
3.4.3 SNOMED-CT cases . . . 68
3.5 Discussion . . . 71
Conclusion . . . 73

Part II The *DyKOSMap* approach 75

4 Interpreting correspondences for mapping adaptation 83
Introduction . . . 83
4.1 Problem statement and definitions . . . 84
4.2 Identifying concept attributes relevant to interpret mappings . . . 86
4.3 The similarity measures investigated . . . 88
4.3.1 Character-based edit-distance similarity . . . 89
4.3.2 Word-based edit-distance similarity . . . 90
4.3.3 Knowledge-based similarity . . . 91
4.4 The potential role of relevant attributes identified . . . 92
4.5 Experimental evaluation . . . 92
4.5.1 Materials and procedure . . . 93
4.5.2 Experimental results . . . 96
4.6 Discussion . . . 99
Conclusion . . . 101

5 KOS change patterns to inform mapping adaptation 103

Introduction . . . 103

5.1 Problem statement and definitions . . . 105

5.2 Change patterns at the level of concept attributes . . . 107

5.2.1 Lexical change patterns . . . 107

5.2.2 Semantic change patterns . . . 109

5.3 Selection of candidate attributes in the context . . . 111

5.4 Lexical change pattern recognition . . . 113

5.5 Semantic change pattern recognition . . . 114

5.6 Experimental evaluation . . . 116

5.6.1 Materials and procedure . . . 116

5.6.2 Experimental results . . . 118

5.7 Discussion . . . 120

Conclusion . . . 123

6 Mapping adaptation 125

Introduction . . . 125

6.1 Problem statement . . . 127

6.2 Mapping adaptation actions . . . 128

6.3 Study of factors influencing mapping adaptation actions . . . 131

6.3.1 Materials . . . 131

6.3.2 Common experimental procedure . . . 133

6.3.3 KOS changes related to revision and removal of concepts . . . 134

6.3.4 KOS changes affecting relevant concept attributes . . . 138

6.3.5 Impact of lexical and semantic change patterns . . . 139

6.3.6 Summary of the findings . . . 143

6.4 Heuristics guiding decisions of mapping adaptation . . . 144

6.4.1 Move and derivation of mappings . . . 145

6.4.2 Modification of semantic relation . . . 147

6.4.3 Removal and no action adaptation . . . 150

6.5 Experimental evaluation . . . 152

6.5.1 Experimental procedure . . . 152

6.5.2 Experimental results . . . 154

6.6 Discussion . . . 159

Conclusion . . . 161

7 The *DyKOSMap* framework **163**
Introduction . . . 163
7.1 The framework description . . . 165
7.2 Mapping adaptation method . . . 166
7.2.1 Adaptation of mappings affected by KOS changes . . . 167
7.2.2 Selection and application of mapping adaptation actions . . . 171
7.2.3 Conflicts and status of mappings under adaptation . . . 172
7.3 The *DyKOSMap* prototype software . . . 174
7.3.1 Architecture . . . 174
7.3.2 Implementation . . . 175
7.4 Experimental evaluation . . . 176
7.4.1 Materials and procedure . . . 176
7.4.2 Experimental results . . . 178
7.5 Discussion . . . 179
Conclusion . . . 181

8 Conclusions and perspectives **183**
8.1 Summary of the contributions . . . 184
8.2 Directions for future work . . . 185
8.3 Final remarks . . . 186

A Publications **187**

Bibliography **191**

List of Figures

1.1 Chapters composing the first part of the thesis 17

2.1 The mapping maintenance problem 23
2.2 Mapping revision category 24
2.3 Mapping calculation category 27
2.4 Mapping adaptation category 29

3.1 Experiments configuration 44
3.2 Scenario involving addition and modification of the *semType* in mapping 59
3.3 Scenario involving removal and addition of mapping 60
3.4 First case of split complex change in ICD9 66
3.5 Second case of split complex change in ICD9 67
3.6 Third case of split complex change in ICD9 68
3.7 Fourth case of split complex change in ICD9 69
3.8 First case of split complex change in SCT 70
3.9 Second case of split complex change in SCT 70
3.10 The *DyKOSMap* approach outlook 78
3.11 Chapters composing the second part of the thesis 79

4.1 Relevant attributes between source and target concepts 87
4.2 Performance of the *topA* method analysing the evolution of SCT 97
4.3 Performance of the *topA* method analysing the evolution of ICD9 98
4.4 Change correlation in the analysis of SCT 99
4.5 Change correlation in the analysis of ICD9 100

5.1 Scenario of the problem for change patterns in concept attributes value 106
5.2 Example of change pattern recognition 115
5.3 Effectiveness of the Lexical Change Pattern identification algorithm 119
5.4 Effectiveness of the Semantic Change Pattern identification algorithm 121

6.1 Scenario of the problem for change patterns-based mapping adaptation 128
6.2 Mapping adaptation actions 130
6.3 Analysis of mapping adaptation actions under revision KOS changes 136
6.4 Behaviour of mappings under revision and removal KOS changes 136
6.5 Analysis of mapping adaptation actions under removal KOS changes 137
6.6 Heuristics for *MoveM* 146
6.7 Heuristics for *DeriveM* 147
6.8 Heuristics for *ModSemTypeM* of an existing established mapping 149

6.9 Heuristics for *ModSemTypeM* between a candidate and the target concept 150
6.10 Heuristics for *RemoveM* 151
6.11 Heuristics for *NoAction* 152

7.1 The *DyKOSMap* framework 165
7.2 Conflicts between KOS changes in source and target concepts 167
7.3 The mapping adaptation method 168
7.4 Adaptation of KOS mappings affected by KOS evolution 170
7.5 *DyKOSMap* component diagram 175
7.6 *DyKOSMap* class diagram 176
7.7 Results of the mapping adaptation quality for SCT-ICD9 and SCT-NCI 179
7.8 Results of the mapping adaptation quality for MeSH-ICD10 and General analysis 180

List of Tables

2.1 Atomic KOS change operations . . . 35
2.2 Complex KOS change operations . . . 36

3.1 Studied biomedical KOSs . . . 44
3.2 Analyzed mappings between biomedical KOS . . . 48
3.3 Proportion of mapping change operations . . . 48
3.4 Relations in mappings with regards to mapping change operations . . . 49
3.5 Changes in SCT entities correlated with mapping change operations . . . 51
3.6 Types of changes in concepts with regard to mapping change operations . . . 52
3.7 Concepts' status changes related to mapping change operations . . . 53
3.8 Concepts' descriptions changes related with mapping change operations . . . 54
3.9 Concepts' relationships changes correlated with mapping change operations . . . 56
3.10 Analyzing the combination of equal sets of mapping change operations . . . 58
3.11 Analyzing the combination of different sets of mapping change operations . . . 59

4.1 Notations for the formalization of the methods . . . 86
4.2 Correlation between relevant attributes and associated mappings . . . 93
4.3 Statistics on the studied datasets . . . 94

5.1 Notations for the formalization of change patterns . . . 106
5.2 Lexical change patterns . . . 108
5.3 Semantic change patterns . . . 110
5.4 Evolution of the studied biomedical KOSs . . . 116

6.1 Statistics on the NCI Thesaurus and ICD10CM . . . 132
6.2 Mappings between biomedical KOSs . . . 132
6.3 KOS change operations identified for the releases of biomedical KOSs . . . 133
6.4 Computed mapping adaptation actions observing mapping evolution . . . 134
6.5 Quantity of analyzed mappings for the types of complex KOS changes . . . 135
6.6 Correlation between KOS changes affecting relevant attributes and MAAs . . . 139
6.7 Correlation between lexical change patterns and mapping adaptation actions . . . 142
6.8 Correlation between semantic change patterns and mapping adaptation actions . . . 143
6.9 Results of the heuristics evaluation for mapping adaptation . . . 154

7.1 Overview on the biomedical KOS entities . . . 177
7.2 Evaluated KOS mapping datasets . . . 177

Chapter 1

Introduction

Contents

1.1 Problem and research question 11

1.2 Objective and contributions 12

1.3 Thesis outline 14

Biomedicine has become a data and knowledge-intensive field. Biomedical research has faced several complex issues that directly impact our lives. For understanding the causes of illness based on molecular and genetic knowledge, biomedical scientists have studied physiological processes, molecular interactions and their effects on diseases [Cases et al., 2013]. The new perspective on biomedical research aims to exchange data and information from the genetic level to the clinical level, to achieve personalized treatments and advanced models of diseases [Burgun and Bodenreider, 2008]. Advancements in the field require relying upon vast arrays of voluminous, dynamic, heterogeneous and complex datasets, resulting in difficulties to use and reuse available data, to interact between different sources, and to understand data as a whole turning it into useful knowledge. This generates an ever greater demand for adequate computer-supported methods for automatically locating, accessing, sharing, analyzing and meaningfully integrating data and information.

Computer scientists have devised approaches allowing software to interpret meanings of data to address open biomedical challenges. Biomedical information systems have intensively relied on semantic technologies that make the semantic of information explicit at different degrees of expressivity [Bodenreider and Stevens, 2006]. Knowledge Organization Systems (KOSs) provide means to explicitly represent the semantics of biomedical data [Hodge, 2000]. KOSs encompass all types of conceptual models for organizing knowledge such as terminologies, classification, taxonomies, thesauri and ontologies that are widely employed to describe biomedical knowledge. They comprise properties that permit software to interpret semantics. Biomedical organisms have extensively built and used several biomedical KOSs like the *Systematized Nomenclature of Medicine-Clinical Terms*[1] (SNOMED CT ®) (SCT for short), *International Classification of Disease - Clinical Modification*[2] (ICD-9-CM) (ICD9 for short), *Gene Ontology*[3] (GO), *NCI The-*

[1] `www.ihtsdo.org/snomed-ct`

[2] `www.cdc.gov/nchs/icd/icd9cm.htm`

[3] `www.geneontology.org`

saurus[4] (NCIt), *Medical Subject Headings*[5] (MeSH), *etc.* Biomedical organisms have also created online repositories, such as the *BioPortal*[6] [Noy et al., 2009] and *CiSMef*[7] [Sakji et al., 2009] that collect and group existing KOSs to make them easily available.

KOSs can bring several benefits in supporting the accomplishment of advanced improvements in biomedicine [Bodenreider and Stevens, 2006] by aiding scientists in fully automatic analyses and correlations to discover implicit knowledge from vast, distributed and complex biomedical datasets. KOSs boost the semantic capabilities of biomedical information systems and have become cornerstones for adequately enabling various data and information-related tasks. In general, KOSs allow accurate management, interpretation, visualization, exploitation, exploration and annotation of data [Bodenreider and Stevens, 2006, Lambrix et al., 2009]. For example, eHealth applications may provide health professionals to access citizens' Electronic Health Records (EHR) by semantically describing the content of EHR to improve information retrieval; large repositories like *PubMed*[8] rely on MeSH for better indexing scientific publications. In addition, KOSs enhance information systems dealing with health management, clinical administration and medical decision support.

The great size of the biomedical domain demands multiple KOSs to obtain the best description of the domain. Moreover, biomedical systems need to rely on various KOSs to ensure an optimal scope of their applications, to support a wide range type of data (*e.g.*, clinical and administrative data, omic data, *etc.*). As biomedical KOSs cover overlapping subjects in part, it requires integrating them. To address this issue aiming to provide the combined use of KOSs by different software, mappings establish semantic relations between entities (more frequently concepts) of different, but interrelated KOSs [Euzenat and Shvaiko, 2007]. For example, the biomedical concept named "*torso*" is an equivalent concept to "*trunk*" of another biomedical KOS.

Mappings allow semantic-enabled software to interpret distributed data annotated using different KOSs, which may help overcoming the semantic heterogeneity commonly encountered in biomedical systems [Fung and Xu, 2012]. For instance, mappings between SNOMED CT and ICD9 can easily support users accessing external data annotated with ICD9 from clinical records annotated with SNOMED CT. In this case, data referring to the same entity may be represented by different concepts' identifiers. Therefore, KOSs interconnected through mappings play a key role in addressing semantic interoperability issues between biomedical information systems, because they allow a shared understanding for communicating information [Burgun and Bodenreider, 2008]. Mappings might also support tasks related to distributed semantic search services, information indexing [Köhler et al., 2006, Kitamura and Segawa, 2008], as well as the integration and exchange of data from various sources [Lambrix et al., 2009]. In biomedical research, mappings especially facilitate integration and in-depth analysis of data collected in research studies with clinical data, linking genomics, transcriptomics and proteomics knowledge with medical knowledge [Burgun and Bodenreider, 2008]. According to Cases *et al.* [Cases et al., 2013], biomedical research will gain several benefits from the integrative analysis of clinical and multi-omics information of large amounts of diverse information.

[4] `www.ncit.nci.nih.gov`
[5] `www.nlm.nih.gov/mesh/meshhome.html`
[6] `www.bioportal.bioontology.org`
[7] `www.chu-rouen.fr/cismef`
[8] `www.ncbi.nlm.nih.gov/pubmed`

In addition to its huge size, the biomedical domain is by nature highly dynamic and its knowledge constantly evolves. According to Baneyx and Charlet [Baneyx and Charlet, 2007], 50% of the knowledge of the biomedical field has been renewed in 10 years. This dynamic nature of the biomedical domain forces knowledge engineers to continuously revise the content of KOSs to assure that they remain useful over time, by modifying and updating entities' definition. This is relevant since the use of outdated KOSs may trigger undesirable results in systems exploiting them due to the incomplete or false concepts' definitions.

Literature has shown high rates of changes in large biomedical KOSs [Spackman, 2005]. As an example, available tools to study the evolution of biomedical KOSs, such as *OnEX*[9] [Hartung et al., 2009] point out that the NCI Thesaurus describing cancer-related content has doubled in size to almost 100 000 concepts between 2006 and 2013, while numerous concepts were revised or declared obsolete. These changes make the frequent releases of new KOS versions inevitable.

1.1 Problem and research question

Changes like the *removal* or *modification* of KOS's entities (*e.g.*, concepts) [Hartung et al., 2013] can potentially impact existing mappings established between interrelated entities, making mappings semantically invalid [Dos Reis et al., 2012]. This problem affects underlying software relying on mappings, which prevents them to fully exploit mappings with accuracy in the mentioned tasks where mappings play a central role. This underlines the relevance of mappings, and shows a clear need to update them every time a KOS evolves to ensure consistency between the cross-referenced KOSs and to keep reliability over time.

To manually keep mappings updated requires huge efforts. Current biomedical KOSs contain a large number of concepts, which reflects on the number of mappings established. Biomedical KOSs usually contain hundreds of thousands of concepts interconnected via mappings. Analysing every single mapping by hand to check their semantic consistency requires enormous time and human resources. Human experts could manually perform this on small KOSs, with a restricted number of mappings, but large and highly dynamic KOSs, like those of the biomedical domain, demand appropriate methods and automatic tools.

To avoid the burden of manual intervention, KOS engineers and experts might delete the whole set of mappings each time a new version of a KOS is released, and re-apply existing matching algorithms [Rahm, 2011] by considering all entities of both aligned KOSs. The periodical and fast release of new KOS versions makes this approach very inefficient because the alignment task remains error-prone and requires tremendous time and efforts to complete and evaluate the resulting mappings due to the large number of possible correspondences between biomedical KOSs. The fact of deleting the whole set of mappings wastes the entire work already performed to validate the created mappings.

To prevent deleting the whole set of mappings requires knowing which mappings are affected by KOS evolution. This may help domain experts only repairing mappings influenced by KOS

[9] `www.izbi.de/onex`

changes. However, this forces identifying changes affecting KOS's entities, determining cases where KOS changes impact mappings, and providing methods for modifying mappings accordingly, which remains unknown in literature. Therefore, to avoid the costly KOS re-alignment process [Euzenat and Shvaiko, 2007] and to guide humans in the laborious task of maintaining mappings up-to-date requires novel supporting tools with semi-automatic approaches to keep mappings semantically valid over time.

This thesis aims to provide an answer to the following major research question:

"How to perform the maintenance of mappings to keep them up-to-date in a semi-automatic way when KOSs evolve, without re-computing the whole set of mappings?"

Coping with the mapping maintenance problem in a fully automatic way entails many open issues (*cf.* Section 2.2.1 that provides a formal definition of this task). We present the specific research questions that this PhD thesis particularly faces when formulating the problem statement throughout the chapters.

Although literature has suggested methods to create and maintain mappings as well as how to manage KOS evolution, the described problematic has motivated us to propose novel alternatives to address (semi-) automatically the reconciliation of dynamic KOSs, which may decrease the time, human efforts and costs to maintain mappings up-to-date.

1.2 Objective and contributions

This thesis originally introduces a novel approach, so-called *DyKOSMap*, for coping with the mapping maintenance problem between dynamic KOSs, like those of the biomedical domain. We address the mapping maintenance by devising new methods suited to adapt mappings impacted by KOS evolution without re-computing the whole set of mappings each time a KOS evolves.

We formulate the main objective of this thesis as follows:

Propose an approach to effectively adapt mappings as automatic as possible under evolving biomedical KOSs.

Mapping adaptation refers to the process where existing mappings are modified, according to changes affecting KOSs, to keep them up-to-date and complete over time [Dos Reis et al., 2013a]. This thesis underlines the possibility of attaining a more reliable and accurate mapping adaptation process by considering relevant information derived from KOS evolution, combined with information coming from existing mappings.

First, as we observe that KOS evolution contributes to the main reason for impacting mappings, understanding and classifying the different aspects of evolution related to KOSs might influence the effectiveness of appropriate methods for adapting mappings. Second, we need to explain established mappings relying on KOS's entities. Our approach assumes that some textual attributes (*e.g.*, name or a synonym term) characterizing concepts of biomedical KOSs, involved within mappings, might play a central role in the interpretation of existing mappings and consequently influence their adaptation. Third, this demands in particular focusing on the

evolution's characterization of these KOS's entities relevant for explaining mappings.

This thesis proposes an empirical approach to reveal the adequate level of granularity with respect to KOS evolution and mapping interpretation that demands a semi-automatic mapping adaptation system. To this end, first of all we conduct in-depth practical experiments with real data as the experimental frame of reference of this thesis. In addition, we clarify the state-of-the-art through a comprehensive literature review that allows to point out the open issues for achieving an automatic approach to mapping maintenance.

The conducted empirical analyses enable us to justify the components in our approach and to demonstrate the key factors to support mapping adaptation. This, combined with the literature review, allows us to affirm the originality of this thesis. The originality relies on the fact that we approach mapping adaptation based on the level of concept attributes according to our observations from the experiments. The proposed approach explains existing mappings via concept attributes, defining the involved concepts in mappings, and characterizes the evolution of these attributes via specific KOS change operations. This provides means to recommend actions suited to adapt mappings with more accuracy. Existing approaches to cope with the mapping maintenance and adaptation fail to understand and define involved factors with accuracy, which forces them to remove and re-match correspondences or to apply simplified but inappropriate adaptation strategies.

To instantiate our approach, we explore the biomedical domain with representative and real KOSs selected to conduct our experiments and validations. In addition to the motivation scenario previously described, we use the biomedical domain as a case study because of multiple aspects: (1) the real usefulness of KOS mappings; (2) the dynamic characteristic of the domain (new KOS versions are periodically released with a high frequency); (3) we find real-world KOSs (and mappings) created and validated by domain experts; (4) the datasets are freely available for research purposes. In principle, we can successfully apply our approach to any other domain where the KOS mappings rely on the level of concept's textual attributes.

We summarize the prime scientific contributions of this thesis as follows:

1. We approach the mapping adaptation based on an empirical basis that uncovers the effects of KOS evolution on the ways that mappings evolve.

   We thoroughly evidenced the way that different types of KOS changes impact the behavior of existing official mappings. This demonstrated the underlying primary factors influencing mapping adaptation.

2. We define the *DyKOSMap* approach to mapping adaptation.

   We determined the necessary aspects for performing (semi-) automatic mapping adaptation. Our research uncovered the elements regarding mappings and KOS evolution that can support mapping adaptation techniques. In particular, we proposed a method to automatically identify concept attributes defining mappings. Our approach suggested the definition of fine-grained change operations suited to characterize the evolution of the relevant KOS's entities to support mapping adaptation, and adequate algorithms to recognize

them between KOS versions. To adapt mappings, we formalized a set of mapping adaptation actions and heuristics accommodated to decide the adequate adaptation actions.

3. We introduce and develop the *DyKOSMap* framework.

   We organized our approach describing the involved components and their relations within a novel framework. This framework deals with the interpretation of mappings and KOS evolution in an original mapping adaptation process. We developed a software prototype implementing the whole process for mapping adaptation in the framework.

4. We thoroughly evaluate all the components of the approach and the framework as a whole.

   The thorough evaluations experimentally validate all the suggested methods deployed within the components of the proposed framework. We separately assessed each suggested component and globally evaluated the performance of the framework. The results empirically achieved from the scholarly experimental evaluations attested the effectiveness of our approach and methods within the proposed framework.

1.3 Thesis outline

We organize this thesis into part I and II and will introduce in more detail the chapters' content description when presenting part I and II, respectively. We devote the first part to an analysis of bibliography and the empirical studies of reference. Chapter 2 introduces basic definitions and presents the state-of-the-art. Chapter 3 describes a set of experiments conducted with biomedical KOSs and mappings to observe the evolution of mappings in a real-world scenario. We rely the proposed approach – to cope with mapping adaptation in the framework – on the results of this empirical frame of reference.

The second part presents our proposed approach in the *DyKOSMap* framework. Chapter 4 introduces the study to define and experiment our method to interpret existing mappings. Afterwards, we present the contribution (including an experimental evaluation) to characterize KOS evolution as specific types of change patterns at the level of concept's attributes (Chapter 5). We devote the chapter 6 to present the proposed techniques to adapt mappings. Chapter 7 introduces the framework connecting the suggested and evaluated components. Chapter 8 closes this thesis by providing a summary of our achievements and highlighting the contributions made throughout the chapters. We present several directions for future work.

Part I

Analysis of bibliography and empirical data

We devote this first part to present the bibliographical and experimental groundwork of the thesis. We conduct a thorough analysis concerning: (1) the literature addressing the specific problems tackled in the thesis (Chapter 1), and (2) our empirical referential through a series of experiments observing the evolution of real-world biomedical KOS and mappings (Chapter 2) (*cf.* Figure 1.1). This paves the foundations for our proposed approach.

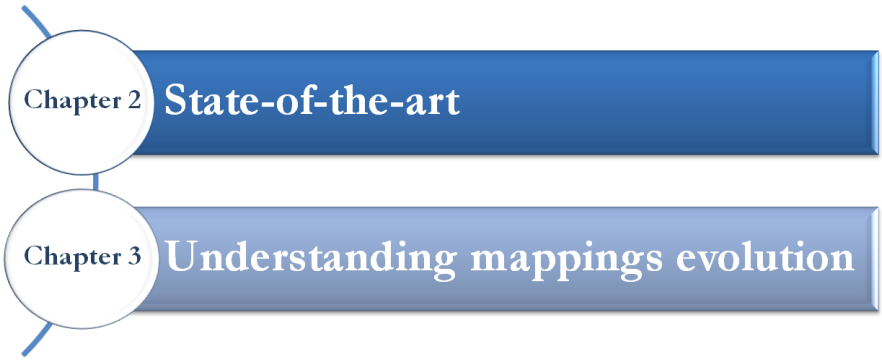


Figure. 1.1: Chapters composing the first part of the thesis

Chapter 2 – State-of-the-art

This chapter begins by formalizing a definition of KOS model and mappings used throughout this thesis. We illustrate and formally define the mapping maintenance problem, then organize and describe an in-depth survey of approaches addressing this problem. We provide an analytical comparison between them. In the sequence, the chapter overviews methods for interpreting established mappings necessary for dealing with our approach. The chapter also briefly presents the literature related with KOS evolution by showing the KOS change operations offered by existing tools. We emphasize investigations that aim to identify change operations between KOS versions and characterize KOS change patterns. Finally, the chapter discusses general open issues that literature fails to address to reach fully automatic mapping maintenance and draws conclusions. The contributions in this chapter have been published in [Dos Reis et al., 2015b].

Chapter 3 – Understanding mappings evolution

This chapter presents a series of descriptive and detailed analyses on the effects of KOS evolution on mappings. We empirically observe the evolution of various real biomedical KOSs as our practical basis. The chapter describes the experimental scenario and we use the official mappings established between SNOMED CT and ICD-9-CM from 2009 to 2011. We investigate whether potential correlations exist between types of changes affecting KOS's entities and the adaptation of mappings associated with KOS's concepts. We present the experimental procedure of the different types of analyzes. Results highlight factors according to which KOS changes in varying degrees influence the evolution of mappings and identify relevant elements involved in mapping maintenance. This chapter's content has been published in [Dos Reis et al., 2014c, Dos Reis et al., 2013b].

We establish and justify our approach in view of this literature survey and empirical findings.

Chapter 2

State-of-the-art

Contents

Introduction **19**
2.1 Definitions **20**
2.1.1 KOS model 20
2.1.2 KOS mappings 21
2.2 Mapping maintenance **22**
2.2.1 The mapping maintenance problem 22
2.2.2 Mapping revision 24
2.2.3 Mapping calculation 27
2.2.4 Mapping adaptation 29
2.3 Mapping interpretation **32**
2.4 KOS evolution characterization **34**
2.4.1 Types of KOS change operations 34
2.4.2 KOS change patterns 36
Summary **39**

Introduction

Addressing the mapping maintenance problem due to KOS evolution via mapping adaptation provides opportunities to intersect two large, but still current largely disjoint fields in the Semantic Web [Berners-Lee et al., 2001] research area. On one hand, the growing development and use of several KOSs has motivated investigations to create KOS mappings via semi-automatic techniques for KOS concepts aligning [Euzenat and Shvaiko, 2007]. On the other hand, the dynamic nature of knowledge-based applications essentially makes KOSs grow and change over time [Shaban-Nejad and Haarslev, 2009]. While mappings have been created without concerning the evolution of KOSs, the latter task has been largely investigated without paying much attention to the consequences of KOS changes on dependent artefacts such as mappings.

We need to investigate how to determine information from established mappings and KOS evolution that can appear useful for mapping maintenance. This chapter faces the following research questions:

1. Which categorization of approaches address the mapping maintenance problem?
2. Are there techniques suited to interpret established mappings?
3. How do existing proposals characterize KOS evolution regarding mapping maintenance?

This chapter provides a thorough survey of the state-of-the-art on mapping maintenance. As we cope with the maintenance via adapting mappings and our approach to mapping adaptation aims to explore information derived from mappings and KOS evolution, we also have an interest in methods that allow the interpretation of mappings and literature concerning KOS evolution characterization. In summary, we make the following contributions in this chapter:

- We systematically review the literature on mapping maintenance problem offering a complete state-of-the-art by presenting, comparing and discussing existing proposals in different suggested categories.
- We present how existing proposals explore and interpret established mappings.
- We demonstrate the types of KOS changes described in literature and attempts for providing KOS change patterns. We study adequate KOS evolution tools that our approach could reuse to determine KOS changes between KOS versions.

We start presenting basic definitions for a KOS model (Section 2.1.1) and KOS mappings (Section 2.1.2). We then formalize, explain and exemplify the mapping maintenance problem (Section 2.2.1). Our careful and articulated analysis of the reviewed literature on mapping maintenance allows to discuss lacks evidencing uncovered aspects (Section 2.2). Additionally, we examine whether existing proposals for mapping interpretation allow providing KOS's entities suited to conceptually justify the established correspondences (Section 2.3). For instance, what information from two linked concepts may explain the relation connecting them. Afterwards, we mainly emphasize the categorization of different types of KOS changes (Section 2.4.1) and briefly describe the major tools to calculate the difference between two versions of a KOS resulting in a set of KOS changes. In this context, we show approaches suggesting change patterns to improve the characterization of KOS evolution (Section 2.4.2).

2.1 Definitions

We provide a definition for a KOS model and mappings that we use throughout in the thesis.

2.1.1 KOS model

In the biomedical domain, different organizations define and maintain KOSs. Hence, biomedical KOSs vary in terms of scope, formats and structures [Vandenbussche and Charlet, 2009]. Usually, they are generically treated as ontologies. Nevertheless, biomedical KOSs rely on slightly different knowledge representation models. The models contain different degrees of expressiveness, especially in how to organize knowledge and possibilities of pre-defined semantic relations [Rees, 2003]. For example, ontologies allow formally describing domain relationships while a taxonomy is mostly restricted to "*is-a*" types of relationship. To understand the range of expressiveness in the biomedical KOSs, Vandenbussche & Charlet [Vandenbussche and Charlet, 2009] conducted a study based on different semantic models and propose a new meta-model for KOS representation based on the observation of different biomedical KOSs. They claim that this

meta-model may offer a more general representation formalism, since this refers to the union of the existing formalism specificities.

Literature has highlighted the differences among such models along with their definitions. Experiments have tried to compare biomedical KOS from different perspectives. Some investigations have focused on the differences in terms of the terminologies' contents [Bodenreider, 2008, Freitas and Schulz, 2009]. Others [Rees, 2003, Breitman et al., 2007] emphasize on the underlying semantic model, by discussing the distinction between the notions of taxonomy, thesaurus and ontology. We do not aim to define and compare KOS models, but we adopt a minimal model to support our approach that we will use as a reference throughout this thesis.

Biomedical KOSs motivate our adopted definition of KOS. These KOSs consist in less logically rigorous ones and instances usually are unavailable. We evoke the classic definition of Gruber [Gruber, 1993] that conceptualizes a domain in terms of concepts, attributes and relationships. A **KOS** K specifies a set of concepts interrelated by directed relationships. We define a set of concepts of a KOS K_x at time j, such that $j \in \mathbb{N}$, as $C(K_x^j) = \{c_i^j | i \in \mathbb{N}\}$. Each concept $c \in C(K_x^j)$ has a unique identifier and is associated with a set of attributes $A(c) = \{a_i | i \in [1..p]\}$ (*e.g.*, label, synonym, definition, *etc.*), where p is the number of attributes characterizing concept c. Furthermore, each attribute is defined for a particular objective, *e.g.*, "*label*" for denoting concept names or "*definition*" for giving the meaning in the context where the concept is used. A relationship rel between two concepts, $c_1 \in C(K_x^j)$ and $c_2 \in C(K_x^j)$ interrelates a particular concept and another one in the same KOS, *e.g.*, $rel(\text{“}is-a\text{”}, c_1, c_2)$, where the label of c_1 refers to "*brain cancer*", and "*cancer*" is the label of c_2, respectively.

2.1.2 KOS mappings

Interconnecting several KOSs is required to globally describe the underlying domain content because KOSs are usually constructed to cover specific subjects. Mappings refer to the materialization of semantic relations between entities of interrelated KOSs [Euzenat and Shvaiko, 2007].

Given two KOSs namely K_S and K_T, we define K_S as the source KOS and K_T the target KOS of mappings. A **mapping** m_{st}^j, established at time j, between two concepts $c_s^j \in C(K_S^j)$ (namely source concept) and $c_t^j \in C(K_T^j)$ (namely target concept) is given by:

$$m_{st}^j = (c_s^j, c_t^j, semType^j) \tag{2.1}$$

where $semType^j \in \{\perp, \equiv, \leq, \geq, \approx\}$ refers to the semantic relation between c_s^j and c_t^j. The $\perp$ stands for *unmappable*, [$\equiv$] *equivalent*, [$\leq$] *more specific than*, [$\geq$] *less specific than* and [$\approx$] *partially matched*, respectively. For instance, concepts can be equivalent (*e.g.*, "*torso*" $\equiv$ "*trunk*"), one concept can be less or more specific than the other (*e.g.*, "*lower limbys*" $\leq$ "*limb segment*") or concepts can be somehow related ($\approx$). We define $\mathcal{M}_{ST}^j = \{(m_{st}^j)_i | i \in \mathbb{N}\}$ as the set of different mappings at time j between KOS K_S and K_T.

Mappings can be manually or (semi-) automatically created with computer-supported alignment tools supervised by knowledge engineers and domain experts. The increasing size of KOSs remains a barrier to the manual performance of this task. This aspect, combined with the growing need of semantic interoperability among heterogeneous systems, has increasingly pushed the

efforts on semi-automatic matching approaches to aligning KOSs, by finding correspondences between concepts. The matching task aims at finding an alignment between a given pair of KOS. The number of matching parameters makes this task very difficult [Euzenat and Shvaiko, 2007].

A large number of researches have investigated approaches to KOS matching to generate KOS alignment. They propose algorithms exploring isolated approaches or using a wide range of hybrid techniques to increase the precision and success of matching algorithms to create mappings between KOSs [Rahm, 2011, Ngo et al., 2013, Tran et al., 2011, Euzenat et al., 2011, Shvaiko and Euzenat, 2013].

2.2 Mapping maintenance

The huge necessity and efforts to provide correct KOS mappings motivate further research to keep them up-to-date over time. We define mapping maintenance as follows:

Mapping maintenance refers to the task aiming to keep existing mappings in an updated and valid state, reflecting changes affecting KOS's entities at evolution time [Dos Reis et al., 2014c].

In the following, we introduce the mapping maintenance problem (Section 2.2.1). A thorough analysis of the literature dealing with mapping maintenance allows us classifying existing approaches to address this problem into three distinct categories:

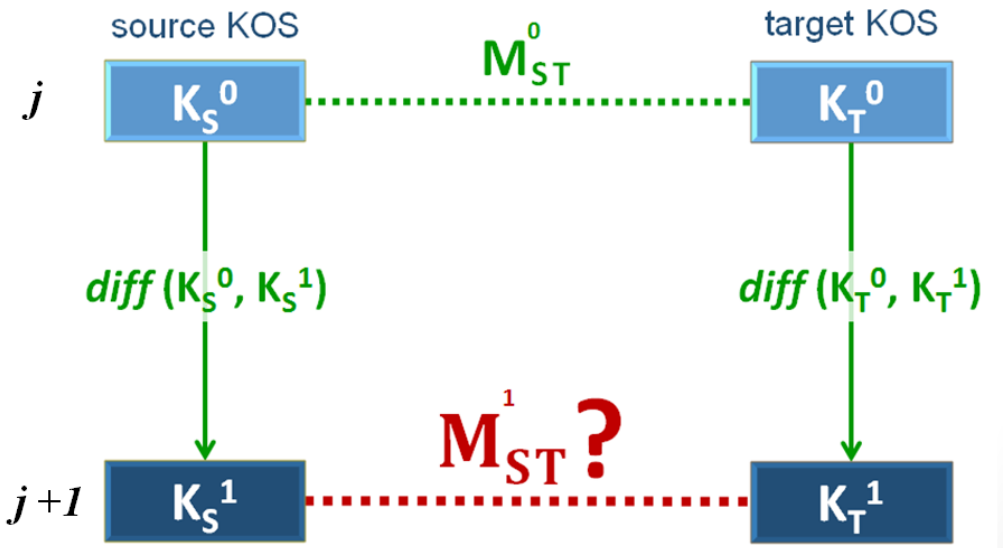
- Mapping revision (Section 2.2.2);
- Mapping calculation (Section 2.2.3);
- Mapping adaptation (Section 2.2.4);

2.2.1 The mapping maintenance problem

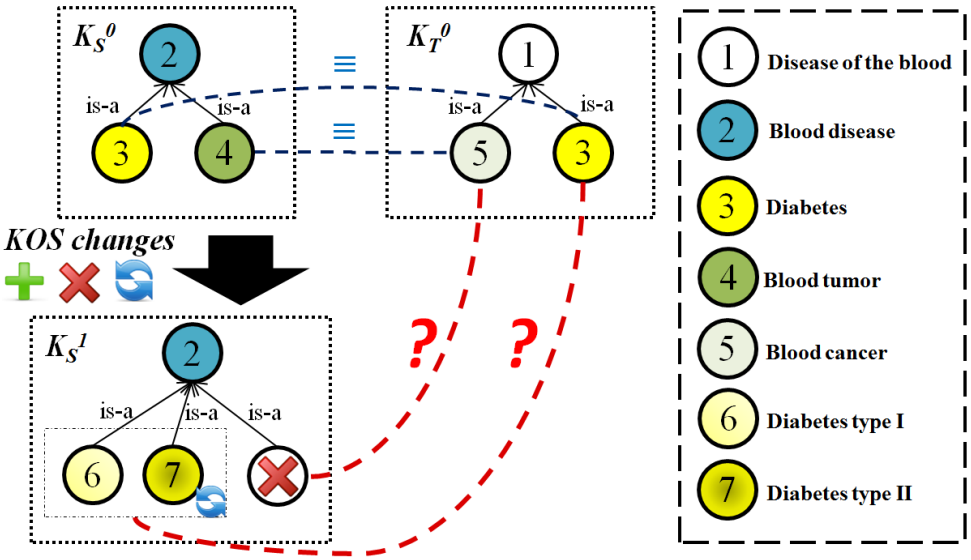
The evolution of a KOS [Klein and Noy, 2003] in terms of atomic or complex changes affecting its entities may invalidate previously determined mappings [Dos Reis et al., 2012]. In other words, given a mapping m^j_{st}, due to the modifications affecting the concept c^j_s or c^j_t, the type of semantic relation $semType^j_{st}$ no longer represents the correct semantic link between c^j_s and c^j_t.

Figure 2.1a presents the general scenario of the investigated mapping maintenance problem. Since we consider KOS evolution, we examine different versions of each KOS. Given two versions of the same source KOS, namely K^0_S at time j and K^1_S at time $j+1$, we always have at least one target KOS K_T and an initial set of valid mappings $\mathcal{M}^0_{ST}$ between K^0_S and K^0_T at time j.

If K_S or K_T evolves, represented by a set of KOS change operations (*diff*), we need to determine the set of updated mappings $\mathcal{M}^1_{ST}$, since the evolution probably impacts mappings in $\mathcal{M}^0_{ST}$. We can have a simplified view of this scenario considering the evolution of only one KOS per time. We assume that the target KOS remains unchanged (*i.e.*, $K^0_T \equiv K^1_T \equiv K_T$) while the source evolves (or *vice-versa*). The results of the mapping maintenance task must consist in a set of up-to-date mappings in $\mathcal{M}^1_{ST}$.



(a) Problem definition



(b) Example of the problem with biomedical KOS

Figure. 2.1: The mapping maintenance problem

Figure (a) shows the mapping maintenance problem. A source and a target KOS interrelated via a sets of mappings $\mathcal{M}_{ST}^0$. New versions of these KOSs may trigger KOS changes that can affect existing mappings. Figure (b) presents an example of the mapping maintenance problem. We depict concepts with their label. The KOS K_S^0 evolves to a new version K_S^1.

Figure 2.1b illustrates an example of a problematic scenario with the evolution of biomedical KOSs. The concept "*diabetes*" in K_S^0 corresponds to the concept named "*diabetes*" in K_T^0. Similarly, "*blood tumor*" and "*blood cancer*" are stated as equivalent. The evolution of K_S^0 affects both concepts in K_S^1. In K_S^1 (*i.e.*, new version of K_S), the concept "*blood tumor*" is deleted and the concept "*diabetes*" is split into "*diabetes type I*" and "*diabetes type II*". We therefore need to

maintain existing mappings affected by KOS evolution whose definition depends on the changed concepts.

Note that the total removal of a concept consists in a relative easy case to handle mapping maintenance, because we can propose to remove mappings. However, scenarios where only some concept attributes are removed or their values are changed, pose a more challenging circumstance to decide a possible action to guarantee an up-to-date state for mappings.

Our adopted definition of KOS mapping (*cf.* Equation 2.1) differs from the one largely accepted by the database community where a schema mapping specifies how data instances of one schema correspond to data instances of another [Velegrakis et al., 2004a]. Mapping maintenance has been historically studied in database schemas. In this context, they represent mappings in a declarative way as queries or view definitions, having a very general form of mapping. They describe a mapping q from a schema S (called the source schema) to schema T (called the target schema), as an assertion of the form: $Q_S \longrightarrow Q_T$, where Q_S consists in a query over S and Q_T refers to a query over T [Velegrakis et al., 2004a]. When handling mappings between XML models they use a similar definition of mapping. Observe that the definition of KOS mappings in terms of correspondences between concepts differs from a schema view or query. A unique query may involve a set of dependent equivalent schema correspondences, which changes the level of abstraction of the mapping definition.

In the following, we present our survey for the different approaches addressing the mapping maintenance problem.

2.2.2 Mapping revision

We consider two dimensions on approaches in this category: the first aims at only identifying invalid mappings; the second detects invalid mappings via debugging methods and repairs them. These approaches are usually implemented to correct the bias induced by matching algorithms at KOS alignment time, but rarely to maintain mappings affected by the evolution of one of the involved KOS. Figure 2.2 presents the mapping revision proposal.

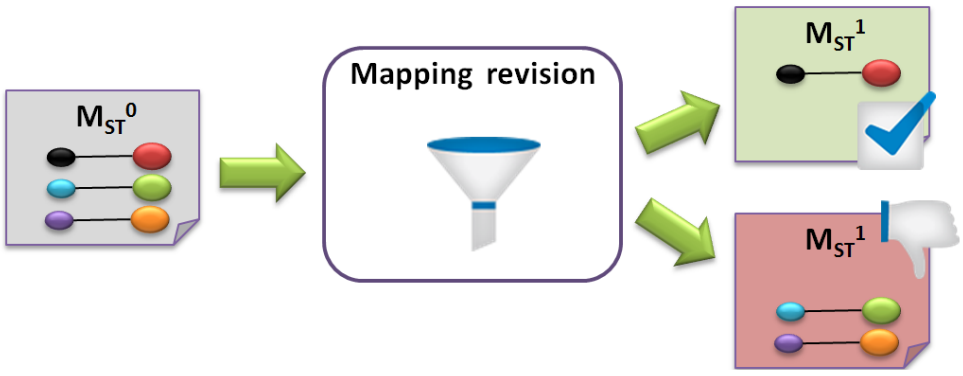


Figure. 2.2: Mapping revision category

This figure shows the proposition of mapping revision, where a subset of mappings are identified and/or repaired.

Identification of invalid mappings

We found different approaches aiming to monitor and identify invalid mappings. They mostly apply to database schemas and rely on different techniques. McCann *et al.* [McCann et al., 2005] proposed monitoring and detecting invalid mappings between dynamic relational schemas in an automatic way, through the *MAVERIC* system. This system periodically checks source schemas and compares query answers to prior known answers. Once query answers differ: the system sends an alert about a potential broken link to the administrator. The approach utilizes mainly three components: sensors, learning and filtering to monitor the characteristics of the schemas, to train the sensors and to filter false alarms of broken mappings while comparing them against known formats and values, respectively.

Similar to *MAVERIC*, Mawlood-Yunis [Mawlood-Yunis, 2008] investigated fault-tolerance techniques to detect temporal semantic mapping inconsistencies in peer-to-peer systems. The author distinguished between permanent and temporary semantic incompatibilities. The temporal dimension refers to the duration in which a mapping fault stays effective. The proposed solution explores a query replication comparing their respective results. Unlike the *MAVERIC* approach, fault-tolerance approaches detect invalid mappings at query time. Motivated by the same problem, Li *et al.* [Li et al., 2010] presented an automatic solution to detect broken mappings based on fuzzy theory. As shown in [Li et al., 2010] this approach allowed better detecting inconsistent mappings in database schemas.

The work of Colazzo & Sartiani [Colazzo and Sartiani, 2009] presented a similar goal, but introduced different properties in the identification method. In their approach, mappings consist in *XQuery* clauses [Chamberlin, 2002] and they identify corrupted ones relying on the comparison of *XQueries* results when applied to instances of the source schema with the instances of the target schema. This induces the notions of mapping validity and correctness in which the system fails to find wrong clauses with respect to the source and target schemas. The proposed technique combines two properties, namely type projection and type inference, independent from queries posed against the integrated database. Based on these properties, the authors proved the decidability of the proposed technique.

Inspired by these investigations on database schemas, recent studies have addressed the broken link problem in the context of open linked data[10]. Popitsch and Haslhofer proposed a change detection framework (*DSNotify*) [Popitsch and Haslhofer, 2011] for dynamic datasets applied to linked data sources to inform actors about various types of events (create, remove, update). This allows actors to maintain links to resources in a distributed environment, fixing broken links in their local data, while preserving the links' integrity. In contrast to previously described approaches, this work focused on identifying events that can lead to broken mappings.

The work of Pathak & Chute [Pathak and Chute, 2009] aims at using reasoning mechanism based on distributed description logics to detect inconsistent mappings between biomedical ontologies. In their approach, the authors consider equivalence and subsumption in mappings and assume that the ontologies are consistent and therefore, semantic inconsistencies obtained when reasoning on it are produced by the mappings. In consequence, they are able to detect inconsistent mappings.

[10] linkeddata.org

Repairment of mappings

Detecting invalid mappings accounts for the first step to maintain them. Only considering the detection remains insufficient for the full maintenance, since this still may require a lot of human intervention to realize how to proceed with invalid mappings. Techniques applied for the identification may support part of a more complete solution for mapping maintenance. Knowledge engineers need additional support to better perform the maintenance in a complete process.

Towards this objective, the first attempts proposed to categorize suspicious mappings suggesting different types of errors and warnings in mapping revision [Wang and Xu, 2008]. Errors consist in the confirmed wrong mappings, but warnings concern the possibly wrong, the right or the imprecise ones. Wang and Xu's work automatically repairs some errors and warnings or presents them to users with revision suggestions. They further refine the proposed categorization including: *redundant mappings* (mappings inferred from existing ones); *imprecise mappings* (inaccurate); *inconsistent mappings* (mappings that may disobey declared axioms) and *abnormal mappings* (mappings having an inappropriate behaviour).

To improve the automated aspect in mapping revision, several proposals have suggested solutions mostly based on logical reasoning methods and conflict detection. Meilicke *et al.* [Meilicke et al., 2007] suggested a debugging method to check the consistence of established mappings. They argue on the need for debugging mappings because of the inefficiency of existing matchers. These authors apply a procedure in which they consider two sets made up of inconsistent and valid mappings, respectively. They addressed the problem of how to determine the intersection between both sets. Since the reference mappings (gold standards) often appear inaccessible or even unknown, they reformulated the problem partitioning the set of mappings into correct and incorrect correspondences with respect to the set of reference. Therefore, repairing a set of correspondences means to determine for each correspondence whether it belongs to the set of correct or incorrect mappings, and eventually deleting the set of incorrect correspondences.

In this context, Meilicke *et al.* [Meilicke et al., 2007] presented an approach that eliminates inconsistencies caused by erroneous mappings through logical diagnostic reasoning, exploiting disjointedness axioms and machine learning techniques [Meilicke et al., 2008]. The underlying assumption affirms that mappings model semantic correspondences between concepts in different ontologies without introducing inconsistencies, *i.e.*, a set of mappings that correctly describes the semantic relations between ontologies should not cause inconsistencies in any of the ontologies. To detect these inconsistencies, the authors explored distributed description logics as a basis for formalizing mappings.

The proposition of Meilicke has been extended by Nikitina *et al.* [Nikitina et al., 2011]. They described consistency criteria for revision states and introduced the notion of revision closure, based on which the revision of ontologies is partially automatized. Moreover, to optimize the revision process, they suggested the notion of axiom impact to determine a beneficial order of axiom evaluation. Similar to the work of Meilicke, Jimenez *et al.* developed the *LogMap* system [Jiménez-Ruiz and Grau, 2011] able to repair, using reasoning techniques, inconsistent mappings. Also in this direction, Qi *et al.* [Qi et al., 2009] proposed the so-called conflict-based mapping revision operator based on the notion of conflict set to detect erroneous mappings.

Aligned with approaches based on reasoning methods, Castano *et al.* [Castano et al., 2008]

suggested a probabilistic reasoning approach to validate mappings with respect to the semantics of the involved ontologies. The particularity of this work considers mappings as probabilistic and hypothetical relations among KOS's entities expressed in description logics. The proposal performs a probabilistic reasoning over mappings and revises them according to the reasoning results. The method discards mappings causing inconsistent reasoning and infers new mappings from the valid ones. More recently, Ivanova & Lambrix [Ivanova and Lambrix, 2013] proposed an unified framework for debugging taxonomies and their alignments, which aims to detect and repair wrong or missing relations and mappings in a network of taxonomies.

Mapping revision approaches provide various benefits for the mapping maintenance required when KOSs evolve. However, the proposed techniques for the reparation mostly rely on reasoning resources which usually only apply for KOSs based on formal logics. These constraints may decrease the usability of mapping revision techniques.

2.2.3 Mapping calculation

In contrast to the previous category of approaches, mapping calculation performs no identification of invalid mappings. This consists in totally, or partially, re-calculating existing mappings to maintain them up-to-date. Figure 2.3 shows the different approaches in the mapping calculation proposal.

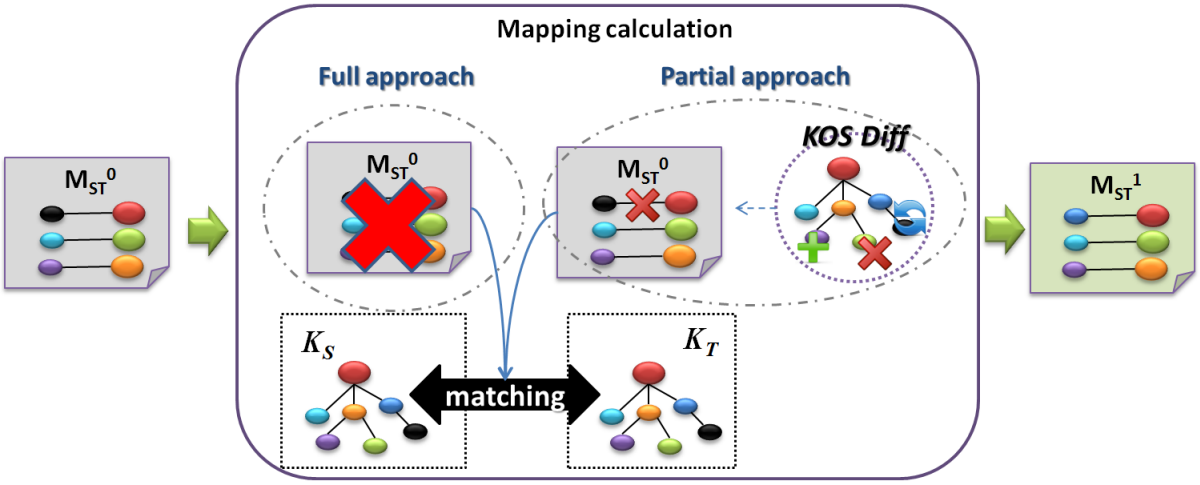


Figure. 2.3: Mapping calculation category

This figure presents the proposition of mapping calculation. The full re-calculation approach removes the whole set of mappings and a matching between source and target KOS happens. The partial re-calculation approach uses the KOS diff to remove the mappings affected by KOS evolution, and applies a matching between the partial part of the source KOS and the whole target KOS.

Full re-calculation

In this approach, all mappings are considered as invalid ones, therefore they are deleted and re-computed from scratch again. Hence, considering the evolution of interconnected KOSs, a naïve approach recalculates the whole set of mappings each time an organization in charge releases a

new version of a source or target KOS. In the context of relational schemas this solution implements schema matching techniques [Gal et al., 2005], while when considering ontologies and taxonomies, knowledge engineers may use ontology matching algorithms and alignment methods [Euzenat and Shvaiko, 2007].

Despite the recent advances on ontology matching [Shvaiko and Euzenat, 2013], the literature highlights that the process of automatic mapping calculation is error-prone and time-consuming with respect to the size of both mapping sets and KOS, and demands a laborious human effort for validation [Velegrakis et al., 2004a]. Yu and Popa [Yu and Popa, 2005] argue that besides the involved costs, these approaches fail to consider the original mappings during the re-calculation, creating a loss of information. In consequence, the re-calculated mappings may differ from the original and desired ones [Kondylakis et al., 2009]. Overcoming these limitations requires novel approaches to reduce human efforts and to improve the quality and reliability of resulting mappings. This has led to proposals focusing on techniques to restrict the number of entity combinations for the re-calculation of mappings.

Partial re-calculation

Proposals following the partial re-calculation of mappings aim at decreasing the reconciliation time of dynamic KOS. Zurawski *et al.* [Zurawski et al., 2008] investigated how to deal with logical consistency of autonomous and dynamic KOSs interconnected in a distributed way. The authors defined and formalized the property of semantic autonomy referring to the ability to support decentralized semantics that can evolve in a consistent way based on their own local needs. According to them, this property might support organizational and distributed knowledge management systems. The authors proposed an ontology-based layered framework suited to deal with mapping maintenance caused by KOS evolution. However, they only consider removing and adding mappings.

The proposition of Khattak *et al.* [Khattak et al., 2012] aims to modify only mappings associated with changed KOS entities (concepts, relationships and attributes) and eliminates the staleness from mappings. According to them, this reduces the time required to maintain mappings up-to-date without affecting their precision. This approach explores an ontology change history (*CHL*) [Khattak et al., 2008] to explore KOS evolution and identifies the subset of mappings associated with changed entities by recovering these entities from the *CHL*. Their solution removes this subset of affected mappings. The solution further adds new mappings calculated using matching techniques considering the whole target KOS.

Although some work in this category have considered KOS evolution, they only dealt with atomic KOS changes in a simple way. They explore KOS changes to identify the modified KOS entities, but fail to use KOS changes information to maintain the mappings. Similarly to the mapping revision approaches, they do not promote the reuse of already determined and curated mappings remaining highly dependent on the performance of matching techniques to recalculate mappings. Furthermore, since proposals remove mappings and add totally new ones, these new mappings still require a human validation.

2.2.4 Mapping adaptation

Proposals in this category take various dimensions of KOS evolution into consideration to tackle the mapping maintenance, including:

- Mapping composition
- Mapping rewriting in database schemas
- Synchronization of models
- Change impact minimization
- Change propagation
- Mapping change strategies

Figure 2.4 represents the mapping adaptation proposal, where approaches mostly explore KOS evolution to change mappings.

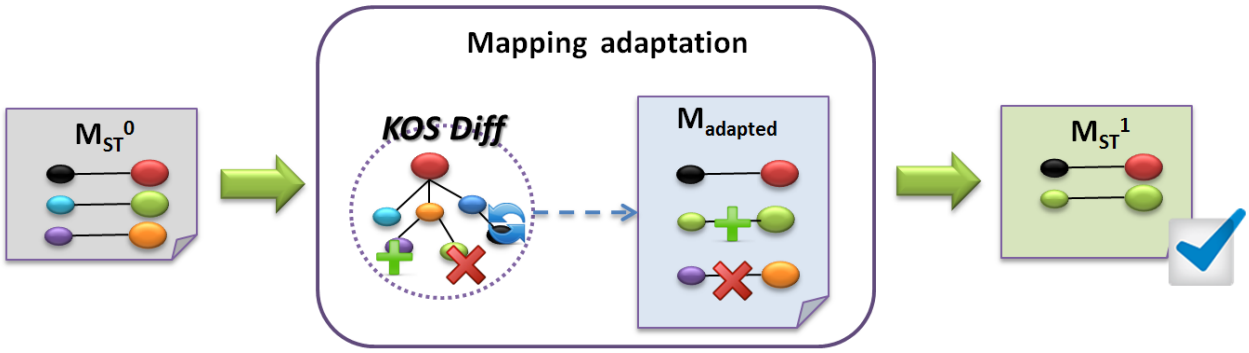


Figure. 2.4: Mapping adaptation category

This figure illustrates the mapping adaptation proposal. The KOS diff representing the KOS evolution is used to change mappings with different strategies.

Mapping composition

The mapping composition principle suggests composing various mappings to generate new ones. Yu & Popa [Yu and Popa, 2005] proposed the *MACES* system to reuse the original mappings and adapt them in a way that: (1) the system preserves the original intention of the original mappings as much as possible, and (2) it takes into account schema changes. They aim to compose two kinds of mappings: the original ones and those that allow moving from a schema to its new version.

More recent work has proposed other types of operators to complement the composition. Fagin *et al.* [Fagin et al., 2011] presented how to explore an inversion operator, together with the composition, to address the mapping maintenance problem using schema evolution. They discuss both operators and use concrete examples of schema evolution to illustrate the main developments and challenges of composition and inversion. Groß*et al.* [Groß et al., 2013] proposed

to enhance the mapping composition principle in the context of large life science ontologies. They completed the resulting set of mappings using matching techniques over newly added concepts.

Results comparing mapping composition with other techniques for mapping adaptation have shown the potentialities of this technique [Groß et al., 2013]. Although the composition principle enhances the reuse of mappings, in some more complex contexts of KOS evolution such as the split of concepts, composition generates many candidate mappings and among them false positives, which decreases the precision of this technique.

Mapping rewriting in database schemas

Unlike previous approaches, the schema mapping rewriting technique performs mapping adaptation by incrementally rewriting the elements in queries representing mappings between database schemas. Velegrakis *et al.* [Velegrakis et al., 2004b] presented the *ToMAS* framework for automatically adapting mappings as database schemas evolve. They argue the need of reflecting schema changes at mapping level in a way to preserve consistency with respect to the semantics of changes and current mappings. In this work, schema mappings, expressed as queries, link data instances of different schemas and mapping maintenance consists in query rewriting. The approach incrementally adapts schema mappings each time a primitive change occurs in the schema. It considers both constraint and structural changes affecting database schemas like adding and removing constraints, expanding, pruning and restructuring schema.

Xue & Kedad [Xue and Kedad, 2007] proposed a three-step incremental mapping adaptation process relying on changes occurring in XML schemas. They express original mappings in *XQuery* [Chamberlin, 2002] following specific patterns, and combine them with schema elements that may become invalidated according to schema changes. Their work considers a pre-defined set of changes that can occur in XML schemas and applies a specific mapping adaptation procedure according to each of these changes (additions, removals and renaming).

Despite improvements in mapping rewriting approaches, literature points out some drawbacks, mainly criticizing two points [Yu and Popa, 2005]: (1) the definition and acquisition of primitive schema changes, and (2) the semantics of the adaptation operations and their impact. Kondylakis *et al.* [Kondylakis et al., 2009] questioned the applicability of such approaches in the context of data integration for highly dynamic ontologies by discussing the main differences between schemas and ontologies. Moreover, the widely used definition of KOS mappings which interrelates entities through a semantic relation symbol raises other specificities that rewriting solutions fail to handle. For instance, this adaptation technique does not allow a modification of the semantic relation type of the mapping to assure that the mapping remains up-to-date due to modifications affecting the involved concepts.

Synchronization of models

Some scenarios require establishing mappings between heterogeneous models like database schemas and ontologies. In this context, the work conducted by Hu *et al.* [Hu et al., 2010] consists in incrementally updating both models and existing mappings as a synchronization process to reflect changes in the schema or ontology. They defined mappings as links between columns of relational tables and properties of concepts in ontology. They represented mappings as formulas relating tables in a schema with a subset of conjunctive formulas encoding a sub-tree in the ontology

graph (s-tree). They characterized the validity of mappings by these formulas, where they associated columns of the tables uniquely with attributes of the concepts in the s-tree. Therefore, a synchronization process occurs between relational tables and the s-tree according to changes.

This approach requests domain experts to declare the similarities between the old version and the new version of KOS, and to select the adaptation strategies. An addition of a column represents an example of the types of changes considered in schemas. In the ontology, they made the distinction between changes that can affect mappings, and those that cannot. To maintain mappings up-to-date, this solution requires adapting at least one of the associated models along with the mappings.

Change propagation

Existing approaches also emphasize the impact of KOS evolution to support the mapping adaptation. Tang & Tang [Tang and Tang, 2010] aimed to find a minimal sequence of changes (*i.e.*, addition and removal of axioms), exploiting the ABox and TBox of ontology, to perform evolution, *i.e.*, to control the impact of ontology change propagation on mappings for which in particular they only consider the removal type of change. Similarly, Chu *et al.* [Chu et al., 2008] proposed a dynamic ontology mapping architecture in the grid. Their approach aims to avoid the impact of KOS changes on mappings. They perform mapping maintenance through contracts with the expiration time from one ontology to another. The authors affirm that the contract for the mappings between a pair of resources guarantees that the resource specified in the ontology will remain unchanged until the expiration time.

Forecasting KOS evolution impact can minimize the difficulties in the task of mapping maintenance. However, the studied approaches insufficiently describe KOS changes affecting KOS's entities, which renders the propagation of changes to mappings yet more complex.

Mapping change strategies

To overcome the lack of change propagation approaches, Martins & Silva [Martins and Silva, 2009] proposed two approaches to adapt mappings impacted by KOS changes. Both approaches are based on the called *Semantic Bridge Ontology* (SBO) [Silva and Rocha, 2005] describing the semantic relations between ontologies entities. The authors assume that all mappings are instances of SBO and only removal change operations in ontology impact such instances. Therefore, for each mapping resolution point, the work proposes one or many elementary mapping adaptation strategies. In the first approach, the user defines these strategies from a list of corrective actions and then the system executes them, while in the second approach the system automatically selects the actions to perform. Although promising, such approaches lack flexibility in terms of KOS changes considered, mappings definition and the techniques of adaptation.

Recently, Groß*et al.* [Groß et al., 2013] proposed an algorithm to semi-automatically adapt mappings. In this proposal, we utilized a *diff* of KOS changes calculated between two ontology versions to determine the set of occurring (complex) change operations (the author of this thesis actively participated in this investigation). The proposal promotes the reuse of unaffected mappings, and maintains out-dated ones according to specific change handlers. This work preliminarily addresses the modification in the relation of mappings by combining, through initial rules, the semantic type of the old correspondences with the semantic relation assumed between

the involved concepts in change. In the adaptation process, the proposed algorithm assigns the different status for the correspondences such as handled or need verification by an expert in function of the adaptation strategy applied.

The mapping adaptation indicates a relevant approach to address mapping maintenance, but a fully automatic mapping adaptation process entails many issues uncovered by the existing approaches. The central drawback is that proposed studies mostly apply the same mapping adaptation strategy to a given type of KOS change operation, without considering further parameters associated to the change operations. This kind of approach makes the adaptation process inflexible and imprecise. In addition, uncovered aspects exist that may provide supplementary information to adapt mappings (*e.g.*, the context of a concept and its evolution in a refined way), which requires further investigations.

2.3 Mapping interpretation

After the creation of mappings, they might require further processing to keep them exploitable by humans and by automatic tools suited to precisely deal with them. As semantic mappings in biomedical domain represent the agreement reached between human specialists who devised them, no explicit and formal computational reference exists to establish the precise meaning of mappings [Cardillo et al., 2008].

Our conducted literature analysis on approaches to mapping maintenance shows that existing techniques to adapt mappings fail to fully explore information expressed by mappings. Explicitly understanding the meanings of established mappings can be useful to perform their maintenance. The approach developed in this thesis aims to further analyze existing mappings to determine information that can be useful for the adaptation. To this end, we survey the literature searching for methods and techniques that can perform mapping interpretation. Although literature has highlighted the relevance of further understanding the meaning of mappings [Cardillo et al., 2008], only very few researches have investigated this issue.

Literature has suggested automated analysis of mappings' meaning by studying how to make a logical encoding in *Web Ontology Language*[11] (OWL) [Antoniou and van Harmelen, 2004] two biomedical classifications and a formalization of existing mappings between them in terms of OWL axioms [Cardillo et al., 2008]. On this basis, they outline semantic techniques for automated mapping analysis via the verification of its coherence using logical reasoning. This technique relies on highly expressive KOSs that decreases their reusability to KOS without well defined semantics.

Other investigations mainly emphasize exploring mappings for some specific application purpose, but not explicitly analyzing them to derive some information. Our survey detected work dedicated to the reuse and enrichment of mappings, and to visualization by human experts.

In the biomedical domain, the *Unified Medical Language System*[12] (*UMLS*) project devel-

[11] www.w3.org/TR/owl-features
[12] `www.nlm.nih.gov/research/umls`

oped by the *U. S. National Library of Medicine*[13] (NLM) stands for a resource where mappings are largely explored to support specific tasks. UMLS refers to one of the most relevant projects addressing mappings between biomedical KOSs [Lindberg et al., 1993, Bodenreider, 2004]. The *UMLS* integrates KOSs in a unique thesaurus called Metathesaurus. This includes over 150 biomedical source vocabularies and contains more than two million of concepts. Nearly 46 million of relations exist among these concepts. Besides the Metathesaurus, a Semantic Network of 133 semantic types organized in a tree structure composes the *UMLS*. Each Metathesaurus concept contains a unique identifier (CUI) and is assigned at least to one semantic type from the Semantic Network.

Existing tools use the semantic network and CUI in *UMLS* to perform specific tasks relying on mappings. For instance, the *MetaMap* tools [McCray et al., 1994] aims to recognize concepts in text via *UMLS* mappings. Aiming to explore semantics in mappings with the purpose of enriching the set of mappings already established, Mougin *et al.* [Mougin et al., 2011] proposed to improve mappings between *MedDRA* and *SNOMED CT* vocabularies using *UMLS* through an automatic lexical-based approach. According to the authors, the proposed method provided several new and correct mappings, which resulted in a more complete alignment between *MedDRA* and *SNOMED CT*.

Emphasizing the reuse of mappings, existing methods aim at supporting new types of automatic inference across the different domains of biomedicine. Rosse *et al.* [Rosse et al., 2005] explored the re-use of existing mappings for the federation of interdependent ontologies. Similarly, Bodenreider [Bodenreider, 2009] reported the use of mappings for enabling interoperability with legacy repositories while Wennerberg [Wennerberg, 2009] proposed a method for aligning medical KOSs for the purpose of extracting clinical query.

In addition to support semi-automatic tools, the meaning of mappings can be used by human experts. This has stimulated researches on mapping explanation and uncovered aspects such as user interaction to handle KOS mappings [Falconer and Noy, 2011]. For example, Tang *et al.* provided a tool named *View graph* for visualizing mappings [Tang et al., 2009] retained the semantics discovered during the process of mapping calculation. Visualizing mappings based on View graph illustrates the semantics involved in mappings, and then can help understanding those mappings for maintaining them.

The investigated literature review shows that different tasks partially explore interpretation of mappings where mappings' meaning is somehow analyzed when reusing, enriching and visualizing mappings. However, existing work performs mappings interpretation in an indirect way where an external (human) agent or software might analyze mappings. This lacks explicitly determining KOS's entities that can explain the reasons for the existence of a correspondence. Very little research exists with dedicated methods to analyze and further interpret mappings in an explicit way. We argue that interpreting already established mappings may play a key role to support their maintenance. For example, an adequate mapping interpretation can generate mappings' meta-data that could provide support and improve the precision of systems performing automatic mapping adaptation.

[13] `www.nlm.nih.gov`

2.4 KOS evolution characterization

Changing requirements in domain knowledge demand updating the content of KOSs over time. After the appearance of Web ontologies, mainly boosted by the Semantic Web initiative [Berners-Lee et al., 2001], the research topic of KOS evolution [Stojanovic et al., 2002] has been widely studied. Various methodologies, approaches and frameworks have been proposed to (automatically) support and manage KOS evolution [Hartung et al., 2011]. Issues raised by this task have also been the source of various research efforts in the biomedical domain community, where KOS evolution remains still under investigation to deal with the increasing number of applications implementing KOSs.

General methodologies guiding the KOS evolution process have been developed. Stojanovic *et al.* [Stojanovic et al., 2002, Stojanovic, 2004] proposed a well accepted methodology through a six-step process consisting in different aspects of KOS changes including: (1) detection, (2) representation, (3) definition of its semantics, (4) implementation, (5) propagation and (6) validation. Literature shows representative contributions addressing specific aspects in each of these phases [Hartung et al., 2011, Flouris et al., 2008, Leenheer and Mens, 2008]. Also, different sources can be considered for KOS evolution, *e.g.*, text, [Sellami et al., 2013].

We observe that, as KOS evolution might influence the reliability of established mappings, we can explore the evolution of KOSs to facilitate and improve mapping adaptation. KOS changes might play a central role because they characterize KOS evolution. Different types of KOS changes affect the understanding of KOS evolution.

In the following, we survey the literature to describe the types of KOS change operations. We distinguish between traditional KOS changes affecting KOS's entities and KOS change patterns built on the top of traditional KOS change operations to improve the KOS evolution characterization and management. We aim at reviewing ways to characterize KOS evolution to reveal possible methods and tools that we could reuse in our context of mapping adaptation. In addition, we aim to examine whether existing types of KOS change patterns can somehow add benefits to support mapping adaptation.

2.4.1 Types of KOS change operations

Knowing changes affecting KOS might provide statements to enable understanding the meaning of the evolution and its effects [Klein and Noy, 2003, Flouris et al., 2008]. A **KOS change operation** (KCO) characterizes the evolution of a KOS in terms of operations applied to concepts, relationships and attributes [Hartung et al., 2013]. This may allow describing in a fine-grained level the way a KOS has changed from one version to another. Moreover, the formal representation of changes as well-defined operations provides the basis for storing them in a machine-understandable format, thus facilitating their propagation [Palma et al., 2009]

Klein & Noy [Klein and Noy, 2003] proposed a first categorization of change operations that can affect KOS to organize and represent the possibilities of change in KOS evolution. This first attempt, aiming at supporting KOS versioning, defined two main categories of changes: *atomic* and *complex*. The former refers to the elementary changes of only a single specific feature of the KOS model (*e.g.*, concepts, attributes or relationships), while the later denotes changes that are composed of multiple atomic ones (a mechanism for grouping atomic change operations

together to form a logical unit). Therefore, each operation in the former category cannot be divided into smaller operations, while each one of the latter category is composed of a combination of multiple atomic operations. For instance, the complex operation "*change attribute value*" is composed of two atomic operations "*delete attribute value*" and "*add attribute value*". Klein & Noy [Klein and Noy, 2003] explained that to define complex operations demands to find useful or common combinations of atomic operations.

We adopt the definitions of atomic and complex change operations from [Hartung et al., 2013] since these definitions do not diverge from existing others [Klein and Noy, 2003]. Table 2.1 and 2.2 present the formalization of atomic and complex KOS change operations, respectively. Our considered change operations stand for a subset of those proposed using meta-models of OWL ontology languages variants to produce a very extensive list of change operations. Klein & Noy [Klein and Noy, 2003] originally suggested an ontology of atomic change operations for the OWL knowledge model. Similarly, Palma *et al.* [Palma et al., 2009] proposed a change ontology for the OWL 2[14] claiming to provide a more fine-grained taxonomy of ontology changes, including atomic and complex changes. This work used a meta-model referring to the OWL 2 elements to develop the suggested change ontology.

Table 2.1: Atomic KOS change operations

	Change operation	**Description**
A	$addC(c)$	Addition of a new concept $c \in C(K_x^1)$
t	$delC(c)$	Deletion of an existing concept $c \in C(K_x^0)$
o	$addA(a, c)$	Addition of a new attribute a to a concept $c \in C(K_x^1)$
m	$delA(a, c)$	Deletion of an attribute a from a concept $c \in C(K_x^0)$
i	$addR(r, c_1, c_2)$	Addition of a new relationship r between two concepts $c_1, c_2 \in$
c		$C(K_x^1)$
	$delR(r, c_1, c_2)$	Deletion of an existing relationship r between two concepts $c_1, c_2 \in C(K_x^0)$

This table presents the description of the considered atomic changes [Hartung et al., 2013].

KOS versioning refers to an approach largely implemented into biomedical KOSs, which requires to automatically determine the changes between two given versions of a KOS to manage KOS evolution. To this end, existing methods and tools aim to identify the difference (*diff*) between KOS versions through *diff* calculation algorithms. The recurring issues in the *diff* computation problem refer to correctly detect the occurred atomic and complex changes only based on the two given KOS versions. Literature provides different techniques and tools for the *diff* calculation.

The most well know tools to calculate *diff* between KOSs include: (1) *PromptDiff*[15], which stands for one of the first *diff* tools [Noy and Musen, 2002] proposed as a plugin in the Protégé[16] platform; (2) Hartung *et al.* [Hartung et al., 2013] proposed the *COnto-Diff*[17] tool to

[14] www.w3.org/TR/owl2-overview
[15] `protege.stanford.edu/plugins/prompt/PromptDiff.html`
[16] `protege.stanford.edu`
[17] `dbserv2.informatik.uni-leipzig.de:8080/webdifftool`

Table 2.2: Complex KOS change operations

	Change operation	Description
	$chgA(c_k, a_i, v_h)$	Change of attribute a_i in concept $c_k \in C(K_x^0)$ with the new value v_h
	$moveC(c_k, p_1, p_2)$	Moving of concept c_k (and its subtree) from concept parent $p_1 \in$
C		$C(K_x^0)$ to concept parent $p_2 \in C(K_x^1)$
o	$substitute(c_k, c_h)$	Replacement of concept $c_k \in C(K_x^0)$ by concept $c_h \in C(K_x^1)$
m	$merge(\zeta_u, c_k)$	Fusion of a set of multiple concepts $\zeta_u \subset C(K_x^0)$ into concept $c_k \in$
p		$C(K_x^1)$
l	$split(c_k, \zeta_u)$	Split of concept $c_k \in C(K_x^0)$ into a set of resulting concepts $\zeta_u \subset$
e		$C(K_x^1)$
x	$toObsolete(c_k)$	Sets status of concept c_k to *obsolete* (c_k is no longer active)
	$delInnerC(c_k, p_l)$	Deletion of concept c_k from $C(K_x^0)$ where p_l refers to a *super-concept* of c_k and $p_l \in C(K_x^0)$
	$delLeafC(c_k, p_l)$	Deletion of leaf concept c_k from $C(K_x^0)$ where p_l refers to a *super-concept* of c_k
	$addInnerC(c_k, p_l)$	Addition of a sub concept c_k to $C(K_x^1)$ under the *super-concept* $p_l \in C(K_x^1)$
	$addLeafC(c_k, p_l)$	Addition of leaf concept c_k to $C(K_x^1)$ where c_k contains no *sub-concepts*
	$revokeObsolete(c_k)$	Revokes obsolete status of concept c_k (*i.e.*, c_k becomes active)

This table presents the description of the considered complex changes [Hartung et al., 2013].

automatically identify the *diff* between biomedical KOS versions; (3) the *OWLDiff* [18] tool [Kremen et al., 2011] focuses on very expressive ontologies specifically described in OWL language; (4) similarly, there is the *Ecco*[19] *diff* tool [Gonçalves et al., 2012].

To characterize the traditional atomic and complex change operations in KOS evolution, we adopted the *COnto-Diff* tool because this has shown useful for biomedical KOSs [Hartung et al., 2013].

2.4.2 KOS change patterns

Many fields have used the notion of patterns as possible "templates" or abstract descriptions encoding best practices. The notion of "pattern" has proved useful in design, as exemplified in software engineering, well known as *Design Patterns* [Gamma et al., 1995]. Recent initiatives have explored the use of patterns in KOS development, since patterns may assist a knowledge engineer in the construction process. In particular, Blomqvist *et al.* [Blomqvist et al., 2009] defined an ontology pattern as a set of ontological elements, structures or construction principles that intend to solve a specific recurring engineering problem. The *Ontology Design Pattern*[20] approach aims to use ontology patterns mainly for reutilization, and for promoting best practices.

Inspired by this perspective, the concept of patterns has also been adopted in KOS evolution. Usually more than one strategy may apply for a given change request, resulting in different ontolo-

[18] `krizik.felk.cvut.cz/km/owldiff`
[19] `rpc440.cs.man.ac.uk:8080/diff`
[20] `ontologydesignpatterns.org`

gies. To cope with different possibilities of applying changes, patterns introduce evolution strategies, which may define steps of complex evolution processes [Rießet al., 2010]. Therefore, change patterns may guide the process of evolution to characterize complex changes within evolution scenarios, simplifying the management of ontologies to control the impact of the evolution. This can ensure consistency and quality in ontology [Shaban-Nejad and Haarslev, 2009]. Moreover, change patterns can provide a way to determine what information to analyze or what constraints are necessary to identify a specific (complex) change. Castano *et al.* [Castano et al., 2007] defined change patterns as the capability of classifying the different situations which trigger KOS evolution by characterizing the results of the semantic interpretation process. According to them, change patterns also consist in a (formal) definition of an appropriate activity to correctly modify the ontology in each specific evolution situation.

Indeed, change patterns refer to more abstract tasks and operations than KOS change operations. They stand for a more high-level classification of changes taking different aspects into account than only KOS's entities, *e.g.*, domain specifities. Our literature review on KOS change patterns investigations points out that the change pattern notion has been appropriated by studies with different objectives. Therefore, we find a set of researches coping with change patterns indirectly connected one with the other. This impedes us determining a finite list of change pattern operations. Our thorough survey analysis allows us suggesting mainly four classifications of work for change patterns as follows:

- Improve complex change identification between KOS versions;
- Describe common change intentions
- Control consistency in KOS evolution
- Characterize domain-related changes

Improve complex change identification between KOS versions

Change patterns may allow identifying complex changes between different versions of the same ontology. Gröner *et al.* [Gröner et al., 2010] addressed the problem of refactoring recognition using reasoning to semantically compare different versions of an OWL DL ontology. To do so, they proposed a high-level categorization of ontology changes like the refactoring patterns in software engineering, and applying this to OWL ontology. A refactoring pattern refers to an abstract description of an ontology change to drive the reconstruction of part of the ontology [Gröner et al., 2010]. Gröner *et al.* applied a semantic comparison using heuristic algorithms to recognize such patterns between ontology versions.

Describe common change intentions

Some approaches define change patterns at the level of *Resource Description Framework*[21] (RDF) data model. Auer & Herre proposed to support KOS evolution by using basic changes and aggregate them into more complex changes in RDF [Auer and Herre, 2007]. Their approach consists in annotating the derived compound changes with meta-information and classifying them as evolution patterns to facilitate KOS evolution. This work classified change operations on

[21] www.w3.org/TR/REC-rdf-syntax

OWL ontologies according to specific patterns reflecting common change intentions. Rieß*et al.* [Rießet al., 2010] proposed a pattern-based approach to data evolution and refactoring RDF knowledge bases. Similarly to [Auer and Herre, 2007], their approach defined basic evolution patterns that can be combined into compound ones. Their work formally specifies modular evolution patterns in a declarative manner, capturing simple evolution and refactoring operations on both data and schema levels. The authors use RDF vocabulary for representing evolution patterns, and expressing pattern behaviours in the form of SPARQL[22] queries.

Control consistency in KOS evolution

Djedidi & Aufaure defined a KOS evolution methodology driven by a pattern-oriented modeling. They proposed the *Change Management Patterns* [Djedidi and Aufaure, 2009] to guide the KOS evolution process by driving and controlling change applications while maintaining consistency of evolving KOS [Djedidi and Aufaure, 2010]. They defined four kinds of consistency concerning the OWL DL language: structural, logical, conceptual and domain modeling consistency [Djedidi and Aufaure, 2010] applicable for ontologies. The solution looks for invariance in change management that repeatedly appears when ontologies evolve. They proposed three types of patterns: change patterns classifying types of changes, inconsistency patterns classifying types of logical inconsistencies, and alternative patterns classifying types of inconsistency resolution alternatives [Djedidi and Aufaure, 2009]. According to the authors, pattern-centered modeling aims to offer, for each of the three dimensions (change, inconsistency and resolution alternative), different levels of abstraction and to establish links between them [Djedidi and Aufaure, 2010]. They formally defined the patterns using OWL DL change operations.

Characterize domain-related changes

Javed *et al.* [Javed et al., 2013] suggested an approach to dealing with ontology evolution through a framework of compositional operators where they represent domain changes as change patterns. They composed different levels of change operators, and empirically studied ontology evolution to investigate the relationships between generic and domain-specific changes to determine common change patterns. The authors identified four levels of change operators and patterns based on the granularity of changes. This approach considers domain-specific change patterns as more general operations than complex changes. Javed *et al.* observed that ontology changes are driven by certain types of common, often frequent changes in the application domain, and argued that change patterns rely on the viewpoints and activities of knowledge engineers who implement KOS changes. Therefore, their approach allows users to define their own change patterns which can be executed many times.

Even though literature has proposed change patterns in an attempt to improve the characterization of KOS changes, it requires in-depth studies to evaluate the adequateness and usefulness of those for mapping adaptation.

[22] www.w3.org/TR/rdf-sparql-query

Summary

This chapter provided an overview of the literature concerning the state-of-the-art on mapping maintenance. We illustrated and formalized the mapping maintenance problem. Approaches addressing as automatic as possible this problem may play a key role in making this task easier, but literature lacked a coherent survey on existing techniques. We thoroughly surveyed related work of the mapping maintenance field through a systematic literature review classifying proposals in different categories of approaches. We analyzed existing investigations elucidating their advantages and drawbacks. Due to our approach exploring information from established mapping and KOS evolution, we provided a survey of investigations aiming to interpret mappings and characterize KOS evolution. We demonstrated the types of KOS changes and researches on KOS change patterns.

This study allowed us to point out open key issues which existing approaches to mapping maintenance have neglected, and to draw some major conclusions that will guide our sequence of study in this thesis. Our main observation relies on the fact that literature lacks a complete and well-studied solution that addresses mapping maintenance when KOS evolves.

Performing mapping maintenance through revisions stands for an incomplete process. This might help, but it remains insufficient since it is restricted to only removing inadequate correspondences. Moreover, existing techniques require logically expressive KOSs with well-defined semantics at a high level of formalization, making this approach unavailable for information systems that rely on semantic resources of low level of formalization such as nomenclatures, thesauri.

Even though mapping calculation for the maintenance task has suggested methods enabled to minimize the comparisons between KOS's entities and consequently reduce time in the matching process, this still remains very costly. The partial re-calculation avoids analysing both KOSs involved in mappings entirely, because this technique applies matching algorithms to perform a new alignment between changed concepts issued from source KOS and the whole target KOS. However, large KOSs like in the biomedical domain, so far represent a big challenge for methods relying on mapping calculation since time and efforts required to generate and evaluate mappings still remain very huge.

Mapping adaptation seems the most adequate approach that might provide more advantages to cope with mapping maintenance according to our findings. However, our results analysis indicates that this still lacks adequate methods to fully perform mapping adaptation in an automatic way according to KOS evolution. Existing approaches fail to accurately use information derived from KOS evolution and mappings, thus containing the following major drawbacks:

1. The semantics of established mappings are poorly interpreted to propose changes in mappings in the adaptation process. Our results reveal that literature lacks adequate methods suited to fully understand and explain correspondences by generating supplementary data on them. This might further support mapping adaptation techniques.

2. Mapping adaptation under evolving KOSs should consider the whole set of possible types of changes rather than only concept removal as the major investigations address. We need to further investigate which types of change patterns might concretely help to specify techniques of mapping adaptation and how to apply them. Existing methods for change

patterns definition seem insufficient because their design fails to consider requirements for adapting mappings. They were conceptualized taking the ontologies in an isolated way.

3. Techniques that explore a bigger set of changes in KOS evolution apply equal adaptation strategies according to each type of considered KOS change, thus lacking flexibility. These methods were not designed to adapt mappings in an individual manner, *i.e.*, analyzing and making adequate decisions for every single mapping.

These issues deserve further investigations to attain tools suited to really reduce the human work in the task of mapping maintenance. To this end, we understand that detailed empirical observations on the evolution of KOS and its impact on mappings will detect and justify the central elements to an automatic approach (subject of the next chapter). This might support the design of methods adequate to the KOSs which we are interested in. Moreover, this can allow elucidating requirements for the adaptation of mappings that can bridge the gap towards a fully automatic approach for mappings maintenance between heterogeneous and dynamic KOSs.

Chapter 3

Understanding mappings evolution

Contents

Introduction **41**
3.1 Experimental scenario **43**
3.1.1 Materials 43
3.1.2 Mapping evolution characterization 45
3.1.3 Experiments organization 46
3.2 Effects of KOS evolution on mapping changes **46**
3.2.1 Experimental procedure 46
3.2.2 General analyses 47
3.2.3 Specific analyses 50
3.3 Interdependencies between mapping change operations **56**
3.3.1 Experimental procedure 56
3.3.2 Results of the quantitative analysis 58
3.3.3 Results of the qualitative analysis 60
3.4 Mapping evolution under complex KOS changes **62**
3.4.1 Experimental procedure 62
3.4.2 ICD-9-CM cases 65
3.4.3 SNOMED-CT cases 68
3.5 Discussion **71**
Conclusion **73**

Introduction

Although significant research efforts have dealt with issues related to KOS evolution (*cf.* Section 2.4), understanding how this evolution affects dependent artefacts, such as mappings, has received very little attention. Search for potential impacts and interdependencies between types of KOS changes and possible changes in mappings can serve as the foundation of innovative techniques for mapping maintenance.

Recent work on KOS change impact analysis mainly concerns internal logical inconsistencies of ontology [Abgaz et al., 2012]. Even though existing researches investigate the evolution

of SCT [Gonçalves et al., 2011, Spackman, 2005], they fail to address the evolution's impact on established mappings. For instance, Gonçalves *et al.* [Gonçalves et al., 2011] analyze occured changes between two KOS versions and use SCT to show the applicability of their approach. Spackman [Spackman, 2005] investigates rates of change in large clinical terminologies using SCT as an object of study, although mappings are not taken into consideration. Closely related to our investigation, Groß*et al.* [Groß et al., 2012] explore how mappings in life science KOSs change. They empirically analyze which KOS changes can lead to the addition or deletion of correspondences between concepts. Using a computed dataset of mappings, the study preliminarily demonstrates how KOS changes can impact mappings and the ratio of changes in mappings according to general categories of KOS change types.

The findings of the aforementioned studies have motivated a new study to gain a more in-depth understanding of how a more fine-grained classification of KOS changes would affect mappings evolution, considering also real-world sets of mappings. It requires better precising KOS evolution impact on how mappings change to adequately propagate KOS changes on mappings, and thus to provide the appropriate use of changes for adapting mappings. Existing studies have analyzed mappings evolution by mainly considering changes affecting concepts of mappings in an isolated way, *i.e.*, without observing the influence of KOS's entities expressing statements that characterize concepts (*e.g.*, attributes) and of neighbour concepts. Therefore, we need to further investigate KOS changes, by observing and measuring in detail which information changes from one KOS version to another and how this influences associated mappings.

This chapter proposes to empirically examine the behaviour of mappings over time. We study changes in the evolution of official releases of SNOMED CT (SCT) and ICD-9-CM (ICD9) with associated and validated mappings established between them over a period of various years. We design a set of experiments with specific objectives highlighting different levels of detail to analyze the evolution of mappings. We expect to understand possible correlations between the evolution of KOS's entities related to mappings and how these mappings behave. In summary, we make the following contributions in this chapter:

- We demonstrate through experiments the relationships between changes affecting entities of biomedical KOSs and the evolution of associated mappings.
- We observe complex behaviours of mapping evolution investigating co-occurrence of different mapping changes.
- We explain the impact of specific complex changes occurring in KOS on mappings via the analysis of different KOS evolution cases.

We start by describing the used materials (*cf.* Section 3.1.1), and proposing means for categorizing mappings evolution to measure the impact of various types of changes in KOS's entities on categorized mappings change (*cf.* Section 3.1.2). We then present the organization of the conducted experiments (*cf.* Section 3.1.3). In the sequence, we present the analyses performed, taking the influence of changes in different KOS's entities into account (*e.g.*, concepts and their attributes and relationships) (*cf.* Section 3.2.3). This has motivated our investigation on interdependencies between mapping changes (*cf.* Section 3.3). Afterwards, we study mapping evolution under complex KOS changes (*cf.* Section 3.4). We provide a comprehensive discussion on our findings (*cf.* Section 3.5). The outcome from this investigation provides empirically justified elements to take into consideration into efficient approaches to keep mappings valid.

3.1 Experimental scenario

We first present the used materials (Section 3.1.1), followed by the designed procedure to characterize mapping evolution (Section 3.1.2). Section 3.1.3 shows the experiments organization.

3.1.1 Materials

The aim of our investigation requires datasets with specific characteristics. We look for biomedical KOSs and mappings considering the following aspects:

1. Availability of several versions of KOS;
2. Substantial number of concepts in each KOS;
3. Availability of several versions of official mappings between considered KOSs;

Experiments conducted in this chapter relies on SNOMED CT[23] (SCT) and ICD-9-CM[24] (ICD9).

SNOMED CT. The *Systematized Nomenclature of Medicine-Clinical Terms* (SNOMED CT) refers to a comprehensive health terminological resource managed by the *International Health Terminology Standards Development (IHTSDO)*[25]. IHTSDO has designed SCT to enable effective representation of clinical information in electronic health, and to cover the whole patient record. SCT also comprises body structures, procedures and relevant health-related aspects, including social context. The SCT model relies on three main entities: *Concepts, Descriptions* and *Relationships*. Concepts are identified by a unique identifier and contain a name and a status. Concepts are not physically removed, but their status can be set to active or inactive. Each concept may be associated with a set of descriptions and relationships. Descriptions refer to special attributes independent of concepts that compose sets of terms that textually describe concepts. Their type denotes either a *preferred term* or a *synonym*, and their associated status follows the same categorization as in the case of concepts. As for Concepts, Descriptions are never removed, but their status can change.

ICD-9-CM. The *International Classification of Diseases, Ninth Revision, Clinical Modification* (ICD-9-CM) refers to the official system of assigning codes to diagnoses and procedures associated with hospital utilization mostly in the United States. ICD-9-CM relies on the *World Health Organization's Ninth Revision, International Classification of Diseases* (ICD-9). ICD-9-CM was originally created for epidemiological purposes, but now it constitutes the most widely used disease encoding system and is globally used as a common basis for health statistics. The U.S. governmental agencies *Centers for Medicare and Medicad Services*[26] and the *National Center for Health Statistics* (NCHS)[27] manage the maintenance of ICD-9-CM. This KOS consists of a tabular list containing a numerical list of the disease code numbers. The ICD-9-CM model is composed of a pre-defined hierarchical structure including *Chapters, Blocks* and *Codes*. Chapters are the most general level of organization. Blocks always belong to a Chapter, and Codes are

[23] www.nlm.nih.gov/research/umls/licensedcontent/snomedctarchive.html
[24] www.cdc.gov/nchs/icd/icd9cm.htm
[25] www.ihtsdo.org
[26] www.cms.gov
[27] www.cdc.gov/nchs

identified by a unique numerical identifier belonging to a unique Block and Chapter. No explicit relationships exist between concepts from ICD9, and mappings are always interrelated with the Codes level [Freitas and Schulz, 2009].

We use SCT as the source KOS (K_S) and ICD9 as the target (K_T). According to the adopted definition of mapping, the source concept stands for a concept in the SCT while the target concept refers to a concept belonging to ICD9. For instance, concept code '*123714004*' with the name "*Ventricular septal defect, spontaneous closure (disorder)*" in SCT is mapped to concept code '*745.4*' with the name "*Ventricular septal defect*" in ICD9. This mapping has the relation type *more specific than* ($\leq$), *i.e.*, the concept in SCT is considered semantically less general than the concept in ICD9. For the experiments in this chapter, we selected six different KOS versions (*i.e.*, four different versions of SCT and two different versions of ICD9). We also selected four official versions of mappings (provided by IHTSDO) between these KOSs (*cf.* Figure 3.1).

Let $\mathcal{M}^j_{ST}$ represent the set of all existing mappings between two different KOSs at time j. The released versions of SCT used to determine each $\mathcal{M}^j_{ST}$ are: Jan/2010 for $\mathcal{M}^1_{ST}$, Jul/2010 for $\mathcal{M}^2_{ST}$, Jan/2011 for $\mathcal{M}^3_{ST}$ and Jul/2011 for $\mathcal{M}^4_{ST}$. Table 3.1 presents some statistics concerning the number of mappings and concepts in the involved KOSs for the different releases.

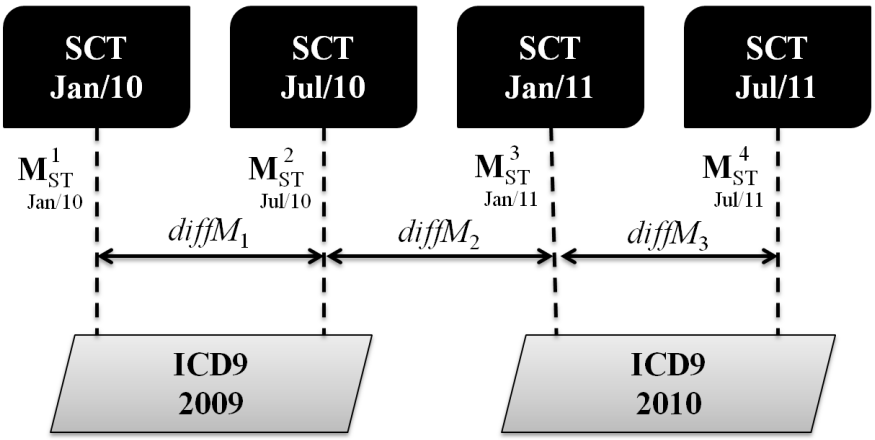


Figure. 3.1: Experiments configuration

This figure shows the different releases of SNOMED CT and ICD-9-CM along with the set of mappings used in our experiments.

Table 3.1: Studied biomedical KOSs

Mappings		$\boldsymbol{K_S}$		$\boldsymbol{K_T}$	
dataset	#mappings	release	#concepts	release	#concepts
$\mathcal{M}^1_{ST}$	100 451	SCT'10-Jan	390 022	ICD9'09	12 734
$\mathcal{M}^2_{ST}$	100 820	SCT'10-Jul	391 170	ICD9'09	12 734
$\mathcal{M}^3_{ST}$	101 215	SCT'11-Jan	393 072	ICD9'10	12 879
$\mathcal{M}^4_{ST}$	102 550	SCT'11-Jul	395 033	ICD9'10	12 879

This table presents statistics on numbers of mappings and concepts in the used KOSs.

3.1.2 Mapping evolution characterization

We establish a system of categorizing mapping evolution to conduct the experiments. The categories correspond to "**mapping change operations**" (MCOs) which lead to mappings evolution from one released version to another. MCOs differ from KOS change operations because they only refer to mappings. For instance, one element of the mapping can change (*e.g.*, the target concept was modified from concept c_1 to c_2, such that these concepts have different identifiers) without implying that the content (*e.g.*, attributes) of these concepts changed in KOS. We aim to identify possible interdependencies between MCOs and KOS change operations, which affect the source concept of mappings. Analyzing two subsequent versions of a set of mappings makes it possible to identify a subset of mappings which differ from one version to another. Given $\mathcal{M}_{ST}^{j}$ and $\mathcal{M}_{ST}^{j+1}$, two successive sets of released mappings (*cf.* Figure 3.1), we define $diffM_j$, the categorization of mappings, according to the following steps:

1. **Unchanged (UNH)**. The first step aims to identify the unchanged mappings. These mappings preserve their values from one version to another. A given mapping m_{st} refers to an unchanged mapping if $m_{st} \in \mathcal{M}_{ST}^{j}$ and $m_{st} \in \mathcal{M}_{ST}^{j+1}$.

2. **Addition and Removal.** The second step identifies all added and deleted mappings by comparing $\mathcal{M}_{ST}^{j}$ and $\mathcal{M}_{ST}^{j+1}$. For instance, if a mapping is added or deleted to/from a new release of mappings, then this mapping is included in the respective $diffM_j$. Given a mapping m_{st}, we search for mappings that were added or removed as follows:

   - **Addition (ADD)**: A given mapping m_{st} is added if $m_{st} \notin \mathcal{M}_{ST}^{j}$ and $m_{st} \in \mathcal{M}_{ST}^{j+1}$;
   - **Removal (REM)**: A given mapping m_{st} is removed if $m_{st} \in \mathcal{M}_{ST}^{j}$ and $m_{st} \notin \mathcal{M}_{ST}^{j+1}$;

3. **Modification.** We identify the modified mappings in the third step. They stand for special cases of Addition/Removal operations and are characterized by changes in their internal elements. Consider the following mappings:

   - $m_{st0} = (c_s, c_t, semType_{st})$;
   - $m_{st1} = (c_s, c_{t1}, semType_{st})$;
   - $m_{st2} = (c_s, c_t, semType_{st1})$;
   - $m_{st3} = (c_s, c_{t2}, semType_{st2})$;

   such that $c_t \neq c_{t1}$, $c_t \neq c_{t2}$ and $semType_{st} \neq semType_{st1}$ and $semType_{st} \neq semType_{st2}$, we search for modified mappings as follows:

   - **Modification in target concept (MOD_t)**: If ($m_{st0} \in \mathcal{M}_{ST}^{j}$ and $m_{st0} \notin \mathcal{M}_{ST}^{j+1}$) and ($m_{st1} \notin \mathcal{M}_{ST}^{j}$ and $m_{st1} \in \mathcal{M}_{ST}^{j+1}$) are fulfilled;
   - **Modification in type of semantic relation (MOD_r)**: If ($m_{st0} \in \mathcal{M}_{ST}^{j}$ and $m_{st0} \notin \mathcal{M}_{ST}^{j+1}$) and ($m_{st2} \notin \mathcal{M}_{ST}^{j}$ and $m_{st2} \in \mathcal{M}_{ST}^{j+1}$) are fulfilled;
   - **Modification in target concept and semantic relation (MOD_{t_r})**: If ($m_{st0} \in \mathcal{M}_{ST}^{j}$ and $m_{st0} \notin \mathcal{M}_{ST}^{j+1}$) and ($m_{st3} \notin \mathcal{M}_{ST}^{j}$ and $m_{st3} \in \mathcal{M}_{ST}^{j+1}$) are fulfilled;

4. **Isolation**. The fourth step consists in selecting the changed mappings affected by the evolution of only one KOS. For the experiments, we selected only mappings whose target concept was unchanged over two successive releases of mappings. The modified concepts

are identified by calculating a simple difference between all concepts of two successive versions of the target KOS (*diffC*). We evaluate one KOS at a time to provide a conclusive analysis from the set of metrics that we calculate in the experiments. Cases where both KOSs evolve at the same time will require a qualitative analysis supervised by domain experts to explain the impact of each changed concept (in the source and in the target) on the mappings. In reality, the organisms managing KOSs publish new releases of KOSs in different periods of time, but mappings are generated only once a year (in the case of SCT and ICD9) several months after the KOS evolution.

3.1.3 Experiments organization

We organize the experiments according to three objectives whose nature differs:

1. We examine the impact that KOS evolution (observing changes of different KOS's entities) has on established mappings (Section 3.2). We characterize such impact as mappings evolution via the different types of designed MCOs that are independently investigated (*i.e.*, one by one). We assume that a mapping is only influenced by changes in the source concept and in its related KOS's entities (attributes, relationships and hierarchical concepts – *i.e.* concepts that are directly linked to the source concept such as *super-concepts* and *sub-concepts*).

2. We investigate possible interdependencies between different MCOs observed in concepts sharing structural properties (parent and sibling concepts)(Section 3.3). This analysis will be useful to detect more complex MCOs.

3. We study the behaviour of mappings evolution under assorted cases of complex KOS changes (Section 3.4). We devote special attention to the split type of change due to the difficulties for adapting mappings in this context.

These experiments address the following research questions:

1. How to determine the impact of KOS evolution on mappings?

2. How to fully take KOS changes into account for performing mapping adaptation?

3. To which extent is it useful to take different types and patterns of KOS changes into consideration?

3.2 Effects of KOS evolution on mapping changes

We first present the experimental procedure conducted (Section 3.2.1). Thereafter, we present the empirical results achieved for the general (Section 3.2.2) and specific (Section 3.2.3) analyses.

3.2.1 Experimental procedure

A general view of mapping behaviour can allow the identification of entities that need deeper investigations. We applied this principle to the experiments and presented in tables 3.2 to 3.4 (*general analyses*). The *specific analyses* were performed to better understand the impact of each entity on the MCOs.

General analyses. They focus on mappings only (*cf.* Section 3.2.2). The experiments provide an overview of the change operations performed in the mappings without considering the potential impacts of KOS changes. We grouped these experiments into two types:

1. Basic analysis of explored mappings presents the absolute values of mappings (Table 3.2) and the relative values of changes in mappings in the MCOs categories (Table 3.3). We can identify the most frequent operations and whether operations exist that in fact never happen, to exclude these from the rest of the study.

2. Analysis of mapping relations focuses on the modifications of types of semantic relations expressed by mappings. We present the relative values of changes in the mappings according to the type of relation (Table 3.4).

Specific analyses. We present the influence of KOS evolution on mapping evolution (*cf.* Section 3.2.3). This differs from the previous analyses by taking into account the various potential types of changes in each entity of the source KOS (*e.g.*, in concepts, relationships, *etc.*). We aim to observe whether changes in KOS's entities have a direct impact on the evolution of mappings and eliminate those that have low or no impact. We started with a general view of changes in KOS's entities with regard to MCOs. This highlights the KOS' most relevant entities, which we will study in depth. Then, we analyze each relevant entity separately.

For instance, we observe the behaviours of mappings when a concept is deleted from KOS or when an attribute is added to a concept. We expect to identify possible correlations between KOS changes and mappings evolution. To do so, we compare the MCOs with the changes in the source KOS. We consider only concepts belonging to the source KOS (*i.e.*, the set of source concepts of mappings) and the set of concepts explored in the experiments refers a subset of the SCT concepts. We consider the influence of *Concepts* (only those that have at least one mapping associated), *Descriptions*, *Relationships* and *Neighbourhood*.

We select KOS's entities to evaluate potential impact on mappings change, not only due to a source concept change, but possibly to a change in different KOS' entities, *e.g.*, a connected relationship. Moreover, the different types of change related to the KOS's entities are taken into consideration. We individually investigated each type of change in KOS to thoroughly evaluate its potential impact on mappings.

3.2.2 General analyses

The following experiments aim to provide an initial overview of mapping evolution. We started with a basic analysis of the categorized mappings and then conducted an analysis of mapping relations.

Analysis of explored mappings

We perform this analysis to gain an overview of the dataset and to evaluate whether mappings creators really need automatic methods and tools to support mappings maintenance activities. The absolute and relative quantity of the categorized mappings are presented in table 3.2 and table 3.3. The MCOs considered are: *Unchanged* (UNH), *Addition* (ADD), *Removal* (REM), *Modification in the target concept* (MOD_t), *Modification in type of semantic relation* (MOD_r)

and *Modification in both target concept and semantic relation* (MOD_{t_r}). This investigation analysed four sets of mappings. From each couple of subsequent mapping versions, the sets of $diffM_i$ were calculated according to the steps described in Section 3.1.2.

Table 3.2 presents total changed and unchanged mappings over $diffM_i$: $diffM_1$ with 100 875 mappings, $diffM_2$ with 101 254 mappings, and $diffM_3$ with 102 601 mappings. Changed mappings are all those affected by the MCOs: *ADD, REM*, MOD_t, MOD_r and MOD_{t_r}. Although unchanged mappings represent 99.10% of the total quantity of mappings, the quantity of mappings that require changes is still substantial, and automatic tools that reduce this laborious human task are fully justified. The obtained results show expressive growth in the quantity of mappings changed during the analysed period. This growth tends to continue with the increasing size of SCT and ICD9.

Table 3.2: Analyzed mappings between biomedical KOS

diffM	**#Total**	**#Unchanged**	**#Changed**
$\boldsymbol{diffM_1}$	100 875	100 394	481
$\boldsymbol{diffM_2}$	101 254	100 281	973
$\boldsymbol{diffM_3}$	102 601	101 076	1525

This table presents the number of changed and unchanged mappings.

Table 3.3 shows the proportions of changed mappings to each mapping change operation. This refers to the ratio between the quantities of mappings categorized with a determined mapping change operation divided by total changed mappings in $diffM_i$ (*cf.* Table 3.2). We first observed modifications in $\mathcal{M}^3_{ST}$ (and included in $diffM_2$), where the three types of modifications (MOD_t, MOD_r and MOD_{t_r}) account for 51% of the mappings in $diffM_2$. One potential explanation can be that $diffM_2$ was calculated after the evolution of ICD9 and SCT. Even applying our procedure to exclude mappings related to concepts that changed in the target KOS, it is possible to have an indirect impact of these KOS changes into the other mappings. For instance, if a *sub-concept* is added to ICD9, the relation of an established mapping can change.

Table 3.3: Proportion of mapping change operations

diffM	**Unit**	**ADD**	**REM**	$\boldsymbol{MOD_t}$	$\boldsymbol{MOD_r}$	$\boldsymbol{MOD_{t\_r}}$
$\boldsymbol{diffM_1}$	%	88.4	11.6	0	0	0
	#	425	56	0	0	0
$\boldsymbol{diffM_2}$	%	44.8	4.2	21	21.4	8.6
	#	436	41	204	208	84
$\boldsymbol{diffM_3}$	%	91.1	3.6	1.4	2.0	1.9
	#	1 389	55	22	30	29

This table shows the absolute numbers and the percentage of mappings for the different mapping change operations.

Analysis of types of mapping relations

We study the evolution of the semantic relations of mappings, *i.e.*, we analyze the MCOs according to the nature of the semantic relation between the source and target concepts. We expect to observe whether the evolution of KOSs tend to increase their "harmony" (*i.e.*, increase of equivalent mappings) or their specialization (*i.e.*, predominance of relation of type $\leq$). With the analysis of relations, we also seek to identify situations in which the relations were impacted by the KOS evolution.

Table 3.4 shows, for each MCO, the frequency of the relations used by the mappings in $diffM_i$. The numbers in Table 3.4 represent the quantity of mappings with a specific relation for one operation divided by the total of mappings from $diffM_i$. For instance, the number 17.55 means that 17.55% of the unchanged mappings from $diffM_1$ have the relation type $\perp$ (*unmappable*). The outcome shows that the number of mappings of type $\perp$ decreased significantly for mappings added during the analysed period, and the number of mappings of this same relation type increased for removed mappings. This can reveal a convergence between the KOSs. However, it does not indicate that both KOSs evolve in the same way. Observing that existing mappings tend not to change when SCT is less general than ICD9 ($\sim$ 60% unchanged), and that mappings with $\leq$ semantic type are most frequently added than removed, while those with relation type $\equiv$ are more frequently removed than added, we can potentially conclude that SCT evolve faster and tend to be more specialized than ICD9.

Table 3.4: Relations in mappings with regards to mapping change operations

MCO \ Relation	diffM	#diffM	$\perp$ (%)	$\equiv$ (%)	$\leq$ (%)	$\geq$ (%)	$\approx$ (%)
Unchanged	$diffM_1$	100 394	17.55	16.82	62.64	0.67	2.32
	$diffM_2$	100 281	17.73	17.13	62.09	0.73	2.32
	$diffM_3$	101 076	18.29	16.76	62.49	0.58	2.20
Addition	$diffM_1$	425	9.41	6.11	83.76	0.72	0
	$diffM_2$	436	4.35	26.62	68.57	0.23	0.23
	$diffM_3$	1389	3.08	5.16	91.50	0.13	0.07
Removal	$diffM_1$	56	8.93	41.07	50	0	0
	$diffM_2$	41	7.32	39.03	51.22	2.43	0
	$diffM_3$	55	32.72	25.45	41.81	0	0
MOD_t when *semType* is	$diffM_2$	204	0	0	100	0	0
	$diffM_3$	22	0	15	85	0	0
MOD_r to	$diffM_2$	208	0	90.45	8.55	1	0
	$diffM_3$	30	0	83.87	16.13	0	0
MOD_t when MOD_r to	$diffM_2$	84	1.19	70.23	26	2.4	0
	$diffM_3$	29	0	86.20	13.8	0	0

This table presents the proportion of different types of relation observed in mappings of different mapping change operations.

3.2.3 Specific analyses

We aim to evaluate how KOS changes impact associated mappings. In the following experiments we started by observing the influence of KOS entity changes on mappings. Note that we only determined if KOS entities changed, not how. Based on the obtained results (Table 3.5), we conducted further experiments to measure in greater detail the impact of different aspects of changes in KOS's entities. The experiments considered the influence of different types of changes affecting the source concept: a change in attributes or the addition of a concept (Table 3.6), a change of status (Table 3.7), a change in descriptions (Table 3.8), and finally a change in relationships (Table 3.9).

Analysis of the KOS's entities

In this analysis we considered changes affecting one of the source KOS's entities (*cf.* Table 3.5). Each column corresponds to an entity (*Concept*, *Descriptions*, *Relationships* and *Neighbours*). The analysis of changes in source concepts and neighbour concepts does not include changes in their descriptions, which are considered distinct entities. In table 3.5, the numbers represent a percentage of total mappings for each MCO. For instance, 0.06% represents the ratio of all unchanged mappings in which the source concept has changed (*e.g.*, in the label). In other words, in 1.71% of the unchanged mappings some neighbour concept of the source concept changed. Note that one concept can have several changes at the same time, *e.g.*, a concept can be "*deleted*" and also has its relationships deleted. In this case, the same concept will be counted in two different columns (*e.g.*, *Concept* and *Relationship*). To better evaluate the impact of changes in each type of KOS entity, we decided to not normalize the values. The sum of the relative values is therefore not necessarily 100%.

The results show that changes in KOS's entities may have no effect on mappings modifications (rows MOD_t, MOD_r and MOD_{t_r} in table 3.5). The small rate observed stands for exceptional cases, which are rare and require a case-by-case analysis. This can be attributable, for instance, to minor corrections in the text (*e.g.*, adding an "*s*" to the end of a name or a capital letter) in the cases of *Concepts* and *Descriptions*, or the addition of a new relationship (implying a new neighbour).

Modifications in the semantic relation of mappings between KOSs are performed without observing changes in KOS's entities. For instance, according to table 3.3, in $diffM_3$, 2% of mappings change had their relation modified, but in table 3.5 (*cf.* row MOD_r in $diffM_3$) these modifications were not associated with changes in KOS. The results obtained so far did not allow any precise explanation about these observations. We can imagine that improvements in mappings from one release to another are potentially the origin of mapping relation modifications.

On the other hand, table 3.5 clearly shows that adding or removing mappings is frequently associated with changes in the source concepts, as well as in concepts' descriptions and relationships. For example, in the row *Addition* and $diffM_1$, all concepts from the source of mappings changed somehow. This fact drove us to deepen the analysis for each type of change in the source concept.

Table 3.5: Changes in SCT entities correlated with mapping change operations

MCO \ Entities	diffM	**Concept** (%)	**Descriptions** (%)	**Relat.** (%)	**Neigh.** (%)
Unchanged	$diffM_1$	0.06	0.25	6.75	1.71
	$diffM_2$	0.02	0.12	3	1.17
	$diffM_3$	0.03	0.11	3.34	1.38
Addition	$diffM_1$	100	100	100	12.93
	$diffM_2$	100	100	100	20.87
	$diffM_3$	66.18	66.18	66.5	28
Removal	$diffM_1$	98.22	98.18	100	39.2
	$diffM_2$	100	100	100	43.9
	$diffM_3$	83.64	83.64	100	47.27
MOD_t	$diffM_2$	1.4	1.4	4.7	0.47
	$diffM_3$	0	0	10	5
MOD_r	$diffM_2$	0.5	0.5	10	3.51
	$diffM_3$	0	0	0	0
MOD_{t_r}	$diffM_2$	1.2	1.2	4.7	2.38
	$diffM_3$	0	0	3.45	3.45

This table shows the percentage of mappings for the mapping change operations where the source concept and related entities are affected by KOS changes.

The influence of changes in concepts

These experiments aim to highlight relationships between different types of changes affecting the source concept and the MCOs. For this purpose, we provide two different analyses. First, we investigate the types of change in concepts according to the taxonomy of concept changes proposed in SCT (Table 3.6). Second, we focus on status change as a specific type of change in concepts (Table 3.7).

The types of change analysed, as shown in table 3.6 are: *Addition, Status change, Minor change and Unchanged*. *Addition* represents a new concept added to KOS; *Status change* means that the property status of the concept has changed; *Minor change* represents that the concept was affected by a small change in its name. Changes in label occur when the text used to name the concept is modified. In this analysis, we also include the type *Unchanged* (last column in table 3.6) to represent a source concept that has not been changed at all.

Table 3.6 shows the results obtained. The numbers in this table represent the ratio of MCOs that are associated with one specific type of concept change. For instance, 98.82% of added mappings are associated with a new concept added to KOS.

A strong correlation exists between removed mappings and changes in concepts' status. Moreover, a mapping is normally added after a source concept is added (in at least 99.54% of the cases). If, from one side, these results show a strong correlation between the addition of concepts and the additions of mappings, from another side, they did not explain why MOD_t, MOD_r and

Table 3.6: Types of changes in concepts with regard to mapping change operations

MCO \ Concept	diffM	**Addition** (%)	**Status** (%)	**Minor change** (%)	**Unchanged** (%)
Unchanged	$diffM_1$	0	0	0.06	99.94
	$diffM_2$	0	0	0.02	99.98
	$diffM_3$	0	0	0.03	99.97
Addition	$diffM_1$	98.82	0.94	0.24	0
	$diffM_2$	99.54	0.46	0	0
	$diffM_3$	66.03	0.15	0	33.82
Removal	$diffM_1$	0	98.22	0	1.78
	$diffM_2$	0	100	0	0
	$diffM_3$	0	83.64	0	16.36
MOD_t	$diffM_2$	0	0	1.4	98.6
	$diffM_3$	0	0	0	100
MOD_r	$diffM_2$	0	0.5	0	99.5
	$diffM_3$	0	0	0	100
MOD_{t_r}	$diffM_2$	0	0	1.2	98.8
	$diffM_3$	0	0	0	100

This table presents the percentage of source concepts affected by different types of KOS change for the mapping change operations.

MOD_{t_r} MCOs are (in most cases) not correlated to changes in source concepts. This requires a case-by-case study.

To show possible relations between MCOs and the status of the source concepts, we compare the value of the status of each concept before and after KOS evolution. We aim at identifying changes in the status of concepts that may indicate the correction of a mistake in the previous version of the KOS (*e.g.*, ambiguous, *etc.*). We search for possible influences that a status change in the source concept can have on changes in mappings. Following are the types of status that can be assigned to a concept in SCT: "$non-existent$" (-1), "*current*" (0), "*inactive with no reason given*" (1), "*duplicated*" (2), "*outdated*" (3), "*ambiguous*" (4), "*erroneous*" (5), "*limited*" (6). The *non-existent* status accommodates those concepts that are added in a later KOS version. The status from 1 to 6 compose the inactive group, and only the reason for the inactivation differentiates them. As the removal of a concept does not exist in SCT, we considered that all inactive status correspond to a concept removal operation. For the purposes of displaying the data, the numbers corresponding to the status are used in table 3.7. This table illustrates only status whose value is not always 0%. For instance, for all concepts analysed before evolution, status 4 (ambiguous) was not found and therefore table 3.7 does not show it. The numbers in this table represent a percentage of total source concepts of each type of MCO that have a given status. For instance, 98.82% of added mappings are associated with concepts whose status before evolution was *non-existent* (-1), whereas 100% of added mappings are associated with active concepts (*i.e.*, after KOS evolution, these mappings' source concepts have the current status).

Considering unchanged mappings, the status of source concepts remains the same before and

Table 3.7: Concepts' status changes related to mapping change operations

MCO \ Status	diffM	Before evolution (%)				After evolution (%)				
		-1	**0**	**1**	**2**	**0**	**1**	**2**	**4**	**5**
Unchanged	$diffM_1$	0	99.80	0.08	0.12	99.80	0.08	0.12	0	0
	$diffM_2$	0	99.78	0.14	0.08	99.78	0.14	0.08	0	0
	$diffM_3$	0	99.85	0.07	0.06	99.85	0.07	0.06	0	0
Addition	$diffM_1$	98.82	0.24	0.24	0.7	100	0	0	0	0
	$diffM_2$	99.54	0	0	0.46	100	0	0	0	0
	$diffM_3$	66.03	33.67	0	0.15	100	0	0	0	0
Removal	$diffM_1$	0	100	0	0	1.78	0	62.6	35.71	0
	$diffM_2$	0	100	0	0	0	0	73.17	26.83	0
	$diffM_3$	0	100	0	0	16.36	9.1	40	27.27	7.27
MOD_t	$diffM_2$	0	100	0	0	100	0	0	0	0
	$diffM_3$	0	100	0	0	100	0	0	0	0
MOD_r	$diffM_2$	0	99.5	0	0.5	100	0	0	0	0
	$diffM_3$	0	100	0	0	100	0	0	0	0
MOD_{t_r}	$diffM_2$	0	100	0	0	100	0	0	0	0
	$diffM_3$	0	100	0	0	100	0	0	0	0

This table shows the percentage of different status for source concepts involved in mappings within the mapping change operations. The status include: "*non – existent*" (-1), "*current*" (0), "*inactive with no reason given*" (1), "*duplicated*" (2), "*ambiguous*" (4) and "*erroneous*" (5).

after evolution, confirming that a strong correlation exists between the changes in the concept status and changes in the mappings. As expected, an addition of mappings is always done between active concepts. However, they are not limited to new concepts. In the row *Addition* and $diffM_3$, it can be observed that 33.67% of the source concepts underwent no status change. We draw equivalent conclusions from mappings removal. Mappings are always removed from active concepts, but the concepts are not always assigned as inactive in the next KOS version.

The impact of changes on concept status is inexpressive (or null) for the three MCOs regarding modification of mappings: MOD_t, MOD_r and MOD_{t_r}.

Analysis of concepts' descriptions

We analyze concepts' descriptions separately to examine correlations between MCOs and changes in the descriptions of the source concept. We assume that concepts' attributes play a central role for establishing semantic correspondences between concepts, and we aim to measure this relevance by verifying whether mappings are sensible to changes in attributes. For this purpose, we analyze *Addition of descriptions* (when a new description is added, with an active or inactive state), *Change of status* (to active state or to inactive state) and *Unchanged*. The numbers in Table 3.8 represent the ratio of total mappings of each MCO for which the source concepts' description underwent a specific change. For instance, 99.05% of added mappings have at least one newly added active description. In the mappings analysed, we did not observe any addition of inactive descriptions; therefore, we decided to exclude the "*inactive*" table column for the added descriptions.

Table 3.8: Concepts' descriptions changes related with mapping change operations

MCO \ Description	diffM	Addition (%)	Change of status (%)		Unchanged (%)
		Active	Active	Inactive	
Unchanged	$diffM_1$	0.13	0.09	0.14	99.92
	$diffM_2$	0.04	0.03	0.09	99.96
	$diffM_3$	0.09	0.01	0.06	99.97
Addition	$diffM_1$	99.05	0.94	0.23	1.17
	$diffM_2$	99.77	0.46	0	0
	$diffM_3$	66.18	0.15	0	33.81
Removal	$diffM_1$	0	0	98.2	23.21
	$diffM_2$	0	0	100	19.51
	$diffM_3$	0	0	86.63	20
MOD_t	$diffM_2$	1.5	1.5	0	98.5
	$diffM_3$	0	0	0	100
MOD_r	$diffM_2$	0	0	0.5	100
	$diffM_3$	0	0	0	100
MOD_{t_r}	$diffM_2$	1.2	0	1.2	98.8
	$diffM_3$	0	0	0	100

This table presents the percentage of source concepts where different changes affect their descriptions for the mapping change operations.

A concept can have a set of descriptions. For example, "*Excision of gallbladder*" and "*Gallbladder excision*" correspond to descriptions of the concept id '*38102005*' with name "*Cholecystectomy*". Each description is considered independently from the others. Therefore, we might find situations where more than one type of change occurs in the same description set at the same time. This explains why the sum of the numbers in the rows of table 3.8 remains not necessarily 100%. For example, in the row *Removal* and $diffM_1$, 98.2% of the source concepts of the removed mappings have at least one description that changed. In the same row, but in the next column (*Unchanged*%), we see that 23.21% of removed mappings have at least one description that did not change. These two columns are not complementary because one concept can have several associated descriptions, and there might exist cases in which a subset of descriptions has changed while another subset has not.

In unchanged mappings, we observe very few cases in which changes in the descriptions occur. Added mappings are frequently related to the addition of descriptions. A correlation exists between additions of descriptions and addition of mappings. Since it is impossible to remove descriptions in SCT, when a description needs to be updated, one possibility is to set its status to inactive and create a new description. These two operations may explain why some descriptions changed to inactive (0.23%) when mappings were added. Furthermore, the results indicate a strong correlation between removal of mappings and changes in the status of the descriptions to inactive (94.94% on average).

Influence of changes in concepts' relationships

We analyze whether information collected from a KOS's structure (relationships) can explain how mappings evolve. This experiment aims to investigate how changes affecting the relationships connected to the source concept of the mappings analysed can have an influence on how these mappings change.

We consider three types of relationships between the source concept and another concept belonging to SCT (*i.e.*, within the same KOS): (1) "*Super*" relationships, those of type "$is-a$" that link the source concept with its "*super concept*(s)" such that the source concept refers to the domain in the relationships; "*Sub*" relationships, the inverse of the previous one; and "*Other*" relationships representing all other types of relationships other than "$is-a$" in which the source concept of mappings stands for the domain or range of the relationship.

We take three types of changes in concepts' relationships (*Addition*, *Removal* and *Unchanged*) into consideration, each represented by a column in table 3.9. Nevertheless, unlike for concepts and descriptions in SCT, no different status for relationships exists. As a consequence, relationships are added or removed, and we consider the removal of relationships to be a particular type of change. Moreover, SCT contains various "*concepts of control*", which have an impact on the analysis of changes in relationships. They are used to assess relationships according to changes in concepts. For instance, when a concept is set as inactive, the set of relationships associated with this concept is removed, and other relationships to the "*control*" concept are created to avoid anomalies in the structures of SCT. For this reason, when analysing the values in table 3.9, we might consider that the removal of a mapping triggered by the inactivation of a concept will lead to both the removal and addition of relationships (at the same time). The values presented in table 3.9 are the ratio of each MCO that has the specific change in the relationship of the source concepts in relation to the total of each MCO. One specific change in the relationship (*e.g.*, *Addition*) has three subgroups of values (*i.e.*, "*Super*", "*Sub*" and "*Other*"). For instance, 66.33% of the mappings' source concept added has at least one new relationship added with a super concept (*cf.* $diffM_3$). We analysed each type of relationship independently.

We describe separately the conducted analysis considering how changes in the relationships affect the MCOs:

- "*Super*" **relationships:** When the majority of the "*Super*" relationships of source concepts remain unchanged, mappings also remain unchanged. As show in table 3.9, in most cases the addition of "*Super*" relationships correlate with the addition of mappings. Only in few insignificant cases can the addition of mappings correlates with the removal of "*Super*" relationships. The ratio of addition and removal of "*Super*" relationships is equal when mappings are removed. As explained before, the inactivation of a concept leads to a structural KOS change. These results reveal that addition of "*Super*" relationships influences the addition of mappings. However, as already noted in the analysis of descriptions, the addition of relationships can also be triggered by the addition of concepts.

- "*Sub*" **relationships:** We cannot observe the addition of "*Sub*" relationships when mappings are removed, while the removal of "*Sub*" is not observed when mappings are added. The values are attributable to the inactivation of a concept (which causes the removal of "*Sub*" relationships) or by the addition of new concepts (which can lead to the addition of "*Sub*" relationships). However, it cannot indicate a clear correlation between "*Sub*" re-

lationships and MCO. The impact of changes on "*Sub*" relationships remains inexpressive for the unchanged and modified mappings categories.

- "*Other*" **relationships:** They behave similarly to "*Super*" relationships.

Table 3.9: Concepts' relationships changes correlated with mapping change operations

MCO	$diffM$	Addition (%)			Removal (%)			Unchanged (%)		
		Super	**Sub**	**Other**	**Super**	**Sub**	**Other**	**Super**	**Sub**	**Other**
UNH	$diffM_1$	1.56	1	3.93	1.38	0.79	4.86	99.60	31.50	95.47
	$diffM_2$	1.01	0.83	1.75	0.8	0.36	1.93	99.78	31.33	95.91
	$diffM_3$	1	0.92	2.18	0.84	0.41	2.29	99.83	31.51	96.18
ADD	$diffM_1$	100	25.4	97.4	1.18	0	1.18	0	0	0
	$diffM_2$	100	29.35	99.7	0.45	0	0.45	0	0	0
	$diffM_3$	66.33	16	66.4	0.14	0	0.29	0	0	0
REM	$diffM_1$	100	0	100	100	34	100	0	0	0
	$diffM_2$	100	0	100	100	29.27	100	0	0	0
	$diffM_3$	83.63	0	83.63	83.63	16.36	83.63	0	0	0
MOD_t	$diffM_2$	1.4	0.94	3.75	0	0	2.34	98.61	36.43	95.87
	$diffM_3$	0	5	0	0	10	0	100	100	100
MOD_r	$diffM_2$	0.5	2.51	5.02	0.5	2.51	5.52	99.58	36.54	96.45
	$diffM_3$	0	0	0	0	0	0	100	100	100
MOD_{t_r}	$diffM_2$	1.2	0	1.2	1.2	0	3.57	98.86	36.28	97.28
	$diffM_3$	0	3.45	0	0	0	0	100	100	100

This table presents an analysis observing the influence on the mapping change operations of changes affecting different categorization of relationships (Super, Sub and Other) connected to the source concept.

3.3 Interdependencies between mapping change operations

We aim to reveal possible interdependencies between mapping change operations. We first present the experimental steps conducted (Section 3.3.1). Results include a quantitative analysis (Section 3.3.2) followed by a qualitative discussion of particular cases (Section 3.3.3).

3.3.1 Experimental procedure

Taking into account the neighbourhood of each concept associated with a mapping, we aim to identify whether other entities (*e.g.*, super/sub concepts and sibling concepts) and KOS's structural changes related to them influence mapping evolution. We conducted this investigation because many matching algorithms use structural properties to create mappings that motivate knowing the exact impact of structural properties on mappings evolution. The structural analysis may allow identifying more complex operations related to mappings (*e.g.*, move or duplication of mappings) and look for correlations with more complex changes in KOS (*e.g.*, substitution, split and merge of concepts). For this purpose, we conduct two types of analyses: quantitative and qualitative.

Quantitative analysis method

We search for interdependencies between MCOs and concepts sharing structural properties (*cf.* Section 3.3.2). We investigate different combinations of MCOs considering mappings that share the same target concepts and whose source concepts have a structural relationship. A structural relationship means that source concepts are interrelated through an "$is-a$" relationship or that they have sibling concepts (*i.e.*, they share the same super concept).

The performed experiments aim to identify a possible influence of the context (concepts sharing structural properties) of the source concept on mapping evolution. There may be potential situations where mappings may be changed due to a combination of KOS changes, regardless of whether there is some kind of influence (dependence) between them. For instance, one concept attribute modified plus one new sibling concept led to a change in the source concept of mapping. We emphasize only observations about the behaviour of couples of mappings. In particular, we search for correlations between MCOs involving two mappings. These mappings might share equal and unchanged target concepts, and their source concepts must share structural properties. We observed concepts having a "*is-a*" relationship between them and sibling concepts in the sense that they share a "*Super*" concept. We expect to understand how KOS changes affecting concepts with these kinds of structural relationships may influence the evolution of associated mapping between the concepts involved.

Based on the categorized sets of mappings $diffM_j$, we define $diffM'_j = diffM_j \setminus \{U \cup MOD_t \cup MOD_{t\_r}\}$. For each different target concept of mappings in $diffM'_j$, we determine three different sets of mappings which must share the same target concept. The "*Added*" (*ADD'*) set is made of new mappings having the considered target concept; the "*Removed*" (*REM'*) set contains the removed mappings; and the "$Modified \quad relation$" (MOD'_r) represents mappings whose relation has changed.

After selecting mappings for each of these sets, we searched for possible structural relationships (super or sibling) between two different source concepts of mappings belonging to two distinct sets. We calculated the possible combinations of source concepts tested according to a *Cartesian* product of sets of mappings: comparing different mappings of equal sets (*first group*), $ADD' \times ADD'$, $REM' \times REM'$ and $MOD'_r \times MOD'_r$; or considering different sets (*second group*), $REM' \times ADD'$, $REM' \times MOD'_r$, $ADD' \times MOD'_r$. In these combinations, all source concepts of one set are compared to all different source concepts of the other set to identify the structural relationships between them. When comparing equal sets we do not consider equivalent concepts. We conducted this procedure considering all $diffM'_j$, and we grouped the results for the entire period considered (*cf.* tables 3.10 and 3.11).

Qualitative analysis method

We select some cases based on the results from the quantitative analysis to examine them in detail (*cf.* Section 3.3.3). We aim to observe the role played by each KOS change and by the involved concepts in the MCO from a qualitative point of view. In particular, we investigate the source concept information that mappings are likely related to and evaluate which of the changes in these concepts may have triggered the need for changes in the associated mappings, *i.e.*, the specific features of changes that could explain the need to adapt mappings.

3.3.2 Results of the quantitative analysis

Table 3.10 shows the results linking the structural properties and the combination of equal sets of MCO (*first group*), while table 3.11 shows results for the combination of different set of operations (*second group*). Considering the *first group*, we found 522 distinct cases with regard to the three sets of $diffM_j$ analysed, while in the *second group* we found a total of 18 distinct cases. The numbers in tables 3.10 and 3.11 represent the percentage of cases found in the different combinations with regard to total cases. For instance, in the combination of equal sets, we found 84.5% of the cases in the combination $ADD' \times ADD'$ between sibling concepts. The percentages are independently calculated for each distinct group of analysis (first and second).

Analyzing the results for the *first group* (Table 3.10), in most cases we observe siblings concepts when mappings are added $ADD' \times ADD'$ compared with the other combinations of this group. This indicates that when two new mappings $m_{st1} = (c_{s1}, c_t, semType_1)$ and $m_{st2} = (c_{s2}, c_t, semType_2)$ are added, c_{s1} and c_{s2} are siblings. Few cases with respect to concepts sharing '*is-a*' relationships are found (6.7%); considering the previous m_{st1} and m_{st2} mappings, in these cases c_{s1} is the "*Super*" concept of c_{s2}. This percentage remains higher than those of the combination $REM' \times REM'$. In this latter combination, both types of structural properties have a slightly higher percentage for the type siblings. The results reveal few cases expressing only the super/sub concept property when analysing MOD_r operations in mappings $MOD'_r \times MOD'_r$.

Table 3.10: Analyzing the combination of equal sets of mapping change operations

Structural property	**Combination of equal sets of operations** (%)		
	$\boldsymbol{ADD' \times ADD'}$	$\boldsymbol{REM' \times REM'}$	$\boldsymbol{MOD'_r \times MOD'_r}$
Siblings	84.5	5.53	0
Parents	6.7	2.87	0.4

This table shows interdependencies between combination of mapping change operations of equal sets and concepts sharing structural properties. Sets ADD', REM' and MOD'_r correspond to a filtered set of mapping change operations where mappings share equal target concepts.

Analyzing the results of the *second group* reveals no cases for the parent structural property (Table 3.11). All cases concern the sibling structural property, and most of them belong to $REM' \times ADD'$, while no case was found when analysing $REM' \times MOD'_r$. We only identified a few for $ADD' \times MOD'_r$ (16.6%). To better understand these cases, we provide a more fine-grained analysis considering the different combination sets $REM' \times ADD'$ and $ADD' \times MOD'_r$, which represent the most noteworthy cases of this study.

A deeper analysis of the behaviour of *added* and *modified* mappings sharing equal target concepts ($ADD' \times MOD'_r$) shows a well-defined transition of relations in mappings. Figure 3.2 provides scenarios presenting the status of concepts and mappings before and after KOS evolution. We represent concepts *sup*, c_s, c_k and c_t, where *sup*, c_s and c_k belong to K_S, while c_t belongs to K_T. Before evolution in Figure 3.2 (a), concept c_s interrelates with concept c_t with the $\leq$ type of relation. After evolution, Figure 3.2 (b) presents that a new concept c_k is added (a concept sibling of c_s) and connected to c_t with a relation *semType*. In a third of the cases of $ADD' \times MOD'_r$, *semType* is of type equivalent $\equiv$, and in two-thirds of the cases *semType*

Table 3.11: Analyzing the combination of different sets of mapping change operations

Structural property	Combination of different sets of operations (%)		
	$ADD' \times MOD'_r$	$REM' \times MOD'_r$	$REM' \times ADD'$
Siblings	16.6	0	81.4
Parents	0	0	0

This table shows interdependencies between combination of mapping change operations of different sets and concepts sharing structural properties. Sets ADD', REM' and MOD'_r correspond to a filtered set of mapping change operations where mappings share equal target concepts.

is of $\leq$ type. When analysing the c_s concept, we notice that this concept had minor changes (*e.g.*, one description changed), and the mapping between c_s and c_t also changed. The value of $semType$ changed from $\leq$ to $\equiv$. We assume that the new mapping $m_{kt} = (c_k, c_t, semType)$ stands for the consequence of adding the new concept c_k. However, further investigation is perhaps warranted to understand the reasons why the mapping $m_{st} = (c_s, c_t, \leq)$ changed its type of semantic relation to $\equiv$. One possible explanation relies on the fact that such mapping was strongly connected to the description that was modified. Thus, after evolution, the semantic type ($semType$) $\leq$ is modified to $\equiv$ to keep the semantic validity of the correspondence between concepts c_s and c_t.

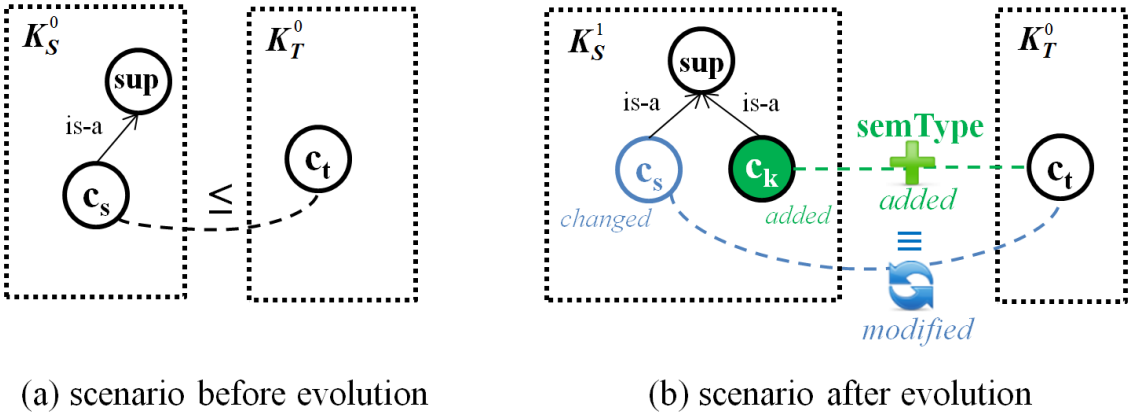


Figure. 3.2: Scenario involving addition and modification of the $semType$ in mapping

This figure shows a scenario before and after evolution illustrating the mapping change operations observed. Analysis of resulting cases of interdependencies between mapping change operations of different sets ($ADD' \times MOD'_r$).

We analyze cases of the combination ($REM' \times ADD'$) which refers to the most frequent in the *second group* (Table 3.11). Figure 3.3 presents two scenarios considering the concepts *sup*, c_s, c_k and c_t similar to Figure 3.2. Figure 3.3 (a) shows the scenario before evolution where the concept c_s is linked to c_t through relation $semType_1$, such that $semType_1 \in \{\perp, \equiv, \leq, \geq, \approx\}$. After evolution, changes affecting concepts and mappings are shown in Figure 3.3 (b). We notice that the concept c_s is assigned to inactive and the mapping $m_{st} = (c_s, c_t, semType_1)$ is removed. The new concept c_k (sibling of c_s) corresponds to c_t with the relation $semType_2 \in \{\perp, \equiv, \leq, \geq, \approx\}$. We detected several cases where $semType_1$ and $semType_2$ were different. Among the 81.4%

of $(REM' \times ADD')$, in 13.3% of cases the relations $semType_1$ and $semType_2$ were of the type *unmappable* $\perp$; in 13.3% of the cases $semType_1$ and $semType_2$ are $\equiv$; in 26.6% they are of the type $\leq$; and in 26.6% of cases, $semType_1$ is $\equiv$ while $semType_2$ consists in $\leq$. The remaining cases, representing 20.2%, are exceptional cases where concepts c_s and c_k are unchanged while the mappings are removed and added (*cf.* Figure 3.3 (b)). In these situations, we also found cases where $semType_1$ and $semType_2$ have distinct types. We deem that these are typical cases where mappings are curated by domain experts.

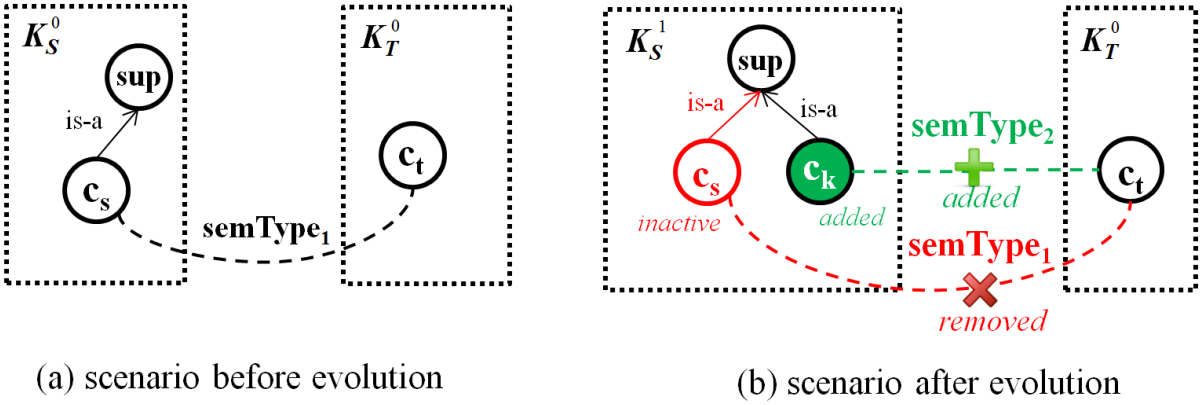


Figure. 3.3: Scenario involving removal and addition of mapping

This figure illustrates a scenario regarding the analysis of resulting cases of interdependencies between mapping change operations of different sets $(REM' \times ADD')$.

Considering cases expressed in Figure 3.3, we face difficulties to identify direct relationships between changes in concepts and changes in mappings. As in the cases studied, concepts had the same change status, while mappings of different relation types were removed and created. It seems that a further analysis of not only the change in a concept's status, but also of detailed KOS changes is required to unfold elements of effects on the evolution of mappings. This is what prompted the analysis presented in the following section. In the case where $semType_1 \neq semType_2$ and c_s and c_k consist of sibling concepts, we must reveal why they have different types of semantic relation, knowing that they are siblings concepts.

3.3.3 Results of the qualitative analysis

We devote this section to a qualitative analysis explaining some relevant cases found. We select some cases of $REM' \times ADD'$ and analyze the content of concepts involved, along with the changes affecting them. This analysis specially aims to explore the existing relations between changes affecting sibling concepts. For instance, we evaluate the content of a concept before and after the applied changes, and search for statements that a transfer from one concept to another was made. These experiments attempt to identify with precision which context content could be used to establish mappings. In particular, we highlight whether mappings follow the related content after evolution. This analysis shows the role played by each concept involved and the details of the changes that occurred in the MCOs. Given their representativeness, we selected three cases and describe them in the following:

Case 1. In this scenario, before SCT evolution, a concept is named "*Primary malignant neoplasm of intrahepatic bile ducts (disorder)*", and a mapping of semantic type ≡ links this concept with the ICD9 concept named "*Malignant neoplasm of Intrahepatic bile ducts*". Analyzing the names, we found a partial match (one is a substring of the other). After evolution, the status of the SCT concept changed to ambiguous (*i.e.*, it remains no longer active), but a new active and sibling concept is added in SCT. We observe that this new sibling concept has the same name as the inactive concept in SCT. In this case, the type ≡ mapping that linked the inactive concept in SCT with the ICD9 concept now links the new sibling concept with the same target ICD9 concept and with the same relation type (*i.e.*, ≡). This case shows that a substitution of concepts occurred and that the associated mapping was transferred to the new sibling concept while maintaining the same type of semantic relation.

Case 2. Before evolution, the SCT concept named "*Cutaneous histiocytoma (disorder)*" is defined as more specific than the ICD9 concept named "*Benign neoplasm of skin, site unspecified*". Note that no exact match occurs between the names, but semantically the concept in SCT can be considered more specific than the one of ICD9. After evolution, the status of the SCT concept changed to duplicated (*i.e.*, it is no longer active), but a new sibling concept named "*Pleomorphic fibroma (disorder)*" is added. Analyzing the content of the descriptions of the inactive concept, we observe that one description exists describing the concept as a "*Pleomorphic fibroma*" (a synonymous term of the inactive concept). This description becomes the name of the new sibling concept. The mapping of semantic type ≤ now links the new sibling concept with the same target concept of ICD9. This case reveals a substitution of concepts, but with a different label (using a synonymous term). As occurred in the other cases, the associated mapping was moved to the new sibling concept while retaining the same type of semantic relation.

Case 3. Before evolution, the SCT concept named "*Chest deformity (finding)*" was equivalent to the ICD9 concept named "*Acquired deformity of chest and rib*". No exact match occurs between both names, but we note some similarity. One description of the SCT concept describe it as "*Deformity of thorax*", which stands for a synonym of the concept. After evolution, the status of the SCT concept changed to ambiguous, but a new sibling concept named "*Deformity of thoracic structure (disorder)*" is added. The description of the inactive concept has a high degree of similarity with the name of the new sibling concept. Particularly in this case, the mapping is moved to the new sibling concept, and its semantic relation is set to type ≤. Thus, "*Chest deformity*" is considered equivalent to the ICD9 concept "*Acquired deformity of chest and rib*", but now the new concept "*Deformity of thoracic structure (disorder)*" is considered semantically more specific than the ICD9 concept. This case shows that substitution between sibling concepts did occur, but modifications in the name affect the evolution of mappings. As in the other cases, the associated mapping is moved to the new sibling concept but with a different semantic relation.

The experiments in this section revealed that interdependencies between different mapping changes operations exist. This can allow providing more complex and adequate mapping adaptation behaviours. Additionally, the qualitative analysis of cases showed details about the mapping evolution. The results underscore the relevance of concepts directly related to the concept involved in mappings (*i.e.*, parent, child and sibling concepts).

3.4 Mapping evolution under complex KOS changes

This experiment aims to analyze mappings evolution in the context of complex KOS changes. In particular, we emphasize split type of change, which consists in recurring complex change appearing in biomedical KOSs. Consequently, the impact of this particular type of change on mappings deserves a closer attention due to the difficulties involved. For instance, when a concept is split into several ones, it is complicated to correctly adapt the early associated mappings, since we can have many ways to change these mappings considering the concepts resulting from the split. This has motivated us to conduct a study on these particular complex changes. We present the experimental procedure performed (Section 3.4.1) followed by the achieved results in ICD9 cases (Section 3.4.2) and SCT cases (Section 3.4.3).

3.4.1 Experimental procedure

We conducted the following four-step procedure for both SCT and ICD9 to investigate cases of split changes impacting mappings: (1) Automatic identification of complex changes; (2) Refinement of the previously identified complex changes; (3) Selection of representative cases impacting associated mappings and (4) Detailed case-by-case analysis of complex changes correlated to the evolution of mappings.

1. **Automatic identification of complex changes.** This is usually referred to as the *diff* calculation problem. In this study, we have not used developed tools (*cf.* Section 2.4.1) since they usually require input files in OWL or OBO formats. Since SCT and ICD9 are unavailable in these formats at the moment of this experiment, we identified the split complex changes by implementing a particular process.

   We first define what is assumed as a split operation. Considering that the K_x^{j+1} is a new version of a K_x^j and that the concept $c_1^{j+1} \in C(K_x^{j+1})$ has the same identifier but at least one attribute changed (*i.e.*, was added, deleted or the value was modified), compared with $c_1^j \in C(K_x^j)$ (*first situation*), or $c_1^{j+1} \notin C(K_x^{j+1})$ (*second situation*). Thus, we detect the split of c_1^j into a non-empty subset of concepts in K_x^{j+1} ($C_r \subseteq C(K_x^{j+1}), C_r \neq \varnothing$) when for each element $c_k \in C_r$, the following conditions are satisfied according to the situation.

   For the *first situation* (c_1^{j+1} modified):

   (a) At least one *common super-concept* exists between c_k and c_1^{j+1}, or c_1^{j+1} is the *super-concept* of c_k (*i.e.*, a "*is-a*" relationship relates them);
   (b) A similarity exists between c_k and c_1^j, greater than a threshold τ.

   For the *second situation* (c_1^{j+1} deleted):

   (a) In the case where c_k refers to a modified concept (*i.e.*, $c_k \in C(K_x^j)$), there must exist at least one *common super-concept* between c_k and c_1^j, or c_1^j stands for the *super-concept* of c_k, otherwise this condition is not considered;
   (b) A similarity exists between c_k and c_1^j, greater than a threshold τ.

   The result of the split is given by $C_r \cup c_1^{j+1}$.

   Based on this definition, we describe the split identification procedure to recognize split change operations between two different versions of a KOS:

- First, we identify all concepts that were affected by the KOS evolution and we group them in the set $diffC$. We calculate a simple *diff* between all concepts of K_x^j and K_x^{j+1}. To this end, we compare the concepts' identifier to determine whether a concept is added or removed. For instance, if the identifier exists in K_x^j and not in K_x^{j+1} then we consider that the concept was removed, and the opposite for added concepts. For modified concepts, we compare the content of concepts that exist in both versions. In the set $diffC$, we associate each concept to one type of change (*add, remove, modify*).
- We filter the set $diffC$ to keep only concepts related with existing mappings to limit our investigation to concepts associated with mappings. In other words, we verify whether each concept from the set $diffC$ is the source concept of a mapping (if analyzing SCT) or the target one (for ICD9). We exclude from $diffC$ concepts that are unrelated to mappings.
- Afterwards, we use the given definition of split to identify concepts belonging to $diffC$ that can be involved in a split change operation. We consider each concept from the set $diffC$ and we start by verifying the first condition (super-concepts or siblings). When two concepts from $diffC$ fulfill the first condition (*i.e.*, we find one added or modified concept from $diffC$ that is sub-concept or sibling of another modified or removed concept of $diffC$), then we verify the second condition (similarity). If the second condition is fulfilled, we group both concepts into a "*pair*". For example, in ICD9 the concept codes '*560.39*' and '*560.32*' (*cf.* Figure 3.4) belong to $diffC$ since the former has been modified and the latter is a new added concept. In this case, both conditions are fulfilled as these consist of sibling concepts and we found a similarity between them. We follow this order of verification of the conditions because it is more likely that neighbor concepts are involved in a split than others, and because this is important for optimization. We decrease the quantity of similarity calculation between concepts. The details on the similarity measures explored will be explained later in this section.
- Finally, we analyze the set of pairs found to identify those concepts from K_x^j that were split into more than one concept in K_x^{j+1}. For this purpose, we group all pairs of similar concepts that share a common concept. In this case, the concept in K_x^j expresses similarity with one or more concept(s) in K_x^{j+1} and we certify that the content of this concept has modified or the concept was removed from one version to another. We associate these pairs to one split operation.

We can calculate a value of similarity between concepts by using different techniques [Cheatham and Hitzler, 2013]. In general, the result expresses a weight of how much both concepts are semantically similar. In this experiment, we utilize a hybrid method considering syntactic and semantic information. Let c_1^j, c_2^{j+1} be two different concepts in two distinct versions of a KOS, where $c_1^j \in C(K_x^j)$ and $c_2^{j+1} \in C(K_x^{j+1})$. The syntactic part of the method compares attribute values of both c_1^j and c_2^{j+1} as strings using the well-known *Levenshtein edit-distance measure* [Levenshtein, 1966] (we provide a normalized definition of this function in Section 4.3.1). Concerning semantics, we used *MetaMAP*[28] [McCray et al., 1994]. If concepts share the same semantic type in *UMLS* [Bodenreider, 2004], then c_1^j and c_2^{j+1} are considered as similar ones.

[28] `metamap.nlm.nih.gov`

Entities used for calculating the similarity differ according to the considered KOS model, as ICD9 and SCT are based on different knowledge models and they do not provide the same type of KOS's entities. *Concepts* in the ICD9, for instance, provide textual statements such as values of titles and of attributes such as *notes*, *includes* and *excludes*. By contrast, we can explore further descriptions and structural information in SCT. Therefore, we define here slightly different approaches to calculate the similarity in ICD9 and SCT based on the proposed hybrid method.

We calculate the similarity between c_1^j and c_2^{j+1} in ICD9 as follows:

- The c_1^j and c_2^{j+1} in ICD9 are considered similar if they have their title attribute considered syntactically or semantically similar, or if they have at least one similar phrase in *notes, includes* and *excludes* attributes.
  - ◦ We compare the titles' value of c_1^j and c_2^{j+1} using both syntactic and semantic methods. If a negative result is found, then we try to compare textual information contained in *notes, includes* and *excludes* attributes in both c_1^j and c_2^{j+1}. For instance, a negative result is detected comparing the value of the title of the concepts '*560.39*' ("*other*") and '*560.32*' ("*fecal impaction*"), but when comparing one of the notes of the former with the value of the title of the latter, we find an exact match.
  - ◦ We compute the *Cartesian* product between these attributes. In this sense, we compare all notes of c_1^j with all *notes* of c_2^{j+1}. We apply a similar approach for *includes* and *excludes*. The value of these attributes is composed of a set of distinct phrases (strings), and each phrase is composed of a set of words. Observing if at least one phrase of c_1^j is similar to a phrase in c_2^{j+1} is made using the syntactic method. We compare all sets of phrases from c_1^j to all set of phrases of c_2^{j+1} for each type of attributes, searching for a "true" similarity.

We calculate the similarity between c_1^j and c_2^{j+1} in SCT as follows:

- To consider that c_1^j and c_2^{j+1} are two similar concepts in the evolution of SCT, one of the conditions must be fulfilled in the following order: (1) Syntactic comparison of name; (2) Semantic comparison of name; (3) Syntactic comparison of descriptions; and (4) Semantic comparison of descriptions.
- Given two sets of descriptions, one belonging to c_1^j and the other to c_2^{j+1} we use the *Cartesian* product between both sets to compare them based on the syntactic and semantic parts of the method.

2. **Refinement of the previously identified complex changes.** We manually refine the identified groups of concepts involved in the split. This step plays a key role due to the possible inaccuracy of similarity measures, and to improve results in a re-organization of splits. In this analysis, we might merge groups of concepts that appeared to belong to the same split operation. We might also identify false positives groups and remove them. For instance, the case of ICD9 presented in figure 3.6 had been firstly automatically identified as different split cases, and by the manual refinement we realized that they concerned the same split operation. We enrich the information about possible concepts involved in a split in adding, *e.g.,* a new sibling concept that should be involved in a split operation and which was not assigned in the automatic step. For example, the concepts '*752.45*', '*752.46*' and

'*752.47*' of ICD9 in figure 3.5 were manually added since it was observed they shared a similarity with the concept '*752.49*'. This step provides cases of split to be analysed.

3. **Selection of representative cases impacting associated mappings.** We associate all mappings with the concepts belonging to cases of split of the latter step. Note that split operations lacking associated mappings are no longer considered. This reduces the quantity of cases we have to manually take into account in the analysis. In this third step, we analyze the evolution of mappings in the context of split operations, *i.e.*, we observe the type of changes occurred in mappings. For instance, we analyze mappings that are adapted to the resulting concepts of the split, those that are removed or with a modification in the type of their semantic relation. Based on this initial analysis, we select the most representative cases for a detailed analysis (next step). For instance, we consider only one case among those containing repeated behaviours. We thus depict the behaviour illustrating a scenario before and after evolution of the selected cases, which shows concepts of the splits, and the changes affecting the associated mappings.

4. **Detailed case-by-case analysis of complex changes correlated to the evolution of mappings.** We analyze the final selected cases of ICD9 and SCT. This consists of observing the types of atomic changes affecting the split concepts. For instance, we observe the value of the attributes shared between the concepts of the splits. We explain the behaviour of the mappings correlating them to the (types of) change(s) affecting the concepts of the split. For instance, we try to understand the modifications of the semantic relations occurred in the mappings. Also, we search for reasons leading mappings to changes toward one or other concept resulting from split. We compare the different cases, searching for contrasts between them. We also relate differences between the cases of SCT and ICD9. The utmost goal in this step is to learn lessons from the selected and analysed cases that all our previous experiments could not reveal.

This experiment allowed the identification of the most representative cases and interesting variations of split changes involved in the evolution of mappings. We mainly underlined six assorted cases of concept splits, of which four affecting ICD9 and two occurring in SCT, having a different impact on the way mappings evolved over time. We demonstrate the found cases through figures 3.4, 3.5, 3.6, 3.7, 3.8 and 3.9. Each figure shows a scenario with concepts and associated mappings before evolution (left part of the figure) and the scenario after evolution with the updated status of concepts and mappings (right part of the figure). We represent concepts in ICD9 as circles while those in SCT are represented as squares. Light blue concepts are modified concepts and green concepts (with larger borders) denote new concepts. Mappings are represented as orange arrow lines connecting the concepts between SCT and ICD9. Blue arrow lines represent a similarity detected between concepts before and after evolution.

3.4.2 ICD-9-CM cases

We observe that no concepts are removed in all split cases found. More generally, the concrete removal of concepts rarely occurs in ICD9 and never occurs in SCT. Addition of concepts remains the most frequent operation, since it is more usual and natural that new knowledge is aggregated into the biomedical KOS over time. The first case (Figure 3.4) highlights the split of the concept '*560.39*'. In this case, the most interesting aspect refers to the attribute notes. Actually, before evolution, it contained three different values "*Concretion of intestine*", "*Enterolith*" and "*Fecal impaction*". After evolution, the latter mentioned value is deleted from the notes of the

concept '*560.39*' and became the title of the new concept '*560.32*'. A closer look at the five mappings linking the SCT concepts to the ICD9 concept '*560.39*' before evolution reveals that two of them have as SCT concept names "*Fecal impaction (disorder)*" and "*Fecal impaction of colon (disorder)*". After KOS evolution, these two mappings are directly moved (without modification of the type of the semantic relation) to the newly created ICD9 concept '*560.32*' that has "*fecal impaction*" as title. This operation means that the mapping has its source or target element changed. This case underlines that attribute values play a relevant role for maintaining mappings valid over time, since mappings follow the evolution of the primary information they are attached to. Besides, the three mappings that remain unchanged involved "*Enterolith (disorder)*", "*Typhlolithiasis (disorder)*" and "*Concretion of intestine (disorder)*" of SCT, three names of concept that correspond to unmodified values of concepts attributes in ICD9.

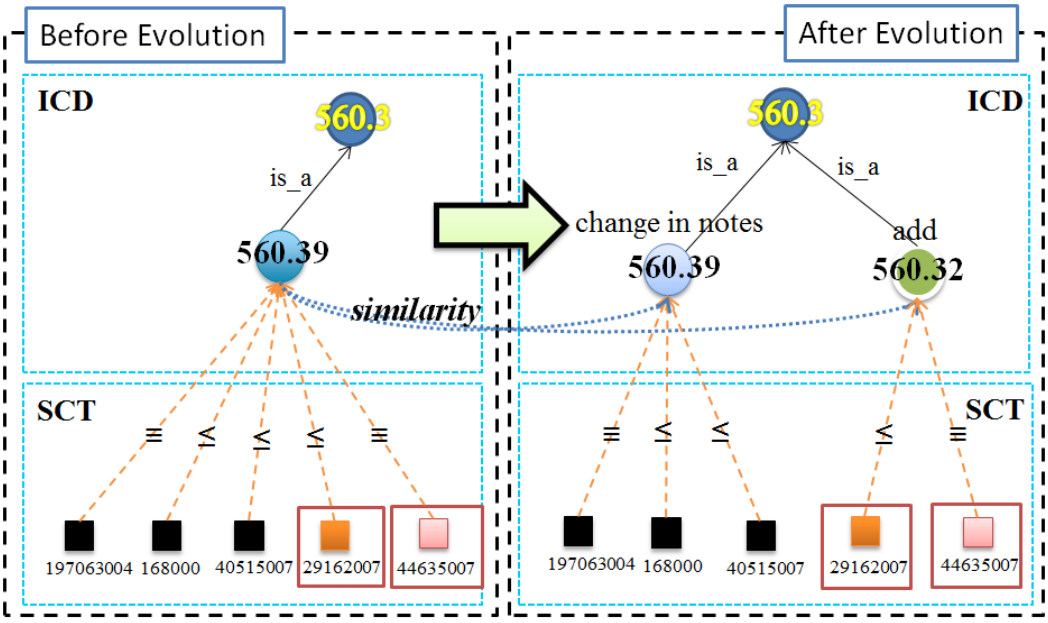


Figure. 3.4: First case of split complex change in ICD9

This figure shows the evolution of ICD9 with associated mappings linked to SCT. The figure presents a scenario with concepts and associated mappings before evolution (left part of the figure) and the scenario after evolution with the updated status of concepts and mappings (right part of the figure). We represent concepts in ICD9 as circles while those in SCT are represented as squares. Light blue concepts are modified concepts and green concepts (with larger borders) denote new concepts. Mappings are represented as orange arrow lines connecting the concepts between SCT and ICD9.

The second case (Figure 3.5) represents a generalization of the split change described in the first case. Note that instead of one new concept in the split we find several of them. We observed that part of the content in note attributes of the initial concept in ICD9 (*i.e.*, '*752.49*' before evolution) is distributed over five newly created concepts. After the evolution, the initial concept becomes semantically more general and the created concepts are semantically more specific. More precisely, information about "*Absence of cervix*" describing the initial concept has been split into two new concepts: '*752.43*' "*Cervical agenesis*" and '*752.44*' "*Cervical duplication*". These modifications caused the move of two established mappings combined with a modification of the type of their semantic relation from (≤) to (≡), since the two new concepts are more

specific than the initial one. Note that three new concepts remain without associated mappings after evolution, and n mappings associated with '*752.49*' before evolution remain unchanged. This indicates that these n mappings are associated with entities of the concept '*752.49*' that did not change. Consequently, in a situation where one of such content is deleted, the adaptation of mappings can consider the removal of directly affected mappings. Therefore, in the context of a split change, mappings can either remain unchanged, or are moved towards a resulted split concept, or are removed. Based on these observations, these aspects of adaptation could be automatically decided according to the KOS' entities content that mappings are associated with, and the flow of content over the concepts belonging to the complex change operation.

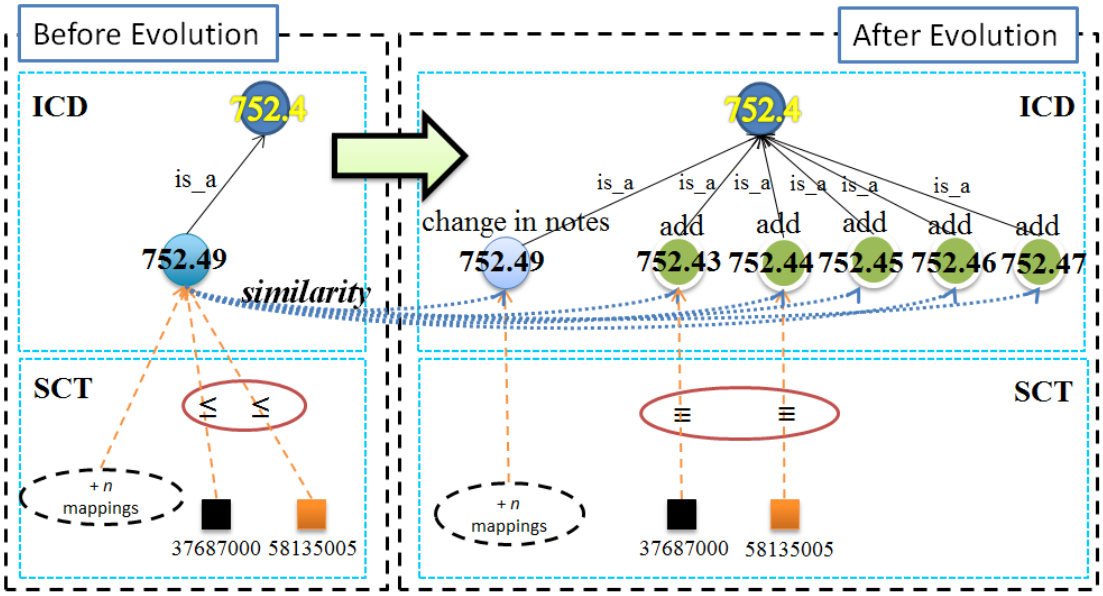


Figure. 3.5: Second case of split complex change in ICD9

The third case (Figure 3.6) differs from the two previous ones since move of mappings may be combined with a change of semantic relation. In the first case studied, any moved mappings changed the type of their semantic relation while in the second one, all moved mappings changed their semantic relation. In the third case, there is a mix of them. A potential explanation relies on the fact that the split change generated new sibling concepts with more semantically specific titles with a close similarity to some (but not all) concepts from SCT, causing the change in the relations. For instance, ICD9 concept '*V13.69*' had the title changed from "*Other congenital malformations*" to "*Personal history of other (corrected) congenital malformations*". After evolution, the concept '*V13.69*' was split into eight concepts (*cf.* Figure 3.6). Associated with this split change, we found the new concept '*V13.68*' whose title "*Personal history of (corrected) congenital malformations of integument, limbs, and musculoskeletal systems*" corresponds, after evolution, to the SCT concept "*History of - congenital dislocation - hip*" (we observe a similarity between "*hip*" and the words "*limbs*" and "*musculoskeletal*"). In this case, the SCT concept still remains more semantically specific than the ICD9, and the type of semantic relation between them kept unchanged. However, the new concept '*V13.64*' (also associated with the same split change operation) with the title "*Personal history of (corrected) congenital malformations of eye, ear, face and neck*" corresponds after evolution to the SCT concept "*History of - cleft lip (situ-*

ation)" (note a better similarity between "*lip*" and the word "*face*"). Consequently, the type of the semantic relation in the adapted mapping needs to reflect this improvement of the similarity, thus the change from *more specific than* (≤) to equivalent (≡) in the semantic relation.

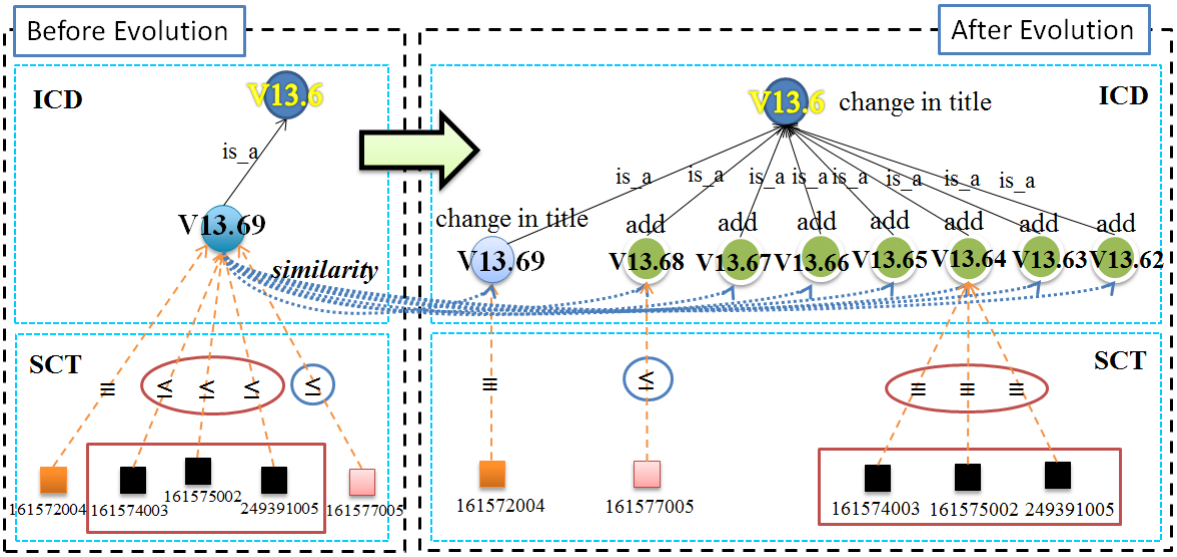


Figure. 3.6: Third case of split complex change in ICD9

The last case in ICD9 (Figure 3.7) describes a structural variant of the split change. Unlike the previously presented cases, the KOS evolution leads to the creation of new *sub-concepts* that describe a refinement of the initial *super-concept*. Actually, the description of the *super-concept* '*752.3*' is made more general from the semantic point of view, and new *sub-concepts* are created based on the content that is removed from this initial concept. This reinforces the idea of a split since an explicit transfer of textual statements occurs from one concept to another. In fact, the title of the new *sub-concepts* is defined based on the content removed from the notes attribute of the concept '*752.3*'. This impacted associated mappings by adapting the type of their semantic relations accordingly. In this case, mappings associated with the *super-concept* before evolution, are duplicated to the new sub-concepts after evolution (*i.e.*, each new sub-concept has a copy of a sub-set of mappings from the *super-concept*). Observe that transfer and duplication of mappings are different since in the latter the original mappings are not deleted. Moreover, some of the duplicated mappings are also affected by a change in the type of their semantic relation from a *more specific than* (≤) to equivalent (≡). However, it might be considered logically inconsistent that two different concepts from ICD9 (connected through an "*is-a*" relationship) link with a mapping of (≡) equivalent type with the same concept (in SCT).

3.4.3 SNOMED-CT cases

The analysis of split cases occurring in ICD9 allows the identification of very interesting aspects for maintaining mappings valid over time. We enrich this study with the analysis of two additional cases of concept splitting that often occur in SCT according to our experiments (*cf.* [Dos Reis et al., 2013b] for the entire description of other SCT cases). As the SCT model is richer than ICD9, it offers more possibilities regarding the behavior of the KOS evolution and

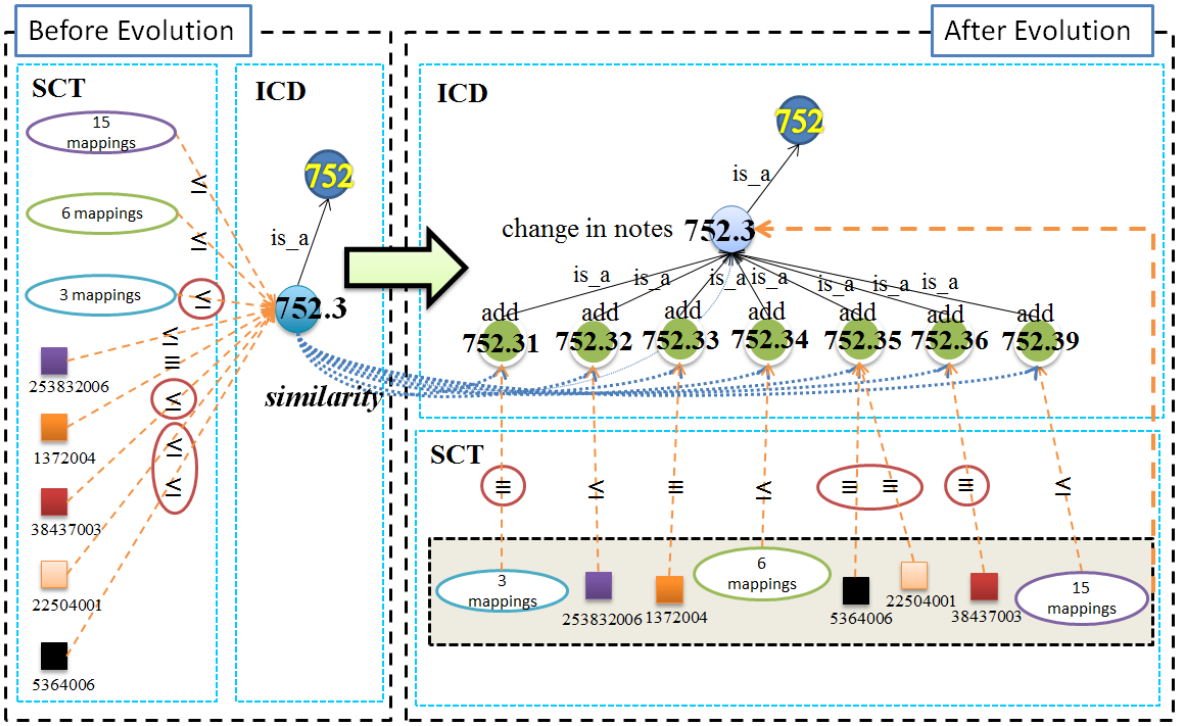


Figure. 3.7: Fourth case of split complex change in ICD9

mappings from the semantic and structural point of views.

Figure 3.8 and Figure 3.9 illustrate the cases of split identified for SCT. The first case consists in a transfer of information contained in the name and description of the concept "*Ventricular septal defect, spontaneous closure (disorder)*" ('*123714004*') to newly created concepts, in which some of them are *sub-concepts* and others are *siblings*. We observe a duplication of the existing mappings to these new concepts without any modification of their semantic relation (*i.e.*, it keeps the type of semantic relation as *more specific than* – $\leq$). From the semantic point of view, we might consider consistent that new more specific concepts have a duplication of a mapping of *more specific than* ($\leq$) type, which remains untrue for other types of semantic mappings. Since the *super-concept* stands for a more specific than the target concept of mappings (in ICD9), *sub-concepts* are naturally even more specific than the target concept in ICD9.

The second case (Figure 3.9) represents a variant of the first case in the sense that the newly created concepts, *siblings* of the initial concept, are related to different *super-concepts*. However, the adaptation of affected mappings expresses the same behaviour as in the first case. Since in both cases mapping adaptation behaves similarly, despite some minor differences in the split of each case. This raises an interesting fact that some aspects of the split cannot deeply affect the evolution of associated mappings, while others are determinant. For instance, new concepts belonging to the split that express relationships to distinct *super-concepts* do not interfere as a determinant factor in mappings adaptation.

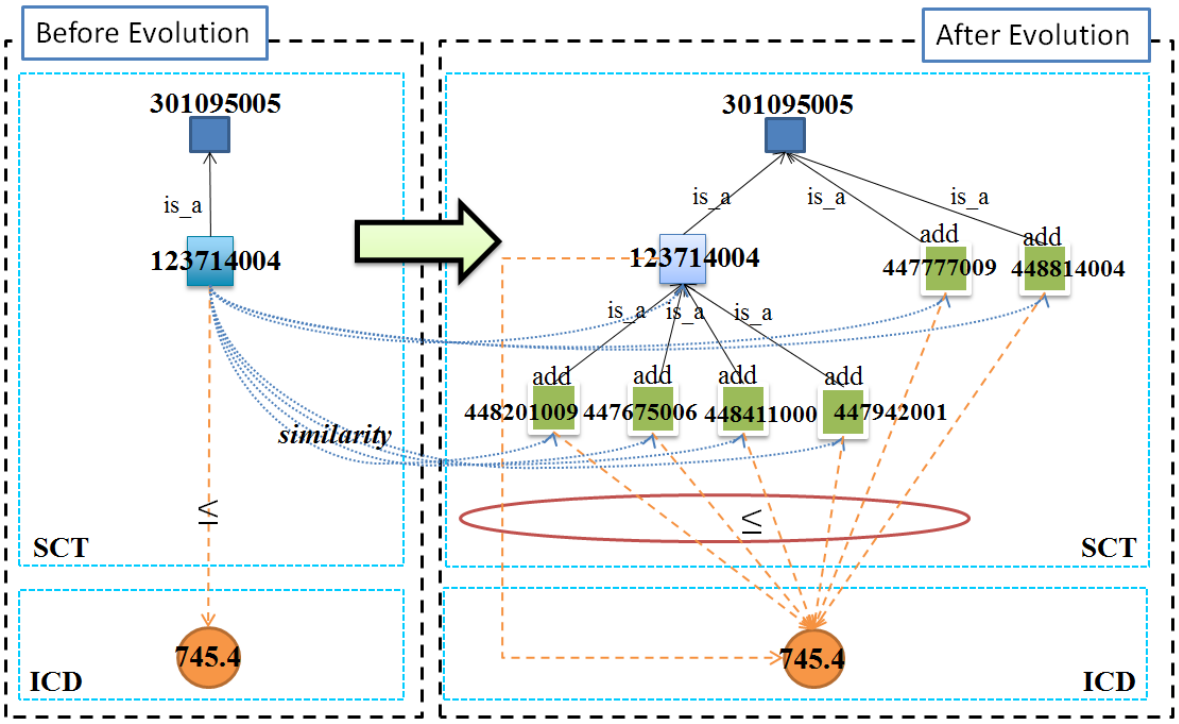


Figure. 3.8: First case of split complex change in SCT

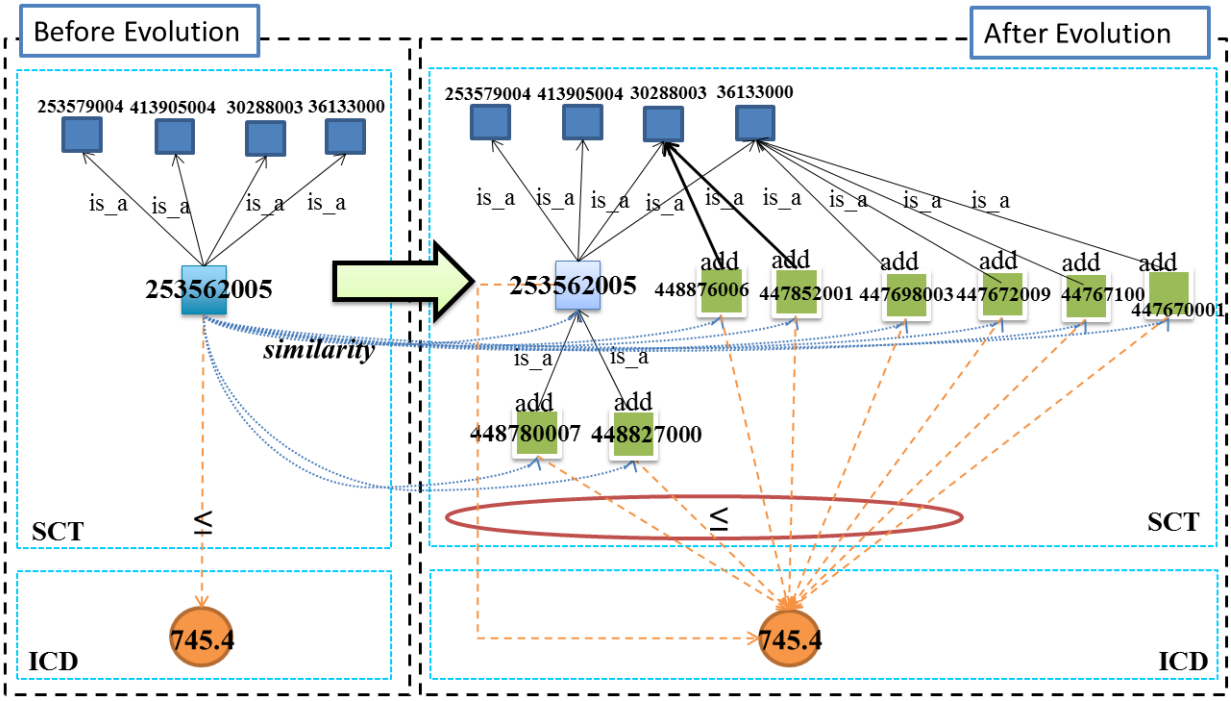


Figure. 3.9: Second case of split complex change in SCT

3.5 Discussion

The obtained results via several analyses uncover major aspects for understanding why mappings have changed, and for the factors that a mapping adaptation system must take into consideration. We demonstrated that we may hardly accomplish this task manually due to the large size and high frequency of new releases of current KOSs. Hence, size, complexity and dynamic of biomedical KOSs avoid a totally manual maintenance of mappings. Our investigation shows the role played by various dimensions discussed as follows:

- Influence of concept attributes in mappings evolution;
- Reasons underlying KOS changes for specific changes in mappings;
- Impact of context and complex KOS changes as well as the combination of mapping change operations;
- Influence of structural relationships organizing the underlying concepts involved in complex KOS changes;
- Influence of concepts' semantic evolution and of similarity between concepts in different KOS versions;

Influence of concept attributes in mappings evolution

Our experiments empirically revealed that although mappings interrelating biomedical KOSs are established between KOS concepts to put them in correspondence in their entirety, the studied cases show that mappings are defined based on information partially described within the concepts (*i.e.*, some concept attributes). Furthermore, the way that mappings evolve after the evolution of a KOS strongly depends on the modifications affecting this specific subset of concept's attributes. For example, a mapping does not require adaptation if it is established based on a specific attribute of a source concept that remains unchanged, even though all other attributes have changed. This finding indicates that adapting mappings requires identifying KOS's entities, which serve to define correspondences between KOS concepts, and to consider these as an additional (meta-) data of mappings. To this end, we suggest the possibility of enriching the well-accepted definition of KOS mappings in literature [Euzenat and Shvaiko, 2007] and also adopted in this thesis (*cf.* Equation 2.1). Considering the obtained results to cope with the mapping maintenance problem, it appears interesting extending this definition by adding information on elements that provide useful statements to explain established mappings between dynamic KOSs.

Reasons underlying KOS changes for specific changes in mappings

Our findings align with related results achieved in literature [Groß et al., 2012]. First, an addition of mappings mainly relates to additions of concepts (*cf.* Table 3.6). We also observe that the addition of attributes such as descriptions also influences the addition of mappings (our study originally demonstrated this finding). Therefore, a high frequency of added concepts reflects the high number of added mappings. Similarly, removal of mappings mainly occurs when the status of concepts or their terms change to inactive (*cf.* Table 3.7). However, we did not observe any clear correlation between changes in SCT entities and modifications in the semantic relation of mappings between KOSs (*i.e.*, MOD_t, MOD_r and MOD_{t_r}). The complexity of current maintenance process may force knowledge engineers to exclude this case. Concepts' relationships also did not indicate any clear influence on how mappings evolve.

Impact of context and complex KOS changes as well as combination of mapping change operations

This study enabled an in-depth understanding of mappings evolution in the scenarios of complex KOS changes by examining qualitative aspects of individual cases. Results pointed out that understanding the evolution of relevant attributes for explaining mappings by considering a set of other closely related concepts (*e.g.*, among concepts composing a complex change) plays a key role in driving how to adapt mappings. This might stand for the cornerstone aspect of an approach to a (semi-) automatic mapping adaptation mechanism.

Considering the context of atomic KOS changes, whenever possible, brings advantage to adequately adapt mappings. Different contextual information regarding KOS changes may have varying influence on how changes occur in mappings. Concepts close to source or target concept involved in mappings provide richer information on how to adapt mappings. For example, sibling concepts can belong to a split change operation. Results showed that only considering traditional simple KOS changes over interrelated concepts does not yield enough information to determine how to apply changes in mappings, because this approach can conceal potential interdependencies between mappings change operations. For example, to verify if one mapping is deleted because another mapping is created (characterizing a move of mapping). We combined different mapping change operations and analyzed if the source concepts of each mapping share some structural property (*cf.* tables 3.10 and 3.11).

The results indicated potential correlations between KOS evolution and mapping change operations. Based on the qualitative analysis of cases, we detected potential influence of KOS changes involving more than one concept. The selected studied cases refer to real examples in which mapping removals and additions are combined into a more complex procedure for adapting mappings in case of a complex KOS change. We demonstrated that in a combined way, removal and addition operations over mappings may belong to a more complex operation (*i.e.*, a move) under a broader scenario of change (*e.g.*, substitution or split of concept).

Influence of structural relationships organizing the underlying concepts involved in complex KOS changes

The expressivity of the knowledge representation model of biomedical KOSs, such as the structural properties of KOS, affects the evolution of KOS and its impact on mappings. This investigation showed that the re-construction of the KOS at evolution time (*e.g.*, creation of new concepts that can be either *siblings*, *children* concepts or both) causes different behaviours in mappings evolution. For example, whether to apply a move or a duplication of mappings seems to depend on the type of structural modification affecting the KOS. This aspect also influences the modification of type of the semantic relation. The obtained findings underscore the need of taking into account the structural organization of the involved concepts when adapting mappings, according to complex change operations or concepts in the neighbourhood.

Influence of concepts' semantic evolution and of similarity between concepts in different KOS versions

Changes interfering in KOS can modify the semantics of its concepts leading either to their domain generalization or specialization. This aspect forces the re-definition of the semantic relations in mappings. On one hand, the values of concepts' names or attributes can suffer lexical changes. For instance, Figure 3.4 shows that the value of the notes attribute "*Agenesis of cervix*"

of the concept '*752.49*' is transformed into a *synonym* "*Cervical agenesis*", which does not really impact an associated mapping. On the other hand, part of the whole content describing a concept (*e.g.*, an attribute) can be transferred to another concept closely related (*e.g.*, a *parent concept*). This can make the initial concept semantically more general and, in consequence, mappings that express an equivalent (≡) type of semantic relation must change to ≤. Moreover, the transferred attribute can change its value turning into a more specific one. Such attribute will impact a mapping if its value is relevant for the definition of this mapping.

These remarks emphasize the similarity shared between concepts of different KOS versions involved in complex change operations or the neighbourhood (that demands further studies). Detecting the different evolution cases requires to understand how one attribute value corresponds to others in the new KOS version, which can rely on the similarity between them. In this context, we deem that the notion of change patterns could specifically apply for concept's attributes that would benefit mapping adaptation. Change patterns at the level of attribute values could delineate the recurring specific types of changes observed for attributes characterizing biomedical concepts. As we observed, attributes can be copied or transferred and their values can become more or less specific. This expresses a real motivation and influence for mapping adaptation according to our experiments. Based on this dimension, change patterns at the level of attributes might make possible better characterization and formalization of the details of complex KOS changes (*e.g.*, all variations of split or merge of concepts), for which adequately adapting mappings requires a further understanding of the underlying KOS changes.

Conclusion

Very few studies attempt to investigate possible correlations between changes affecting KOS's entities and changes in associated mappings. This chapter studied the impact of KOS evolution on mappings, by examining the evolution of real mappings between biomedical KOSs, to understand the central elements in this phenomenon through practical experiences. We conducted a set of experimental analyses observing official mappings between SNOMED CT and ICD9. We characterized the impact via potential interdependencies between how the KOS evolved, and the consequent changes observed in mappings.

Firstly, this investigation studied quantitative analyses of different KOS's entities focused on their atomic changes. Secondly, we further performed qualitative analyses to explain additional facts uncovered by the former analyses. Results revealed the key role played by individual concept attributes to adapt mappings. We demonstrated the utmost factors that shall be explored. They mostly rely on the understanding and characterization of KOS evolution, and especially the complex changes affecting KOS's entities, taking several aspects of the underlying changes and established mappings into account. In particular, the studied real cases of split of concepts highlighted that a fine-grained definition of changes, categorizing as possible change patterns recurring scenarios of modifications at the level of concept attributes, might provide further support for the update of mappings according to KOS evolution.

This chapter provided the basis for defining (semi-) automatic mechanisms to keep mappings up-to-date at KOS evolution time (the next part of this thesis). The conducted research allows to conclude the key fact that to adequately adapt mappings requires to identify relevant concept attributes for a given mapping and to accurately characterize their evolution.

Part II

The *DyKOSMap* approach

We dedicate the second part of this thesis to present both our approach and contributions to mapping adaptation. Firstly, we show the intuition behind the proposed mapping adaptation approach. We then justify our original proposal in view of the conducted experiments and literature review. In the sequence, we describe the chapters that compose this second part.

The proposed solution analyzes established mappings and KOS evolution, by taking into account changes affecting attributes of interrelated concepts, to inform a mapping adaptation component that accounts for changing mapping's elements.

To determine an updated set of mappings as a final outcome, the defined approach requires as input the set of initial and likely affected mappings by KOS evolution. As we assume that only source or target KOS can affect mappings per time (*i.e.*, not simultaneously a priori), our approach requires two versions of the evolving source/target KOS. The first version K_S^0 interrelates the initial out-dated mappings $\mathcal{M}_{ST}^0$ with K_T^0, and the second version stands for the new up-to-date KOS version K_S^1. To determine the $\mathcal{M}_{ST}^1$ updated in accordance with K_S^1, our approach will rely on three main concepts (*cf.* Figure 3.10).

The *mapping interpretation* component ([A] in Figure 3.10) processes the input mappings to better understand the correspondences. The *KOS changes* component ([B] in Figure 3.10) handles the KOS changes between the KOS versions (*diff*). Both [A] and [B] provide information to the *mapping adaptation* component ([C] in Figure 3.10) that accounts for changing mappings' elements. These modules compose the resulting *DyKOSMap* framework.

When some KOS changes really affect a source concept c_s of a mapping, the adaptation approach aims at deciding which action from a set of pre-defined mapping adaptation actions to apply for adapting the impacted mapping. We rely on the mapping's elements c_s, c_t and $semType$ to perform the adaptation. First, our adaptation assumption states that the concept c_t in mapping will always remain fixed in adaptation (*i.e.*, we never replace the concept target). The proposed technique re-defines the type of semantic relation $semType$ of the original mapping, if possible. Otherwise, and if needed, it selects a new source concept and defines an updated semantic relation connecting the new adequate source concept and the original target concept. Finally, if we cannot indicate the proper changes regarding the c_s and $semType$, the defined technique suggests removing the affected mapping.

The major research challenges rely on this decision of adequate actions to apply and on the proposition of adequate candidates of source concepts and/or semantic relations to guarantee that each adapted mapping remains accurate. As an example, given a mapping $m_{st}^0 = (c_s^0, c_t^0, semType^0)$ the suggested possibilities of adaptation are:

- $m_{st}^1 = (c_s^1, c_t^1, semType_{st1}^1)$ [change semantic relation];
- $m_{st}^1 = (c_{s1}^1, c_t^1, semType_{st}^1)$ [change source concept];
- $m_{st}^1 = (c_{s1}^1, c_t^1, semType_{st1}^1)$ [change both relation and source concept];
- $m_{st}^1 = \varnothing$ [remove];

such that $c_s^0 \neq c_{s1}^1$ (*i.e.*, different identifiers) and $semType_{st}^0 \neq semType_{st1}^1$.

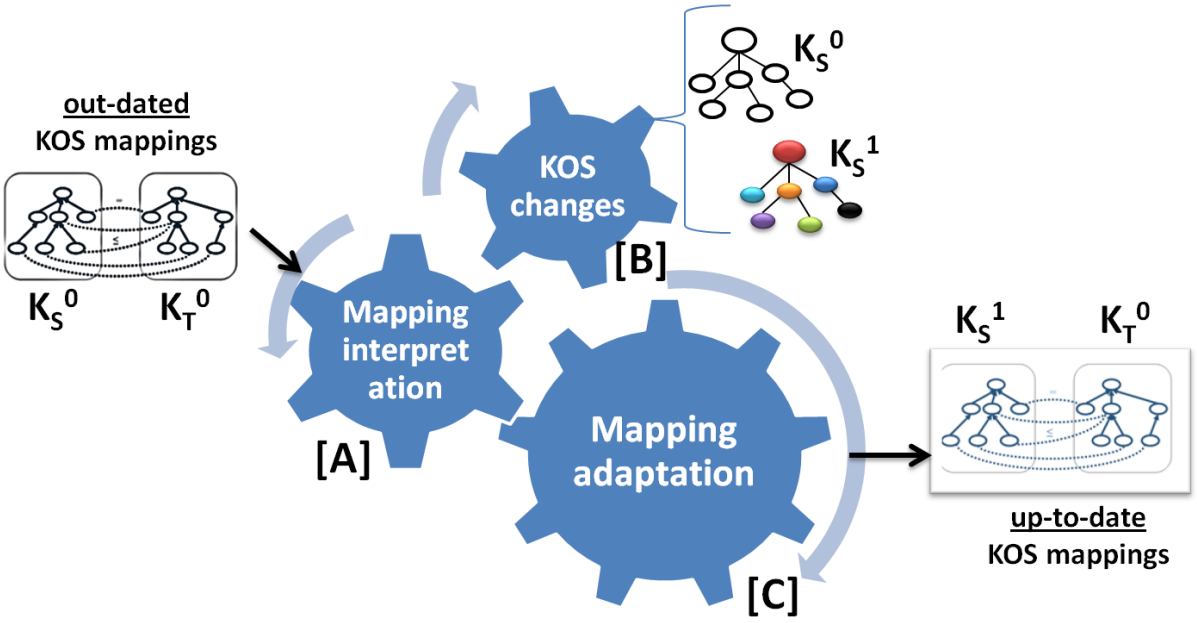


Figure. 3.10: The *DyKOSMap* approach outlook
This figure shows an outlook on the approach with the main components.

We evidence our thesis' prime scientific originality and innovations in view of the conducted literature survey analysis (*cf.* Chapter 2). This enables us to define a set of open issues that existing mapping maintenance investigations fail to tackle, and that we address in this thesis.

Despite the different approaches to deal with mapping maintenance, and the possible benefits of mapping adaptation, literature still lacks a flexible adaptation approach. Existing studies fail to reuse information expressed by mappings and to accurately understand KOS evolution affecting mapping. This avoids proposing richer mapping adaptation strategies like modifying the type of semantic relation in mapping. In general, existing approaches only explore primitive and simple axioms of changes. As a result, for instance, adapting the semantic relations in mappings creates difficulties and forces re-aligning the involved KOSs. This explains the reasons why most of the existing approaches only remove out-dated mappings.

The investigation conducted in this thesis answers several lacks of mapping adaptation literature, providing a more complete solution able to adapt mappings with more precision. In contrast to literature, we originally propose adapting mappings, providing ways to change mappings' elements in function of the evolution of relevant information explaining mappings. The empirical experiments of reference, observing characteristics of the evolution of biomedical KOS and mappings, allow us to situate this work relying on the interpretation of concepts' attributes.

This dimension gives an opportunity to adapt mappings with a higher level of precision because it enables understanding on which KOS's entities mappings rely their existence on. Therefore, the proposed approach emphasizes the detection of concept attributes adequate to explain mappings and the evolution's characterization of these attributes from different viewpoints (*e.g.*, lexical, semantic) to support mapping adaptation. In this context, no work aiming to interpret mappings enables detecting KOS's entities that justify established mappings (*cf.* Section 2.3),

which is required in the mapping adaptation approach. Regarding KOS evolution, while existing tools allow describing and identifying traditional KOS changes, and change patterns mostly apply to manage KOS evolution consistency, our mapping adaptation approach requires understanding the evolution with specific change operations at the level of concept attributes.

The experiments in chapter 3 justify the defined approach and suggested components in the framework. First, the findings revealed the relevance of interpreting mappings (component [A]) that we further study in chapter 4. Second, we demonstrated the role of KOS evolution and in particular the relevance of characterizing the evolution of textual statements explaining mappings (component [B]). In addition to existing tools to calculate KOS changes, we present in chapter 5 an original proposal with respect to KOS evolution to specifically support mapping adaptation. Third, supplementary to atomic changes of mapping adaptation (*e.g.*, a simple mapping removal), the results from the experiments showed complex behaviours of changes in mapping's elements, *e.g.*, the replacement of a source concept to another one from the new KOS version. This aspect aggregates more refined adaptation operations, which can express a move of mapping in its adaptation (component [C]). Chapter 6 introduces the designed novel techniques to adapt mappings. We integrate all proposed components in the *DyKOSMap* framework performing the whole process to adapt mappings (Chapter 7).

The chapters composing the second part of this thesis detail the propositions and evaluations that materialize and assess our approach (*cf.* Figure 3.11).

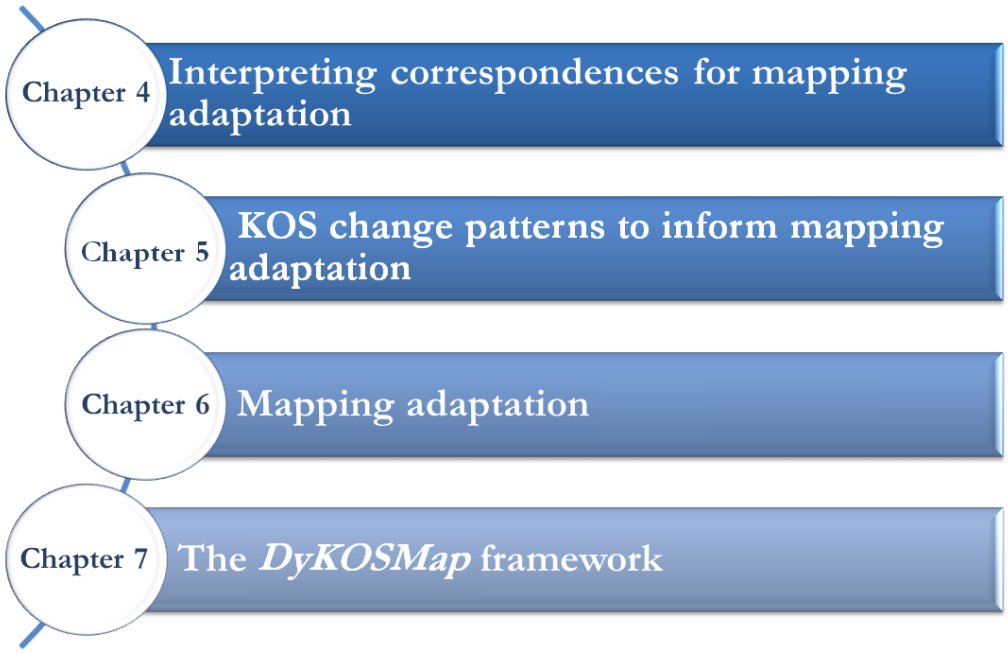


Figure. 3.11: Chapters composing the second part of the thesis

Chapter 4 – Interpreting correspondences for mapping adaptation

This chapter proposes a method to identify relevant KOS's entities that might define established mappings. We assume that the most relevant entities stand for concept attributes that express the closest similarity between involved concepts in mapping. The proposed method, expressed in an algorithm underlay by similarity measure, automatically detects the concept attributes suited to explain a given mapping. In particular, we examine the influence of different similarity metrics in the identification method to compute the relatedness between the attributes' value of interrelated concepts. The chapter presents the proposed algorithm and the adapted similarity metrics studied. We discuss the potential role of the identified attributes for mapping adaptation. In the suggested approach to adapt mappings, we use the results yielded by the identification method (*i.e.*, a minimal set of concept attributes for a given mapping) as an information to support the techniques of adaptation. We also investigate the role played by concepts nearly related with concepts in mappings (parents, children and sibling concepts). The chapter experimentally evaluates the proposed identification method by calculating correlations between KOS changes affecting detected attributes with the observed evolution of associated mappings. The material presented in this chapter was published in [Dinh et al., 2014b].

Chapter 5 – KOS change patterns to inform mapping adaptation

This chapter provides methods to further support characterizing KOS evolution, investigating adequate and fine-grained change operations suited to characterize the evolution of relevant KOS's entities to inform mapping adaptation. We propose two categories of change patterns at the level of concept attributes, formally defining a set of *change patterns* to express different behaviours of the evolution of concepts' attributes. The defined change patterns specifically describe the diffusion of attribute values between concepts in different KOS versions, and how such attributes become more or less semantically specific in their evolution. This chapter conceptualizes methods to automatically recognize instances of the proposed change patterns by comparing successive KOS versions. We inquire whether techniques based on linguistic characteristics of textual values, combined with similarity measures, play a role in supporting automatic change patterns detection at the level of attributes. The chapter experimentally assesses the change pattern recognition methods over a dedicated gold standard extracted from real-world biomedical KOSs. This chapter's content has been partially published in [Dos Reis, 2013, Dinh et al., 2014a].

Chapter 6 – Mapping adaptation

This chapter presents the proposed techniques to adapt mapping in the *DyKOSMap* approach. We investigate, define and formalize a set of mapping adaptation actions as the framework basis to change mapping's elements. To know the necessary conditions to apply the mapping adaptation actions, we describe a series of experiments to study specific factors that may influence applying one action or other. On this ground, we design and formalize so-called *heuristics* with techniques that model rules, as a combination of the studied conditions to guide and take semi-automatic decisions on mapping adaptation. The heuristics take into consideration the evolution of KOS's entities that allow to explain the definition of mappings, based on KOS changes and patterns affecting such entities. The chapter validates the proposed heuristics by evaluating the correctness of the suggested mapping adaptation actions in mapping adaptation. We measure to which extent the proposed heuristics impact the quality of the results yielded by mapping adaptation. The publication [Dos Reis et al., 2013a, Dos Reis et al., 2015a] refers to the basis of this chapter, and part of this chapter's content has been submitted to an international journal.

Chapter 7 – The *DyKOSMap* framework

This chapter introduces the *DyKOSMap* framework in detail. We highlight the integration of the suggested components illustrating the whole mapping adaptation process. The chapter proposes specific algorithms able to handle revision and removal of concepts and attributes from KOS evolution, to semi-automatically adapt mappings based on the proposed techniques in the framework. First, we assume that by accurately identifying and characterizing the evolution of relevant KOS's entities defining mappings might enable attaining a more automated, adequate and reliable mapping adaptation system. Second, taking decisions of adaptation, considering mappings in an individual manner, may improve the mapping adaptation accuracy (*i.e.*, individually analyze each correspondence with the affecting KOS evolution operations to apply the adequate mapping adaptation actions). For example, even though two mappings are under the influence of the same KOS complex change (*e.g.*, a split of concepts), actions applied for modifying these mappings can differ. We describe the implementation aspects of the framework within the developed software prototype. The chapter presents a final experimental evaluation to show a global validation of the framework, which demonstrates that the framework produces meaningful results in mapping adaptation. The publications [Dos Reis et al., 2012, Dos Reis et al., 2014a] are related to this chapter.

Chapter 4

Interpreting correspondences for mapping adaptation

Contents

Introduction **83**
4.1 Problem statement and definitions **84**
4.2 Identifying concept attributes relevant to interpret mappings . **86**
4.3 The similarity measures investigated **88**
4.3.1 Character-based edit-distance similarity 89
4.3.2 Word-based edit-distance similarity 90
4.3.3 Knowledge-based similarity 91
4.4 The potential role of relevant attributes identified **92**
4.5 Experimental evaluation **92**
4.5.1 Materials and procedure 93
4.5.2 Experimental results 96
4.6 Discussion **99**
Conclusion **101**

Introduction

The conducted empirical analyses (*cf.* Chapter 3) have shown that although concepts are considered in their entirety, mappings established between biomedical KOSs reveal that only attributes (*i.e.*, textual statements) characterizing concepts are used to define the semantic correspondences. However, when matching systems semi-automatically create mappings, they fail to keep the KOS's entities used to justify such mappings in their definition, thus preventing any future use for maintenance purpose.

Literature offers very little studies with the aim of interpreting mappings (*cf.* Section 2.3). Despite the fact that some surveys have reported on the performance of string-based [Bethea et al., 2006, Lambrix et al., 2008, Cheatham and Hitzler, 2013] and semantic-based [Pesquita et al., 2009] similarity metrics in the course of developing KOS alignment systems, literature requires further studies to understand to which extent similarity measures of different

natures may contribute when underlying a method to interpret mappings.

In this chapter, we address this issue by investigating techniques suited to identify textual statements in concepts that might represent the most meaningful attributes for a given correspondence. We assume that adequately supporting the mapping adaptation task requires the correct identification of these statements. In summary, we make the following contributions in this chapter:

- We propose a method named *topA* (Algorithm 1) to identify a sufficient **S**ubset of **C**oncept **A**ttributes (SCA) relevant for the interpretation of mappings.
- We suggest the adaptation of various similarity measures that might support our method. The measures target the following level of information: the *lexical* level [Levenshtein, 1966], the *syntactic* level [Maedche and Staab, 2002] and the *semantic* level [Jiang and Conrath, 1997]. We study their influence on the performance of our *topA* Algorithm.
- We conduct a set of experiments to assess the quality of the results yielded by the identification method using two biomedical KOSs (SCT and ICD9) and their associated mappings. In particular, we measure interdependencies between KOS changes affecting the identified attributes and adaptation of associated mappings.

We first state the problem, research questions and necessary definitions (Section 4.1). Afterwards, we present our approach to identify relevant attributes for mappings (Section 4.2) and the formalization of the similarity metrics studied with the method (Section 4.3). We show the potential role of detected attributes for mapping adaptation (Section 4.4). Section 4.5 presents the experimental evaluation while section 4.6 discusses the obtained results.

4.1 Problem statement and definitions

We define the **context** of a particular concept $c_i \in C(K_x)$ as the union of the sets of $sup(c_i)$ (direct super concepts), $sub(c_i)$ (direct sub concepts) and $sib(c_i)$ (sibling concepts) of c_i:

$$CT(c_i) = sup(c_i) \cup sub(c_i) \cup sib(c_i) \tag{4.1}$$

where

$$\begin{cases} sup(c_i) & = & \{c_h | c_h \in C(K_x), c_i \neq c_h \wedge \ c_i \sqsubset c_h\} \\ sub(c_i) & = & \{c_h | c_h \in C(K_x), c_i \neq c_h \wedge \ c_j \sqsubset c_i\} \\ sib(c_i) & = & \{c_h | c_h \in C(K_x), c_i \neq c_h \wedge \ sup(c_h) \cap sup(c_i) \neq \varnothing \wedge c_i \notin sup(c_h)\} \end{cases} \tag{4.2}$$

In equation 4.2, the notation $c_i \sqsubset c_h$ stands for "c_i is more specific than concept c_h", *e.g.*, "*hypotension*" is more specific than *"vascular disease"*. These definitions rely on the "is-a" relationship that forms the hierarchical structure of KOSs. Moreover, this structure consists in a *Directed Acyclic Graph* which prevents circular definition of concepts.

The context of a particular concept c_i, denoted as $CT(c_i)$, refers to the set of concepts in the neighborhood of c_i, *i.e.,* direct parents, direct children and sibling concepts. This excludes concepts linked to c_i by other relationships than "is-a" relationship. Indeed, our previous investigations pointed out that concepts outside this context are much more unlikely to impact mappings evolution and are thus less relevant for our investigation in this chapter.

Given a mapping m_{st} between two concepts $c_s \in C(K_S)$ and $c_t \in C(K_T)$, we investigate the task of determining a set of meta-data containing a sufficient number of attributes of c_s or those from its context $CT(c_s)$ to explain the semantic correspondence between c_s and c_t. We call this set of best attributes **topA** (*i.e.*, *top attributes*). We will further use this set for supporting mapping adaptation (*cf.* Section 4.4).

Understanding established mappings which requires maintenance remains an aspect ignored by existing approaches. The description of input mappings lacks data about methods used to generate them, and what conceptual information (characteristics or concepts' attributes) mostly served to define each mapping individually. Despite the importance of explicitly considering this subset of textual statements in mapping adaptation according to our previous experiments, mappings are released along with the source and target KOS without these meta-data defining mappings (*i.e.*, most similar attributes according to our assumption). Therefore, given a mapping, we need to (re-)identify the most relevant attributes of c_s and/or c_t that optimize its semantic confidence.

To adequately address this open issue, we firstly focus on determining attributes belonging to a source concept that could be used as relevant concept attributes for interpreting mappings associated with this concept. This objective requires to investigate the similarity between attributes in the context of the source concept and the ones belonging to the target concept. We will consider the calculated similarity values as a criteria for ranking candidates of relevant attributes. Since similarity measures aim at computing the degree of relatedness between a given pair of objects (*e.g.*, two attribute values), we judge relevant to explore them in this task. This chapter addresses the following open research questions:

1. Is it possible to identify a significant and sufficient subset of attributes of a source concept that served to define a mapping with respect to a target concept and justify the link?
2. Is it useful to consider the context (neighbour concepts) of the source concept in this identification?
3. Is it possible to benefit from existing similarity measures?
4. How and which benefits may yield if explicitly and concretely exploring the identified attributes to support mapping adaptation?

We define the **set of attributes in the context** of a concept c_i as $A_{ct}(c_i)$. Formally:

$$A_{ct}(c_i) = \bigcup_{l=1..n} A(c_l), c_l \in CT(c_i) \tag{4.3}$$

where n refers to the number of concepts in the context of c_i.

We also denote the set of all attributes of concept c_i and the ones in its context as follows:

$$A_{all}(c_i) = A(c_i) \cup A_{ct}(c_i) \tag{4.4}$$

where $A(c_i)$ stands for the set of attributes of concept c_i and $A_{ct}(c_i)$ refers to the set of attributes from each concept in the context of the concept c_i (*cf.* Equation 4.3).

Table 4.1 presents the notations used in the formalization of the similarity measures (*cf.* Section 4.3) and in our algorithm.

Table 4.1: Notations for the formalization of the methods

Notation	Description
Ω	a finite alphabet (universe of characters or symbols)
Υ	a non-empty universe over Ω
x_i	the i^{th} symbol or character ($x_i \in \Omega$)
X	the string $X = x_1x_2...x_n$ of length n ($X \in \Upsilon$)
$\|X\|$	the length of string X in terms of symbols or characters
$X_{w..z}$	a substring of X including characters from x_w to x_z ($x_w, x_z \in \Omega$) where $z < n$
L_X	the set of words/tokens (substrings) of X
λ	the null symbol
a_i	attribute a_i denoted by a string
$a_i.name$	attribute name (string)
$a_i.value$	attribute value (string)
$sim(a_i, a_h)$	similarity between attributes a_i and a_h
w^j_{ki}	single word/token w_k from attribute value $a_i.value$ at time j
$C(K^j_x)$	set of concepts of KOS K_x at time j
$A(c^j_k)$	set of attributes of the concept c_k at time j
$R(c^j_k)$	set of relationships of the concept c_k at time j
$\mathcal{M}^j_{ST}$	set of mappings between KOS K_S and K_T at time j

This table presents the notations relevant for this chapter and their descriptions.

4.2 Identifying concept attributes relevant to interpret mappings

We present our identification algorithm aiming to detect the relevant attributes to interpret a mapping (Algorithm 1). We assume that mappings are established according to the degree of similarity between a subset of attributes' values, belonging to the interrelated concepts or their context. Thus, the set of most similar concept attributes between source and target concepts defines a given mapping. For this purpose, we use similarity measures to quantify the semantic relatedness between values of concepts' attributes.

On this ground, we aim to empirically observe whether a KOS change, affecting any of those identified attributes in **topA**, leads to the adaptation of mapping (Section 4.5). To examine possible interdependencies between KOS changes specifically affecting attributes and mapping adaptation, we suppose that c_t remains unchanged while c_s evolves.

We define $top_A(c_s, c_t, n)$ as the set of top n attributes that may come from concept c_s or its context, and that are the more similar to those of c_t. Thus, this set corresponds to the subset of concept attributes (**SCA**) defining a mapping. Each attribute $a_p \in A(c_s)$ can have a particular similarity value with each attribute $a_q \in A(c_t)$. We compute the similarity between each attribute value $a_p \in A_{all}(c_s)$ and $a_q \in A(c_t)$ exploring a similarity function. We let the type of the considered attributes (*e.g.*, label, synonym) as a parameter in the system, and we are interested in the attributes' value. Therefore, when we refer to an attribute a_i, we mean its value. The similarity function $sim(X, Y)$ calculates the semantic relatedness between two given strings X and Y. This returns a value ranging from 0 to 1 and as higher this value is higher is the similarity between the given strings. Figure 4.1 illustrates a scenario of identifying concepts'

attributes, where we highlight a few relevant attributes.

We denote (a_p, s_{a_p}, ct_{a_p}) as the set of parameters related to attribute $a_p \in A_{all}(c_s)$. The similarity value s_{a_p} refers to best similarity of attribute a_p with the most similar attribute in concept c_t, when the target concept c_t (all its attributes) remains unchanged while the source concept c_s (at least one attribute) changes. The argument ct_{a_p} accounts for the context where we find the attribute a_p. We set ct_{a_p} to $NOCT$ (not in the context) if $a_p \in A(c_s)$, *i.e.*, attribute a_p belongs to concept c_s, but not to its context.

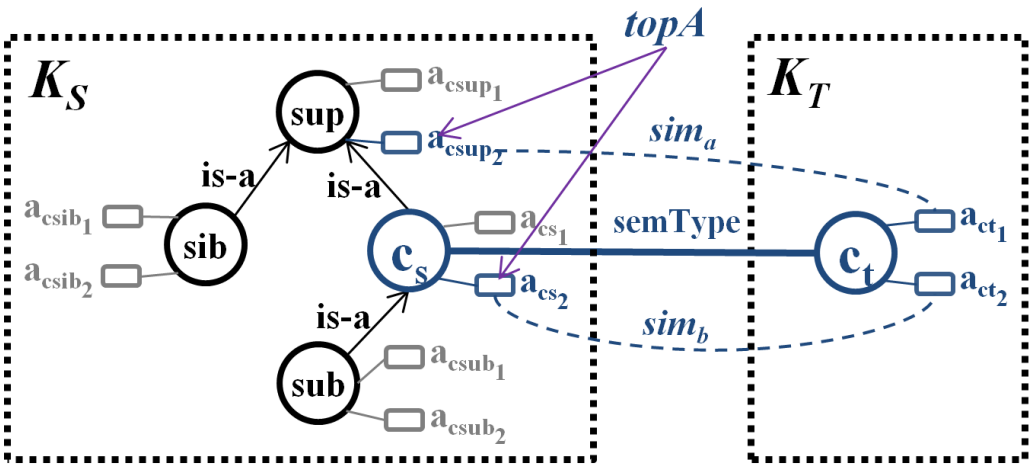


Figure. 4.1: Relevant attributes between source and target concepts

This figure presents the identification of relevant concept attributes defining the mapping between concept c_s and c_t. Candidate attributes may come from concept c_s or its context. The most similar candidate attributes constitute $topA$, *e.g.*, $topA(c_s, c_t, 2) = \{a_{csup2}, a_{cs2}\}$ because sim_a and sim_b refer to the best similarity values reflecting the similarity between c_s and c_t, hypothetically in this illustration.

Algorithm 1 describes the designed method for selecting the top attributes of the source concept c_s or its context $CT(c_s)$. After calculating the similarity between each attribute of c_s and the ones in c_t (lines 3-12) through the similarity function $sim(X, Y)$, the algorithm tries to find the best candidate attribute in c_s. We formally define the strings X and Y as follows:

$$\begin{cases} X = a_p.value, a_p \in A_{all}(c_s) \\ Y = a_q.value, a_q \in A(c_t) \end{cases} \tag{4.5}$$

In the algorithm, if there is no exact match (*i.e.*, $maxSim < 1$), then it calculates the similarity between each attribute in the context of c_s and each attribute of c_t (lines 14-21). Finally, the algorithm sorts the set S according to the calculated similarity values and returns the top n attributes.

To examine the most adequate similarity measures to use in the proposed algorithm, we selected, adapted and evaluated different well-known measures. Note that the possible similarity measure to use is customizable in our algorithm. We consider three different similarity measures: **character-based**, **word-based** and **semantic-based** similarity. We discuss the motivations leading to the chosen measures (*cf.* Section 4.3).

Algorithm 1 Select top n attributes defining a given mapping

Require: $m_{st} = (c_s, c_t, semType) \in \mathcal{M}^0_{ST}; c_s \in C(K^0_S); c_t \in C(K^0_T), n \in \mathbb{N}$
Ensure: $S = \{(a_1, s_{a_1}, ct_{a_1}), (a_2, s_{a_2}, ct_{a_2}), ..., (a_n, s_{a_n}, ct_{a_n})\}$

```
S ← ∅; {Initialize the final result set}
{Compute similarity between attributes in c_s and c_t}
for all a_p ∈ A(c_s) do
    maxSim ← 0;
    for all a_q ∈ A(c_t) do
        s_p ← sim(a_p, a_q);
        S ← S ∪ {(a_p, s_{a_p}, NOCT)};
        if maxSim < s_p then
            maxSim ← s_p;
        end if
    end for
end for
{Select attributes in context if exact matches are not found}
if maxSim < 1.0 then
    for all a_k ∈ A_ct(c_s) do
        for all a_q ∈ A(c_t) do
            s_k ← sim(a_k, a_q);
            S ← S ∪ {(a_k, s_{a_k}, ct_{a_k})};
        end for
    end for
end if
S ← sort(S, n); {Select top n attributes}
```

4.3 The similarity measures investigated

We first investigate the edit-distance between attributes' value characterizing concepts. We observe the value of p^{th} attribute a_p (denoted by string X) in concept c_s or its context and the value of q^{th} attribute a_q (denoted by string Y) issued from concept c_t (*cf.* Equation 4.5). For example, given two attribute's values X="*tracheal stenosis following tracheostomy*" and Y="*tracheal stenosis due to tracheostomy*", the similarity function expresses to which extent these values are similar or different in terms of lexical similarity. The more similar they are, the more they are considered as related in the sense that we can account a semantic relation between the two underlying concepts. For this purpose, we use the well-known *string edit-distance measure*, but at the level of characters [Levenshtein, 1966] (Section 4.3.1).

Second, the similarity function must also take into account the level of words in addition to the differences at the level of characters. For example, we must consider the following attribute's values "*skin cancer*" and "*cancer of the skin*" as equivalent ones. Indeed, the edit distance at the level of characters fails to allow coping with the word order issue in those terms. Therefore, we should consider the distance at the level of words to provide a summarizing figure for the lexical level [Maedche and Staab, 2002] (Section 4.3.2).

Third, we also evaluate the impact of using semantic similarity for identifying relevant attributes related to each mapping, by considering the semantic information in common between the attributes' value. This refers to the conceptual similarity between strings, to which Resnik [Resnik, 1995a] proposed to determine by using the information content approach. For example, the two following attribute's values "*bone of the extremity*" and "*limb bone*" semantically refer to the same concept and shall be considered as equivalent because "*extremity*" and "*limb*" are semantically related. To compute this semantic distance, we further need an external lexical knowledge resource, *e.g.*, *WordNet*[29] [George, 1995] and an annotated corpus such as *SemCor* [Miller et al., 1993] (Section 4.3.3).

4.3.1 Character-based edit-distance similarity

Several related applications require to determine the similarity between two strings, such as pattern recognition, information retrieval, ontology alignment, *etc.* A widely-used notion of string similarity stands for the edit distance or the Levenshtein [Levenshtein, 1966]. It corresponds to the minimum number of edit operations (insertion, deletion, and substitution of individual symbols) required to transform one string into the other. Mathematically, the string-based or character-based edit distance between two strings X and Y, denoted as $ED(X,Y)$, can be defined as the minimum weight of transformation through a sequence of weighted edit operations.

A normalised $ED(X,Y)$ aims at ensuring that it provides values in the interval [0..1]. We adopt the normalization of the character-based edit distance similarity measure between two strings X and Y, denoted as $sim_{CED}(X,Y)$, defined as follows:

$$sim_{CED}(X,Y) = \frac{2 * ED(X,Y)}{(\|X\| + \|Y\|) + ED(X,Y)} \tag{4.6}$$

The advantage of such normalisation relies on the fact that it allows to measure the similarity between strings using a genuine metric [Marzal and Vidal, 1993] with a degree of similarity between 0 and 1, where 1 stands for an exact match. We formally define the edit distance between two strings X and Y as follows:

$$ED(X,Y) = min\{\gamma(T_{X,Y})\} \tag{4.7}$$

where

- $T_{X,Y} = T_1 T_2 ... T_n$ is an edit path, *i.e.*, the sequence of atomic edit operations transforming X into Y;
- $\gamma(T_{X,Y}) = \sum_{i=1}^{n} \gamma(T_i)$ corresponds to the cost function of an edit transformation $T_{X,Y}$ from X to Y, which is the sum of the individual cost of each edit operation $\gamma(T_i)$ [Marzal and Vidal, 1993].

An atomic edit operation T_i can be among the three following ones: *insertion, deletion* and *substitution.* Formally:

$$\begin{array}{lll} \lambda \longrightarrow a & (\forall a \in \Omega): & insertion \\ a \longrightarrow \lambda & (\forall a \in \Omega): & deletion \\ a \longrightarrow b & (\forall a, b \in \Omega): & substitution \end{array} \tag{4.8}$$

[29] `wordnet.princeton.edu`

where

- Ω is a finite set of characters or symbols, *e.g.*, $\Omega = \{a, b, c, ..., 1, 2, 3, etc.\}$.
- a, b are characters or symbols belonging to Ω.
- λ is the null symbol.

$ED(X, Y)$ becomes a metric if the following conditions are satisfied:

$$\begin{array}{l} \forall a, b \in \Omega, \gamma(a \longrightarrow a) = 0 \\ \gamma(a \longrightarrow b) > 0 \; if \; ((a \neq b) \wedge (\gamma(a \longrightarrow b) = \gamma(b \longrightarrow a))) \end{array} \tag{4.9}$$

4.3.2 Word-based edit-distance similarity

The *character-based edit distance* (*cf.* Section 4.3.1) provides an effective way to compute the semantic relatedness between similar strings, especially for single-word strings. However, it fails to take into account the syntactic information in compound terms. For example, it results in a very low degree of similarity between two synonymous terms, but with a different word order, *e.g.*, "*cancer of the skin in the face*" *vs.* "*face skin cancer*".

To cope with this issue, we examine another similarity function that determines the extent to which the symbols in each substring $w_i \in L_X$ in the first string X are similar to the symbols in each substring $w_j \in L_Y$ of the second string Y. The underlying assumption states that: if the first string contains all substrings of the other, it is likely that they are semantically related [Maedche and Staab, 2002]. Originally proposed in [Maedche and Staab, 2002], we adapt the word-based edit-distance similarity measure, by extending the character-based similarity. This enables us to evaluate the performance of our Algorithm 1, by comparing these similarity measures underlying the algorithm.

We define the *word-based edit-distance similarity measure* (namely syntactic measure) between two strings X and Y, denoted $sim_{WED}(X, Y)$, as follows:

$$sim_{WED}(X, Y) = \frac{1}{\|L_X\|} \sum_{w_i \in L_X} \max_{w_j \in L_Y} sim_{CED}(w_i, w_j) \tag{4.10}$$

where

- $sim_{CED}(w_i, w_j)$ refers to the normalised character-based edit-distance between single words (substrings or tokens) in w_i and w_j in strings X and Y respectively (*cf.* Equation 4.9);
- L_X and L_Y are the sets of words in X and Y, respectively;
- $\|L_X\|$ is the length of string X in terms of tokens.

Note that in equation 4.9, the normalization by $\|L_x\|$ aims to ensure that $sim_{WED}(X, Y)$ provides values in the interval [0..1]. The *max* function ensures that w_i and w_j constitute the couple of most similar substrings in X and Y.

4.3.3 Knowledge-based similarity

Using lexical knowledge resources to compute the similarity can help to determine the semantic relatedness in a semantic network, constructed by a taxonomy of classes. Combining corpus statistical information can enhance the resources, because the computational evidence derived from a distributional analysis of corpus data can better quantify the semantic distance between nodes in the semantic network.

We propose to adapt the original Jiang-Conrath similarity [Jiang and Conrath, 1997] measure, by normalizing the similarity with the length of the source string, so that the similarity values stay between 0 and 1.

We define the semantic similarity measure between two strings X and Y, denoted as $sim_{SEM}(X,Y)$, as follows:

$$sim_{SEM}(X,Y) = \frac{1}{\|L_X\|} \sum_{w_i \epsilon L_X} \max_{w_j \epsilon L_Y} sim_{JCN}(w_i, w_j) \tag{4.11}$$

where $sim_{JCN}(w_i, w_j)$ stands for the Jiang & Conrath's knowledge based similarity between two words w_i and w_j referring to two concepts in the same ontology. Formally:

$$sim_{JCN}(w_i, w_j) = max\{[IC(w_i) + IC(w_j) - 2 * IC(lcs(w_i, w_j))]^{-1}\} \tag{4.12}$$

where

- $IC(w_k)$ refers to the information content of word w_k denoting a particular concept in the semantic network, computed as follows:

  $$IC(w_k) = -logP(w_k) \tag{4.13}$$

  where $P(w_k)$ calculates the probability of encountering an instance of a concept denoted by the single-word term w_k in a corpus [Jiang and Conrath, 1997].

- $lcs(w_i, w_j)$ stands for the lowest common subsumer, *i.e.*, the lowest concept in the hierarchy subsuming[30] both concepts denoted by w_i and w_j.

Computing the information content in equation 4.13 [Resnik, 1995b] demands having a large corpus. In the biomedical domain, there exists an annotated corpus namely *GENIA* corpus [Kim et al., 2003]. However, the semantic annotation is only limited to entities of interest in molecular biology such as proteins, genes and cells constituting the *GENIA* ontology. In the conducted experiments, we accept *SemCor* as the underlying corpus because it has been widely used for computing the similarity measures based on external resources [Li et al., 2006], and can be suited to our need. *SemCor* consists of a text corpus which has been semantically annotated with information about *Part-Of-Speech* (*i.e.*, noun, verb, adjective and adverb), lemma and *WordNet* synset [Miller et al., 1993]. This is composed of 352 texts and includes more than 200.000 sense-tagged words.

[30] subsume means contain, comprise or include, *e.g.*, lcs("car", "bicycle") = "vehicle"

4.4 The potential role of relevant attributes identified

To the best of our knowledge, no approach in literature has asserted the relevance of considering conceptual information explaining established mappings to support and increase the precision of mapping adaptation. Additional research efforts improving this aspect can bring novel insights on techniques to automatically tackle mapping adaptation.

Identified attributes can play a major role for mapping adaptation. An important issue of mapping adaptation concerns determining in which situation a mapping should be adapted. The set of identified attributes provided by our algorithm may play a role to decide whether the mapping remains semantically valid or not.

In some cases, the mapping adaptation decisions could easily be made, *e.g.*, when a source concept is totally removed, we can suggest removing associated mappings. Although one could judge trivial to identify the cases where no KOS changes affect the concepts involved in mappings, and would recommend the direct reuse of such mappings, other scenarios pose much more complicated situations to decide the impact of KOS changes with difficulties to determine the effects on the associated mappings, *e.g.*, when concepts are only affected by changes in attributes. Let us imagine a source concept of a mapping containing a set of five attributes describing it. Even if we could compute that a user has deleted two attributes among them, from one KOS version to another, it is still required to determine whether such deletion of attributes impacts this particular mapping and semantically invalidates it. We can find situations where even when changes affect a source concept, these changes fail to really impact associated mappings. In these situations mappings should not be removed.

The SCA relevant attributes identified might help to adapt mappings, since we can exploit changes in attributes' values as well as their similarity with attributes in the corresponding target concept. We will consider as unaffected mappings those with relevant attributes that remain unchanged from one KOS version to another (*cf.* Section 6.4). On the other hand, it is more likely to adapt a mapping when a change affects an identified attribute for such mapping.

To better illustrate the role of the attributes and possible interdependencies between KOS changes affecting relevant attributes and changes in mappings relating to the corresponding source and target concepts, Table 4.2 provides an example where an attribute in SCA of the source concept changes from one KOS version to another (it is removed) and in consequence the corresponding mapping changes.

Our experimental evaluation in this chapter (next section) will show the benefits of considering such attribute-based approach to support mapping adaptation.

4.5 Experimental evaluation

We present the experiments conducted to empirically evaluate the proposed approach to identifying concepts' attributes defining mappings. We compare the performance of similarity measures implemented in our algorithm examining the context of concepts in mappings, *i.e.*, attributes denoting concepts in the neighbourhood of a mapped concept. We study the impact of the different similarity measures (*cf.* Sections 4.3.1, 4.3.2, 4.3.3) considered through the correlation

Table 4.2: Correlation between relevant attributes and associated mappings

Before evolution of source concept	
$topA$(SCT:422338006^0)	m^0_{st}
a_1 = Senile macular retinal degeneration$_{\#[0.5493]}$ a_2 ← Degenerative disorder of macula (disorder)$_{\#[0.3714]}$ a_3 ← Degenerative disorder of macula$_{\#[0.3714]}$	(422338006, 362.50, '≤')
After evolution of source concept	
$topA$(SCT:422338006^1)	m^1_{st}
a_1 ← ~~Senile macular retinal degeneration~~$_{\#[0.5493]}$	
a_2 ← Degenerative disorder of macula (disorder)$_{\#[0.3714]}$ a_3 ← Degenerative disorder of macula$_{\#[0.3714]}$	(422338006, 362.50, '≡')

This table shows an example of the potential relevance of exploring the relevant attributes. A change in attribute a_1 of concept source "422338006" in SCT (***delA***) triggers a ***change in the relation*** between source and target concepts in their mapping because attribute a_1 has been used to define the mapping ($\#topA = 1$). In this example, no changes exist in target concept ($A(362.50^0) = A(362.50^1) =$ {"**Macular degeneration (senile)**, unspecified"}).

between changes in mappings and modifications in candidate attributes (*i.e.*, KOS change operations affecting attributes – Table 2.1). Our experiments evaluate to which extent changes affecting relevant attributes, susceptible for defining mappings, influence their adaptation.

We first describe the used dataset and the experimental procedure (Section 4.5.1) followed by the results (Section 4.5.2).

4.5.1 Materials and procedure

Similar to our previous experiments (*cf.* Section 3.1.1), the conducted evaluation consider two biomedical KOSs namely SCT and ICD9, and the official mappings between them.

From available data, we use the first release and the last, respectively. In contrast to our previous experiments, we do not use every single release to run our evaluation in this chapter to simplify results analysis. We consider the following versions of SCT and ICD9: SCT released in January 2010 and in January 2012, and ICD9 released in 2009 and in 2011. Therefore, our experiments are based on the two sets of mappings between SCT and ICD9, which have been established by experts at *IHTSDO*: $\mathcal{M}^0_{ST} = (SCT\ Jan.\ 2010, ICD\ 2009)$ and $\mathcal{M}^1_{ST} = (SCT\ Jan.\ 2012, ICD\ 2011)$. Table 4.3 presents some statistics about the dataset.

In the conducted experiments, we study the evolution of both KOSs. In particular, our focus is on source concepts involved in mappings. We conducted the following steps to select the adequate set of mappings:

1. Based on the two releases of the same KOS, we identify the set of concepts that change from one version to another. We name this subset of concepts as $diff(K^0_x, K^1_x)$ that is obtained using the *Conto-Diff* tool [Hartung et al., 2013]. To this end, we needed to develop a script that transforms the original KOS sources into OBO format to enable the use of *Conto-Diff* tool. We are particularly interested in exploiting the two KOS change operations namely

Table 4.3: Statistics on the studied datasets

Characteristics \ Mappings	$\mathcal{M}^0_{ST}$		$\mathcal{M}^1_{ST}$	
	SCT'10	**ICD9'09**	**SCT'12**	**ICD9'11**
Nr. of concepts	390 022	12 734	395 346	13 059
Nr. of $\sqsubset$ relationships	530 433	11 619	567 719	11 962
Nr. of attributes	1 547 855	34 046	1 570 504	34 944
Nr. of mappings	100 451		102 703	

This table shows statistics about the dataset including KOSs and associated mappings.

delA and *chgAttValue* because they directly impact on the values of attributes that can serve for defining mappings. (*cf.* Table 2.1)

2. From the $diff(K^0_x, K^1_x)$, we remove all unassociated concepts with a mapping, since we consider that only concepts associated with mappings can impact the validity of mappings.

3. We further remove from the $diff(K^0_x, K^1_x)$ added and removed concepts. The correlation between these two change operations and mapping adaptation was already studied in our previous experiments (*cf.* Chapter 3). We focus on analysing mapping adaptation particularly on those mappings associated only with concepts that have theirs contents somehow modified, *i.e.*, affected by the change operations concerning addition and removal of attributes.

4. Finally, to avoid misunderstandings on the results, we consider only those mappings where the target concepts remained unchanged from one version to another. Therefore, we removed from the analysis mappings where source and target concepts simultaneously change. This results in a final subset of modified concepts from $diff(K^0_x, K^1_x)$ containing mappings associated with them. These mappings are represented by the set $\mathcal{M}_{affected}$.

We conduct these steps with both KOSs in an isolated way. We present the achieved results in section 4.5.2. From now on in this thesis, we only refer to the set of mappings $\mathcal{M}_{affected}$ associated with the concepts of the calculated and filtered $diff(K^0_x, K^1_x)$ for both KOSs. We analyze a total number of 6 672 mappings in the resulted set of mappings regarding SCT and 3 788 for ICD9, respectively. Note that no intersection exists between these sets of mappings.

We aim to examine the way that we can adapt mappings under evolving KOS, by considering changes affecting relevant attributes (SCA). To this end, we measure the correlation between changes detected in identified attributes (*i.e.*, *delA*, *chgAttValue*) and the adaptation of the associated mapping. We also assume as a correlation when both the identified attributes and the associated mapping remain unchanged. The objectives of our experiments are two-fold:

1. We evaluate the utility of using different similarity measures to identify relevant attributes (SCA) for supporting mapping adaptation.

2. We evaluate the impact of using the context (*cf.* Equation 4.2) of source concepts on the yielded correlations.

The experiments may allow us to observe several aspects:

1. The quality of the identified SCA by the proposed method (*cf.* Algorithm 1);
2. A comparison among the three types of similarity measures studied and the two KOSs of the experimental dataset;
3. Most importantly, to which extent mapping adaptation should rely on the identification of relevant attributes.

We calculate the different types of correlations between changes affecting the attributes in SCA and mapping adaptation. To this end, we count the number of mappings that changed when at least one attribute of SCA has changed (*i.e.*, its value has been modified or deleted), in addition to the number of mappings that remained unchanged when particularly any attribute of SCA changed. Note that the source concept of all the analyzed mappings belongs to the subset $diff(K_x^0, K_x^1)$, which means that at least one attribute of such concept has been changed. However, this attribute may belong to SCA or not, which justifies the different types of correlations proposed. We characterize a mapping change when a mapping is removed from one release to another, or when the source concept is replaced by another one, or either the semantic relation in the mapping is modified. Formally, we define the different types of correlations as follows:

$$\begin{aligned} \#\textit{change-correlation} \quad &= \|\{m_{st}^0 \in \mathcal{M}_{ST}^0 | (\forall m_i^1 \in \mathcal{M}_{ST}^1, m_i^1 \neq m_{st}^0) \wedge \\ &(\exists a_k \in topA(c_s^0, c_t^0, n), a_k \notin A_{all}(c_s^1))\}\| / \|\mathcal{M}_{affected}\| \end{aligned} \tag{4.14}$$

$$\begin{aligned} \#\textit{unchange-correlation} \quad &= \|\{m_{st}^0 \in \mathcal{M}_{ST}^0 | (\exists m_{st}^1 \in \mathcal{M}_{ST}^1, m_{st}^1 = m_{st}^0) \wedge \\ &(\forall a_k \in topA(c_s^0, c_t^0, n), a_k \in A_{all}(c_s^1))\}\| / \|\mathcal{M}_{affected}\| \end{aligned} \tag{4.15}$$

$$\#\textit{total-correlation} = \#\textit{change-correlation} + \#\textit{unchange-correlation} \tag{4.16}$$

The *change-correlation* in equation 4.14 corresponds to the percentage of adapted mappings that have at least one modified attribute in SCA. Similarly, the *unchange-correlation* in equation 4.15 corresponds to the percentage of unchanged mappings because no changes occurred in attributes identified in **topA**. Finally, the *total-correlation* in equation 4.16 calculates the sum of the two previous types of correlation.

For each one of the considered KOSs, we first compute the *total-correlation* values for mappings of $m_{st} \in diff_M$ without using the contextual information of source concepts (denoted as ***NOCT***). Afterwards, we measure the correlations obtained by using solely the different concepts from the context (denoted as **SUP** for *super concepts*, **SUB** for *sub concepts*, and **SIB** for *sibling concepts*). For the *unchange-correlation*, we measure the contextual information considering the retrieved attribute with the highest similarity value. For instance, if our algorithm 1 finds such attribute in a super concept, then we compute the correlation for this type of concept. We compare the contributions of the different types of context (SUP, SUB, SIB) separately, and with the obtained results without using the context (***NOCT***). Finally, we combine all conceptual information together to build the SCA as a whole (denoted as **ALL**). This allows evaluating the influence of the context for supporting mapping adaptation. The size of the set of relevant attributes in SCA examined (denoted as $\#topA$) refers to a parameter taking integer values ranging from 1 to 10 (n).

4.5.2 Experimental results

We first present the results concerning the analysis of *total-correlation* (Figure 4.2 and Figure 4.3). Afterwards, we present the results with respect to a more detailed analysis on the *change-correlation* (Figure 4.4 and Figure 4.5).

Figure 4.2 depicts the results obtained by computing the ***total-correlation*** (*cf.* Equation 4.16) with respect to SCT and figure 4.3 presents the results for ICD9. More specifically, figure 4.2.1 and figure 4.3.1 present the results without considering the context (NOCT) by comparing the different similarity measures. This allows us to point out three main aspects:

1. The similarity measures perform differently for SCT and ICD9. If we do not consider the information about the context of mapped concepts (without CT, *cf.* Figure 4.2.1), we observe that the *Character-based ED* outperforms the *Word-based ED* and *Knowledge-based* similarity for SCT while we observe the inverse findings for ICD9, but with a slight difference between the measures (*cf.* Figure 4.3.1).

2. If we consider the parents of source concepts (SUP), the *Knowledge-based* similarity outperforms the two other similarity measures with an improvement rate of ~9% for SCT while no improvement is observed for ICD9. In addition, children (SUB) and sibling (SIB) concepts are not useful for identifying SCA. This indicates the usefulness of considering the context, more precisely the super concepts of mapped concepts in the mapping adaptation process.

3. The percentage of found correlations with respect to the number of analyzed mappings differs comparing SCT (~25% without context and ~90% with context in addition) and ICD (~3% without context and ~5% with context in addition). This observation is coherent with the statistics about SNOMED-CT and ICD9-CM. Indeed, SNOMED-CT stands for a much larger KOS than ICD9-CM in terms of number of attributes and concepts, number of associated mappings (*cf.* Table 4.3).

The results indicate several interesting facts by analysing the performance of the similarity measures, considering the different concepts retrieved from the context for the source concepts in mappings. Figure 4.2.2 and figure 4.3.2 show that by considering ALL attributes from concepts involved in mappings and their context, we obtain the best correlation counts for both SCT and ICD9.

These results also show that while the super concepts play a major role in SCT (*cf.* Figure 4.2.2), in ICD the correlations regarding the sibling concepts dominated, but they still remain less expressive than correlations found only considering the source concept (NOCT) for ICD9 (*cf.* Figure 4.3.2).

At this level, we aim to observe the results only for the *change-correlation* in SCT (*cf.* Figure 4.4) and ICD9 (*cf.* Figure 4.5), respectively. We skip the *unchange-correlation* analysis because this seems a more trivial scenario for us that, if detected attributes denoting a concept involved in a mapping remain unchanged, the associated mappings might also remain unchanged.

In contrast to the results on the analysis of *total-correlation* for SCT, results in *change-correlation* illustrate that the *Word-based ED* performs better than the *Character-based ED*, but

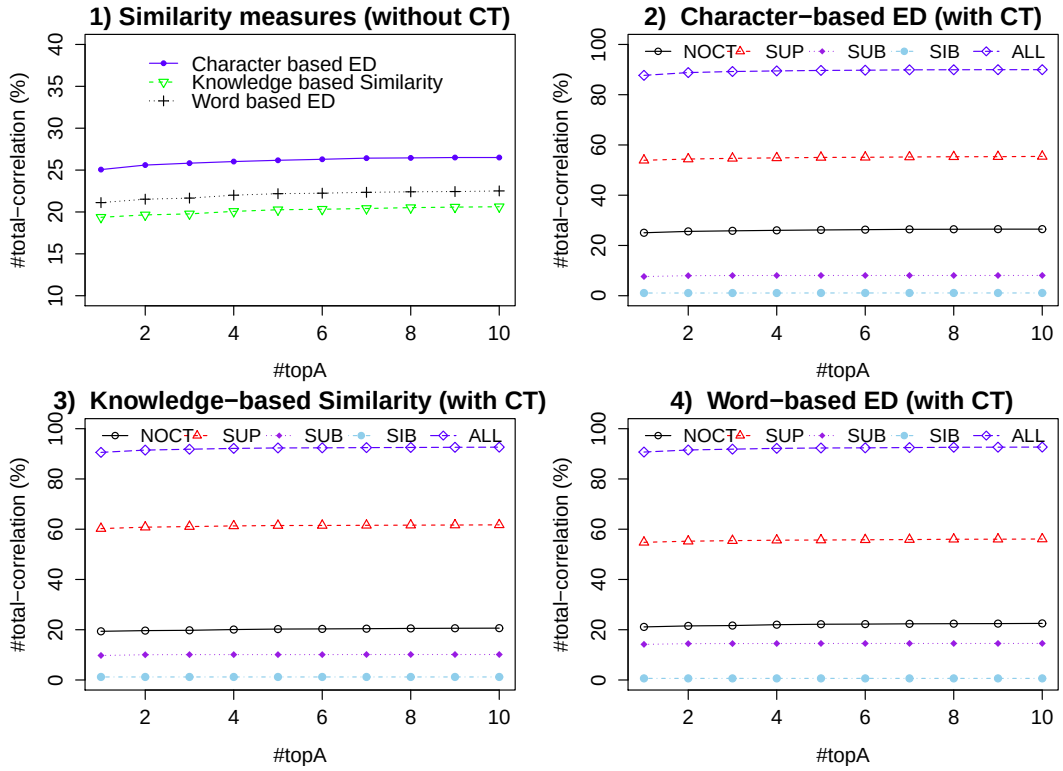


Figure. 4.2: Performance of the *topA* method analysing the evolution of SCT

This figure presents results of applying algorithm 1 using three different similarity measures and the impact of using context (CT) for understanding the **correlation** of **mappings evolution** with the **evolution of SNOMED-CT** based on relevant attributes. The *#topA* corresponds to the number of candidate relevant attributes observed. The *#total-correlation* refers to the percentage of correlations between the number of changed and unchanged mappings and the number of changed and unchanged candidate attributes in mapped concepts and/or their context (*cf.* Equation 4.16). The plot on the top left (1) compares the performance of the three similarity measures, while the remainder plots (2, 3, 4) show the performance of each similarity measure individually by comparing the different types of context where the candidate attribute is found (**SUP**, **SUB**, **SIB** concept or from the source concept without using its context (**NOCT**) or they can come from both (**ALL**)).

slightly different to the *Knowledge-based* similarity measure. The *Word-based ED* potentially leverages the number of *change-correlation* because the considered attribute values of changed attributes are processed by the *diff* tool and some stopwords are not considered. However, similar to the results on the analysis of *total-correlation* regarding ICD9, results illustrate that the *Word-based ED* and *Character-based ED* perform slightly different.

Analyzing the influence of context for identifying relevant attributes in *change-correlation*, we observe that attributes in the source concepts yield the best correlations rather than the ones in their context. This observation is coherent for the three similarity measures used as depicted in Figures 4.4.2, 4.4.3 and 4.4.4 for SCT. Indeed, the *NOCT* curve remains always above other curves corresponding to each type of context (*i.e.*, SUP, SUB, SIB). We observe the same results on ICD9 with the *Character-based ED* similarity (*cf.* Figure 4.5.2) with the exception that *NOCT* and *SUP* tend to perform similarly both for *Knowledge-based* and *Word-based ED*

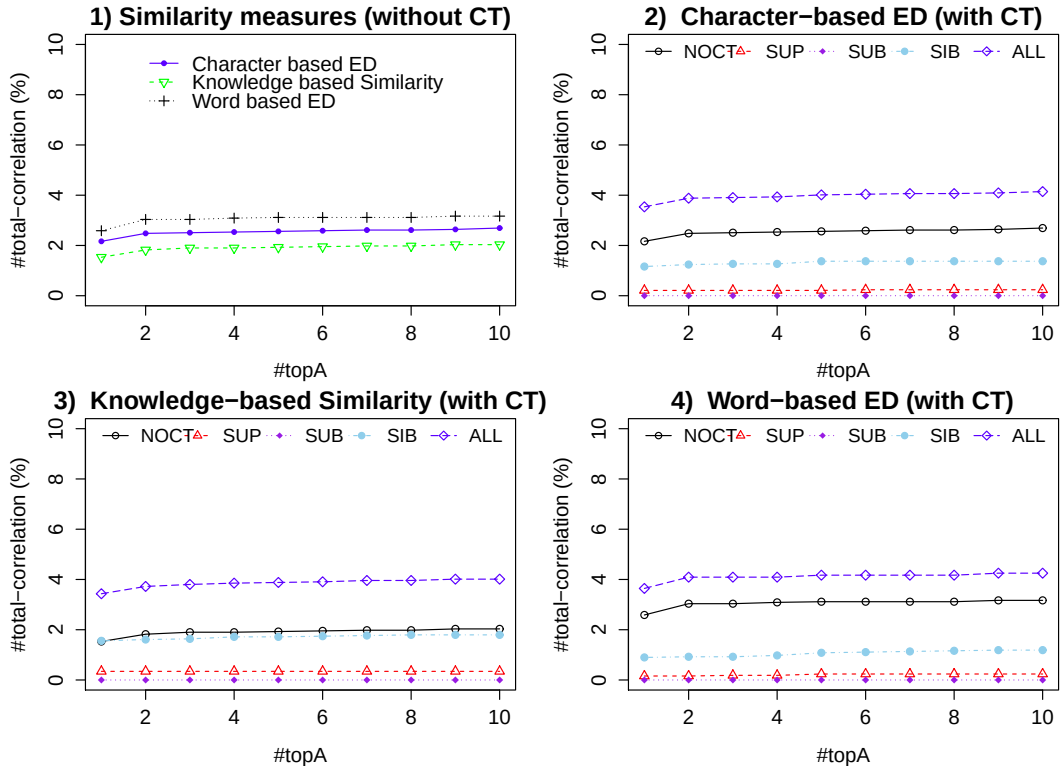


Figure. 4.3: Performance of the *topA* method analysing the evolution of ICD9

This figure presents results of applying algorithm 1 using three different similarity measures and the impact of using context (CT) for understanding the **correlation** of **mappings evolution** with the **evolution of ICD9-CM** based on relevant attributes. The *#topA* corresponds to the number of candidate relevant attributes observed. The *#total-correlation* refers to the percentage of correlations between the number of changed and unchanged mappings and the number of changed and unchanged candidate attributes in interrelated concepts and/or their context (*cf.* Equation 4.16). The plot on the top left (1) compares the performance of the three similarity measures, while the remainder plots (2, 3, 4) show the performance of each similarity measure individually by comparing the different types of context where the candidate attribute is found (**SUP**, **SUB**, **SIB** concept or from the source concept without using its context (**NOCT**) or they can come from both (**ALL**)).

similarity measures (*cf.* Figure 4.5.3 and Figure 4.5.4).

Overall results indicate that both lexical and syntactic measures tend to outperform the *Knowledge-based similarity*. The latter mainly relies on the similarity between senses of concepts in *WordNet*, and the exploitation of the term co-occurrence in the *SemCor* corpus. The low performance of the *Knowledge-based similarity* measure quality is probably because of the lack of domain information, *i.e.*, the biomedical knowledge available in the training data as well as in the underlying knowledge source. This underscores the relevance of using several combined similarity measures to leverage our identification method.

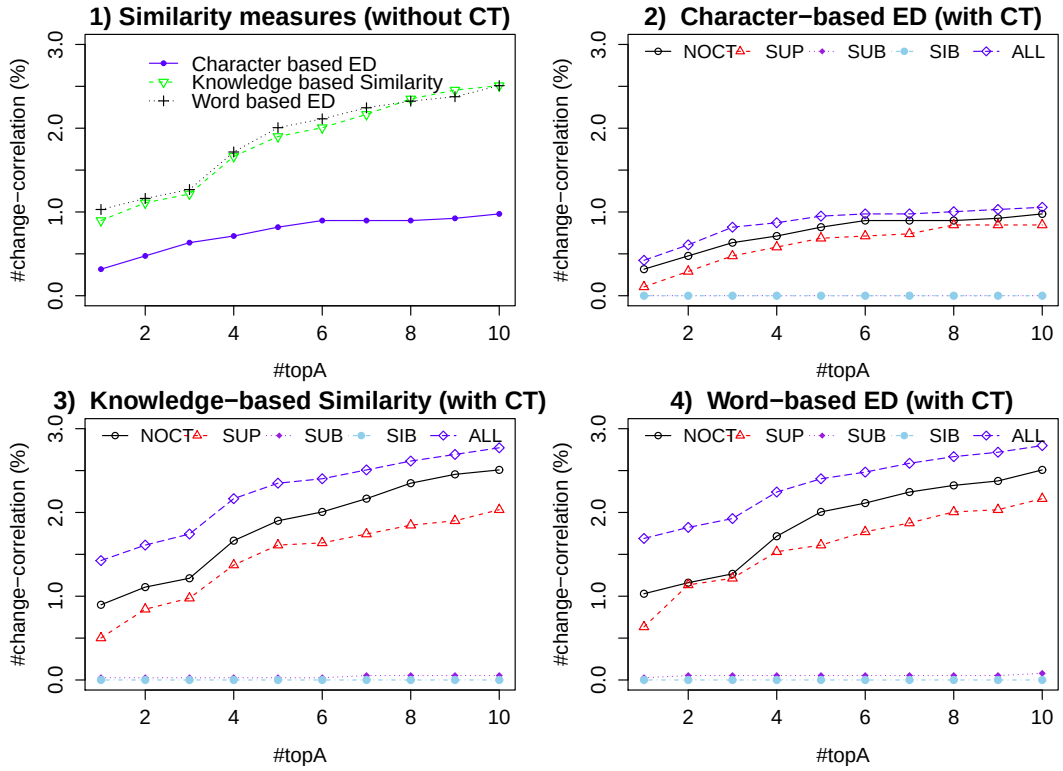


Figure. 4.4: Change correlation in the analysis of SCT

This figure presents the analysis of *change-correlation* (*cf.* Equation 4.14) with respect to **SCT**. These plots are similar to the ones in figure 4.2 with the exception that we only plot the percentage of *change-correlation*.

4.6 Discussion

This chapter proposed a novel method (Algorithm 1), investigating the influence of different similarity measures, at three different linguistic levels, to identify concepts' attributes that served to define existing mappings. It explored similarity between concept attributes of source and target concepts in existing mappings. Through the reported experiments, we investigated the impact of KOS changes affecting detected concept attributes on mappings adaptation. We claimed that identifying a minimal set of relevant candidate attributes (in mappings' source concept or its context) might play a central role for adapting mappings, because the identified attributes might provide evidences for a better interpretation of mappings.

The achieved results underscored the quality of the yielded relevant attributes identified by our algorithm and the usefulness of exploiting them for supporting mapping adaptation. We found correlations between changes affecting identified attributes and modifications in the corresponding mapping interconnecting source and target concepts. This allowed further understanding the influence of KOS changes specifically affecting concept attributes on mapping adaptation.

According to our obtained results, we found it relevant to combine the attributes from the concepts of the context with the source concept attributes for boosting the identification of SCA. Using the context revealed more relevant to find *unchange-correlation* than *change-correlation*.

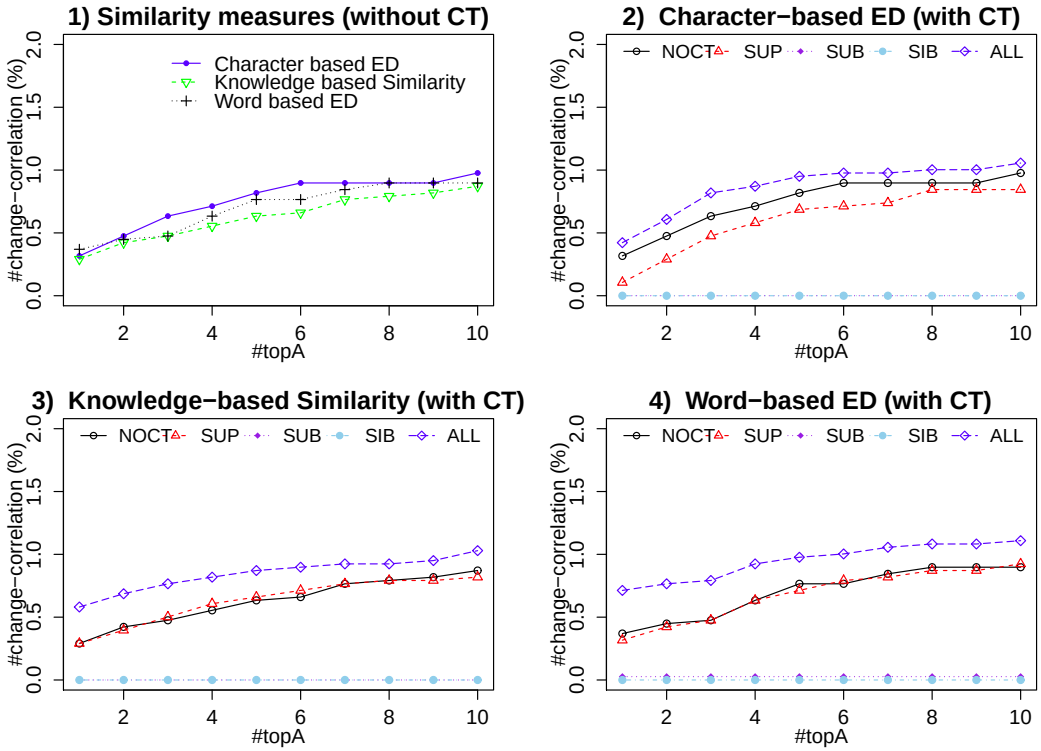


Figure. 4.5: Change correlation in the analysis of ICD9

This figure presents the analysis of *change-correlation* (*cf.* Equation 4.14) with respect to **ICD9**. These plots are similar to the ones in figure 4.3 with the exception that we only plot the percentage of *change-correlation*.

Our method performed well without using context relying either on *Character-based ED* or *Word-based ED*. Indeed, in some cases, the *Word-based ED* allows to efficiently cope with the problem of word order in biomedical terms such as "*skin cancer*" and "*cancer in the skin*" in comparison to the *Character-based ED*.

When considering the context of concepts involved in mappings for studying mapping adaptation, results demonstrated the convenience of exploiting the similarity based on the background knowledge such as *WordNet*, even though this consists in a general resource lacking domain information. To improve these results, we judge it necessary to investigate integrating domain-specific resources in our background knowledge similarity, but reliable semantically annotated corpus underlying such resources remain so far scarce and need more support from the community. One possibility refers to an in-depth study concerning semantic similarity exploring semantic distances and linguistic approaches such as the *MetaMap* [McCray et al., 1994] with the use of external resources such as *UMLS* and *BioPortal*. This might complement and improve our knowledge-based similarity approach.

Furthermore, we observed that results vary according to the studied KOSs. We should particularly chose the adequate similarity metrics according to characteristics of the involved KOSs, but this requires further research. Therefore, we judge more appropriate having the similarity metric as a parameter in the proposed method.

Conclusion

The approach developed in this thesis relies on the concept attributes to adapt mappings. This chapter demonstrated the effectiveness of our proposed method (Algorithm 1) for identifying the most relevant concept's attributes (namely **topA**), which are susceptible for explaining existing mappings. We aim to facilitate and improve the mapping adaptation based on topA. The obtained experimental results conducted with large and real biomedical KOSs reveal that our proposed technique allows to identify the adequate attributes, relevant for mapping adaptation, since we found correlations between KOS changes affecting the identified attributes and mappings change.

We provided empirical evidences for the importance of identifying the relevant attributes to support mapping adaptation. In particular, the achieved results showed to which extent the similarity values appear valuable between source and target concepts involved in mappings, especially at the level of concept attributes, to support mapping adaptation.

The next chapter seeks to understand and precisely recognize the evolution of the relevant attributes from one KOS version to another. We will further use this set of attributes and their evolution's characteristics for concretely making decisions on the adaptation of mappings.

Chapter 5

KOS change patterns to inform mapping adaptation

Contents

Introduction 103
5.1 Problem statement and definitions 105
5.2 Change patterns at the level of concept attributes 107
5.2.1 Lexical change patterns 107
5.2.2 Semantic change patterns 109
5.3 Selection of candidate attributes in the context 111
5.4 Lexical change pattern recognition 113
5.5 Semantic change pattern recognition 114
5.6 Experimental evaluation 116
5.6.1 Materials and procedure 116
5.6.2 Experimental results 118
5.7 Discussion 120
Conclusion 123

Introduction

The dynamic aspect of knowledge in various domains requires that knowledge engineers apply changes to KOS entities by adding, removing and revising them (*cf.* Section 2.4). This leads to new KOS versions periodically, which ensures that semantic-enabled information systems use the most up-to-date representation of the domain knowledge. However, KOS changes potentially impact mappings which rely on these KOSs, as we have demonstrated in this thesis. In the previous chapter, we provided a way to identify attributes defining mappings; in this chapter we address their evolution.

Changes applied to generate new KOS versions are not always fully documented, which impedes the minimization and handling of their impact. To this end, we need methods to automatically identify KOS change operations (KCOs) in an explicit way, given two versions of the same KOS [Hartung et al., 2013]. Our previous studies have underlined the need of precisely

characterizing the evolution of attributes describing concepts for maintaining mappings valid over time (*cf.* Chapter 3).

An example of a real situation observed in our previous studies refers to the transfer of information between concepts. When analyzing two consecutive versions of the same KOS, we found cases where textual statements, which are values of attributes describing concepts, are completely transferred from one concept to its siblings. This affected the associated mappings, since their definition relies on such textual information. For example, we observed this case with the concept '*560.39*' of the ICD9. Such concept contains three attributes and one of them has as value "*Fecal impaction*" (release 2009). Five mappings are defined with this concept as domain, and one of these mappings has a range called "*Fecal impaction (disorder)*", from SCT. After evolution (*i.e.*, ICD9 release 2010), the attribute value "*Fecal impaction*" is no longer associated with the ICD9 concept '*560.39*' and the previously mentioned mapping has been removed. Moreover, the concept "*Fecal impaction*" has been newly created in ICD9 (release 2010) and is reconnected to "*Fecal impaction (disorder)*" of SCT.

Literature has highlighted challenges related to KOS changes' management and has proposed change patterns to improve the KOS evolution process (*cf.* Section 2.4.2). Although useful tools exist to identify the most traditional and frequent KCOs between two KOS versions (*cf.* Section 2.4), taking into account the nature of changes (*e.g.*, atomic or complex) and the type of changes (*e.g.*, addition, removal, split and merge of entities), these tools fail to automatically identify KOS modifications at a finer level of detail, which is required for supporting tasks dependent on KOS changes (*e.g.*, mapping adaptation as elucidated in chapter 3). Our literature review clearly highlighted that although existing techniques perform somehow efficiently for characterizing KOS evolution, they lack important aspects with respect to the adaptation of mappings. This remains an open issue that requires further research.

First, even though literature has proposed change patterns in an attempt to improve the characterization of KOS changes, it requires in-depth studies to evaluate the adequateness and usefulness of that for mapping adaptation. While existing change patterns seem sufficient to identify a set of inconsistencies, they might remain inefficient for dealing with the impact of KOS evolution on dependent artifacts because their design fails to consider requirements for adapting mappings. In fact, KOS modifications that impact mappings must enable a more adequate specification of the evolution of each entity describing a concept (*e.g.,* attributes) to remain useful for maintaining mappings valid over time. Therefore, we need to further investigate which types of change patterns might concretely help to inform techniques of mapping adaptation, and how to correctly apply them.

Second, our empirical studies have shown a need of characterizing the evolution of concepts from the semantic viewpoint. This means to determine from one KOS version to another if a concept becomes more or less specific. This requires performing a comparison between different KOS versions. Approaches aiming at aligning concepts via the subsumption relationships ([≤] or [≥]) deserve our attention because the discovery of these relationships can be especially relevant to help calculating advanced change operations between KOS versions, with evolutionary mappings between KOS versions that characterize the semantic evolution of concepts. However, very few investigations appear in literature with this purpose [Giunchiglia et al., 2004, Spiliopoulos et al., 2010, Arnold and Rahm, 2013]. In addition to the lack of conclusive work, existing approaches are only tailored to KOS alignment purposes. Nevertheless, in a KOS evolu-

tion perspective, we must reconsider design choices to handle the specificities of our investigated scenario of biomedical KOS evolution and mapping adaptation.

This chapter addresses change patterns at the level of concept attribute values. We use linguistic-based features for identifying the diffusion of textual values between concepts from one version of the KOS to another, and how such attributes become more or less specific. Given an attribute a_i from a concept c_x at time j, we investigate means to characterize the way this attribute evolves by considering the context of c_x at time $j+1$ (*cf.* Equation 4.2).

Complementary to existing approaches, our study inquires whether techniques based on linguistic characteristics of textual values, combined with similarity measure value, play a role in supporting automatic change patterns recognition at the level of concept attributes. In particular, to inform mapping adaptation decisions, we are interested in the evolution of the relevant concept attributes detected by **topA** algorithm (*cf.* Chapter 4). In summary, we make the following contributions in this chapter:

- We formally define a set of *KOS change patterns* to express different behaviours of the evolution of attributes relevant for supporting automatic mapping adaptation techniques. We distinguish between lexical and semantic change patterns.
- We introduce a novel approach implementing methods to automatically identify instances of the proposed change patterns by comparing successive KOS versions. Our systematic study provides useful tools to precisely characterize KOS evolution at the level of concept attributes.
- We experimentally assess our approach by using real-world biomedical KOSs. We investigate the influence of different aspects in the performance of the proposed methods and the obtained results show innovative findings.

We first present the problem statement and definitions (Section 5.1). In the sequence, we define the change patterns (Section 5.2). We introduce the algorithms designed to recognize instances of the suggested change patterns between KOS versions (Sections 5.3, 5.4 and 5.5). Section 5.6 presents the experimental evaluation while section 5.7 discusses the obtained results.

5.1 Problem statement and definitions

Analyzing changes on attributes of interest might help in the mapping adaptation task. Complementary to the existing literature (*cf.* Section 2.4), characterizing KOS changes must include the identification of the flow of attributes between concepts of different KOS versions (as demonstrated in our experiments in chapter 3). We need to detect the explicit share and/or transfer of attribute values between different concepts in successive KOS versions.

Figure 5.1 depicts the investigated scenario. Given an attribute value $as_i.value$ from a concept c_s at time j, we investigate ways for characterizing how such attribute evolves by considering the context of the concept c_s^1 at time $j+1$ (*i.e.*, in the new version of KOS K_x) (*cf.* Equation 4.2 for the definition of context). The evolution of KOS entities usually remains restricted to its context [Dos Reis et al., 2014c]. We focus on $a_i.value$ to identify useful behaviours of evolution concerning attributes and search for describing these behaviours as well-delineated change patterns. We face issues to determine which attribute at time $j+1$ represents the most adequate

candidate in the recognition process to identify change pattern occurrences. We apply syntactic analysis techniques to recognize textual values of attributes in different versions of the same KOS.

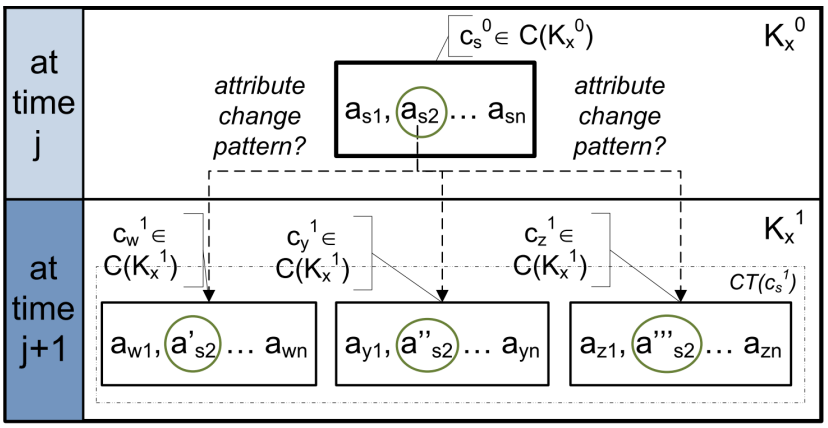


Figure. 5.1: Scenario of the problem for change patterns in concept attributes value

This figure presents the problem statement. Rectangles represent concepts that traditional KOS change operations might affect. Given an attribute value $as_i.value$ that can be relevant for a mapping m^0_{st}, the problem concerns the identification of lexical (LCP) and semantic (SCP) change patterns related to such attribute, considering changed attributes in the context, *i.e.*, parents, children and sibling concepts at time $j+1$.

This chapter copes with the following open research questions:

1. Is it possible to define change patterns at the level of the concept attributes?
2. How to explicitly recognize the defined change patterns between different KOS versions?
3. Is it possible to observe the defined change patterns in real cases of biomedical KOS evolution based on the proposed methods?

We present the notations used in this chapter in table 5.1. Notations previously presented in chapter 4 complement it (*cf.* Table 4.1).

Table 5.1: Notations for the formalization of change patterns

Notation	Description
a_i^j	attribute a_i at time j
$W(a_i^j)$	set of words/tokens from attribute value $a_i.value$
$Ord(w_{ki})$	position of word w_k from attribute value $a_i.value$
$Ch(w_{ki})$	set of characters in word w_{ki}

This table presents the notations relevant for this chapter and their descriptions.

We define that two attributes a_i and a_h are considered as totally "equivalent", denoted as $a_i = a_h$, if and only, if they satisfy the following conditions:

$$a_i = a_h \Leftrightarrow \begin{cases} \forall w_{ki} \in W(a_i), w_{ki} \in W(a_h) \wedge Ord(w_{ki}) = Ord(w_{kh}) \\ \forall w_{kh} \in W(a_h), w_{kh} \in W(a_i) \wedge Ord(w_{kh}) = Ord(w_{ki}) \end{cases} \tag{5.1}$$

Equation 5.1 expresses that two given attributes are only considered totally equivalent if they contain exactly the same words in a ordered manner (exact match). Otherwise, the attributes are different, denoted as $a_i \neq a_h$.

5.2 Change patterns at the level of concept attributes

Considering **change patterns** (CPs) as means to deal with KOS entity changes, we focus on change patterns related to concept attribute values. We distinguish two independent types of change patterns namely: **lexical** and **semantic**. While the former relates to the modification characteristics of the content value of attributes, the latter concerns their semantics, *i.e.*, whether they have become more or less specific. Given a concept c_k in KOS K_x, we define a change pattern between an attribute a_p^0 of concept $c_k^0 \in C(K_x^0)$ and an attribute a_q^1 of concept $c_{cand}^1 \in C(K_x^1)$ ($k \neq cand$), such that $c_{cand}^1 \in CT(c_k^1)$ (*cf.* Equation 4.1). In addition, we suppose that any change pattern must satisfy the constraint 5.2, which states that the attribute a_q is new (it does not exist at time j) or its value differs at time $j+1$.

$$a_q^0 \notin A(c_{cand}^0) \vee a_q^0 \neq a_q^1 \tag{5.2}$$

5.2.1 Lexical change patterns

We define lexical CP (LCP) types as "*Total Copy*" (TC), "*Total Transfer*" (TT), "*Partial Copy*" (PC), and "*Partial Transfer*" (PT). Table 5.2 illustrates lexical change patterns in KOS.

We formalize each type of LCP between attributes $a_p^0 \in A(c_k^0)$ and $a_q^1 \in A(c_{cand}^1)$, if any, in the following. Note that the definitions are only valid when a candidate attribute exists. The used similarity measure indicates the degree of relatedness between two given textual values. We use the γ parameter to control the overlap in terms of words between two attribute values.

- **Total Copy.** A *Total Copy* of content occurs between attribute a_p^0 in concept c_k and attribute a_q^1 in concept c_{cand} if, and only if, a minimal degree γ of words in a_p appears in attribute a_q and a minimal similarity value τ exists between them. Formally:

$$TC(a_p^0, a_q^1) \Leftrightarrow \begin{cases} a_p^0 \in A(c_k^0) \\ c_k^1 \in C(K_x^1) \\ a_p^1 \in A(c_k^1) \\ sim(a_p^0, a_q^1) \geq \tau \\ \|W(a_p^0) \cap W(a_q^1)\| / \|W(a_p^0)\| \geq \gamma \end{cases} \tag{5.3}$$

- **Total Transfer.** A *Total Transfer* of content occurs between attribute a_p^0 in concept c_k and attribute a_q^1 in concept c_{cand} if, and only if, a minimal degree γ of words in a_p appears in attributes a_q and a minimal similarity value τ exists between them, while the original

Table 5.2: Lexical change patterns

LCP	**Attribute**	**Pattern**		**Type**	**Example**	
		j	$j+1$		j	$j+1$
TC	a_p	**ABC**	ABC	*Total Copy*	"*portal systemic encephalopathy*"	"portal systemic encephalopathy"
	a_q	~~ABC~~	**ABC**(D)		∅	"*portal systemic encephalopathy*"
TT	a_p	**ABC**	~~ABC~~	*Total Transfer*	"*fecal impaction*"	∅
	a_q	~~ABC~~	**ABC**(D)		∅	"*fecal impaction*"
PC	a_p	**AB**C	ABC	*Partial Copy*	"*familial* hyperchylomicronemia"	"familial hyperchylomicronemia"
	a_q	~~ABC~~	**AB**(D)		∅	"*familial* chylomicronemia"
PT	a_p	ABC	~~ABC~~	*Partial Transfer*	"eye *swelling*"	∅
	a_q	~~ABC~~	**AB**(D)		∅	"head *swelling*"

This table presents description and examples of lexical change patterns (LCP) from an attribute a_p^0 to an attribute a_q^1. The type of LCPs include: *Total Copy* (TC), *Total Transfer* (TT), *Partial Copy* (PC), *Partial Transfer* (PT). The ∅ symbol means that the corresponding attribute does not exist.

attribute a_p^0 is removed from $c_k^1 \in C(K_x^1)$. Note that in *Total Copy* (*cf.* Equation 5.3) $a_p^1 \in A(c_k^1)$ while in *Total Transfer* (*cf.* Equation 5.4) $a_p^1 \notin A(c_k^1)$, which states the main difference between them. Formally:

$$TT(a_p^0, a_q^1) \Leftrightarrow \begin{cases} a_p^0 \in A(c_k^0) \\ a_p^1 \notin A(c_k^1) \\ sim(a_p^0, a_q^1) \geq \tau \\ \|W(a_p^0) \cap W(a_q^1)\| / \|W(a_p^0)\| \geq \gamma \end{cases} \tag{5.4}$$

- **Partial Copy.** A *Partial Copy* of content occurs between attribute a_p^0 in concept c_k and attribute a_q^1 in concept c_{cand} if, and only if, a partial overlap exists between words constituting attribute a_p^0 and attribute a_q^1 superior to 0 and inferior to γ, while respecting

a minimal similarity value τ. Formally:

$$PC(a_p^0, a_q^1) \Leftrightarrow \begin{cases} a_p^0 \in A(c_k^0) \\ c_k^1 \in C(K_x^1) \\ a_p^1 \in A(c_k^1) \\ sim(a_p^0, a_q^1) \geq \tau \\ 0 < \|W(a_p^0) \cap W(a_q^1)\| / \|W(a_p^0)\| < \gamma \end{cases} \tag{5.5}$$

- **Partial Transfer.** A *Partial Transfer* of content occurs between attribute a_p^0 in concept c_k and attribute a_q^1 in concept c_{cand} if, and only if, a partial overlap exists between words constituting attribute a_p^0 and attribute a_q^1 superior to 0 and inferior to γ, while respecting a minimal similarity value τ, and the original attribute a_p^0 is removed from $c_k^1 \in C(K_x^1)$. Formally:

$$PT(a_p^0, a_q^1) \Leftrightarrow \begin{cases} a_p^0 \in A(c_k^0) \\ a_p^1 \notin A(c_k^1) \\ sim(a_p^0, a_q^1) \geq \tau \\ 0 < \|W(a_p^0) \cap W(a_q^1)\| / \|W(a_p^0)\| < \gamma \end{cases} \tag{5.6}$$

The definition of LCPs included the threshold γ mostly to differentiate *Total Copy* and *Total Transfer* from *Partial Copy* and *Partial Transfer*. The threshold γ contributes to make our defined CPs more flexible. Practical examples can illustrate the advantages of assigning the value γ different of 1. For instance, the attribute "*Diabetes type **1***" can be considered as a *Total Copy* of the attribute "*Diabetes type **I***", even though they are not totally exact from the lexical point of view, thus reducing possible cases of false-negative in the identification method.

5.2.2 Semantic change patterns

We define semantic CP (SCP) types as "*Equivalent*" (≡) (EQV), "*More Specific*" (<) (MSP), "*Less Specific*" (>) (LSP) and "*Partial Match*" (≈) (PTM). Table 5.3 shows examples of semantic change patterns. These change patterns aim to capture the evolution from the semantic point of view of an attribute value. This creates a much greater challenge for the recognition method because situations not fully corresponding to the intersection of words or characters might exist. For example, "*hypotension*" is more specific than "*vascular disease*". Our previous research with biomedical KOS evolution shows that these cases are less frequent in observed evolution scenarios [Dos Reis et al., 2013b], so we do not emphasize them in this investigation. Our proposed semantic change patterns still allow, for instance, the detection that "*cerebral hypoxia*" is a more general term than "*cerebral anoxia*".

We formalize each type of SCP between attributes $a_p^0 \in A(c_k^0)$ and $a_q^1 \in A(c_{cand}^1)$, if any, in the following. Aside from the EQV (*cf.* Equation 5.7), we use the statement $sim(a_p^0, a_q^1) \geq \tau$ to filter cases of low similarity between attributes.

- **Equivalent.** We consider an attribute a_p^0 as semantically equivalent to an attribute a_q^1 (*cf.* Equation 5.7) if, and only if, we observe all words from a_p^0 in a_q^1 and *vice-versa* (*cf.* Equation 5.1) or the overlap in terms of words between them must be equal to or greater than a threshold γ. In the latter case, to harmonize the rate of the words intersection between

Table 5.3: Semantic change patterns

SCP	Attribute	Pattern j	Pattern $j+1$	Type	Example j	Example $j+1$
EQV	a_p	**ABC**	ABC	$a_p^0 \equiv a_q^1$	"*phimosis*"	"phimosis"
	a_q	~~ABC~~	**ABC**		∅	"*phimosis*"
MSP	a_p	ABC**D**	ABCD	$a_p^0 < a_q^1$	"kappa *light* chain disease"	"kappa light chain disease"
	a_q	~~ABC~~	ABC~~D~~		∅	"kappa chain disease"
LSP	a_p	ABC	ABC	$a_p^0 > a_q^1$	"cerebral *hypoxia*"	"cerebral hypoxia"
	a_q	~~ABC~~	ABC**D**		∅	"cerebral *anoxia*"
PTM	a_p	**AB**C	ABC	$a_p^0 \approx a_q^1$	"*focal* atelectasis"	"focal atelectasis"
	a_q	~~ABC~~	**A**C**B**		∅	"*helical* atelectasis"

This table presents description and examples of semantic change patterns (SCP) from an attribute a_p^0 to an attribute a_q^1. The types of LCPs include: *Equivalent* (EQV), *More Specific* (MSP), *Less Specific* (LSP), *Partial Match* (PTM). The ∅ symbol means that the corresponding attribute does not exist. The symbols $\equiv, <, >$ and $\approx$ mean that attribute a_p^0 is equivalent, more specific, less specific and is partially matched to attribute a_q^1, respectively.

the attributes, we multiply the number of intersection by 2 over the total number of words of the two attributes. We also take into account the hierarchy of concepts where they are found. Based on the analyses conducted in our previous studies [Dos Reis et al., 2013b], we exclude from this SCP all concepts c_{cand}^1 with hierarchical relationship with c_k^1. We observed that a conceptual error can exist if at time j a mapping between two concepts exists with an equivalent relation $(c_s^0, c_t^0, \equiv)$ and at time $j+1$ a mapping between two concepts consists of an equivalent relation $(c_{cand}^1, c_t^1, \equiv)$ in which $c_t^0 = c_t^1$, $c_s^1 \in sup(c_{cand}^1)$ or $c_s^1 \in sub(c_{cand}^1)$. Moreover, the change pattern in equation 5.7 does not explore the similarity function because we perform the matching directly on the words comparison. Indeed, we use the similarity function in the other SCPs as a filter (*cf.* equations 5.8, 5.9 and 5.10). We formally define the *Equivalent* as follows:

$$EQV(a_p^0, a_q^1) \Leftrightarrow \begin{cases} c_k^1 \in C(K_x^1) \\ c_k^1 \notin sub(c_{cand}^1) \wedge c_k^1 \notin sup(c_{cand}^1) \\ \left[\begin{array}{l} a_p^0 = a_q^1 \\ \vee \\ (2 * \|W(a_p^0) \cap W(a_q^1)\|)/(\|W(a_p^0)\| + \|W(a_q^1)\|) \geq \gamma \end{array} \right. \end{cases} \tag{5.7}$$

- **More Specific.** We define an attribute a_p^0 as *More Specific* (less generic) than an attribute a_q^1 if, and only if, we observe all words from a_q^1 in a_p^0, but not the opposite, *i.e.*, a_q^1 consists of a subset of words of a_p^0. We introduce the parameter γ to make more flexible whether

a_q^1 has a declination of at least one word from a_p^0 (*i.e.*, this can accommodate cases where a prefix or a suffix of a_p^0 slightly differs from a_q^1). If we want to consider that the whole subset of words (intersection) remains strictly exact, we only need to set $\gamma = 1$. We also observe whether words exist in a_p^0 containing the whole set of characters of a word from a_q^1 in addition to other different characters. Moreover, concept c_k cannot be located at a higher position in the hierarchical structure than concept c_{cand}. Formally:

$$MSP(a_p^0, a_q^1) \Leftrightarrow \begin{cases} c_k^1 \notin sup(c_{cand}^1) \\ sim(a_p^0, a_q^1) \geq \tau \\ \left[\begin{array}{l} (\|W(a_p^0)\| > \|W(a_q^1)\|) \wedge (\|W(a_p^0) \cap W(a_q^1)\| / \|W(a_q^1)\| \geq \gamma) \\ \vee \\ \exists w_p^0 \in W(a_p^0), w_q^1 \in W(a_q^1) \mid Ch(w_q^1) \subset Ch(w_p^0) \end{array} \right. \end{cases} \tag{5.8}$$

- **Less Specific.** We consider attribute a_p^0 as *Less Specific* (more generic) than attribute a_q^1 applying the inverse proposition suggested for the *More Specific*. Formally:

$$LSP(a_p^0, a_q^1) \Leftrightarrow \begin{cases} c_k^1 \notin sub(c_{cand}^1) \\ sim(a_p^0, a_q^1) \geq \tau \\ \left[\begin{array}{l} (\|W(a_p^0)\|) < (\|W(a_q^1)\|) \wedge (\|W(a_p^0) \cap W(a_q^1)\| / \|W(a_p^0)\| \geq \gamma) \\ \vee \\ \exists w_q^1 \in W(a_q^1), w_p^0 \in W(a_p^0) \mid Ch(w_p^0) \subset Ch(w_q^1) \end{array} \right. \end{cases} \tag{5.9}$$

- **Partial Match.** Attribute a_p^0 is partially matched to attribute a_q^1 if, and only if, we find some common words between them and if the similarity value between them is higher than a threshold τ.

$$PTM(a_p^0, a_q^1) \Leftrightarrow \begin{cases} \exists w_p^0 \in W(a_p^0), w_p^0 \in W(a_q^1) \\ \exists w_p^0 \in W(a_p^0), w_p^0 \notin W(a_q^1) \\ sim(a_p^0, a_q^1) \geq \tau \end{cases} \tag{5.10}$$

5.3 Selection of candidate attributes in the context

In our approach to recognize change patterns, we first determine a candidate attribute a_q^1 in the context of a concept c_k^1. This candidate refers to a changed attribute at time $j+1$ related to the attribute a_p^0 in concept c_k^0 that we use to identify occurrences of lexical CPs (Section 5.4) and semantic CPs (Section 5.5).

We design algorithm 2 that explores textual attributes from a given concept at time j. In particular, given an attribute $a_p^0 \in A(c_k^0)$ from K_x^0, the algorithm courses the whole set of changed attributes of the context of c_k at time $j+1$ by calculating the similarity to retrieve candidate attributes. It aims to detect the most adequate attribute in the context of the given one from $A(c_k)$, which we will use in algorithms 3 and 4 to identify change patterns.

We consider the types of comparable textual attributes as a parameter in our approach. For example, we can take only attributes of type "*name*" and "*synonym*" into consideration when

comparing the attribute values (*i.e.*, strings denoting concepts). Our methods exclude all types of attributes out of the comparable set of attributes defined beforehand. The function $sim(a_i^0, a_j^1)$ computes the similarity between two given attribute values. It returns a value ranging from 0 to 1. The higher the result is, the more similar these attributes are. We explore traditional string-based similarity metrics when calculating the similarity between attribute values in the algorithm 2. Since the performance of the similarity measure is not the focus of this investigation, we keep it generic in our CPs definition and recognition methods so that we can choose it as a parameter in our experiments.

Algorithm 2 generates a list of candidate attributes which is denoted as $S_{cand}(a_p^0) = \{(a_{q_1}, c_1^1, sim_{pq_1}), (a_{q_2}, c_2^1, sim_{pq_2}), ..., (a_{q_m}, c_m^1, sim_{pq_m})\}$, where $a_{q_i} \in A_{ct}(c_k^1)$ (*cf.* Equation 4.3), $c_i \in CT(c_k^1)$ and $sim_{pq_i} = sim(a_p^0, a_{q_i}^1)$. In fact, $S_{cand}(a_p^0)$ stores the candidate attributes along with their similarity with the attribute $a_p^0 \in A(c_k^0)$. This algorithm uses a ranking function to determine the best candidate attribute as a result.

Algorithm 2 Find candidate attribute in the context

Require: $a_p^0 \in A(c_k^0); CT(c_k^1) \subset C(K_x^1)$

```
1: sim ← ∅; a_q^1 ← ∅; S_cand ← ∅;
2: for all c_i^1 ∈ CT(c_k^1) do
3:    for all a_i^1 ∈ A(c_i^1) do
4:       if a_i^0 ∉ A(c_i^0) ∨ a_i^0 ≠ a_i^1 then
5:          sim ← sim(a_p^0, a_i^1);
6:          S_cand ← S_cand ∪ {(a_i^1, c_i^1, sim)};
7:       end if
8:    end for
9: end for
10: return S_cand ← rank(S_cand).first;
```

The candidate attribute may have a strong influence on the CP identification methods, which emphasizes the relevant of methods for selecting these candidates. We explore a ranking relying on the optimum similarity value considering the whole context, where the best candidate attribute a_q^1 (found at time $j+1$) refers to the one that has the highest similarity with a given attribute $a_p^0 \in A(c_k^0)$. Formally:

$$rank(S_{cand}) \leftarrow \underset{a_{q_i}^1 \in A_{ct}(c_k^1)}{\arg\max} \{sim(a_p^0, a_{q_i}^1)\} \tag{5.11}$$

We have preliminarily investigated other ranking functions to determine the best approach of selection. Experimental results indicated the approach expressed in equation 5.11 as the most adequate one [Dinh et al., 2014a]. Therefore, in this chapter we only report our experiments and results using this ranking.

5.4 Lexical change pattern recognition

Algorithm 3 describes the designed method to identify lexical CPs. The best candidate c^1_{cand} refers to the concept denoted by attribute a^1_q, retrieved with algorithm 2. The algorithm checks whether the similarity value of the attribute candidate a^1_q with attribute a^0_p is greater or equal to a threshold τ, and the conditions for applying each type of lexical change pattern on the couple of attributes a^0_p and a^1_q.

Algorithm 3 calculates the number of common words between a^0_p and a^1_q by removing stop words from the original attributes. Non-stopwords are stemmed using the Porter stemmer algorithm [Porter, 1997]. Algorithm 3 also explores whether attributes $a^0_p \in A(c^0_k)$ and $a^1_q \in A(c^1_{cand})$ remain at time $j+1$ (*i.e.*, it is not deleted). We explore the KOS change operations calculated by *COnto-Diff* tool [Hartung et al., 2013] for this purpose. According to the LCPs definitions, algorithm 3 assigns the adequate LCP. Given two versions of the KOS, we can apply algorithm 3 to concepts involved in KOS mappings or to all concepts placed in KOS regions affected by traditional change operations. Figure 5.2 presents an example of change pattern recognition where a *Total Transfer* LCP of the attribute value "*kappa chain disease*" is detected in a concept at time $j+1$.

Algorithm 3 Lexical change pattern recognition

Require: $a^0_p \in A(c^0_k); c^0_k \in C(K^0_x); CT(c^1_k) \subset C(K^1_x)$
1: $LCP \leftarrow \varnothing; sim \leftarrow 0; nbEqWords \leftarrow 0$
2: $a^1_q; sim \leftarrow getCandAttribute(a^0_p; CT(c^1_k))$; (*cf.* Algorithm 2)
3: **if** $a^1_q \neq \varnothing$ **then**
4: **if** $sim \geq \tau$ **then**
5: $nbEqWords \leftarrow \|W(a^0_p) \cap W(a^1_q)\|$
6: **if** $nbEqWords/\|W(a^0_p)\| < \gamma \wedge nbEqWords > 0 \wedge nbEqWords < \|W(a^0_p)\|$ **then**
7: **if** $a^1_p \in A(c^1_k)$ **then**
8: $LCP \leftarrow PC$;
9: **else**
10: $LCP \leftarrow PT$;
11: **end if**
12: **else**
13: **if** $nbEqWords/\|W(a^0_p)\| \geq \gamma$ **then**
14: **if** $a^1_p \in A(c^1_k)$ **then**
15: $LCP \leftarrow TC$;
16: **else**
17: $LCP \leftarrow TT$;
18: **end if**
19: **end if**
20: **end if**
21: **end if**
22: **end if**
23: **return** (a^0_p, a^1_q, LCP);

5.5 Semantic change pattern recognition

Algorithm 4 describes the designed method to identify semantic change patterns. This algorithm aims to determine the adequate semantic change pattern given an attribute from $A(c_k)$. If a candidate attribute exists in the context, which is retrieved with algorithm 2, algorithm 4 proceeds by verifying conditions to assign one of the four semantic CPs defined (*Equivalent*, *More Specific*, *Less Specific* or *Partial Match*).

Firstly, if no *Equivalent* SCP is found (lines 6), the algorithm 4 verifies a minimal threshold τ regarding the similarity between the attributes for which a *Less Specific* or a *More Specific* SCP can be assigned (line 8). The algorithm also calculates the number of subwords from one attribute to the other (lines 9 - 11) by using algorithm 5, which checks whether the set of characters of one word corresponds to a subset of characters of another word in a different attribute.

Therefore, the algorithm checks the conditions related to the set of words or characters (number of common words and subwords calculated) in addition to the hierarchy of concepts to assign a *More Specific* or *Less Specific* SCPs (lines 12 - 15). If it fails to fulfill the required conditions for determining EQV, MSP or LSP, it finally checks the conditions for the *Partial Match* (line 16).

Algorithm 4 Semantic change pattern recognition

Require: $a_p^0 \in A(c_k^0); c_k^0 \in C(K_x^0); CT(c_k^1) \subset C(K_x^1)$

1: $SCP \leftarrow \varnothing; sim \leftarrow 0;$
2: $nbEqWords \leftarrow \|W(a_p^0) \cap W(a_q^1)\|$
3: $nbSubWordsAP \leftarrow 0; nbSubWordsAQ \leftarrow 0;$
4: $a_q^1; sim; c_{cand}^1 \leftarrow getCandAttribute(a_p^0; CT(c_k^1) \cup c_k^1);$ (*cf.* Algorithm 2)
5: **if** $a_q^1 \neq \varnothing$ **then**
6: **if** $(a_p^0 = a_q^1) \vee ((2 * \|nbEqWords\|)/(\|W(a_p^0)\| + \|W(a_q^1)\|) \geq \gamma) \wedge (c_k^1 \notin sup(c_{cand}^1) \wedge c_k^1 \notin sub(c_{cand}^1))$ **then**
7: $SCP \leftarrow' \equiv';$
8: **else if** $sim \geq \tau$ **then**
9: $len_p \leftarrow \|W(a_p^0)\|; len_q \leftarrow \|W(a_q^1)\|;$
10: $nbSubWordsAP \leftarrow getSubWords(len_p; W(a_p^0); len_q; W(a_q^1));$ (*cf.* Algorithm 5)
11: $nbSubWordsAQ \leftarrow getSubWords(len_q; W(a_q^1); len_p; W(a_p^0));$
12: **if** $(nbSubWordsAP > 0 \wedge nbSubWordsAQ = 0) \vee ((\|W(a_p^0)\| > \|W(a_q^1)\|) \wedge (\|nbEqWords\| \geq \|W(a_q^1)\| * \gamma) \wedge (c_k^1 \notin sup(c_{cand}^1))$ **then**
13: $SCP \leftarrow' <';$
14: **else if** $(nbSubWordsAQ > 0 \wedge nbSubWordsAP = 0) \vee ((\|W(a_p^0)\| < \|W(a_q^1)\|) \wedge (\|nbEqWords\| \geq \|W(a_p^0)\| * \gamma) \wedge c_k^1 \notin sub(c_{cand}^1)$ **then**
15: $SCP \leftarrow' >';$
16: **else if** $(\|nbEqWords\| > 0) \wedge (nbEqWords \neq \|W(a_q^1)\|) \wedge (nbEqWords \neq \|W(a_p^0)\|)$ **then**
17: $SCP =' \approx';$
18: **end if**
19: **end if**
20: **end if**
21: **return** $(a_p^0, a_q^1, SCP);$

Note that for SCP identification, the concept c_k^1 may be considered as candidate in addition to concepts in $CT(c_k^1)$ (line 4), because the whole set of concepts from $CT(c_k^1)$ can remain unchanged, and the attribute a_p^0 can evolve becoming more or less specific in concept c_k^1. However, for LCP identification, only concepts in $CT(c_k^1)$ are considered as candidates, because we consider a copy or a transfer only between different concepts (*i.e.*, containing different identifiers).

Algorithm 5 Calculate number of subwords

Require: $len_x \in \mathbb{N}; W(a_x); len_y \in \mathbb{N}; W(a_y);$

```
1:  i ← 0; h ← 0; nbSubWords ← 0;
2:  while i < len_x do
3:    h ← 0; found ← 0;
4:    while h < len_y ∧ found = 0 do
5:      if Substring(w_ix, w_hy) then
6:        nbSubWords ← nbSubWords + 1;
7:        found ← 1;
8:      end if
9:      h ← h + 1;
10:   end while
11:   i ← i + 1;
12: end while
13: return nbSubWords
```

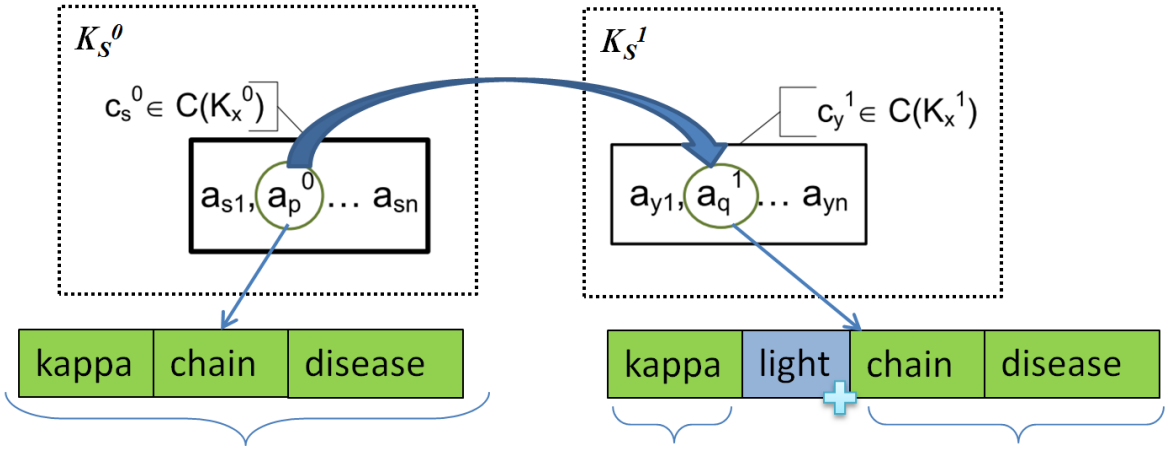


Figure. 5.2: Example of change pattern recognition

This figure shows an example of change pattern recognition where our algorithms can recognize a *Total Copy* LCP and a *Less Specific* SCP.

Figure 5.2 shows an example of a *Less Specific* SCP detected in a new attribute value at time $j+1$. The fact that the set of words "*kappa chain disease*" is completely found in the new attribute in addition to a new word "*light*" indicates that the attribute a_q^1 is more specific than attribute a_p^0. In addition to these word-level types of patterns, algorithm 4 also enables to detect cases of "subword-level". For instance, in the case of attributes "*bacterial encephalitis*" and "*bacterial meningoencephalitis*", we find no total intersection between word-level, but the word "*meningoencephalitis*" contains the word "*encephalitis*" which allows us to detect SCPs.

5.6 Experimental evaluation

We evaluate the effectiveness of the proposed methods for recognizing change patterns based on exploiting lexical features of attributes. We present the used materials followed by the conducted experimental procedure (Section 5.6.1). Section 5.6.2 reports on the obtained results.

5.6.1 Materials and procedure

In the conducted experiments we used various versions of three biomedical KOSs: SCT, ICD9 and MeSH. In addition to SCT and ICD9 that chapters 3 and 4 have studied (*cf.* Section 3.1.1), in this chapter we added the MeSH.

MeSH. The *Medical Subject Headings*(MeSH)[31] consists of a controlled vocabulary mainly used for indexing and retrieving health related documents and life science literature. For example, *PubMed* articles database is annotated using MeSH. The MeSH's content covers a broad aspect in biomedicine including anatomy, organisms, diseases, chemicals, drugs, *etc.* The *U.S. National Library of Medicine* (NLM) manages the maintenance of MeSH. This KOS proposes a particular data structure organization. The main element refers to descriptors. A set of concepts compose a descriptor. Each concept in MeSH is denoted by an attribute (preferred term), which is also the name of the concept. A concept can have one or more synonyms (non-preferred terms). Terms in one concept are not strictly synonymous with terms in another concept. Concepts in a descriptor are linked by *is-a* relationships.

Table 5.4 presents statistics regarding the number of concepts, attributes and of direct subsumption relationships between concepts, since this study focused on exploiting the hierarchical structure of KOSs. SCT contains a much higher number of concepts than MeSH and ICD9. Table 5.4 also depicts the evolution rate of the KOS entities for the three studied biomedical KOSs in a combined way over the last years.

Table 5.4: Evolution of the studied biomedical KOSs

KOS	**Release**	**#Concepts**		**#Attributes**		**#Subsumptions**	
ICD-9-CM	2009	12 734		34 065		11 619	
	2011	13 059	(+2.55%)	34 963	(+2.64%)	11 962	(+2.95 %)
SNOMED-CT	2010	390 022		1 531 288		523 958	
	2012	395 346	(+2.12%)	1 570 504	(+2.50%)	539 245	(+2.83%)
MeSH	2012	50 367		259 565		59 191	
	2013	50 971	(+1.18%)	264 783	(+1.97%)	59 844	(+1.09%)

The numbers between parentheses represent the change rate between two releases of the same KOS.

Reference change patterns

To evaluate the effectiveness of our proposed recognition methods, we manually defined a set of reference change patterns as our gold standard. Since no reference dataset (gold standard)

[31] `www.nlm.nih.gov/mesh`

exists for the defined patterns, the building of our own reference change patterns which allowed performing the validation was required. To this end, we conducted the following steps:

- We combined the considered biomedical KOSs (*cf.* Table 5.4) and we randonly selected 1 000 couples of attributes. We found samples in regions of the KOSs (by examining the raw log of brut changes) that have changed the most, since unchanged concepts did not allow observing our proposed change patterns. We defined the size of our sample in accordance with the involved experts, taking into account their availability, and seeking scientific consistencies for our experiments. One attribute of each couple comes from a concept in a concept at time j and the other attribute of the couple comes from a concept in the context of the concept at time $j+1$. We chose these couples based on the similarity between the attribute values, excluding attributes with very low similarity and unchanged attributes at time $j+1$.
- We invited three KOS engineering experts familiar with the several biomedical subfields to evaluate all selected attribute couples to assign their answer regarding LCP and SCP. They have been working in the biomedical field for 5 to 10 years and have thorough experience with terms in biomedical KOSs. We supported them with a software tool suited to present additional information regarding each attribute. This tool presents the couple of attributes along with concepts in the context, the attributes denoting concepts as well as the changes affecting them. We gave instructions on the purpose of the different patterns, and recorded the answers for each evaluator separately.
- The experts performed one evaluation round and we merged the agreement answers. The biomedical domain experts collaborated and re-evaluated a second round with the disagreement part of couples only. We merged the final agreement couples for both LCP and SCP with the respective correct answers according to the evaluators. We achieved an average agreement rate of 86% for LCP and SCP. Finally, in our experiments, we retained 675 pairs of attributes which had the consent of all evaluators for either the same LCP or SCP.

We computed the standard metrics of *Precision, Recall* and *F-measure* based on the reference change patterns as input. Therefore, given a pair of attributes from the reference set, we used our recognition algorithms to identify lexical and semantic change patterns for the attribute at time j. We compared the outcome with the adequate answer in the gold standard (type of CPs and attribute at time $j+1$), calculating the evaluation metrics. In this way, we involved human experts in the evaluation only once, when constructing the reference change patterns.

Specifically, we computed the *Precision* as the number of CPs correctly identified by the algorithms, in contrast with the expected ones evaluated in the set of reference change patterns, over the total number of identified CPs:

$$Precision = \frac{\#correctly\ identified\ CP}{\#identified\ CP} \tag{5.12}$$

Recall was computed as the number of correctly identified CPs over the total number of relevant CPs (expected) in the set of reference:

$$Recall = \frac{\#correctly\ identified\ CP}{\#relevant\ CP} \tag{5.13}$$

F-measure was computed as the harmonic mean of precision and recall.

$$F\text{-}measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{5.14}$$

We measured these metrics separately for each type of lexical and semantic change patterns. This evaluation explored traditional string-based similarity metrics (the *bi-gram* measure) when calculating the similarity between attributes in algorithm 2. We selected this metric as the default similarity in these experiments because it performs well on ontology matching as demonstrated in recent literature [Cheatham and Hitzler, 2013], but we kept the similarity measure as a parameter in the prototype. We calculated the bi-gram similarity by using the *Dice coefficient* [Dice, 1945], which is twice the number of matching bi-grams between two words w_x and w_y divided by the sum of the number of bi-grams in word w_x and the number of bi-grams in word w_y.

During our research, we tested other string-based similarity measures, in particular the well-known *levenshtein edit-distance* measure (*cf.* Section 4.3.1) and the word-based similarity measure (*cf.* Section 4.3.2). The performance of our algorithms to recognize change patterns underlaid by *bi-gram* measure slightly outperformed the other string-based similarity measures evaluated, so we only report our results with *bi-gram* in this chapter.

In addition, we investigated the influence of the thresholds of the similarity in the CP recognition algorithms. For this purpose, we analyzed the CP identification performance by varying the thresholds from 0 to 1 with a step of 0.05 to observe the performance of our algorithms, and we set $\tau = \gamma$. Note that this experiment does not aim to determine specific values for τ and γ, but rather to observe the behaviour of the algorithms under the modification of these parameters' value.

5.6.2 Experimental results

We present the results on the effectiveness of the proposed algorithms to recognize change patterns. We first show the findings concerning LCPs identification and afterwards SCPs identification.

Lexical change patterns

Figure 5.3 presents the effectiveness of the LCP recognition algorithm in terms of *Precision, Recall* and *F-measure* by varying the thresholds.

The performance of the LCP algorithm varies according to the threshold values. Overall, the *F-measure* is greater than 0.60 for all types of LCP. We observe that the similarity threshold plays a relevant role in LCP identification because its performance dramatically changes when the threshold is set too low (*e.g.,* $\tau < 0.5$). We notice that our LCP identification algorithm reaches the best performance with thresholds ranging from 0.7 to 0.9. This points out the necessity of having a minimal similarity between the attribute values to boost the identification results.

As an example, we show the role played by the thresholds involved in our recognition algorithm. Before evolution, a SCT concept had an attribute value equal to "*Deformity of thorax*". The whole concept was assigned as ambiguous after evolution, and a new sibling concept named

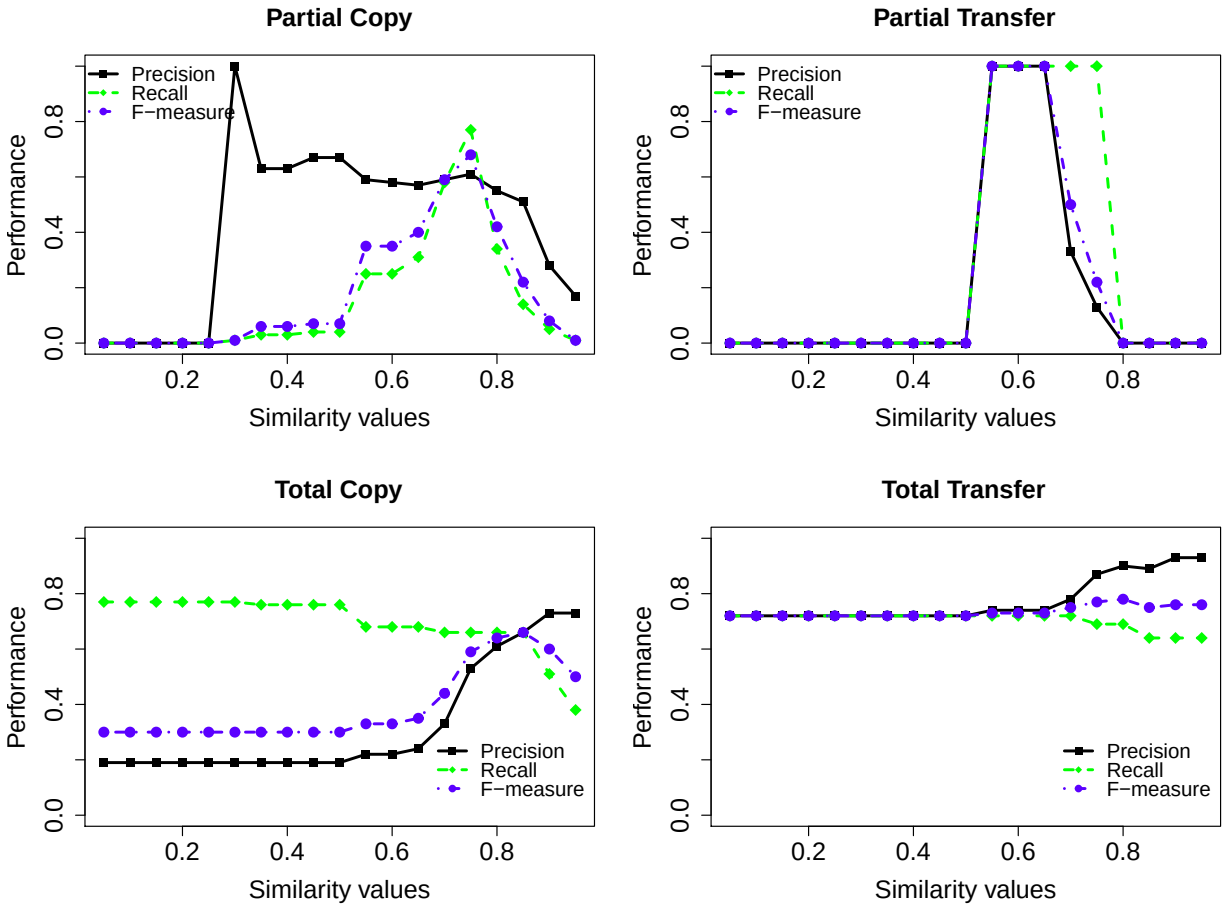


Figure. 5.3: Effectiveness of the Lexical Change Pattern identification algorithm

The graphics show the results for the performance (y-axis) of the metrics of *Precision, Recall* and *F-measure* (*cf.* equations 5.12, 5.13 and 5.14) for different threshold values (x-axis). We evaluate the recognition algorithm against the reference change patterns (gold standard) for the different defined *lexical change patterns*: **Partial Copy** (plot on the top left); **Partial Transfer** (top right); **Total Copy** (bottom left); **Total Transfer** (bottom right).

"*Deformity of thoracic structure (disorder)*" is added in the new version of the KOS. Our algorithm can determine a *Total Transfer* or a *Partial Transfer* of the original attribute "*Deformity of thorax*" considering the new sibling concept. After removing stopwords and stemming, we have two sets of tokens: $tk^0 = \{deformity, thorax\}$ and $tk^1 = \{deformity, thorax, structure\}$. The similarity value and the proportion of words intersection calculated are 0.67 and 1.0, respectively. For instance, if we set the thresholds in algorithm 3 as $\tau = \gamma = 0.6$ then the algorithm will determine a *Total Transfer* for this particular case, because both values for τ and γ are higher than 0.6. However, if we augment both τ and γ to a threshold higher than 0.8, algorithm 3 will not assign a *Total Transfer* even if $\gamma = 1$ since the similarity value criteria is not respected. This highlights that both τ and γ thresholds impact the algorithm performance.

By observing the results for each type of LCP, we found that the identification of *Partial Copy* LCP reaches the highest *F-measure* of 0.68 (*Precision*=0.61, *Recall*=0.77) at $\tau = 0.75$.

This remains similar to the case of *Total Copy* LCP, where the highest *F-measure* is 0.66 (*Precision*=0.66, *Recall*=0.66) at τ = 0.85. Moreover, the *Recall* for *Total Copy* LCP tends to be higher than the *Precision* for $\tau < 0.85$, but we observed the contrary phenomenon for *Partial Copy* LCP. We potentially explain this by the fact that for correctly identifying *Total Copy* CPs require higher similarity between attributes, while for *Partial Copy*, the higher the similarity value, the lower the number of *Partial Copy* CPs correctly identified.

Regarding *Total Transfer* LCP, algorithm 3 reaches the best *F-measure* at 0.78 (*Precision*=0.90, *Recall*=0.69) for τ = 0.80. The algorithm performs better on identifying *Total Transfer* than on *Partial Transfer*. We notice that *Partial Transfer* LCP seems to be a particular case (not frequently found) because evaluators assigned only one case in the reference change patterns.

The performance on identifying *Total Copy* or *Total Transfer* cases remains better than for *Partial Copy* or *Partial Transfer*, because recognizing partial cases involves more difficulties on analyzing and interpreting the textual value of the attributes.

Semantic change patterns

Figure 5.4 presents the effectiveness of the proposed algorithm 4 for SCP identification. Overall evaluation points out that LCP identification method outperforms SCP identification. Similar to LCP, results concerning the different types of SCP vary and the threshold also plays a determinant role. Lower thresholds yields the worst results while higher thresholds tend to yield better results. According to algorithm 4, the threshold rather influences one type of SCP than another. However, we found the best result scenario with a similarity threshold of 0.8. Apart from the results concerning *Partial Match* SCP, the algorithm for SCP identification reaches a minimal *F-measure* of 0.53.

By individually analysing each type of SCP, we observe that algorithm 4 performs best overall results for identifying *Equivalent* change patterns, with a *Precision* of 0.78 and *Recall* of 0.61 and yields a *F-measure* of 0.69. The algorithm overperforms this *Precision* for *Less Specific* SCPs (0.96), but affects the *Recall* while keeping a *F-measure* of 0.75. Results regarding *More Specific* SCPs show medium performance with a highest *F-measure* (0.53). We found a less consistent performance for identifying *Partial Match* SCPs. Possible explanations either lie on the evaluators' misunderstanding of such change pattern while building the reference change patterns, or are due to the proposed definition of *Partial Match*. A domain knowledge-based approach would improve this performance, which demands further studies.

5.7 Discussion

The investigation in this chapter claimed that finely describing changes that affect attributes of concepts involved in mappings, in addition to existing KOS change operations, may provide the necessary statements to make the appropriate decisions on mappings adaptation. We found that the suggested types of change patterns at the level of attributes can be observed in real cases of KOS evolution. These change patterns refine the traditional ones at a finer level of granularity to characterize KOS evolution. Overall results pointed out the effectiveness of the proposed recognition methods underlaid by similarity measure and intersection of words between

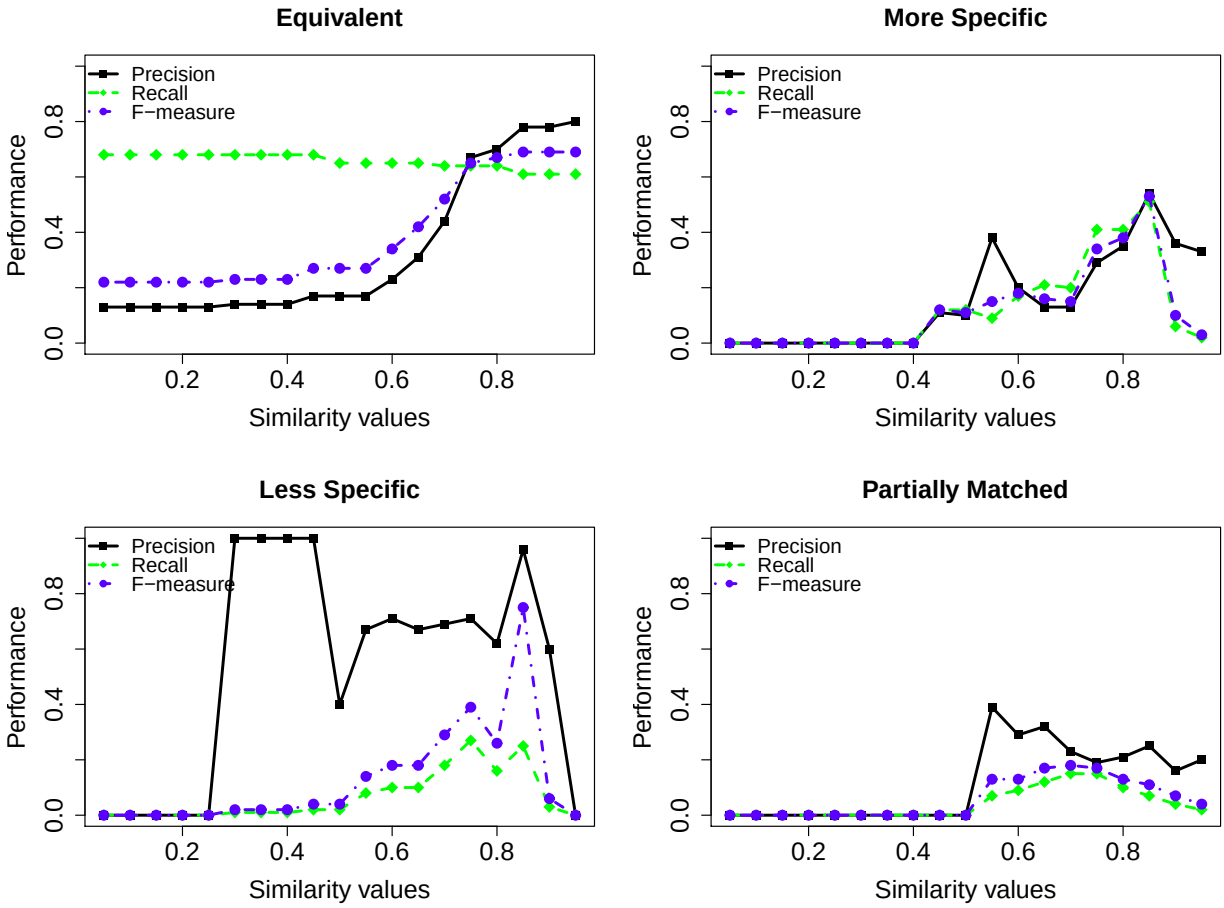


Figure. 5.4: Effectiveness of the Semantic Change Pattern identification algorithm

The graphics show the results for the performance (y-axis) of the metrics of *Precision, Recall* and *F-measure* (*cf.* equations 5.12, 5.13 and 5.14) for different threshold values (x-axis). We evaluate the recognition algorithm against the reference change patterns (gold standard) for the different defined *semantic change patterns*: **Equivalent** (plot on the top left); **More Specific** (top right); **Less Specific** (bottom left); **Partial Match** (bottom right).

attribute values to identify change patterns between KOS versions.

We demonstrated that the explored thresholds (τ and γ) play an important role in the quality of the outcome. We explain this by the fact that our approach selects candidate attributes based on the similarity that is proportional to the degree of relatedness between the analyzed attributes. Results indicated that the adequate values for τ and γ (for words intersections and attribute value similarity, respectively) seem to range from 0.7 and 0.9. However, to determine the real impact that each threshold isolatedly performs on the effectiveness of the change pattern recognition algorithms demands further experiments. We can conduct analyses in a two-step procedure to evaluate the effects of the threshold values, observing the influence of each threshold per time. This can allow us to observe more exact values for these variables that can boost the performance of our algorithms as well as their potential impact in context of mapping adaptation.

The findings revealed evidences of the quality of the outcome yielded by the proposed methods, relying on standard evaluation metrics. We conducted experiments using real biomedical KOSs which strengthen our results. We could improve the obtained results by investigating the combination of the proposed techniques with other approaches based on background knowledge, especially to infer the semantic relations between successive versions of concepts, thus improving the recognition of our semantic change patterns. In this context, the use of string-based similarity metrics may create issues, which could be minimized by using domain-specific background knowledge. This can favor the cases where no inteserction of words exist.

Furthermore, the sequential approach of the algorithm for SCP recognition impacts in particular the results for *Partial Match* because this remains the last option that can be assigned by the algorithm. Also, in the reference change patterns, the pre-selection of attribute couples containing a higher similarity values favors the identification of *Equivalent* SCP type, which naturally will decrease the performance of the *Partial Match*.

We observe that the approach implemented in algorithm 2 for the selection of candidate attributes can influence the results. We could further study additional approaches to select these candidate attributes. Other factors than the similarity values could be explored in this perspective as well as the possibility of considering more than one candidate during the change pattern recognition.

Although existing approaches exploit change patterns to deal with KOS evolution, their definitions frequently rely on ontology meta-models and languages (*e.g.,* OWL or RDF) and are conceptualized taking KOS in an isolated way. We judged this insufficient in the conducted investigation context because their design fails to consider requirements for adapting mappings. Additionally, even though popular initiatives have begun to add more logical formalisms to biomedical KOSs in the last years, *e.g.*, *Gene Ontology*[32], most of the existing biomedical KOSs are rarely fully expressed in standard formats. This makes existing patterns hardly usable because the most disseminated biomedical KOSs (*e.g.,* MeSH, ICD, LOINC, SNOMED-CT, *etc.*) still rely on simple formalisms.

To address this drawback and cope with the special requirements for mapping adaptation according to our previous experiments, we tackled change patterns at the level of attribute values, using linguistic-based features to identify the diffusion and semantic state of textual values between concepts over time. Addressing the change patterns specifically at the level of attributes' values referred to a key lack in KOS evolution literature. While existing approaches and frameworks emphasized change operations mainly at the structural level (which is really important as well), the linguistic and textual level remained unaddressed. This was fulfilled by the investigation conducted in this chapter.

[32] www.geneontology.org

Conclusion

The proposed approach to address mapping adaptation requires understanding KOS evolution in a fine-grained way at the level of attribute values, but existing solutions to calculate KOS changes and patterns fail to entirely support this task. This might play a relevant role in controlling the impact of KOS changes on dependent artefacts. In addition to existing and traditional KOS change operations, the contribution in this chapter originally allowed to characterize KOS evolution by means of change patterns at attribute level.

This chapter proposed and defined change patterns of concept attributes to characterize the evolution of their textual values. This enables to further handle KOS evolution with the specific objective of facilitating adaptation of mappings associated with concepts affected by KOS changes. We specifically designed and implemented novel methods to recognize the proposed change patterns between KOS versions, and validated our proposition by observing the evolution of real biomedical KOSs. On this ground, we are able to detect the adequate evolution of relevant attributes that define mappings (topA), which will stand for the basis of our mapping adaptation method.

In addition, this research empirically evaluated the effectiveness of the proposed methods, and studied the influence of different aspects for the change pattern identification on the quality of the outcome. Our prime finding included a reliable performance of the recognition methods with respect to standard metrics of *Precision* and *Recall*, by assessing the achieved results against a set of manually created reference change patterns specially built for this purpose.

In the next chapter, we will investigate to which extent the different types of change patterns may influence the way KOS mappings evolve, and how to use them to inform which mapping adaptation actions to apply. This will further evidence the benefits and utility of exploring the suggested change patterns, and recognition methods to guide decisions on mappings adaptation.

Chapter 6

Mapping adaptation

Contents

Introduction 125
6.1 Problem statement 127
6.2 Mapping adaptation actions 128
6.3 Study of factors influencing mapping adaptation actions 131
6.3.1 Materials 131
6.3.2 Common experimental procedure 133
6.3.3 KOS changes related to revision and removal of concepts 134
6.3.4 KOS changes affecting relevant concept attributes 138
6.3.5 Impact of lexical and semantic change patterns 139
6.3.6 Summary of the findings 143
6.4 Heuristics guiding decisions of mapping adaptation 144
6.4.1 Move and derivation of mappings 145
6.4.2 Modification of semantic relation 147
6.4.3 Removal and no action adaptation 150
6.5 Experimental evaluation 152
6.5.1 Experimental procedure 152
6.5.2 Experimental results 154
6.6 Discussion 159
Conclusion 161

Introduction

In our attribute-based mapping adaptation approach, we have suggested methods to identify the most relevant attributes for a given mapping (Chapter 4), and methods to characterize the evolution of the attributes from one KOS version to another (Chapter 5). The results yielded by these proposals aim to support and inform our approach to adapt mappings, providing the required elements to address mapping adaptation. However, this demands additional experiments to further understand the behaviors of mapping evolution and to determine how we can apply these techniques to adapt mappings. Existing studies in literature mostly emphasize the consequences of changes at the level of concepts, remaining inefficient for handling mapping adaptation when

changes solely occur at the level of attributes. In this chapter, our research contribution covers this gap.

This chapter investigates different actions suited to adapt mappings and show how to apply them, considering mainly modifying the source concept of mappings and their semantic relation to assure that mappings remain up-to-date when KOSs evolve. We suggest heuristics modeling conditions to trigger one action or other for mapping adaptation. We present ways to take advantage of the proposed KOS change patterns (*cf.* Chapter 5) to adapt mappings and a rule-based technique for modifying the type of semantic relation in mappings. Our solution enables to select and automatically apply a set of predefined mapping adaptation actions (MAAs) to change mappings in an efficient and flexible way.

In particular, we propose and formalize the MAAs, and conduct a series of empirical analyses to observe factors influencing how to perform the selection of the most adequate MAAs. We judge relevant to empirically investigate more fine-grained aspects regarding mapping evolution to justify and propose refined behaviours of mapping adaptation. The experiments reported in this chapter go beyond our previous studies on the evolution of mappings (*cf.* Chapter 3), allowing us to attain a more in-depth understanding of the phenomenon.

We study several aspects correlated with MAAs including: (1) revision and removal KOS change operations; (2) KOS changes affecting relevant concept attributes; and (3) lexical and semantic change patterns. These experiments aim to uncover lessons that can help us to model the necessary conditions to apply the MAAs. On the basis of these empirical observations, we design and formalize heuristics to adapt mappings. In summary, we provide the following contributions in this chapter:

- We define and formalize a set of MAAs expressing different behaviours of mapping changes that are used in our semi-automatic mapping adaptation approach.
- Aiming at adequately applying these actions for each individual mapping affected by KOS evolution, we investigate via experiments the most suited action(s) according to different conditions. We suggest factors that might influence applying these actions and refine our initial experiments with deeper empirical analyses. More specifically, we investigate the impact of relevant attributes as well as lexical and semantic change patterns.
- Inspired by the findings from the analyses, we model a set of heuristics expressing conditions suited to guide decisions on mapping adaptation according to the proposed MAAs.
- Our research experimentally validates the heuristics for mapping adaptation. We evaluate the quality of the mapping adaptation decisions yielded by the proposed heuristics. We use the new mapping releases as our gold standard to measure the obtained results via standard metrics of *Precision*, *Recall* and *F-measure*.

We start by describing and elucidating the problem, which includes the addressed research questions in this chapter (Section 6.1). In the sequence, we present the mapping adaptation actions (Section 6.2) and the set of experiments studying factors that might help us to adapt mappings via the suggested actions (Section 6.3). On this basis, we present the techniques to guide automatic mapping adaptation (Section 6.4). We report on the experimental evaluation in section 6.5 and provide a discussion on the obtained findings (Section 6.6).

6.1 Problem statement

A closer analysis of existing work in literature reveals that removing out-dated mappings accounts for the strategy mostly performed in mapping adaptation. Although this operation avoids having inconsistent mappings as a result, this shows many drawbacks. For instance, only removing mappings results in a less rich final set of adapted mappings. Applications relying on the initial mappings may require the removed correspondences in a near future. Therefore, mapping adaptation should perform in a way that a removal of mappings occurs as the last option of adaptation, *i.e.*, to have additional, sufficiently rich mapping changes to apply more complex operations in mapping adaptation.

The proposed approach aims to replace a deleted concept involved in a mapping by another adequate current concept whenever possible. In addition, our technique aims to redefine the type of semantic relation in mapping adaptation. Existing approaches addressing this aspect remain rare and superficial, which deserves further research efforts to understand how the mapping adaptation process might take the different types of semantic relations of mappings into consideration. Indeed, the majority of the approaches only consider equivalent correspondences, which delimit their power of expressiveness and of adaptation. The kind of adaptation strategies intended by the *DyKOSMap* approach can lead to a more complete set of updated mappings, but entail several difficulties. This chapter addresses the following open research questions:

1. What mapping elements might change in mapping adaptation and how to express meaningful and complex behaviours of modifications performed on mappings?

   ◦ We need to determine the adaptable elements involved in mappings and investigate adequate methods that enable replacing them. This requires investigating techniques to design adaptation actions.

2. Is it possible to use derived information from KOS changes and KOS mappings in a combined way to correlate with adequate adaptation actions suited to correctly adapt mappings on an individual basis?

   ◦ This issue requires studying underlying empirical facts and methods that allow taking decisions on modification of mapping elements.

3. How to determine the most appropriate parameters for the mapping adaptation actions?

   ◦ The proposed approach faces a particular challenge in determining the new value of each element of adapted mappings. For example, in which way can mapping adaptation determine and replace one of the concepts involved in mapping with another adequate concept from a close neighborhood (context), to maintain source and target concepts correctly interrelated.

4. Is it possible to adapt mappings by taking different types of semantic relations in mappings into consideration and enabling their modification?

   ◦ The defined techniques aim at allowing to semi-automatically modify the semantic relation of a mapping (*e.g.*, change an equivalent relation to a subsumption relation) to provide adequate correspondences from the semantic point of view. We must further investigate what information from established mappings and KOS evolution supports determining the new type of semantic relation. We also need to achieve concrete and well evaluated experimental results concerning how to determine a new type of semantic relation at mapping adaptation time.

Figure 6.1 presents the investigated scenario. Given a mapping $m^0_{st} = (c^0_s, c^0_t, semType^0_{st})$ at time j, where a KOS change affects the concept c^0_s, we specifically examine the influence of lexical and semantic change patterns with concepts in the context for mapping adaptation. We name as the c^1_{cand} the concept in the context where our methods identify lexical and semantic change patterns, and c^1_{obs} as the concept where we observe a mapping modification involved by analysing official mapping releases. The main challenge refers to determining the cases where the source concept c^0_s of mapping remains inadequate due to KOS changes, and to selecting a possible adequate candidate concept at time $j+1$ that can replace the old one. Related to this issue, we will need to determine the new adequate type of semantic relation connecting c^1_s or c^1_{cand} with c^0_t.

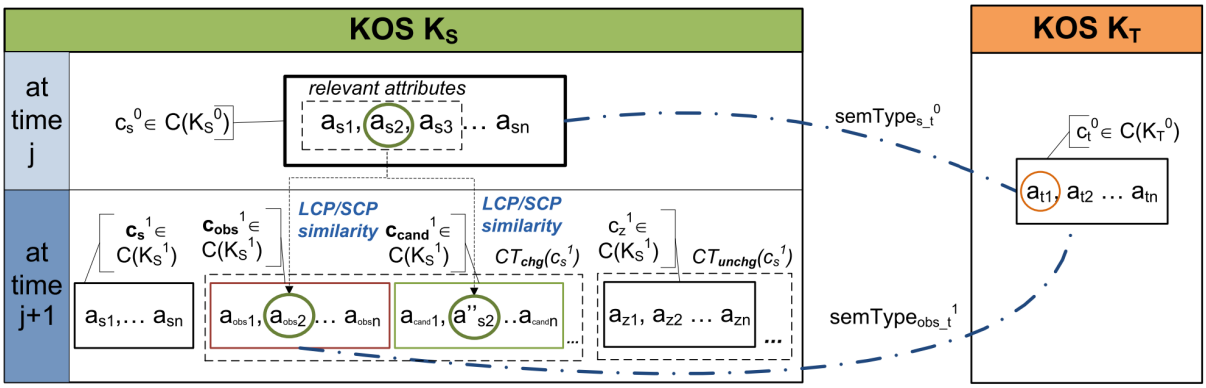


Figure. 6.1: Scenario of the problem for change patterns-based mapping adaptation

This figure presents a scenario for studying positive and negative impact of change patterns on mapping adaptation. Two different KOS are interrelated via a mapping between the concepts c^0_s and c^0_t. Rectangles depicts concepts with their attributes inside and we express two moments in time, which represent the evolution of KOS K_S. The original mapping $m^0_{st} = (c^0_s, c^0_t, semType^0_{st})$ may have possible different modifications (states) at time $j+1$ due to specific KOS changes related to attributes of the source concept c^0_s. Considering relevant attributes selected from all attributes in concept c^0_s for mapping m^0_{st} (circles represent attributes with highest similarity), the problem concerns the determination of correct mapping adaptation actions to assign the adequate state to the affected mapping. To this end, we study the influence of several factors including lexical (LCP) and semantic (SCP) change patterns on modifications observed in the original mapping.

To design and formalize our experiments and methods in this chapter, we rely on definitions and notations already presented in previous chapters (*cf.* Tables 4.1 and 5.1). In the following, we present the MAAs.

6.2 Mapping adaptation actions

Our previous experiments on observing the evolution of mappings (*cf.* Chapter 3) allow us to indicate some well-delineated behaviours of mapping changes that can be applicable for mapping adaptation. The empirical observations have inspired us, and we propose to express behaviours of mapping adaptation as ***Mapping Adaptation Actions*** (MAAs).

We model six distinct actions that represent different possibilities for adapting mappings: *AdditionM*, *RemoveM*, *MoveM*, *DeriveM*, *ModSemTypeM* and *NoAction*. We explore these ac-

tions throughout in this chapter to conduct our experiments and in our proposed method for adapting mappings.

In the following, we formally describe each action. To this end, let $m^0_{st} \in \mathcal{M}^0_{ST}$ (resp. $m^1_{st} \in \mathcal{M}^1_{ST}$) be the mapping between two particular concepts $c^0_s \in C(K^0_S)$ (resp. $c^1_s \in C(K^1_S)$) and $c^0_t \in C(K^0_T)$ (resp. $c^1_t \in C(K^1_T)$) at time j (resp. $j+1$). Moreover, we suppose that c_t remains totally unchanged, while the concept c_s evolves from one KOS version to another. We use the combination of identifiers from source and target concepts to recognize each distinct mapping.

Add mapping. This stands for an atomic action through which a new mapping m^1_{st} is added to $\mathcal{M}^1_{ST}$:

$$AdditionM(m_{st}) \longrightarrow \begin{cases} m^0_{st} \notin \mathcal{M}^0_{ST}, \\ m^1_{st} \in \mathcal{M}^1_{ST} \end{cases} \tag{6.1}$$

Remove mapping. This refers to an atomic action through which a mapping m^0_{st} is deleted from $\mathcal{M}^0_{ST}$:

$$RemoveM(m_{st}) \longrightarrow \begin{cases} m^0_{st} \in \mathcal{M}^0_{ST} \\ m^1_{st} \notin \mathcal{M}^1_{ST} \end{cases} \tag{6.2}$$

Move mapping. In *MoveM* the source concept c_s of the mapping is replaced by another concept c_q. This refers to a composed action for which an existing mapping from $\mathcal{M}^0_{ST}$ is reallocated in $\mathcal{M}^1_{ST}$, thus the source concept is different. This action plays a central role for adapting mappings, by reusing an existing mapping which can be considered invalid in $\mathcal{M}^1_{ST}$ due to KOS changes affecting c_s. The mapping is thus adapted considering its $CT(c^1_s)$. When concept c^1_s does not exist in $C(K^1_S)$, we get its context at time $j+1$ via concepts in $CT(c^0_s)$.

$$MoveM(m_{st}, c^1_q) \longrightarrow \begin{cases} m^0_{st} \in \mathcal{M}^0_{ST}, \\ m^1_{st} \notin \mathcal{M}^1_{ST}, \\ \left[\begin{array}{l} \exists c^1_q \in CT(c^1_s), c^1_s \in C(K^1_S), \\ \vee \\ \exists c^1_w \in C(K^1_S) \wedge \exists c^1_q \in CT(c^1_w), c^1_s \notin C(K^1_S) \wedge c^0_w \in CT(c^0_s) \end{array} \right. \\ \exists m^1_{qt} \in \mathcal{M}^1_{ST} \end{cases} \tag{6.3}$$

Derive mapping. In *DeriveM*, the original mapping remains in $\mathcal{M}^1_{ST}$ and a new mapping appears connecting a concept c_q with c_t. Similar to *MoveM* action, $c_q \in CT(c^1_s)$. Therefore, this refers to a composed action for which an existing mapping in $\mathcal{M}^0_{ST}$ has a modified copy in $\mathcal{M}^1_{ST}$ with a different source concept, which belongs to the context of the original source concept. This action plays a role for reusing an existing mapping, which can be still considered as correct in

$\mathcal{M}^1_{ST}$. Note that for a given mapping, several *DeriveM* actions can be applied.

$$DeriveM(m_{st}, c^1_q) \longrightarrow \begin{cases} m^0_{st} \in \mathcal{M}^0_{ST}, \\ m^1_{st} \in \mathcal{M}^1_{ST}, \\ c^1_s \in C(K^1_S), \\ \exists c^1_q \in CT(c^1_s), m^1_{qt} \in \mathcal{M}^1_{ST} \end{cases} \tag{6.4}$$

Modify semantic relation of mapping. This action consists of modifying the type of semantic relation. This refers to a composed action in which the type of the semantic relation of a given mapping is modified by a different one. We propose this action for supporting the adaptation of mappings with different types of semantic relations, rather than only considering the type of equivalence relation (≡).

$$ModSemTypeM(m_{st}, newSemType_{st}) \longrightarrow \begin{cases} m^0_{st} \in \mathcal{M}^0_{ST} \\ \exists semType^1_{st} \in \{\perp, \equiv, \leq, \geq, \approx\} \\ \exists m^1_{st} \in \mathcal{M}^1_{ST}, semType^1_{st} = newSemType_{st} \end{cases} \tag{6.5}$$

No action. The *NoAction* refers to the cases where mappings remain unchanged. Formally:

$$NoAction(m_{st}) \longrightarrow \begin{cases} m^0_{st} \in \mathcal{M}^0_{ST}, \\ m^1_{st} \in \mathcal{M}^1_{ST}, \\ semType^0_{st} = semType^1_{st} \end{cases} \tag{6.6}$$

We can apply the action for the modification of semantic relation in combination with the actions of move or derivation of mapping. When moving/deriving a mapping, we make the modification of semantic relation type of such mapping possible. Figure 6.2 presents an illustration of the proposed mapping adaptation actions.

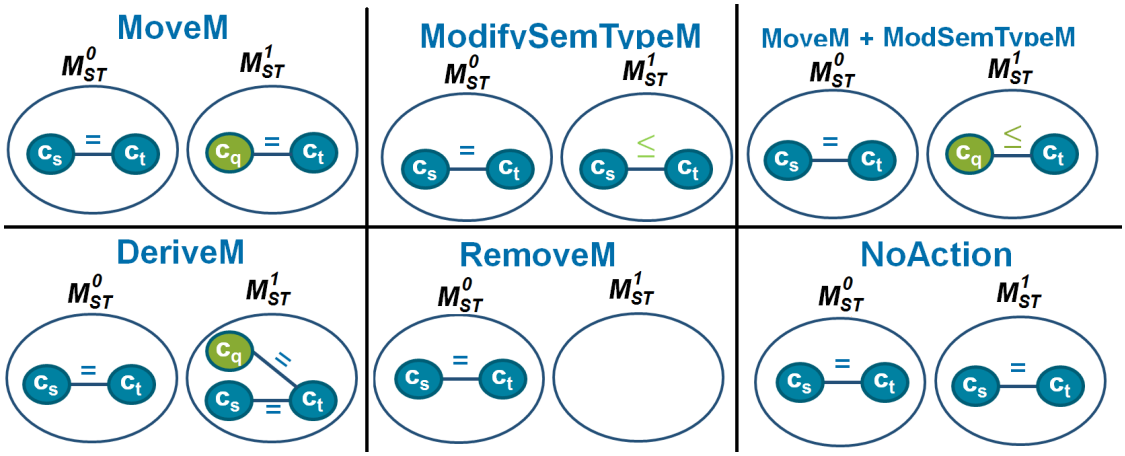


Figure. 6.2: Mapping adaptation actions

This figure presents examples of mapping adaptation actions illustrating their behaviour.

6.3 Study of factors influencing mapping adaptation actions

We conduct a set of experiments to understand the influence of several aspects for applying MAAs. We aim to take into consideration the resulted empirical observations from these experiments to define heuristics that support the selection and the execution of MAAs. (*cf.* Section 6.4). In summary, we propose three analyses:

1. We measure the frequency of occurence of the proposed MAAs by observing mapping evolution and the correlation to KOS change operations that revise concepts (split, merge and substitution), and specific KOS change operations that remove concepts. We aim to observe the mappings' behaviors after executing KOS change operations, and classify these behaviours according to the MAAs. Section 6.3.3 reports on the results of this analysis.

2. We calculate the number of different KOS changes specifically affecting relevant attributes computed with our method (*cf.* Algorithm 1) by correlating that with MAAs. This analysis intends to observe whether the KOS changes affecting these relevant attributes can influence the observed MAAs. Section 6.3.4 describes the obtained results.

3. We aim to examine the influence of our lexical (*cf.* Section 5.2.1) and semantic (*cf.* Section 5.2.2) change patterns. We measure to which extent these patterns can play a role for helping deciding one action or other in mapping adaptation. Section 6.3.5 presents the results for this experiment.

In the following, we first describe the used materials (Section 6.3.1) followed by the experimental procedure in common for all conducted analyses (Section 6.3.2). This stands for basic steps applied for all implemented experiments. In addition, each experiment contains specific procedures that we describe in the respective subsections. After describing the obtained results in the series of experiments, section 6.3.6 provides a summary of the findings and briefly discusses how they can inform the heuristics modeling.

6.3.1 Materials

The experimental analyses in this section and the conducted validation (Section 6.5) are based on several datasets regarding biomedical KOSs and their associated mappings. Similar to previous chapters, experiments force us to rely on mapping datasets with several versions of validated mappings that are available for research purposes.

In addition to SCT, MeSH and ICD9 already considered in previous chapters (*cf.* Table 5.4), we include the use of ICD-10-CM[33] (ICD10) and NCI Thesaurus[34] (NCI). Table 6.1 presents statistics regarding the number of concepts, attributes denoting concepts, and of direct subsumption relationships between concepts for NCI and ICD10. As the characteristics of ICD10 are similar to ICD9 already described (*cf.* Section 3.1.1), we provide only a brief description of NCI.

NCI Thesaurus. The *National Cancer Institute Thesaurus* (NCI) refers to a terminology covering areas of basic and clinical science. NCI has a cancer-centric focus content, and the

[33] www.cdc.gov/nchs/icd/icd10cm.htm
[34] ncit.nci.nih.gov

National Cancer Institute[35] has originally designed this thesaurus as a key element of the cancer common ontological representation environment (caCORE). The basic unit of meaning in the NCI Thesaurus stands for concepts. NCI organizes them in an unlimited hierarchy level and also enables multiple inheritance. The concepts contain properties as their unique code and textual definitions. Moreover, each concept has a preferred name and may have different explicit synonyms relations.

Table 6.1: Statistics on the NCI Thesaurus and ICD10CM

KOS	**Release**	**#Concepts**	**#Attributes**	**#Subsumptions**
NCI	2009	77 448	282 434	86 822
	2012	94 732	365 515	105 406
ICD10CM	2011	43 351	87 354	40 330

This table presents the number of concepts, attributes and subsumption relationships of the NCI Thesaurus for releases of 2009 and 2012; and ICD10CM release 2011. See table 5.4 for other biomedical KOS considered.

We take NCI and ICD10 into account in this chapter because we include two supplementary mapping datasets. In addition to the official mappings[36] established between SCT and ICD9 provided for each release of SCT by the IHTSDO[37] organization, we include mappings established between MeSH and ICD10, and between SCT and NCI. Mappings between MeSH and ICD10 were established by the *CisMef team*[38] involving biomedical experts, while the mappings linking SCT and NCI were extracted using the UMLS (*cf.* [Jiménez-Ruiz et al., 2011] for extraction details). We refer to these latter mappings as silver-standards and they only express equivalent relations. Exploring this material allows us studying diversified data in experiments and evaluations in addition to SCT-ICD9. Table 6.2 shows the quantity of mappings established between various mapping releases of SCT-ICD9, SCT-NCI and MeSH-ICD10.

Table 6.2: Mappings between biomedical KOSs

KOS mapping	**Release**	**#Mappings**
SNOMEDCT-ICD9CM	2010-2009	100 451
	2012-2011	102 703
SNOMEDCT-NCI	2009-2009	19 971
	2012-2012	22 732
MeSH-ICD10CM	2012-2011	4 631
	2013-2011	5 378

This table shows statistics for the several releases of studied KOS mappings.

Except for the experiment described in section 6.3.3 (Analysis 1), where we use the mapping

[35] www.cancer.gov
[36] www.nlm.nih.gov/research/umls/mapping_projects/icd9cm_to_snomedct.html
[37] www.ihtsdo.org
[38] www.cismef.org

datasets SCT-ICD9 and SCT-NCI, for all other experimental analyses, we use the mapping datasets SCT-ICD9 and MeSH-ICD10 to make observations. This is due to the fact that we conducted these experiments at different moments. We take into account all the datasets to conduct the experimental validation (Section 6.5).

6.3.2 Common experimental procedure

We apply the following procedure as a first step for all conducted analyses:

- Calculate KOS *diff*;
- Select mappings impacted by KOS changes;
- Calculate observed mapping adaptation actions (MAAs);

Calculate KOS *diff*. Firstly, we identify a set of KOS changes (namely $diff(K_x^1, K_x^2)$) using two different releases of the same KOS K_x. Table 6.3 presents the number of changes for the different KOS change operations calculated with *COnto-Diff* tool [Hartung et al., 2013] between the two releases of SCT, ICD9, NCI and MeSH, respectively. The *diff* helps identifying the mappings impacted by KOS evolution (next step).

Table 6.3: KOS change operations identified for the releases of biomedical KOSs

	Number of changes					Number of changes			
KCO	**SCT**	**ICD9**	**MeSH**	**NCI**	**KCO**	**SCT**	**ICD9**	**MeSH**	**NCI**
merge	0	0	0	35	addLeaf	3 649	140	372	347
split	380	45	0	225	addSubGraph	0	0	0	9
substitute	0	0	0	3	addR	4 327	110	23	541
toObsolete	794	0	0	0	delR	7 210	39	33	155
revokeObsolete	20	0	0	0	move	7 140	31	0	91
delInner	0	0	8	99	addA	7 720	79	1 817	3 156
delLeaf	10	0	0	286	delA	4 003	25	192	80
delSubGraph	0	0	0	0	chgAttValue	950	155	119	185
addInner	1348	26	291	13	**Total changes**	37297	624	2 564	5 225

This table shows the number of different KOS change operations (KCOs) calculated between two different versions of biomedical KOSs considered in the experiments by using the *COnto-Diff* tool [Hartung et al., 2013]. See tables 2.1 and 2.2 for the formalization of the change operations.

Select mappings impacted by KOS changes. Given all mappings between source and target KOS (SCT and ICD9 (*resp.* between MeSH and ICD10, and SCT and NCI), we select only those affected by some KOS changes in *diff*. From the analyses, we removed unaffected mappings (*i.e.*, the source and target concepts are unchanged) because we aim to study the influence of KOS changes on how mappings evolve. We study the evolution of both source and target KOSs involved in mappings, but we consider that only one KOS evolved at time j to assure consistence in our experiments, and to avoid misunderstandings on the results analysis to understand the researched interdependencies between KOS changes and MAAs. We must make sure that a modification applied to a mapping is only motivated by changes in one of the two concepts (source or target). Therefore, we also remove mappings where the *diff* affects both

source and target concepts of mappings at the same time.

Calculate observed mapping adaptation actions (MAAs). For each mapping impacted by KOS changes, we determined a list of MAAs for such mapping. For instance, a mapping can be removed while for another mapping a *MoveM* with a modification of relation occurs. We defined a method that calculates the MAAs given a mapping at time j. The method compares the elements of the mapping (identifier of source and target concept, and relation) with the elements of the set of mappings at time $j+1$. Whether our method founds all elements of the given mapping at time $j+1$ (all elements remain equal), the method sets no MAAs (*NoAction, i.e., unchanged mapping*). Note that having both versions of the mapping releases enables us to determine whether a mapping was moved or derived instead of only being removed or added. The method points out a *MoveM* action if it detects a new mapping connecting the target concept of the given mapping and a source concept coming from the context of the original source concept. Additionally, the original source and target concepts must not remain connected at time $j+1$ (*i.e.,* the original mapping is removed), otherwise the method points out a *DeriveM* action (*i.e.*, the original mapping remains unchanged, but one or more new concept is/are linked with the target concept). The method sets a *RemoveM* if, and only if, a new mapping does not exist at time $j+1$, interrelating a concept from the context of the source concept of the given mapping with the target concept. For instance, in the first release of mappings between SCT and ICD9 we find the mapping between "*Primary malignant neoplasm of intrahepatic bile ducts (disorder)*", with the semantic type (≡) linking to the ICD9 concept named "*Malignant neoplasm of Intrahepatic bile ducts*". In the second release this mapping is no longer found, which indicates a *RemoveM* action. Table 6.4 presents the total number of MAAs observed in the used datasets. We use these actions as our references in the experimental validation (Section 6.5).

Table 6.4: Computed mapping adaptation actions observing mapping evolution

MAA	**#MAA observed**		
	SCT-ICD9CM	SCT-NCI	MeSH-ICD10
MoveM	635	4	0
DeriveM	747	14	21
ModSemTypeM	383	0	0
RemoveM	187	316	6
NoAction	9 031	5 595	249

This table shows the numbers of identified MAAs calculated between the KOS mapping releases related to those source/target concepts in mappings affected by KOS changes.

6.3.3 KOS changes related to revision and removal of concepts

This analysis aims to examine to which extent the suggested MAAs are observed, where the source concept suffers at least one KOS change operation. In particular, we emphasize the KOS change operations related to the revision and removal of concepts.

We first apply the common procedure that calculates the *diff*, select the mappings affected by KOS changes from the *diff* and calculate the MAAs for these mappings. We use the two versions of SCT, ICD9 and NCI as well as the mapping datasets SCT-ICD9 and SCT-NCI (*cf.* Table 6.2).

We generate the *diff* for each KOS and filter the KOS change operations focused on this experiment. We retain six KOS change operations at the level of concepts as presented in table 6.5.

For all concepts identified in the KOS *diff*, we have the corresponding impacted mappings in $\mathcal{M}^0_{ST}$. This produces for each type of KOS change a particular subset of mappings that contains the MAAs calculated for them. To analyze the results with a global view, we group all instances of changes identified for each type of KOS change in the calculated *diffs* of all observed KOSs (SCT, ICD9 and NCI). Within this assumption, we also group the subset of associated mappings accordingly. Table 6.5 shows the total number of analyzed mappings for each type of KOS change.

We select the calculated MAAs for the filtered mappings associated with the KOS change operations studied. This experiment evaluates the behaviour of each individual mapping according to the proposed MAAs for each type of KOS change operation. To this end, for each subset of mappings under analysis, related to the different KOS change operations, we measure the proportion of each type of MAA correctly identified in the evolution of mappings from one mapping release to another.

Table 6.5: Quantity of analyzed mappings for the types of complex KOS changes

		#KOS changes			#Mappings
	KOS / **KCO**	**ICD**	**NCI**	**SCT**	**ICD+NCI+SCT**
Revision	$substitute(c_i, c_h)$	0	3	0	3
	$merge(C_r, c_h)$	0	35	0	59
	$split(c_i, C_r)$	45	225	380	858
Removal	$delInner(c_i)$	0	99	0	48
	$delLeaf(c_i)$	0	286	10	74
	$toObsolete(c_i)$	0	0	794	1704

This table shows the number of KOS change operations studied and the number of analysed mappings associated with them.

We classify the results according to the nature of the KOS change operations. We first analyze KOS change operations related to revision. Figure 6.3 depicts the percentage of MAAs applied for the studied mappings in context of revision change operations (*substitute*, *split* and *merge*). We notice that the $MoveM(C_r)$ MAA is applied with a percentage of 65% for the *split* operation in source concepts. C_r refers to the set of resulting concepts of the split change operation. We observe that in some cases, two MAAs like $MoveM(C_r)$ and $ModSemTypeM$ can be simultaneously applied with a percentage of 10.26%. For *merge* change operation, mappings are adapted by applying $MoveM(C_r)$ (22.04%) or *RemoveM* (19%). In case of *substitute* concepts, mappings are more frequently removed (67%) than moved. We explain the latter by the fact that whenever a relevant concept attribute relating to the source concept of mapping is not detected in the resulting concept (that replaces the old one), then the mapping is removed. This highlights the relevance of observing the ***topA*** attributes, identified from existing mappings impacted by KOS changes, in the resulting concepts of changes like *substitute*, *split* and *merge*.

The *NoAction* in figure 6.3 indicates that the concept affected by *split* or *merge* remains in the resulting set of concepts, and the associated mappings keep linked with the same concept at time $j+1$.

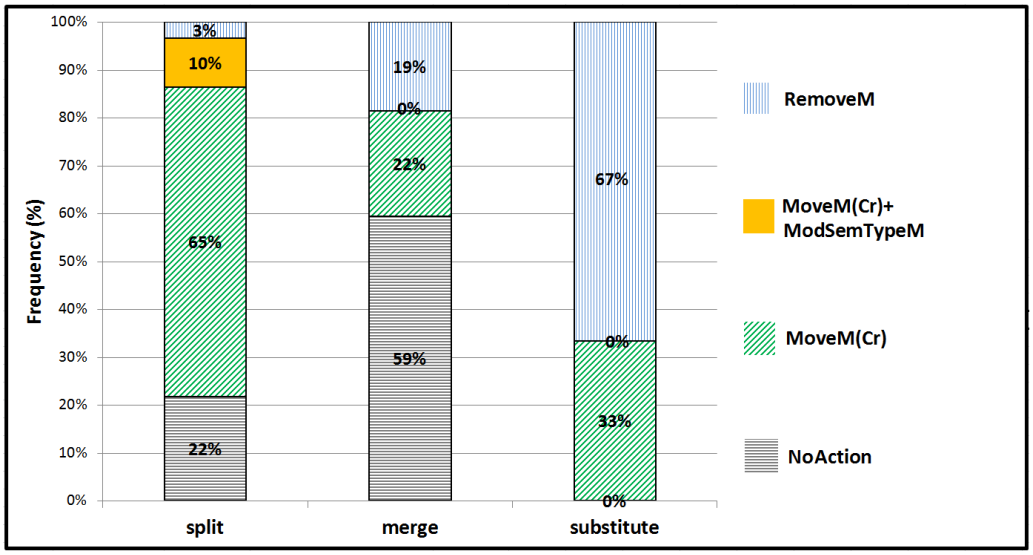


Figure. 6.3: Analysis of mapping adaptation actions under revision KOS changes

This figure presents the proportion of MAAs observed when concepts in mappings are affected by *substitute*, *split* and *merge* KOS changes.

Figure 6.4 (left side) presents the general proposal of adapting mappings according to the revision KOS change operations (split in this case). In this figure, we represent the evolution of the concept $c_{s_1} \in C(K_S^0)$, three mappings are associated with this source concept at time j. After evolution, a split of the concept c_{s_1} occurs with two sibling concepts on which a mapping m_1 remains attached to c_{s_1}, and two other mappings (m_2' and m_3') are moved to resulting concepts accordingly.

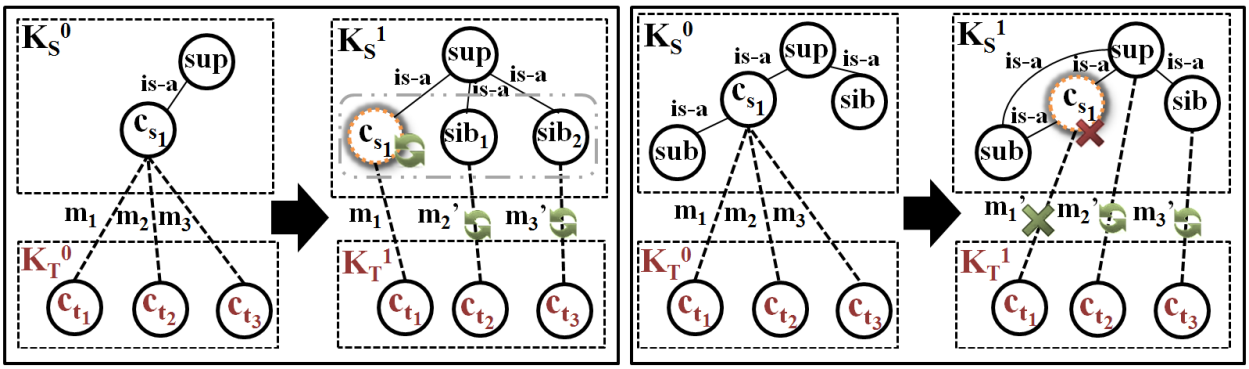


Figure. 6.4: Behaviour of mappings under revision and removal KOS changes

This figure illustrates mapping behaviours when revision (left side) and removal (right side) KOS changes affect involved concepts of mappings.

We analyse the MAAs in the context of KOS change operations regarding removal of concepts (*cf.* Figure 6.5). This examines whether mappings evolve by removing and also by moving the source concept to a concept in its context. When a concept is assigned to *toObsolete*, several MAAs are observed, but at different frequencies: $MoveM(sup)$ – move the source concept to a

super concept (65%), $MoveM(sib)$ – move the source concept to a sibling concept (23%) and $RemoveM$ – remove the mapping (9%). Note that mappings are mostly moved to a super concept rather than removed. Mappings are mainly moved to sibling concepts ($MoveM(sib)$) when the occurring KOS change operation is $delLeaf$ (50%). Similarly, the same type of MAA is also applied for a significant percentage of mappings (38%), where the source concept is affected by the KOS change operation $delInner$, but mappings are more frequently removed (48%) when inner concepts are deleted.

These results highlight the importance of considering the context with hierarchical relationships for mapping adaptation. In the context of concept deletion KOS operation, results indicate that the $RemoveM$ action is applied, but with a low percentage than $MoveM$. This underscores that for reaching a high quality of mapping adaptation, we need to carefully search for the relevant conceptual content related to existing mappings in the context of source concepts, where our change patterns must be useful (*cf.* Section 6.3.5), and consider adequate hierarchical relationships between evolving concepts. This reinforces the findings obtained in chapter 3.

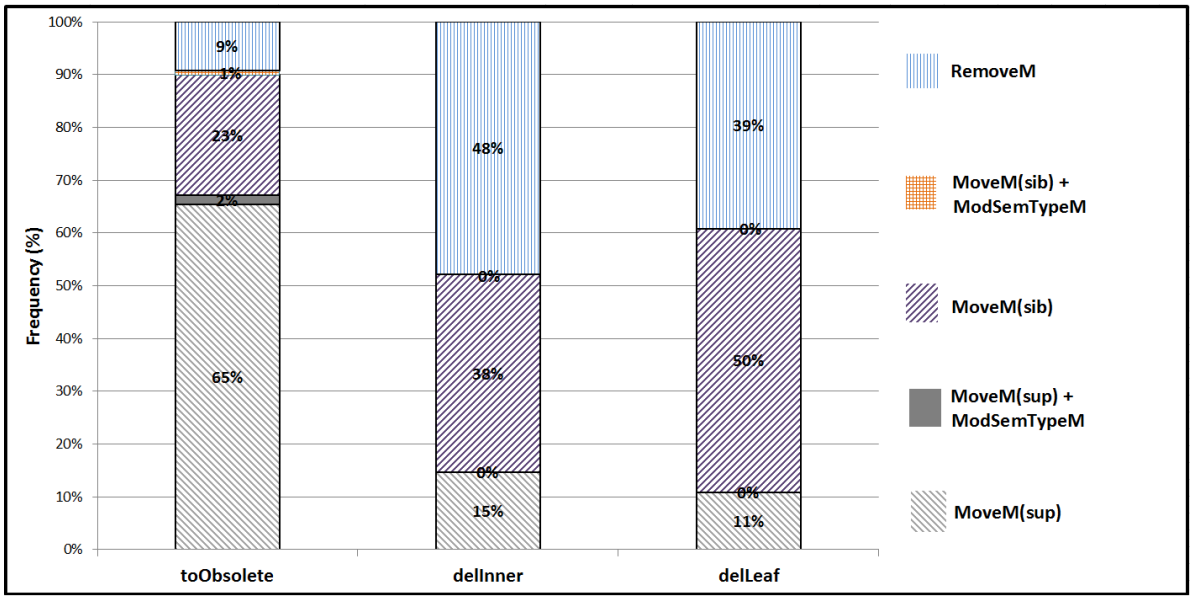


Figure. 6.5: Analysis of mapping adaptation actions under removal KOS changes

This figure presents the proportion of MAAs observed when source concepts are affected by *toObsolete*, *delInner* and *delLeaf* KOS changes.

Figure 6.4 (right side) presents a general proposal of adapting mappings according to the removal KOS change operations. Given the evolution of the concept $c_{s_1} \in C(K_S^0)$, three mappings are associated with this concept at time j (m_1, m_2, m_3). After evolution, some attributes belonging to c_{s_1} are deleted or the whole concept is deleted. In consequence, one mapping is removed (m_1') while two others (m_2' and m_3') are moved to concepts in the context of c_{s_1} at time j+1.

Overall, the obtained results reveal that considering the delimited context of KOS complex changes for handling revision KOS change operations (*i.e.*, the resulting set of concepts of a split for instance), and the context (CT) for removal operations appears a relevant approach to adapt-

ing mappings. Furthermore, the proposed mapping adaptation strategies based on the MAAs according to the KOS change operations can be mainly applied in cases of revision and removal of concepts in KOS evolution. Results point out the relevance of mapping maintenance considering the adaptation decisions of mappings on an individual basis. We demonstrated evidences that mappings associated with the same type of KOS change behave differently. Therefore, our technique must carefully analyze and determine candidate actions for each mapping.

6.3.4 KOS changes affecting relevant concept attributes

This experiment emphasizes the level of concept attributes and aims at observing whether KOS changes, specifically affecting the relevant concept attributes for mappings, show evidences of usefulness in helping to decide on adequate MAAs in mapping adaptation. Firstly, we apply the common procedure (Section 6.3.2) to calculate the KOS *diff*, to select the mappings and to calculate MAAs. We conduct this experiment with the mapping releases regarding SCT-ICD9 and MeSH-ICD10. Table 6.4 presents the total number of MAAs observed, summing the results of these mapping releases. We identify the relevant attributes as follows.

Identify relevant attributes. For each source concept of a mapping impacted by KOS evolution, we use our designed and evaluated method for identifying the most relevant concept attributes (*cf.* Chapter 4). Given a mapping m_{st} between two concepts $c_s \in C(K_S^0)$ and $c_t \in C(K_T^0)$, the proposed algorithm 1 retrieves a set of source concept attributes, consisting of the most similar to the ones in the target concept. For the experiments conducted in this chapter, we set the number of relevant attributes to three, which represents the most meaningful attributes for a given mapping. The results in chapter 4 motivate the use of this number for relevant attributes. Since a concept may be the source concept of more than one mapping, the relevant attributes are specific to one mapping and can therefore differ from one mapping to another. For example, among all attributes characterizing the concept identified by the code identifier '*422338006*' in SCT, the attribute with value "*Senile macular retinal degeneration*" stands for the attribute with the highest similarity with those attributes in the concept of ICD9.

We observe three types of KOS changes (unchanged, deleted and modified) which affects the relevant attributes correlating them with the different MAAs. For this purpose, for each type of mapping adaptation action, we calculate the number of unchanged, deleted and modified attributes in three perspectives: (1) we count the absolute number of observed change type affecting every single relevant attribute (named #single); (2) we count when the same type of KOS change simultaneously affects all relevant attributes of the concept (named #whole); (3) we count the number of change types affecting the best relevant attribute (named #best) (*i.e.*, the most similar source concept attribute to an attribute of the target concept). The latter perspective allows us to observe the role played by the most similar attribute identified among the relevant attributes for mapping adaptation.

Table 6.6 presents the obtained results. We observe that when mappings evolve with *RemoveM* action, the relevant attributes are deleted more frequently; compared with the respective total number of action type (*cf.* Table 6.4 – SCT-ICD9 and MeSH-ICD10, which shows a total of 187 + 6 *RemoveM*). Observe that the best attribute plays a relevant role because the number of *#best* is higher than the number of *#whole*. This indicates that when the best relevant attribute is deleted, associated mappings are removed. When observing *MoveM*, *DeriveM*, *ModSemTypeM* and *NoAction*, relevant attributes remain mostly unchanged. Our experiment cannot indicate

any clear influence of modified attributes on mapping adaptation actions.

Table 6.6: Correlation between KOS changes affecting relevant attributes and MAAs

MAA \ Changes	Unchanged			Deleted			Modified		
	#single	#whole	#best	#single	#whole	#best	#single	#whole	#best
MoveM	1 084	156	443	382	41	123	101	0	69
DeriveM	2 271	738	741	21	1	16	11	0	11
ModSemTypeM	1004	294	331	78	8	33	21	0	19
RemoveM	45	2	15	495	137	169	18	0	9
NoAction	24 038	7 205	9 068	190	0	57	300	0	155

This table shows the change behaviours of relevant attributes identified (*cf.* Algorithm 1) for the MAAs. We use $topA(c_s, c_t, n)$ with $n = 3$ (*i.e.*, the 3 most relevant source concept attributes for a given mapping) and calculate the number of types of changes (unchanged, deleted, modified) considering: (1) every single relevant attribute (#single); the number of times that the same type of KOS change simultaneously affects the whole set of relevant attributes (#whole); (3) changes affecting only the most similar attribute with the target concept (#best).

The achieved results point out an interesting behaviour that, combined with other factors studied in this section, can help us modeling the heuristics.

6.3.5 Impact of lexical and semantic change patterns

We analyse the influence of change patterns on mapping evolution. This experiment aims to assess the utility of considering lexical and semantic change patterns (*cf.* Chapter 5) for supporting mapping adaptation under evolving KOS. To this end, we measure the potential of using these patterns to help us deciding MAAs in mapping adaptation. Our investigation expects observing the correlation between modifications occurred in mappings, between the source and target concepts (*i.e.*, MAAs), and the change patterns identified in the evolution of the source concepts.

We first apply the common experimental procedure (*cf.* Section 6.3.2), combined with the procedure to identify relevant attributes (*cf.* Section 6.3.4). In this analysis, we set the number of relevant attributes to three, representing the most meaningful attributes for a given mapping. We propose to recognize change patterns for those relevant attributes detected. Similar to the previous experiments, we use the mappings between SCT-ICD9 and MeSH-ICD10. Table 6.4 presents the number of observed MAAs in this study. For each different expected behaviour of mapping adaptation (*i.e.*, observed MAAs), we compute the frequency of each type of lexical (*cf.* Section 5.2.1) and semantic (*cf.* Section 5.2.2) change patterns identified. We aim to detect possible interdependencies between the types of MAAs and the types of change patterns. We conduct the following procedure to recognize the change patterns.

Recognize change patterns. For each retrieved relevant attribute from a source concept c_s of impacted mappings, we identify lexical and semantic change patterns using algorithms 3 and 4, considering the candidate attribute a_q^1 of the concept c_{cand}^1 from the context $CT(c_s^1)$. We calculate the total number of identified LCPs and SCPs for each type of MAAs. When no (lexical or semantic) change pattern is detected for the whole set of relevant attributes, we count *no-lcp* and *no-scp*, respectively. In these cases, we also count the frequency of no changes in

diff affect the whole set of concepts in the context. This allows us to observe cases where the algorithms fail to identify change patterns because the source concept context of the mapping remains unchanged. This experiment enables us to thoroughly understand the updating process of mappings and to identify potential correlations between MAAs and the change patterns. We defined the threshold of τ and γ based on the results of our previous experiments. Analysing results of the evaluation conducted in section 5.6, we found a compromise regarding the threshold value of $\gamma = 0.9$ (for the word comparison threshold) and $\tau = 0.7$ (similarity value threshold) for the change pattern recognition algorithms. We use these thresholds to conduct the experiments.

Measure positive and negative impact. We intend to analyse in more detail two MAAs: *MoveM* and *DeriveM*. This step aims to investigate whether our change pattern recognition algorithms allow us to correctly identify the concepts involved in these actions. For *MoveM* and *DeriveM* actions, a concept c^1_{obs} in the context of the source concept in the original mapping plays a major role in mapping adaptation (*cf.* equations 6.3 and 6.4 where we can consider c^1_q as equivalent to c^1_{obs} here). Indeed, c^1_{obs} becomes the source concept of the adapted mapping. We calculated the frequency of cases where the algorithms identify LCPs and SCPs with attributes issued from c^1_{obs}. More specifically, we considered a positive impact of change pattern if the candidate concept c^1_{cand} in context (to which CP is recognized) is equal to c^1_{obs}. Otherwise, we judge a negative impact if the candidate concept c^1_{cand} differs from c^1_{obs} (*i.e.*, the algorithm recognizes the right CP, but with a wrong candidate concept according to the observations on mapping evolution). The negative impact means that the identified change patterns fail to influence the mapping adaptation action applied for a mapping. In this case, we could find a change pattern, but relating to a concept that differs from the concept involved in a *MoveM* or *DeriveM* observed (*i.e.*, the most adequate one). Figure 6.1 presents a scenario for studying positive and negative impact of change patterns on mappings adaptation with both concepts c^1_{obs} and c^1_{cand}. The figure highlights that a *MoveM* happens with the concept c^1_{obs} and CPs can be identified with attributes issued from concept c^1_{cand} (*i.e.*, a negative impact).

In the following, we present the achieved results for the influence of both lexical and semantic change patterns.

Lexical change patterns in correlation with mapping adaptation actions

This experiment investigates correlations between MAA and LCPs. In particular, we evaluate the capacity of the LCP algorithm to correctly identify to which concept the mapping will be associated when the *MoveM* and *DeriveM* actions are applied. To specifically present these results, table 6.7 includes two indicators "*positive impact*" (represented by the symbol ▲), which indicates that our algorithm recognized the adequate concepts for *MoveM* and *DeriveM* and "*negative impact*" (represented by the symbol ▽), which indicates that the algorithm fails to find the adequate concept, *i.e.*, c^1_{obs}. Note that we conducted this analysis by considering the whole set of interrelated concepts as initial input (datasets SCT-ICD9 and MeSH-ICD10), and did not constrain the set by the number of cases in the constructed reference change patterns (as conducted in section 5.6).

Table 6.7 presents the obtained results when running algorithm 3 with $\gamma = 0.9$ and $\tau = 0.7$. This highlights that correlations exist between LCPs and MAAs. However, these correlations are too complex to accurately describe, due to the fact that we observe each LCP type at least once for each MAA. Despite this fact, these results globally allow highlighting interesting aspects.

When analyzing MAAs, we observe the following findings:

- *MoveM* is more frequently associated with *Total Transfer* and *Total Copy*.
- *DeriveM* is more frequently associated with *Total Copy* and *Partial Copy*.
- *ModSemTypeR* is more frequently associated with *Partial Copy*.
- *RemoveM* and *NoAction* are more frequently associated with *noLCP*.

When analyzing the different types of LCPs, we point out the following facts:

- *Total Transfer* has a strong correlation with *MoveM*.
- *Partial Transfer* remains well distributed between MAAs, but the predominance is *MoveM* and *NoAction*.
- *Total Copy* is strongly associated with *MoveM*, *DeriveM* as well as *NoAction*.
- *Partial Copy* is the most frequent case and is well distributed between MAAs with predominance to *DeriveM*.

Overall, LCP identification algorithm yields positive impact more frequently than a negative one for *MoveM* and *DeriveM*. This suggests that identifying LCPs appears useful to determine a *MoveM* or a *DeriveM* in mapping adaptation.

The number of *RemoveM* and *ModSemTypeR* actions decreases compared to the number of *MoveM* and *DeriveM*. When adapting mappings with *RemoveM* and *ModSemTypeR* actions, the algorithm 3 identifies very few LCPs. This suggests that if no LCP is recognized, it can be more convenient to adapt mappings by removing them or by modifying the semantic relation instead of applying *MoveM* or *DeriveM* actions.

When we observe *NoAction* for the impacted mappings, a similar scenario occurs compared to *RemoveM* and *ModSemTypeR* actions. However, MAAs other than *NoAction* are applied in a smaller part of the mappings impacted by KOS change operations, *i.e., NoAction* occurs much more frequent (*cf.* Table 6.4). The high number of *NoAction* occurrence (~95%), when *noLCP* cases are observed, shows the relevance of our approach to reduce human effort for mapping maintenance. The algorithm positively fails to identify LCPs in most of the cases where the corresponding mappings remain unchanged.

Obtained results demonstrate that the proposed change patterns provide useful statements to support actions for adapting mappings, since the defined change patterns influence the way mappings evolve. Our analysis of positive and negative impact indicated, especially for *MoveM* and *DeriveM* actions, that LCPs show an evidence for adapting mappings using these actions. Similarly, when our method fails to identify change patterns in the source concept of a mapping, it may suggest applying a mapping adaptation action such as *RemoveM* or *NoAction*. This study indicates that LCPs remain useful in supporting the decision process of MAA selection to maintain mappings up-to-date.

Table 6.7: Correlation between lexical change patterns and mapping adaptation actions

LCP / MAA	transfer of attributes		copy of attributes		no-lcp	
	#TT	#PT	#TC	#PC	#MAA	#noChgCT
MoveM	182▲;188▽	25▲;28▽	135▲;35▽	73▲;54▽	221	2
DeriveM	5▲;6▽	2▲;7▽	139▲;124▽	470▲;277▽	264	24
ModSemTypeM	44	7	45	89	211	41
RemoveM	14	1	7	8	173	140
NoAction	28	51	101	468	8 865	5 482

This table presents the number of LCPs identified for each type of MAA. This shows the numbers of LCPs that have a positive impact (denoted as ▲), *i.e.*, concept where the algorithm recognizes the LCP is equal to the concept where a *MoveM* or *DeriveM* occurs ($c^1_{cand} = c^1_{obs}$) and negative impact (denoted as ▽), *i.e.*, concept where the algorithm recognizes a LCP differs to the concept where a *MoveM* or *DeriveM* occurs [$c^1_{cand} \neq c^1_{obs}$], on each MAA. $\#MAA$ expresses the number of MAAs where no LCP is detected, while $\#noChgCT$ presents the number of MAAs where no changes occurs for the whole set of concepts in the context of the source concept in mapping.

Semantic change patterns in correlation with mapping adaptation actions

Table 6.8 presents the number of SCPs identified for the different types of MAAs. In general, the performance of the algorithm 4 to recognize accurate SCPs in the identified correlations remains good. The number of SCPs with a positive impact is always higher than negative impact (for *MoveM* and *DeriveM* actions), except for the correlation between *MoveM* and *Less Specific* (MSP in table 6.8). In future work, we can combine the results of the correlations between lexical and semantic change patterns with respect to MAAs in order to try increasing the performance of our algorithms to detect more positive cases.

Similar to the LCP impact analysis, the number of SCPs found is smaller for the *RemoveM* and *ModSemTypeR* actions. This confirms a potential sign of applying these actions when we are not able to identify SCPs. We expected to observe a higher influence of SCP on the *ModSemTypeR* action. We would like to observe whether every time the type of a mapping's semantic relation changes, at least one SCP is recognized. This reinforces the findings of our previous studies, which pointed out the non-correlation between modifying the semantic relation of mappings and KOS changes, mainly because this action is applied to correct erroneous mappings in the studied datasets [Dos Reis et al., 2014c].

Although used datasets were extremely useful for finding potential correlations between change patterns and mapping evolution, providing a richer gold standard could point out ways to improve the obtained results. Mappings between biomedical KOSs mainly contain equivalent correspondences between concepts, which compromises the analysis of the *ModSemTypeR* action as well as the analysis of the SCP recognition algorithm. The reference mappings did not allow specifically observing how SCPs influence the way mappings change their semantic relation. This could further explain the results of the correlation between some mapping adaptation actions and change patterns in concept attributes, especially for semantic change patterns.

Overall results achieved in this research indicate that the proposed set of change patterns refers to a relevant evidence to guide the adaptation of mappings.

Table 6.8: Correlation between semantic change patterns and mapping adaptation actions

MAA \ SCP	#EQV	#MSP	#LSP	#PTM	no-scp #MAA	no-scp #noChgCT
MoveM	284▲;223▽	1▲;3▽	22▲;15▽	47▲;16▽	221	2
DeriveM	143▲;56▽	32▲;16▽	34▲;33▽	96▲;77▽	476	24
ModSemTypeM	73	12	19	47	0	0
RemoveM	20	0	3	5	173	140
NoAction	223	77	29	173	8 951	5 482

This table presents the number of SCPs identified for each type of MAA. This shows the numbers of SCPs that have a positive impact (denoted as ▲), *i.e.*, concept where the algorithm recognizes the SCP is equal to the concept where a *MoveM* or *DeriveM* occurs ($c^1_{cand} = c^1_{obs}$) and negative impact (denoted as ▽), *i.e.*, concept where the algorithm recognizes the SCP differs to the concept where a *MoveM* or *DeriveM* occurs [$c^1_{cand} \neq c^1_{obs}$], on each MAA. $\#MAA$ expresses the number of MAAs where no SCP is detected, while $\#noChgCT$ presents the number of MAAs where no changes occurs in concepts in the context of the source concept in mapping.

6.3.6 Summary of the findings

The conducted experiments examined several aspects, which can influence decisions on mapping adaptation actions. This is not an exhaustive list of possible analyses, but represents a step to understand in-depth, with fine-grained elements, the evolution of mappings. We designed the conducted analyses according to our previous experiences and our needs for elucidating uncovered elements to help understanding mapping adaptation actions.

We provide a summary of the major points and conclusions revealed by each one of the analyses. We aim at combining the observations from the different analyses to guide the design of heuristics (*cf.* Section 6.4).

Analysis of revision and removal KOS change operations (Section 6.3.3):

- Mappings associated with the same type of change operation can behave differently. This suggests that we need to take one individual mapping adaptation decision for each mapping even though mappings belong to an equal complex KOS change operation.
- Results showed that mappings frequently evolve considering a delimited set of concepts. Mapping adaptation can use the resulting concepts of complex KOS changes (*split*, *merge* and *substitution*) as a delimited context for mapping adaptation. This reduces the space of candidate concepts and can help minimizing adaptation errors.
- When concepts are deleted or assigned to obsolete, different from existing solutions that by default remove mappings, our experiments showed that associated mappings are more frequently moved to concepts in the context.

Analysis of changes in relevant attributes (Section 6.3.4):

- When relevant attributes are removed, this can indicate the application of a *RemoveM* action.

- The best relevant attributes showed a central role in this analysis for *RemoveM*. We mostly observed that when the best relevant attribute is deleted, the mapping is removed.
- When relevant attributes remain unchanged, this can indicate that we can apply a *DeriveM*, a *MoveM* or *NoAction*.

Analysis of lexical and semantic change patterns influence (Section 6.3.5):

- When we can recognize a lexical change pattern for any of the relevant attributes, this can indicate that we apply a *MoveM* or a *DeriveM* action.
- When no lexical change pattern is recognized, this can indicate that we should apply a *RemoveM* or *NoAction*.
- Despite the difficulties of observing clear findings with respect to semantic change patterns effects on mapping evolution, we believe that this can be useful for supporting a modification of the type of semantic relation in mappings (*i.e.*, apply the *ModSemTypeM* action). We can combine a possible recognized semantic change pattern with the old semantic type of the mapping to derive a new adequate semantic relation (*cf.* Section 6.4).

We rely on these empirical facts as an initial step to model the heuristics to guide mapping adaptation (next section) in addition to our previously achieved experimental analyses. However, due to the complexity of the studied phenomenon, we deem that for refining and completing the modeling, supplementary assumptions, uncovered by the current empirical observations, will be required. Further empirical experiments have been subject to investigation in the frame of this thesis, and which help justifying the decision choices in the heuristics, but we do not report the obtained results in this manuscript [Dos Reis et al., 2014b].

6.4 Heuristics guiding decisions of mapping adaptation

Based on empirical observations from the experimental analyses conducted throughout this thesis, we define and formalize conditions to trigger the different MAAs modeling heuristics. We take into account several aspects such as lexical and semantic change patterns to design the conditions.

We follow some basic principles to delineate the behaviour of the heuristics. We explore relevant attributes that we are able to identify with algorithm 1. The adaptation decisions will rely on this set of attributes. Moreover, LCPs related to the relevant attributes must trigger *MoveM* and *DeriveM* actions. SCPs are used to guide the modification of the type of semantic relation in mappings. When no change patterns are found, indicating that neither *MoveM* nor *DeriveM* are applicable, we analyze whether a *RemoveM* or a *NoAction* can be applicable based on aspects related to the similarity between attributes and their deletion.

We first present the necessary definitions for the formalization. Section 6.4.1 presents the heuristics related to *MoveM* and *DeriveM* actions. We present how to adapt the type of semantic relations in mappings according to heuristics related to the action *ModSemTypeR* in section 6.4.2. Section 6.4.3 shows the heuristics concerning *RemoveM* and *NoAction*.

For the design of the heuristics, given two attribute values a_1 and a_2, the function LCP indicates whether any type of LCP exists. Similarly, the SCP function returns a possible type

of semantic change pattern if it exists. Formally:

$$\begin{aligned} LCP : A(c_i^0) \times A(c_q^1) &\longrightarrow \{TRUE, FALSE\} \\ (a_1, a_2) &\longrightarrow \begin{cases} TRUE & \text{if } TT(a_1,a_2) \oplus TC(a_1,a_2) \oplus PT(a_1,a_2) \oplus PC(a_1,a_2) \\ FALSE & otherwise \end{cases} \end{aligned}$$

$$\begin{aligned} SCP : A(c_i^0) \times A(c_q^1) &\longrightarrow \{\equiv, <, >, \approx, \varnothing\} \\ (a_1, a_2) &\longrightarrow \begin{cases} \equiv & \text{if } EQV(a_1,a_2) \\ < & \text{if } LSP(a_1,a_2) \\ > & \text{if } MSP(a_1,a_2) \\ \approx & \text{if } PTM(a_1,a_2) \\ \varnothing & otherwise \end{cases} \end{aligned}$$

6.4.1 Move and derivation of mappings

Let m_{st}^0 be a mapping at time j to be adapted, so c_s^0 and c_t^0 correspond to the source and target concepts, respectively. Also, let $a_i \in topA(c_s^0, c_t^0, n)$ (*i.e.*, the set of relevant attributes selected with algorithm 1 for a given mapping), we define the set $Cand$ as all those concepts at time $j+1$ (belonging to the context of c_s^1) where we can recognize at least one LCP between one relevant attribute from c_s^0 and one attribute from a concept in its context. Simultaneously, for such relevant attribute that a LCP was identified in, SCPs are **NOT** detected with attributes in c_s^1. Formally:

$$Cand = \left\{ c_{cand}^1 \;\middle|\; \begin{array}{l} \exists c_{cand}^1 \in CT(c_s^1), \\ \exists a_q^1 \in A(c_{cand}^1), \\ \forall a_i \in topA(c_s^0, c_t^0, n), \forall a_z^1 \in A(c_s^1) | LCP(a_i, a_q^1), SCP(a_i, a_z^1) = \varnothing \end{array} \right\}$$

Move mapping. We associate the *MoveM* with the fact of observing LCPs between attributes of different concepts. More specifically, there exists *one and only one* relevant attribute of the source concept c_s^0, where we identify a LCP with an attribute from one concept of the context of c_s^1. Our early intuition relied on the fact that *MoveM* could mostly be related to *Total Transfer* or *Partial Transfer*. Indeed, results in section 6.3.5 have shown that this really occurs (especially for *Total Transfer*), but also revealed the influence of the copy type of LCP (*cf.* Table 6.7). Therefore, we chose to design the conditions to apply *MoveM* or *DeriveM* actions not based on the type of LCPs, but on the frequency of LCPs that we can observe concerning relevant attributes in a source concept. This choice has allowed us to attain better results (*cf.* Section 6.5) when testing other options during the heuristics design process. Therefore, when we are able to observe only one candidate c_{cand}^1, *i.e.*, we found only one attribute with LCP, we apply a *MoveM* action. Moreover, it must also satisfy the constraint of **NOT** having SCPs between **ALL** relevant attributes (from topA) and the attributes from the source concept c_s^1. Formally:

$$\left.\begin{array}{l} \|Cand\| = 1, \\ \forall a_w \in topA(c_s^0, c_t^0, n), \\ \forall a_z^1 \in A(c_s^1), SCP(a_w, a_z^1) = \varnothing \end{array}\right\} \Longrightarrow c_{cand}^1 \in Cand, MoveM(m_{st}, c_{cand}^1) \tag{6.7}$$

For identifying SCPs with attributes from the c_s^1, it was necessary to make our algorithm 4 more flexible, by not considering the hierarchical aspect. This also remains relevant for the heuristics expressed in equation 6.9. The heuristics for *MoveM* in equation 6.7 can be more frequently applicable for cases when the source concept has been removed, but we find an adequate candidate in the context. In cases where the source concept is totally deleted, we get the context of c_s^1 via the concepts that determine its context at time j. Figure 6.6 illustrates a scenario representing the heuristics for applying *MoveM*. Note that only a unique LCP is presented in the context at time $j+1$.

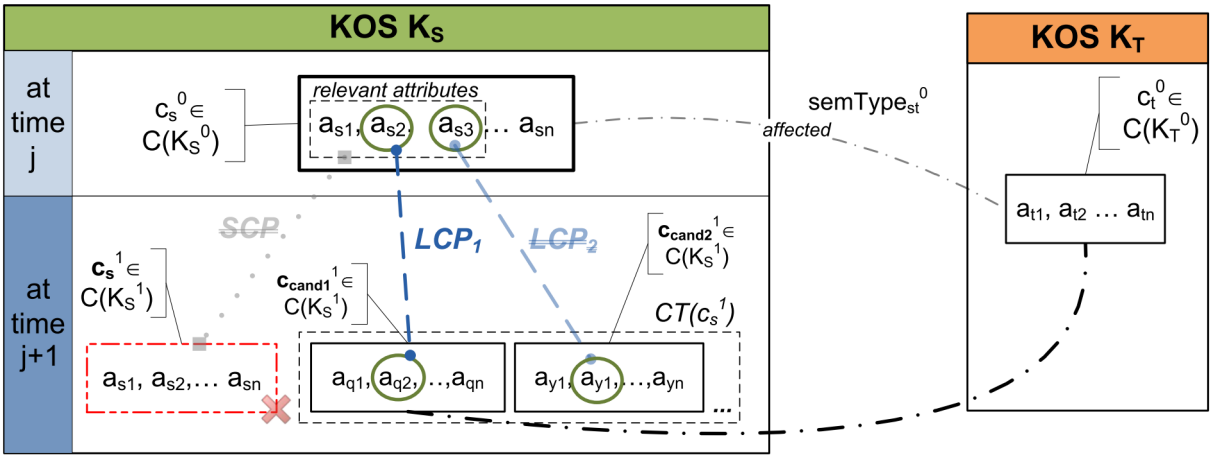


Figure. 6.6: Heuristics for *MoveM*

This figure presents a scenario where a *MoveM* action must be applied. Rectangular represents a concept with its denoting attribute values inside. The figure illustrates the both source K_S and target K_T KOS and the respective evolution of a concept source $c_s \in C(K_S^0)$. We delimitate the hypothetical relevant attributes of such source concept c_s^0. We illustrate possible LCPs related to the relevant attributes with respect to concepts at time $j+1$. Only one LCP is detected with a relevant attribute of c_s^0 and an attribute of a concept c_{cand}^1.

Derive mapping. The *DeriveM* action suggests a modified copy of the original mapping. We observed that this action has been mostly related to the scenario where several LCPs are simultaneously found (independently of their type) concerning distinct attributes. Therefore, this heuristics relies on the fact of having several concept candidates at time $j+1$. Formally:

$$\left.\begin{array}{l} \|Cand\| > 1, \\ c_s^1 \in C(K_S^1) \end{array}\right\} \Longrightarrow \forall c_{cand}^1 \in Cand, DeriveM(m_{st}, c_{cand}^1) \tag{6.8}$$

Equation 6.8 requires that $c_s^1 \in C(K_S^1)$ because the *DeriveM* action conserves the original mapping, so that it needs to make sure the concept c_s^1 still exists at time $j+1$. Figure 6.7 depicts a scenario of applying *DeriveM* actions. In contrast to the *MoveM* action, the original mapping remains related to the target concept and new adapted mappings are added with the respective candidates identified. The heuristics presented in the section 6.4.2 determine the type of semantic relation of these derived mappings.

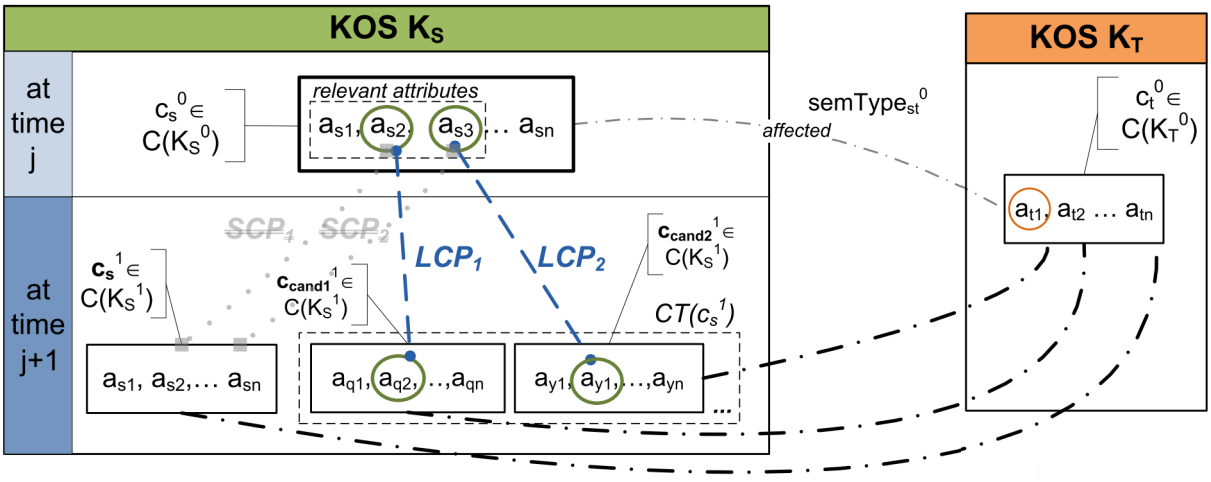


Figure. 6.7: Heuristics for *DeriveM*

This figure presents a scenario where *DeriveM* actions are applied. Several LCPs are detected with relevant and distinct attributes of c_s^0 and attributes of candidate concept $c_{cand_i}^1$ in the context of the source concept.

6.4.2 Modification of semantic relation

Applying the *ModSemTypeR* action relies on SCPs detected between attributes from different KOS versions. We propose two scenarios for modifying the type of semantic relation in mappings. First, we design the situation where the relation of the original mapping m_{st}^0 is modified (*cf.* Equation 6.9). In this case, the source concept remains the same (*i.e.*, same concept identifier number between source and target concepts), but the semantic relation between source and target changes at time $j+1$. Second, we define the semantic relation between a candidate concept c_{cand}^1 and the target concept for cases where a *MoveM* or a *DeriveM* actions are applied (*cf.* Equation 6.10). In the first case, the new semantic type connects the evolved source concept at time $j+1$ (in terms of content) and the target concept, while in the second situation, the source concept is replaced by a candidate from the context (*i.e.*, c_s^0 and c_{cand} has a distinct CUI).

We determine the type of semantic relation by combining the original $semType^0$ and the type of SCP recognized between involved attributes. We define the function *getSemType* that aims at composing a given *semType* and *SCP* according to their nature, and returns a resulting *semType*. For example, for a given mapping to which c_s^0 is equivalent to c_t^0 (x equal to $\equiv$ in *getSemType* function), and a SCP of type *Less Specific* ($>$) (value of y), detected between an attribute in c_s^0 and another attribute in c_s^1, the new *semType* as outcome between c_s and c_t will be more specific than ($\leq$) according to the *getSemType*. In this case, since the evolved attribute

in c_s^1 is more specific than the original attribute in c_s^0, and given that $c_s^0 \equiv c_t^0$, according to our heuristics we imply that $c_s^1 \leq c_t^0$. Formally:

$$
\begin{aligned}
getSemType : semType \times SCP &\longrightarrow \{'\equiv', '\leq', '\geq', '\approx'\} \\
(x, y) &\longrightarrow \begin{cases} '\equiv' & \text{if } x = '\equiv' \wedge y = '\equiv' \\ '\leq' & \text{if } (x = '\leq' \wedge y = '>') \vee (x = '\leq' \wedge y = '\equiv') \vee (x = '\equiv' \wedge y = '>') \\ '\geq' & \text{if } (x = '\geq' \wedge y = '<') \vee (x = '\geq' \wedge y = '\equiv') \vee (x = '\equiv' \wedge y = '<') \\ '\approx' & otherwise \end{cases}
\end{aligned}
$$

Equation 6.9 formalizes the first scenario for applying the $ModSemTypeR$. This covers the case where SCPs are found with c_s^1. As we have already mentioned, we adapted the semantic change pattern algorithm to not consider the hierarchy between concepts when identifying SCP between attributes from c_s^0 and c_s^1.

There exists an attribute $a_z^1 \in A(c_s^1)$ (*i.e.*, c_s^1 remains valid at time $j+1$) to which a relevant attribute $a_i \in topA(c_s^0, c_t^0, n)$ is involved in some type of SCP with a_z^1. We must identify the attribute a_z^1 as a new one to consider it in the pattern recognition. Moreover, the heuristics requires to check that combining the original $semType_{st}^0$ with the resulting symbol from $SCP(a_i, a_z^1)$, according to the $getSemType$ function, we attain a new $semType_{st}^1$, which differs from the $semType_{st}^0$. In this case, we modify the relation of m_{st} with $semType_{st}^1$. We also make sure that m_{st} exists at time $j+1$, indicating that it is not influenced by a $MoveM$ action, which would imply removing the original mapping.

$$
\left.
\begin{array}{l}
\exists a_i \in topA(c_s^0, c_t^0, n), \\
\exists c_s^1 \in C(K_S^1), \\
\exists a_z^1 \in A(c_s^1), SCP(a_i, a_z^1) \neq \varnothing, \\
\exists m_{st}^1 \in \mathcal{M}_{ST}^1, \\
\exists semType_{st}^1 \in \{\equiv, \leq, \geq, \approx\}, \\
semType_{st}^1 \in getSemType(semType_{st}^0, SCP(a_i, a_z^1)), \\
semType_{st}^0 \neq semType_{st}^1
\end{array}
\right\} \Longrightarrow ModSemTypeR(m_{st}, semType_{st}^1) \tag{6.9}
$$

Figure 6.8 shows the scenario for the heuristics expressed in equation 6.9, where a SCP is detected with the concept c_s^1, but no LCP is found with the context, which is mostly applicable (appropriated) when the whole CT remains unchanged.

Equation 6.10 presents the formalization of the heuristics that enables modifying the type of semantic relation considering a concept from the context c_{ct}^1 and c_t. In this heuristics, we observe a given concept $c_{ct}^1 \in CT(c_s^1)$ and an attribute $a_q^1 \in A(c_{ct}^1)$, to which exists a relevant attribute $a_i \in topA(c_s^0, c_t^0, n)$ involved in some type of LCP and SCP between a_q^1 and a_i. We make sure that it exists the new mapping (m_{ct}) connecting the concept from the context and the target. Implicitly, this means that a $MoveM$ or a $DeriveM$ action is applied with respect to the c_{ct}^1, so by definition a LCP will exist. Moreover, similar to the heuristics in equation 6.9, we need to observe that combining the $semType_{st}^0$ of the original mapping with the symbol resulted

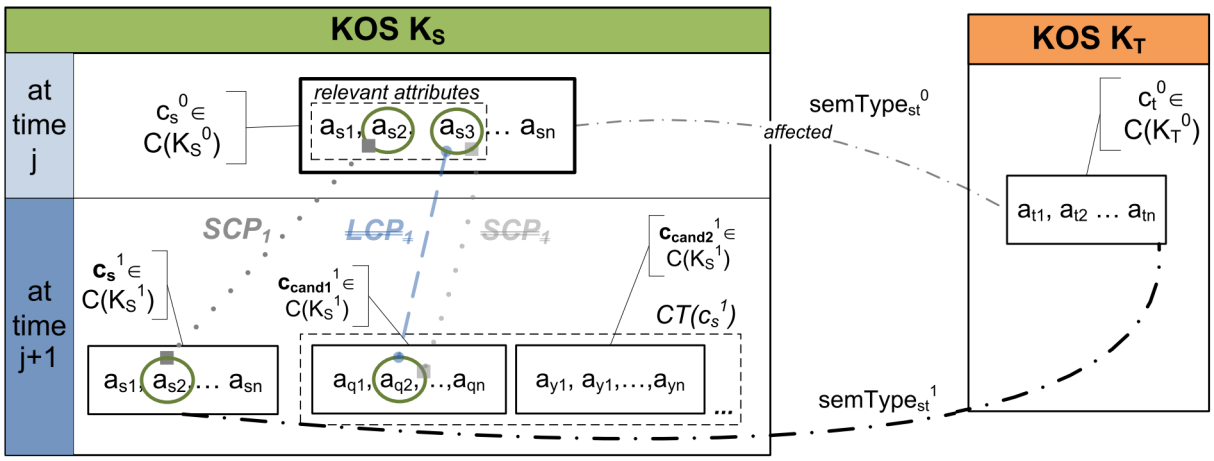


Figure. 6.8: Heuristics for *ModSemTypeM* of an existing established mapping

This figure presents a scenario where a *ModSemTypeR* action is applied for the original mapping m_{st} that interrelates the concepts c_s^0 and c_t^0. A SCP is detected between a relevant attribute of c_s^0 and an attribute in c_s^1. The $semType_{st}^1$ refers to the updated type of semantic relation between c_s^1 and c_t.

from $SCP(a_i, a_q^1)$, relying on the *getSemType* function, we attain a $semType_{ct}^1$ (*i.e.*, between the concept from the context and the target concept), which differs from the original type of semantic relation $semType_{st}^0$ of the mapping. In this case, we define the relation of the mapping m_{ct} as $semType_{ct}^1$ (*cf.* Equation 6.10).

$$\left.\begin{array}{l} \exists a_i \in topA(c_s^0, c_t^0, n), \\ \exists c_{ct}^1 \in CT(c_s^1), \\ \exists a_q^1 \in A(c_{ct}^1), SCP(a_i, a_q^1) \neq \varnothing, \\ \exists m_{ct} \in \mathcal{M}_{ST}^1, \\ \exists semType_{ct}^1 \in \{\equiv, \leq, \geq, \approx\}, \\ semType_{ct}^1 \in getSemType(semType_{st}^0, SCP(a_i, a_q^1)), \\ semType_{st}^0 \neq semType_{ct}^1 \end{array}\right\} \Longrightarrow ModSemTypeR(m_{ct}, semType_{ct}^1) \tag{6.10}$$

Figure 6.9 presents an illustrative scenario for the heuristics in equation 6.10, where several LCPs and SCPs are found with concepts from the context of c_s^1. For each different candidate concept, we determine the respective *semType* connecting to c_t.

The way that the semantic change patterns were designed, taking into account the hierarchy between concepts, allow us to combine the type of semantic relation of the original mapping, with the recognized type of SCP. This may occur even though the mapping links the concept level, and the SCP the attribute level. We rely on the fact that the hierarchical aspect between the concepts, where the involved attributes of a SCP happen, respects the conceptual level. For example, the designed patterns avoid the situation that between two attributes a_i and a_h, the LSP(a_i, a_h) type of pattern happens (*i.e.*, $a_i > a_h$) whether $a_i \in A(c_i)$ and $a_h \in A(c_h)$ to which $c_i \sqsubset c_h$.

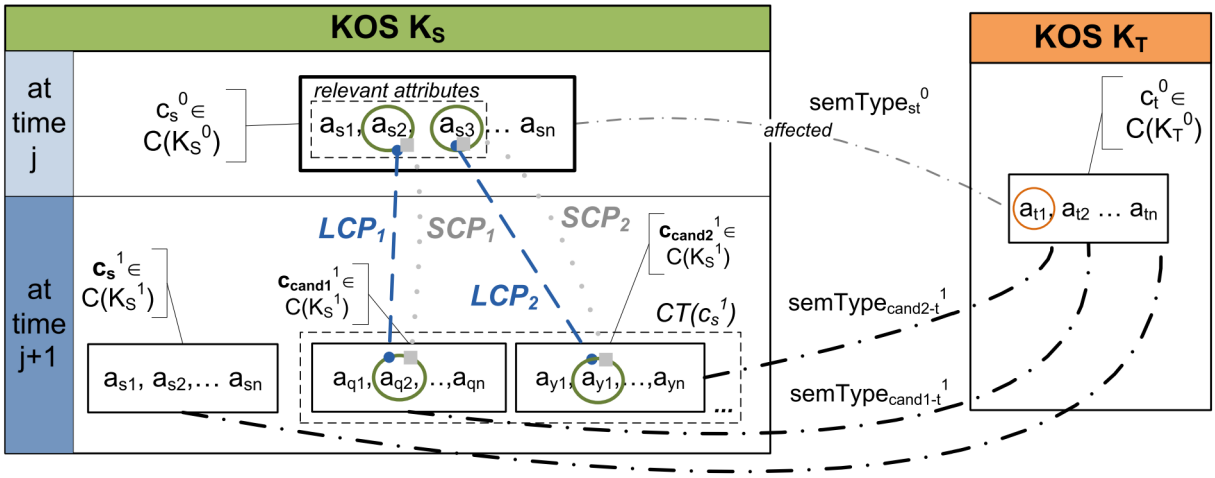


Figure. 6.9: Heuristics for *ModSemTypeM* between a candidate and the target concept

This figure presents a scenario where *ModSemTypeR* action are applied for mappings influenced by *MoveM* or *DeriveM* action. SCPs are detected between relevant attributes of c_s^0 and attributes from concepts of the context of c_s^1. The $semType_{cand1-t}^1$ refers to the updated type of semantic relation between c_{cand}^1 and c_t.

6.4.3 Removal and no action adaptation

We present heuristics that must trigger the removal of a mapping or the case where *NoAction* is applied for a given mapping (*i.e.*, the mapping remains unchanged from one version to another).

Remove mapping. The proposed mapping adaptation method applies *RemoveM* when for all relevant attributes identified, no change pattern is detected with the context. Equation 6.11 models the heuristics for *RemoveM*. In this equation, $A_{ct}(c_s^1)$ refers to the set of attributes from all concepts of the context (*cf.* Equation 4.3). Also, the best relevant attribute is deleted from one KOS version to another (this choice relies on the experiments in section 6.3.4). This equation also demonstrates that for all relevant attributes, no SCP must be recognized with the c_s^1, if this concept exists.

When $c_s^1 \notin C(K_S^1)$ or c_s is assigned to obsolete, we consider that all attributes belonging to c_s are deleted. Our experiments observed the similarity between relevant attributes with concept attributes in the context [Dos Reis et al., 2014b]. In particular, we calculated the similarity when the mappings were removed and the results revealed that the similarity remains very low. Therefore, we introduce the parameter (α) related to the similarity in equation 6.11. We formalize the heuristics for *RemoveM* as follows:

$$\left.\begin{array}{l} \forall a_i \in topA(c_s^0, c_t^0, n), \\ \forall a_q^1 \in A_{ct}(c_s^1), \neg LCP(a_i, a_q^1), \\ SCP(a_i, a_q^1) = \varnothing, \\ \exists \alpha \in \mathbb{R}_{>0}, sim(a_i, a_q^1) \leq \alpha, \\ \forall a_z^1 \in A(c_s^1), SCP(a_i, a_z^1) = \varnothing, \\ \exists a_i \in topA(c_s^0, c_t^0, 1), a_i \notin A(c_s^1) \end{array}\right\} \Longrightarrow RemoveM(m_{st}) \quad (6.11)$$

Figure 6.10 depicts a scenario of applying the *RemoveM* action. In this figure, the concept c_s^1 is removed and no LCPs nor SCPs are detected.

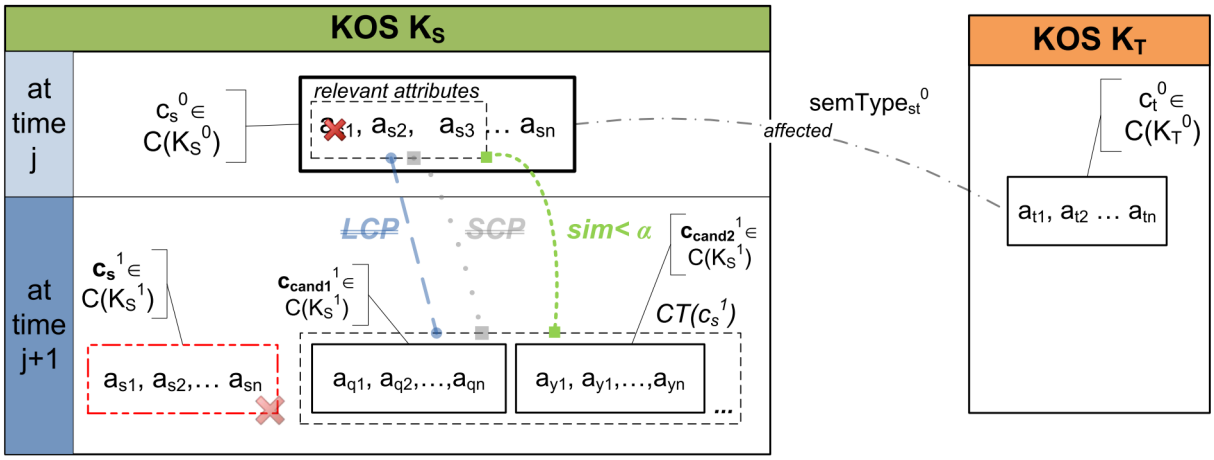


Figure. 6.10: Heuristics for *RemoveM*

This figure presents a scenario where a *RemoveM* action is applied for a mapping. We consider the role played by the best relevant attribute in case it is deleted.

NoAction. Similar to the heuristics for *RemoveM* (Equation 6.11), the heuristics for *NoAction* also relies on the fact that we are unable to detect adequate LCPs and SCPs with the context, as well as with the source concept at time $j+1$. This heuristics addresses the situations where a KOS change affects the source concept involved in a mapping, but relevant attributes remain unchanged or the similarity with new attributes in context remains low [Dos Reis et al., 2014b]. We inspired this heuristics mostly from the experimental results conducted in section 4.5, combined with the analysis of the change patterns' influence on the mapping adaptation actions (Section 6.3.5). We formalize the heuristics as follows:

$$\left.\begin{array}{l} \forall a_i \in topA(c_s^0, c_t^0, n), \\ \forall a_q \in A_{ct}(c_s^1), \neg LCP(a_i, a_q^1), \\ SCP(a_i, a_q^1) = \varnothing, \\ \exists c_s^1 \in C(K_S^1), \\ \forall a_z^1 \in A(c_s^1), SCP(a_i, a_z^1) = \varnothing, \\ \left[\begin{array}{l} a_i \in A(c_s^1) \\ \vee \\ \exists \alpha \in \mathbb{R}_{>0}, sim(a_i, a_q^1) \leq \alpha \end{array}\right. \end{array}\right\} \Longrightarrow NoAction(m_{st}) \tag{6.12}$$

Figure 6.11 presents an illustrative scenario where the required conditions to apply *NoAction* are fulfilled. The mapping m_{st}^0 remains totally equal to the mapping m_{st}^1.

We have experimentally observed in less frequent cases that the designed heuristics can lead to some conflicts in the adaptation process. After presenting the mapping adaptation method (Section 7.2), we describe and discuss possible scenarios of conflict that will require user intervention (Section 7.2.3).

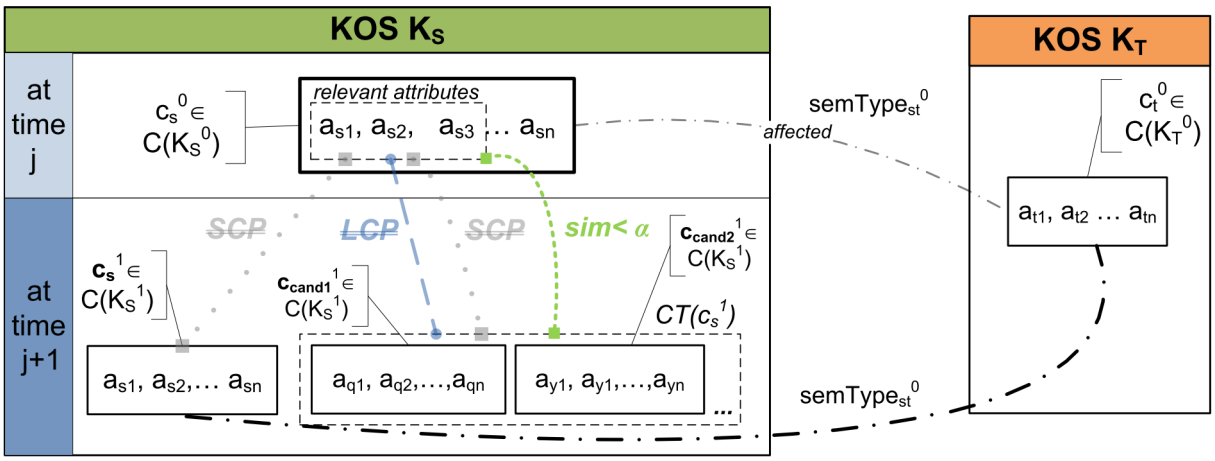


Figure. 6.11: Heuristics for *NoAction*

This figure presents a scenario where no action (*NoAction*) is applied for a mapping. This heuristics considers that when occurring unchanged relevant attributes with no identification of change patterns, we must keep the correspondence between source and target (even thought the source concept is affected by some KOS change).

6.5 Experimental evaluation

We conduct experiments to evaluate the proposed heuristics on mapping adaptation. This validation uses the materials presented in section 6.3.1. Therefore, this evaluation examines the mapping adaptation behaviours, according to our suggested heuristic techniques based on three mapping datasets (SCT-ICD9, SCT-NCI and MeSH-ICD10). We present the experimental procedure in section 6.5.1 followed by the obtained results (Section 6.5.2).

6.5.1 Experimental procedure

We aim to analyze to which extent the adaptation of mappings relying on MAAs, and their suggested decision, based on the conditions modeled in the heuristics, may correspond to the evolution of real-world mappings (*i.e.*, we observe the evolution of mappings by analyzing two version releases as conducted for other experiments in this thesis). To this end, we performed the following procedure to validate the proposed heuristics:

1. For each dataset, we first apply the common experimental procedure described in section 6.3.2. This allows us to have the KOS *diff*, to select the mappings impacted by KOS evolution (only source or target concepts affected per time), and for the affected mappings, to calculate the sets of observed mapping adaptation actions. Indeed, we use these observed actions as our gold standard. Table 6.4 presents the number of MAAs observed. Our evaluation does not consider those mappings where source and target concepts simultaneously change in KOS evolution.

2. We adapt the affected mappings according to the modeled heuristics. Each mapping leads to a nonempty set of MAAs. In fact, depending on the MAA, a unique action is possible (*e.g.*, when a mapping is removed). However, one mapping can lead to several MAAs, for instance, if it applies a *MoveM* and a *ModSemTypeR* action. Chapter 7 presents the proposed framework with the mapping adaptation mechanism implemented in detail. This

shows the way that an entire mapping adaptation process explores the heuristics. In particular, this experiment aims at examining the proposed heuristics for mapping adaptation. On this basis, we analyze the performance of each type of MAA separately. When conducting this experiment, we use the following values for the thresholds: $\tau = 0.7$, $\gamma = 0.9$ and $\alpha = \tau$. We also set the number of relevant attributes to three.

3. Given that we have sets of expected MAAs (calculated observing the evolution) and sets of proposed MAAs (resulted from adapting mappings relying on the heuristics), we measure standard metrics of *Precision*, *Recall* and *F-Measure* for each different type of MAA as follows:

   We computed the *Precision* as the number of MAAs correctly proposed by the adaptation mechanism in contrast to the expected MAAs (observing mapping evolution), over the total number of proposed MAAs:

$$Precision = \frac{\#correctlyProposedMAA}{\#proposedMAA} \tag{6.13}$$

   Recall was computed as the number of correctly proposed MAAs over the total number of expected MAAs:

$$Recall = \frac{\#correctlyProposedMAA}{\#expectedMAA} \tag{6.14}$$

   The *F-measure* refers to the harmonic mean of *Precision* and *Recall*.

$$F\text{-}measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{6.15}$$

4. To rigorously evaluate the proposed actions, we suggest two types of measures denoted by the symbols ▲ and ▽ (*cf.* achieved results in table 6.9). This distinction remains especially relevant for evaluating the *MoveM*, *DeriveM* and *ModSemTypeR* actions. The symbol ▲ expresses the *Precision*, *Recall* and *F-Measure* when for a given mapping, the proposed MAA and its argument are both correct compared to the expected MAAs (*i.e.*, the mapping adaptation proposes the adequate MAA type in addition to the correct candidate concept or semantic relation). On the other hand, the symbol ▽ represents cases in which for a given mapping, only the type of MAA is correct, but not its argument. Note that there might exist a tendency of having lower values of ▲ compared to ▽, because the measure ▲ requires taking into account more constraints. For example, for a given mapping that we observe a *MoveM* with a concept c_{obs} in context, if for this mapping, the adaptation mechanism allows proposing a *MoveM* for the concept c_{obs}, we thus measure *Precision*, *Recall* and *F-Measure* related to ▲ (*i.e.*, right action and argument). Otherwise, if for this mapping, only the *MoveM* is correctly proposed (instead of any other MAA), but with a different concept than c_{obs}, we assign this case to ▽ (*i.e.*, right action, but wrong argument according to the evolution of the given mapping). Similarly, we calculate the cases of ▲ and ▽ for the *ModSemTypeR*, to which the type of semantic relation proposed and observed is referred to by the argument.

5. We calculate the *Global* and *General* aspects in the evaluation (*cf.* Table 6.9). The *Global* refers to the *Precision*, *Recall* and *F-Measure* for each dataset, independent of the type of MAAs. The *General* stands for the measure concerning all datasets.

6.5.2 Experimental results

Table 6.9 presents the obtained results on adapting mappings based on the heuristics for each type of actions and datasets.

Table 6.9: Results of the heuristics evaluation for mapping adaptation

MAA	Dataset	Precision	Recall	F-Measure
MoveM	SCT-ICD9	0.47▲;0.55▽	0.47▲;0.56▽	0.47▲;0.55▽
	SCT-NCI	0.10▲;0.13▽	0.75▲;1.0▽	0.18▲;0.23▽
	MeSH-ICD10	*N/A*	*N/A*	*N/A*
DeriveM	SCT-ICD9	0.41▲;0.54▽	0.40▲;0.52▽	0.40▲;0.53▽
	SCT-NCI	0.0▲;0.0▽	0.0▲;0.0▽	0.0▲;0.0▽
	MeSH-ICD10	0.30▲;0.40▽	0.75▲;1.0▽	0.43▲;0.58▽
ModSemTypeM	SCT-ICD9	0.27▲;0.45▽	0.31▲;0.53▽	0.29▲;0.49▽
	SCT-NCI	*N/A*	*N/A*	*N/A*
	MeSH-ICD10	*N/A*	*N/A*	*N/A*
RemoveM	SCT-ICD9	0.56▲;0.56▽	0.76▲;0.76▽	0.65▲;0.65▽
	SCT-NCI	0.69▲;0.69▽	0.67▲;0.67▽	0.68▲;0.68▽
	MeSH-ICD10	0.75▲;0.75▽	0.50▲;0.50▽	0.60▲;0.60▽
NoAction	SCT-ICD9	0.94▲;0.94▽	0.94▲;0.94▽	0.94▲;0.94▽
	SCT-NCI	0.98▲;0.98▽	0.84▲;0.84▽	0.90▲;0.90▽
	MeSH-ICD10	0.77▲;0.77▽	0.96▲;0.96▽	0.86▲;0.86▽
Global	SCT-ICD9	0.85▲;0.87▽	0.86▲;0.88▽	0.86▲;0.87▽
	SCT-NCI	0.84▲;0.84▽	0.83▲;0.83▽	0.83▲;0.83▽
	MeSH-ICD10	0.72▲;0.73▽	0.95▲;0.95▽	0.82▲;0.82▽
General		0.85▲;0.86▽	0.85▲;0.86▽	0.85▲;0.86▽

This table presents the results of applying MAAs based on the designed heuristics for three mapping datasets (SCT-ICD9; SCT-NCI; MeSH-ICD10). We use metrics of *Precision*, *Recall* and *F-Measure* between expected MAAs (*cf.* Table 6.4) and proposed sets of MAAs running our mapping adaptation mechanism (*cf.* equations 6.13, 6.14, 6.15). We calculate cases where the type of MAA and its argument (candidate concept for *MoveM* and *DeriveM* or *semType* for *ModSemTypeR* action) are correct compared with the expected ones (denoted with symbol ▲); as well as cases where only the type of MAA is correct (denoted with symbol ▽). The *Global* presents results for each dataset considering all types of MAA, and *General* for all datasets.

Move mapping. For the *MoveM* action, table 6.9 shows that results remain reasonable for the dataset SCT-ICD9, but decrease for SCT-NCI even though the *Recall* remains high. However, the expected cases of *MoveM* in SCT-NCI are much fewer than in SCT-ICD9. This result reveals that the heuristics for *MoveM* can work better for one dataset than for other in terms of *F-Measure*. We could not observe *MoveM* in the dataset MeSH-ICD10. Due to the fact that the difference of results between ▲ and ▽ stands for a relatively low value, we notice that mostly when the adaptation system proposes *MoveM* actions, this leads to an adequate candidate concept in the context. We present examples of right *MoveM* actions and a wrong one extracted by observing the execution of the experiment.

We first present a scenario of correct proposition of *MoveM* by the modeled heuristics. In the dataset SCT-NCI, we observe the mapping between '*128829008*' - "*Acute myeloid leukemia, 11q23 abnormalities (morphologic abnormality)*" and '*C6924*' - "*Acute_ Myeloid_ Leukemia_ with_ 11q23_ MLL_ Abnormalities*". When analyzing the evolution of the concept '*C6924*' (in NCI), the mapping adaptation method identifies a *Total Transfer* regarding the best relevant attribute ("*Acute Myeloid Leukemia with 11q23 Abnormalities*") and a *Total Copy* of another relevant attribute. Both LCPs relate to the same candidate concept '*C82403*' - "*Acute_ Myeloid_ Leukemia_ with_ t_ 9_ 11_ p22_ q23_ MLLT3-MLL*" in the new KOS version. The adaptation correctly proposes a *MoveM* action towards this candidate concept according to the expected MAAs. Note that two LCPs are detected, but they refer to the same candidate concept.

We show a scenario where a *MoveM* action is expected, but the mapping adaptation proposes a different MAA. Analysing the dataset SCT-NCI, we observe a mapping between'*C71477*' - "*Usage*" and '*277889008*' - "*Usage*". The mapping adaptation applies a *RemoveM* action, since neither lexical nor semantic change patterns are recognized, and the concept '*C71477*' is deleted according to the KOS *diff*. However, for this mapping a *MoveM* is expected. We observe that this concept is substituted by the concept '*C95018*' - "*Use_ Action*" (according to the KOS *diff*), which explains the *MoveM*. A closer analysis on the attributes values, the deleted attribute label of '*C71477*' denotes "*Usage*" and the involved attributes belonging to '*C95018*' describe "*Use_ Action*", "*Use*" and "*Employ*". Due to the syntactical differences, the change pattern recognition algorithm fails to identify a possible relation among the involved attributes. In another substitute KOS change involving '*C25399*' - "*Pelvic*" and '*C94249*' - "*Intrapelvic*", the mapping adaptation successfully identifies change patterns and also proposes the modification of semantic relation because "*Intrapelvic*" is more specific than "*Pelvic*". Note that in the dataset *SCT-NCI*, only an equivalent type of semantic relations exists; so that a possible modification of the type of semantic relation leads to the application of *ModSemTypeR* action, which will be considered incorrect according to the reference mapping actions.

Analysing the mapping in the dataset SCT-NCI between '*87148003*' - "*Amphetamine sulfate (substance)*" and '*C28822*' - "*Amphetamine*", affected by the evolution of the NCI, the mapping adaptation identifies a *Total Copy* between the attribute "*Amphetamine Sulfate*" in '*C28822*' and a candidate attribute "*L-Amphetamine Sulfate*" in concept '*C95316*', belonging to the new version of the NCI. Through the KOS *diff*, it also detects that other synonym attributes are deleted. On this basis, the mapping adaptation proposes a *MoveM* action to this mapping (mainly due to the *Total Copy* LCP), but a *NoAction* is expected according to the reference actions. We could find other examples where a LCP involving a *Total Copy* led to a *MoveM* action, which does not appear in this case. We observe that much more frequently, the *MoveM* correlates to the *Total Transfer* type of LCP.

Derive mapping. The achieved results for the *DeriveM* action remain slightly lower, compared to the *MoveM*. The *DeriveM* action involves more difficulties because for a given mapping, the adaptation mechanism proposes not only a unique *DeriveM*, but several ones. We could observe cases of *DeriveM* for the three datasets, but the number of cases changes, which also refers to the proportional difference in terms of size of the studied datasets (*cf.* Table 6.4), *i.e.*, we might observe more *DeriveM* actions in a huger dataset. In particular, while we observe 747

cases of *DeriveM* for SCT-ICD9, this number remains 14 and 21 for SCT-NCI and MeSH-ICD10, respectively. The results for SCT-ICD9 present a reasonable quality, similar to *MoveM*, as well as for the measures ▲ and ▽. For SCT-NCI, while the adaptation system could propose *DeriveM* actions, it failed to determine correct cases. Therefore, similar to the *MoveM* action, we observe that conditions for the actions modeled into the heuristics partially diverge according to the studied datasets. This aspect may be related to the maintenance and matching process when creating the mappings as well as the granularity of involved KOSs for each dataset, which we will further discuss in section 6.6. We present an example from the experiment of adequate *DeriveM* applied and another considered incorrect according to the gold standard actions.

We observe a mapping between concept '*30288003*' - "*Ventricular septal defect (disorder)*" in SCT described as equivalent (≡) to concept '*745.4*' - "*Ventricular septal defect*" in ICD9. When analysing the evolution of the concept '*30288003*', the mapping adaptation recognizes a *Total Copy* between "*Ventricular septal defect*" and "*Residual ventricular septal defect*" in the candidate concept '*447941008*' as well as a semantic change pattern of type LSP(>). Simultaneously, it also identifies a *Partial Copy* between "*Interventricular septal defect* " and "*Subarterial ventricular septal defect*" in the candidate concept '*448876006*' as well as a semantic change pattern of type LSP(>) between these attributes. Based on the heuristics, the *DeriveM* actions as well as the *ModSemTypeR* are successfully proposed with respect to the candidate concepts '*447941008*' and '*448876006*'.

We describe a scenario where *DeriveM* is incorrectly proposed. We observe a mapping between '*41452004*' - "*Uterus acollis (disorder)*" in SCT defined as more specific than concept (≤) '*752.3*' - "*Other anomalies of uterus*" in ICD9. Analysing the evolution of the concept in ICD9, our proposal recognizes several LCPs with distinct candidate concepts and their respective SCPs. This identifies a *Total Copy* involving the relevant attribute "*Other anomalies of uterus*" and an attribute of the concept '*752.33*' as well as a EQV(≡) SCP. Simultaneously, a *Total Transfer* is recognized for the relevant attribute "*Bicornuate uterus*" with respect to concept '*752.34*' and a SCP of type EQV(≡). On this basis, the heuristics propose *DeriveM* actions towards the two candidate concepts detected related to LCPs and keep the same type of semantic relation of the original mappings (due to the combination as presented in equation 6.10). The expected actions refer to a *MoveM* for concept *752.31*' - "*Agenesis of uterus*" and a *ModSemTypeR* action modifying the type of semantic relation from more specific than (≤) to equivalent (≡).

Modify type of semantic relation. The results for the *ModSemTypeM* type of action refer to the hardest for evaluating. In fact, table 6.9 shows that we could not observe this type of action in two out of the three datasets. The SCT-NCI stands for a dataset containing only equivalent relations in mappings.

In SCT-ICD9, where we find some cases of *ModSemTypeM* MAAs, obtained results did not favor observing that our proposed heuristics, to adapt the type of semantic relation, can frequently suggest the adequate *semType* in mapping adaptation.

The involved difficulties of adaptation refer to the fact that we have two situations to propose the adequate type of action and type of semantic relation for (*cf.* equations 6.9 and 6.10). One related to the original mapping (*cf.* Equation 6.9), and another referred to an adapted mapping proposed in correspondence with a candidate concept at time $j+1$ (*cf.* Equation 6.10), *i.e.*, in addition to proposing the adequate candidate, the *semType* connecting this concept with the tar-

get concept must be correct. Therefore, if the mapping adaptation fails to determine the correct candidate, the system will also incorrectly imply the *semType*. This poses a more challenging scenario of adaptation.

Moreover, our series of experiments throughout this thesis allow observing that in several cases where the action *ModSemTypeM* occurs in the dataset SCT-ICD9, it is not related to the evolution of the source or target concept. When this is not the case, new attributes are added to a source or target concept, and the *semType* changes. These facts suggest that the changed *semType* results from a new alignment between source and target KOS (in the maintenance process), but it is not the consequence of the evolution of the interrelated concepts.

Therefore, taking for granted that the type of semantic relation in mappings only changes due to the evolution of source or target (assumption explored in this thesis) seems not to happen very frequently in the studied datasets. In addition, considering the new mapping release version as a real gold standard can also pose difficulties in our evaluation. The new mapping version is an evolution, but not a real designed gold standard reference. In consequence, although a possible *semType* proposed via the heuristics can be different from the one observed in the new mapping version, it could still be correct from the semantic point of view by analysing source and target concepts. This must require the intervention and supplementary evaluations involving domain experts, which is further discussed in section 6.6. We present examples of applied *ModSemTypeM* actions that can help us illustrating the difficulties of adaptation and of the validation.

We describe some examples where we detect a SCP with the new version of the source concept, which leads to proposing a *ModSemTypeR* action, but the expected action differed. Observing the dataset SCT-ICD9, there exists a mapping between '*233945002*' - "*Drug-induced pulmonary hypertension (disorder)*", which is more specific than (≤) '*416.8*' - "*Other chronic pulmonary heart diseases*" in ICD9. Analysing the evolution of the ICD9 concept, the adaptation method recognized a PTM (≈) between the best relevant attribute "*Pulmonary hypertension, secondary*" and a new attribute in the source concept "*Pulmonary hypertension NOS*". This led the system to propose a *ModSemTypeR* action based on the heuristics expressed in equation 6.9, but a *NoAction* was expected for this mapping according to the reference actions. In this case, we observed no deletion involving the relevant attributes, which may explain a *NoAction*. However, in similar scenarios (where no deletion is observed with respect to the relevant attributes) and the mapping adaptation recognizes a SCP that leads to the *ModSemTypeR* action, the mappings are removed. This stands for very hard situations, where similar scenarios lead to different expected actions. This suggests that some of these observed cases are not fully related to the KOS evolution, but to a mapping maintenance process involving human experts to correct possible errors.

In the dataset SCT-ICD9, we observe a mapping between '*268838005*' - "*Neonatal dacryocystitis or conjunctivitis due to Escherichia coli (disorder)*" which is more specific than (≤) '*041.4*' - "*Escherichia coli [E.coli]*". The mapping adaptation identifies a *Total Copy* between the attributes "*Escherichia coli [E.coli]*" and "*Escherichia coli [E. coli] NOS*" (attribute newly added in the concept '*041.49*' from the context). A SCP of type LSP (>) is also recognized, which led the system to propose a *MoveM* and a *ModSemTypeR* action. According to the expected actions for this mapping, the *MoveM* was correctly identified to this concept in the context. However, even though a *ModSemTypeR* is also observed from the reference actions, the proposed type of semantic relation differed. Our heuristics proposed a *Partial Match* (≈), because it combined the semantic type ≥ from the original mapping (inverting source and target concepts because we

analyse the evolution of ICD9) with the LSP(>) of the semantic change pattern, which led to a *Partial Match* (≈) (*cf.* Equation 6.10). The expected type of semantic relation observed is the equivalent (≡) (*i.e.*, '*268838005*' - "*Neonatal dacryocystitis or conjunctivitis due to Escherichia coli (disorder)*" is judged equivalent to '*041.49*' - "*Other and unspecified Escherichia coli [E. coli]*" in the evolved mapping of reference).

Remove mapping. We can observe acceptable results related to the *RemoveM* action (*cf.* Table 6.9). Results do not vary strongly according to the datasets, and we achieve a minimal *F-Measure* of 0.60. This underscores that taking the removal of the best relevant attribute into account plays a relevant role in the heuristics for *RemoveM* action.

In the following, we present an example of *RemoveM* extracted from the experiment. In the dataset MeSH-ICD10, we observe a mapping between '*M0007301*' - "*Migration*" and '*Z60.3*' - "*Acculturation difficulty*". In the execution of the mapping adaptation method, neither lexical nor semantic change patterns are detected, and the concept '*M0007301*' is removed according to the KOS *diff* statements. In this scenario, the heuristics correctly propose the *RemoveM* action compared to the expected action. In another scenario between two interrelated concepts '*M0011288*' - "*Infants, Premature*" and '*O48*' - "*Prolonged pregnancy*" (in the dataset MeSH-ICD10), by analyzing the evolution of MeSH, the adaptation method does not identify neither lexical nor semantic change patterns for the relevant attributes. Observing the KOS *diff*, only new synonym attributes are added to the concept '*M0011288*'. On this basis, the method proposed a *NoAction*, but the expected action for this mapping is *RemoveM* (according to the reference actions), even though neither single attributes nor the concept are entirely deleted. This highlights an example that mapping adaptation may be uncorrelated with KOS evolution, which affects our obtained results.

No action applied. The *NoAction* MAA refers to the type of action with the highest quantity of expected actions, according to the observed MAAs (*cf.* Table 6.4), and compared to other types of MAAs. Results point out a great effectiveness of the heuristics related to this action for all datasets. In accordance to previous experiments, the obtained findings reveal that even though source or target concepts are affected by some KOS change, if relevant attributes keep unchanged, associated mappings will also stay unchanged. The good results with respect to this heuristics allow domain experts to concentrate only on a very small portion of mappings (compared to the whole set of initial mappings), which refers to more difficult adaptation scenarios that may request human intervention.

Global and general results. To have an overall sense of the mapping adaptation quality based on the heuristics, we analyse the results by combining all types of actions and the datasets. The obtained results indicate that *F-measure* for all datasets remain relative high in the global analysis effectiveness (minimum of 0.82 in the dataset MeSH-ICD10). The *NoAction* has a high impact on these results, since this is the dominating action set with the highest number of expected ones. These results highlight that despite the difficulties of applying some types of actions, the overall results remain acceptable and promising. Overall, this emphasizes that the proposed heuristics are suited to model the adequate conditions to apply the adequate actions. To improve the results related to some specific actions requests investigating supplementary unknown elements, that still may influence the MAAs, as well as how to combine these elements.

6.6 Discussion

This chapter proposed mapping adaptation actions suited to perform changes in mappings. We demonstrated the complexity of the problem of making decisions over these actions. Our research conducted several experiments to understand factors that help making proper decisions on the actions. As a result, we modeled heuristics representing case scenarios for applying the suggested actions.

The evaluation reported on the results of the effectiveness of the suggested heuristics to adapt mappings affected by KOS evolution. We could evaluate the proposed actions according to the heuristics in contrast to the evolution of real-world mappings (our considered gold standard) for several datasets. This experimental validation allowed us to point out the strengths and limitations of the approach and of the proposed heuristics. First, we showed the viability of adapting mappings based on the MAAs and modeled heuristics. Second, we experimentally demonstrated to which extent our approach enables to automatically adapt mapping with quality. The obtained findings underscored that despite the acceptable results, we face difficulties in some scenarios of adaptation and we still have room for improvements. We discuss some of these major difficulties and aspects that help us to better analyse the achieved results.

***MoveM* and *DeriveM*.** Modeling the heuristics for *MoveM* and *DeriveM* required making further choices uncovered by observations from our conducted experiments in this thesis. The difficulty lay on the fact of discovering the key aspects that may differ a *MoveM* and *DeriveM*. Despite our heuristics modeling the cases for these two actions (*cf.* equations 6.7 and 6.8), we still could observe a few cases mostly in the dataset SCT-ICD9, where a single LCP led to a *DeriveM* and multiple LCPs led to a unique *MoveM*. Also for cases that our modeled heuristics determined as *NoAction*, we could observe as *DeriveM* actions. We explain this fact by the precision of underlying methods required in the heuristics. For instance, the algorithm for lexical change pattern can fail to recognize pattern instances, which would assign a *DeriveM*, instead of a *NoAction*. This might request further researches in both heuristics refinement and the improvement of the dependent methods like change patterns recognition.

Moreover, we could observe that some cases that were assigned as expected *DeriveM* actions when calculating the observed MAAs, maybe cannot represent "real" cases of *DeriveM*. In these cases, although KOS changes simultaneously affected the source concept in mapping and the c_{obs} (the concept connected to concept target at time $j+1$), *i.e.*, it characterizes a *DeriveM*, we cannot explicitly realize how the c_{obs} can refer to the evolution of the source concept. Therefore, we can conclude that the new mapping between c_{obs} and c_t results of an alignment process, but not as the evolution consequence of the source concept. In addition, we could observe expected *DeriveM* actions when the whole context remained unchanged. In this situation, our algorithms cannot identify change patterns (no new added attributes exist from concepts of the context). These cases impact the performance of our obtained results, and pose difficulties to automatically select and exclude them.

Influence of KOS evolution viewpoint. We notice that the way the studied datasets are maintained (without considering KOS evolution) affects the obtained results concerning the *ModSemTypeR* action. In fact, there exists a difference between our viewpoint of adaptation, such that changes in mappings must be the consequence of KOS evolution, and the maintenance process of the datasets, which applies changes on mappings independently of KOS evolution.

This prevented us from adequately evaluate to which extent the heuristics for *ModSemTypeR* can leverage mapping adaptation. Moreover, the fact of conducting the assessment over not successive mappings releases can also slightly influence the characterization of KOS evolution and consequently, the adaptation of mappings.

Further experimental evaluations must require more specialized gold standards and involve domain experts to obtain a more refined validation of heuristics for *ModSemTypeR* actions. Nevertheless, the obtained results allow us to affirm that our proposal remains able to adapt mappings when mapping evolution follows KOS evolution. The treatment of cases out of this assumption demands future studies.

Influence of dataset characteristics. We considered three different mapping datasets between biomedical KOSs. These datasets represent set of mappings of different sizes and between biomedical KOSs of different characteristics and granularity. This influenced the number of expected MAAs for each dataset. For example, we could proportionally observe much more *MoveM* and *DeriveM* actions in the dataset SCT-ICD9. The difference of characteristics of the interrelated KOSs posed difficulties to attain the optimal heuristics that could accommodate the best results for all datasets. For example, ICD9 is a much less granular KOS than SCT, which also changes when aligning SCT and NCI. Therefore, the characteristics and goals of the mappings in these two datasets differ. In fact, in the evaluation process we experienced that some heuristic decisions favored better results for one dataset than for another. To discover the adequate compromise between the different datasets (to attain improved experimental results) poses a great challenge and we judge that we can devote future research in this direction.

Other factors. In addition to the modeled heuristics, we observe that several additional factors can influence the mapping adaptation results. As one of these factors, the used thresholds in the lexical and semantic change patterns may impact obtained instances of change patterns, and consequently, the decisions on mapping adaptation. We deem that further studying the influence of the different used thresholds can help improving and refining the obtained results. Another aspect relies on the used criteria to chose candidate concepts during the change pattern recognition. At this stage, we investigated the similarity value obtained between attribute values from different concepts, but we judge that observing and considering further criteria can improve the proposed candidates. Consequently, it might ameliorate the change patterns recognition and the mapping adaptation.

This work provides the following answers to the research questions raised:

1. Our investigation proposed the mapping adaptation actions suited to change mapping elements and to express complex behaviours of mapping adaptation.
2. We demonstrated that the heuristics refer to a way allowing to combine statements from KOS evolution and mapping interpretation in order to make decisions on the MAAs for a given mapping.
3. We showed that change patterns at the level of attributes can stand for a way of indicating candidate entities to replace concepts involved in mappings affected by KOS evolution.
4. The proposal revealed that we can adapt mappings by taking different types of semantic relations into consideration and we can also modify them.

Conclusion

This chapter reached the core problem of this thesis concerning specific research questions of mapping adaptation. We originally showed that adapting mappings based on the concept attribute level may enable fine grained decisions of mapping adaptation. However, it can also lead to issues related to adaptation conflicts that might require human intervention. We demonstrated that it is possible to combine KOS evolution (in particular our change patterns proposed in chapter 5) and the interpretation of mapping relied on attributes defining mappings (Chapter 4) to adapt KOS mappings.

We proposed ways to adapt KOS mappings affected by KOS evolution via mapping adaptation actions, and we formalized heuristics that model conditions to trigger the different types of actions. We inspired the modeling of our heuristics based on supplementary experiments conducted to observe factors influencing mapping adaptation actions. In particular, we investigated the influence of: (1) KOS changes related to removal and revision of concepts (2) types of changes affecting relevant attributes identified; and (3) the defined types of change patterns.

We conducted experiments to evaluate the quality of mapping adaptation according to the proposed techniques. We examined the adaptation individually evaluating each type of mapping adaptation action. The findings showed that our proposal performs better for some types of actions than to others, and overall results slightly vary according to the different studied datasets. More importantly, our proposal can still adapt mappings largely automatic and can decrease the human intervention, facilitating therefore the mapping maintenance task. We also discussed different factors impacting the obtained results and pointed out possible studies that can lead to the improvement of results.

In the next chapter, we present the *DyKOSMap* framework to adapt mappings that put all designed components of our approach together. We will describe the whole process for mapping adaptation.

Chapter 7

The *DyKOSMap* framework

Contents

Introduction **163**
7.1 The framework description **165**
7.2 Mapping adaptation method **166**
7.2.1 Adaptation of mappings affected by KOS changes 167
7.2.2 Selection and application of mapping adaptation actions 171
7.2.3 Conflicts and status of mappings under adaptation 172
7.3 The *DyKOSMap* prototype software **174**
7.3.1 Architecture 174
7.3.2 Implementation 175
7.4 Experimental evaluation **176**
7.4.1 Materials and procedure 176
7.4.2 Experimental results 178
7.5 Discussion **179**
Conclusion **181**

Introduction

Throughout the chapters in this thesis, we have conducted in-depth experiments to thoroughly understand the evolution of KOSs and mappings. On this ground, we proposed original methods to interpret established mappings and to further characterize KOS evolution in order to leverage and rely our adaptation techniques on. We have also experimentally validated all these components in a partial way. However, so far we have presented the designed components in a separated way, and our proposed mapping adaptation techniques have studied the adaptation decisions at the level of a single mapping.

To provide a complete solution requires combining the studied components of our approach in an integrated way. We aim to implement an entire workflow handling methods of adaptation for a whole set of mappings and KOSs as input. For this purpose, we propose a mapping adaptation framework that assembles and organizes all the designed components of the *DyKOSMap* approach. This chapter addresses the following specific research questions:

1. How to cope with the revision and removal of concepts and attributes involved in mappings when both source and target KOSs evolve for mapping adaptation?
2. What are the benefits of a mapping adaptation method where decisions are individually taken for each correspondence affected by KOS evolution?
3. Is it possible to integrate the designed components of the *DyKOSMap* approach in a framework and implement a software prototype?
4. Does an original mapping adaptation workflow proposed in the framework yield meaningful and useful validation results for mapping maintenance?

This chapter reports on the *DyKOSMap* framework for mapping adaptation. We present how the framework integrates the components of the approach developed in this thesis. We study and explain an entire process for a largely automatic adaptation of KOS mappings. In particular, our solution avoids the expensive redetermination of the complete set of mappings, and reuses all stable parts from input mappings. The framework describes the way KOS evolution and change patterns, which characterize the evolution of concept attributes, are grabbed and combined to determine those affected mappings and to adapt them. We further present the way that the framework extracts action arguments and predicates to perform the selection and application of mapping adaptation actions based on the modeled heuristics (*cf.* Chapter 6). This chapter provides the following contributions:

- We present the general schema of the framework delineating the integration of the designed components of the approach.
- We propose novel algorithms implemented in the framework on top of the proposed mapping adaptation techniques (actions and heuristics). The framework introduces techniques to handle two evolving KOSs. The designed workflow allows to handle a large aspect of KOS evolution with specific methods to deal with the revision and removal of concepts and attributes. Moreover, we describe observed scenarios leading to conflicts in the adaptation, which requires human intervention.
- This chapter presents the developed software prototype consisting in the implementation of the *DyKOSMap* framework. Furthermore, our research globally evaluates the performance of the proposed framework by adapting mappings between biomedical KOSs over the studied datasets of biomedical KOS mappings.

We start presenting the *DyKOSMap* overview by assembling the components already designed and evaluated (Section 7.1). We then go into detail of the workflow to adapt mappings, showing the different algorithms and their relationships. These algorithms explore the relevant attributes, change patterns, heuristics and mapping adaptation actions (Section 7.2). Section 7.3 briefly presents the developed prototype that largely implements the framework. We then report on the final experimental evaluation to validate the *DyKOSMap* approach to mapping adaptation (Section 7.4), and discuss the achieved results in section 7.5.

7.1 The framework description

Figure 7.1 illustrates the *DyKOSMap* framework integrating the proposed components in the frame of this thesis to adapt KOS mappings. This framework provides the relationships between the modules from the three major aspects considered in the approach: KOS evolution, mapping interpretation and mapping adaptation.

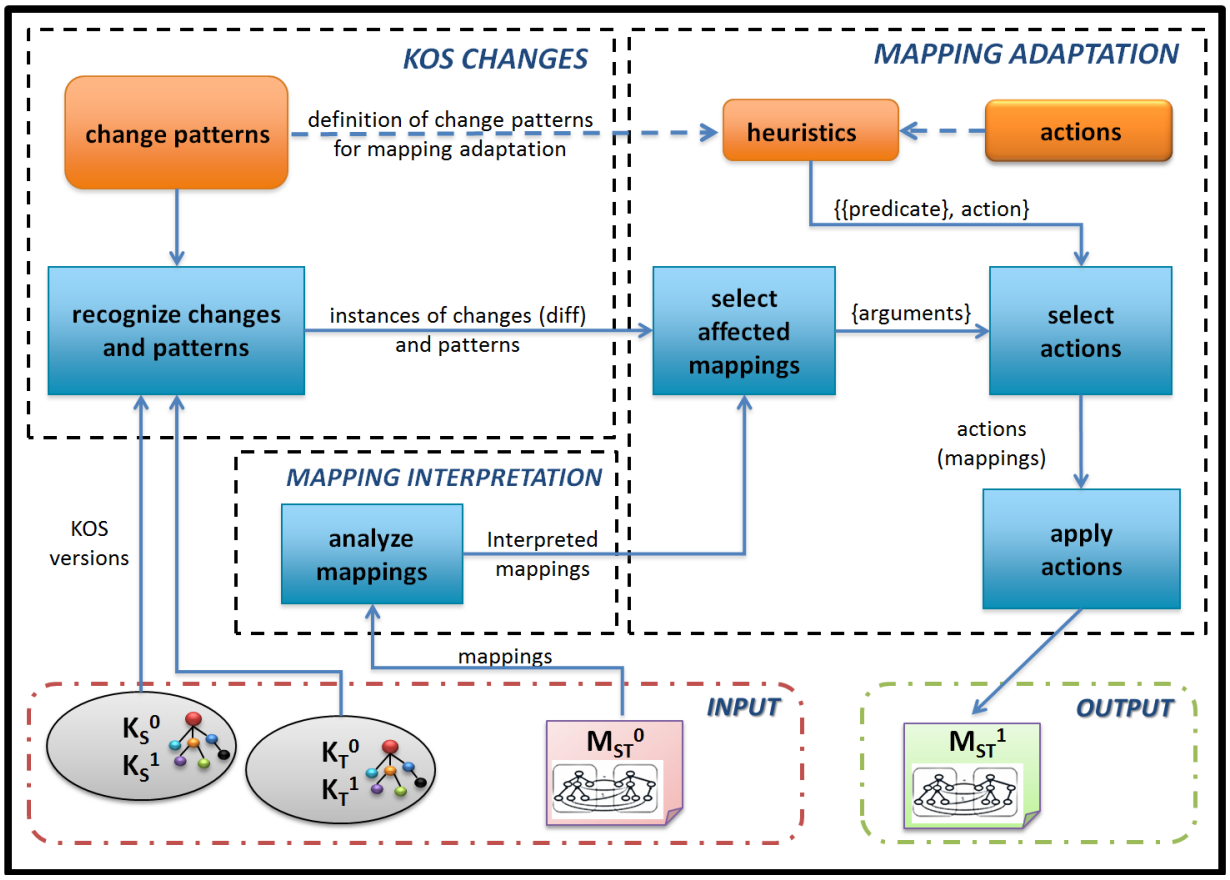


Figure. 7.1: The *DyKOSMap* framework

This figure shows the relationships among the components composing the *DyKOSMap* framework [Dos Reis et al., 2012, Dos Reis, 2013]. This presents the main elements involved with KOS changes, mapping interpretation and mapping adaptation.

The framework foresees as input the set of mappings $\mathcal{M}_{ST}^0$ that requires maintenance. In addition, the source and target KOSs (K_S^0 and K_T^0) interrelated by the input mappings are used. The framework also demands the new release versions of source and/or target KOS (K_S^1 and K_T^1).

KOS changes. The KOS changes component accounts for recognizing instances of KOS changes and patterns given two KOS versions. In this component, we integrate the *COnto-Diff* tool [Hartung et al., 2013] with the change pattern recognition algorithms proposed in this thesis (*cf.* Chapter 5). Given two versions of the same KOS (*e.g.*, K_S^0 and K_S^1), this component based on the *COnto-Diff* tool provides a KOS *diff*. This consists in a set of KOS changes according to

the tables 2.1 and 2.2. The mapping adaptation method explores the KOS *diff* to determine the unaffected part of mappings. In our studies, the designed heuristics in chapter 6 explore the defined change patterns (*cf.* Chapter 5).

Mapping interpretation. This component stands for the analysis of mappings to interpret them and to detect the relevant KOS entities (*cf.* Chapter 4). The adaptation workflow demands this analysis for all mappings that require maintenance.

Mapping adaptation. The proposed approach to adapt mappings requires both interpreted mappings and instances of KOS changes (*diff*) and patterns. The mapping adaptation module relies on the designed heuristics to select and apply the mapping adaptation actions. Chapter 6 investigated the techniques suited to adapt mappings according to the mapping adaptation actions based on the information from KOS changes and interpretation of mappings. Firstly, the mapping adaptation component selects the affected mappings. This extracts further statements related to KOS changes and patterns generating action arguments and predicates. The selection of actions relies on the modeled heuristics and the extracted information to determine the mapping adaptation actions and thus to apply them. Sections 7.2.1 and 7.2.2 provide detail of this process.

The output from the framework refers to a set of up-to-date mappings $\mathcal{M}^1_{ST}$ and a set of conflict mappings filtered by the adaptation method. The latter set of mappings needs the validation by domain experts. Section 7.2 reports on the way mappings are assigned as conflict based on source and target KOS evolution, and section 7.2.3 describes evolution scenarios that can also generate conflicts in the adaptation workflow. The following sections give additional detail on the designed algorithms expressing the entire mapping adaptation workflow.

7.2 Mapping adaptation method

The mapping adaptation workflow aims to separate the input mappings in different subsets and performs the adaptation accordingly. Algorithm 6 presents the main mapping adaptation method. Firstly, it determines the set of unaffected mappings $\mathcal{M}_{unaffected}$ (line 1). This set refers to those mappings where KOS evolution does not impact both source and target concepts in mappings. To this end, the procedure checks whether some KOS changes exist in the respective KOS *diff* related to source or target concepts. We assume that these unaffected mappings remain stable and $\mathcal{M}^1_{ST}$ receives them (line 2).

When source and target KOS change, KOS evolution can simultaneously affect both source and target concepts in mappings. For example, if both concepts of a mapping are split into several concepts, independently handling these changes one after the other can yield wrong adaptation results. Figure 7.2 presents a possible problem scenario [Groß et al., 2013], where the concept named "*extremities*" and "*limbs*" are split and pose difficulties for mapping adaptation. To deal with such situations when both KOS have evolved, we propose to firstly identify correspondences involved in conflicts and isolate these mappings before applying the adaptation method. Domain experts must handle this set of $\mathcal{M}_{conflict}$ (line 3). In particular, we recommend to check conflicting change combinations as *split-split*, *merge-split*, *substitute-split* as well as cases where attribute changes affect source and target concepts.

Algorithm 6 Mapping adaptation

Require: $\mathcal{M}^0_{ST}, n \in \mathbb{N}, K^0_S, K^1_S, K^0_T, K^1_T, diff_{K^0_S,K^1_S}, diff_{K^0_T,K^1_T},$

Ensure: $\mathcal{M}^1_{ST} = \{(m_{st1}), (m_{st2}), ..., (m_{stz})\}$

1: $\mathcal{M}_{unaffected} \leftarrow getUnaffectedMappings(\mathcal{M}^0_{ST}, diff_{K^0_S,K^1_S}, diff_{K^0_T,K^1_T})$
2: $\mathcal{M}^1_{ST} \leftarrow \mathcal{M}_{unaffected}$
3: $\mathcal{M}_{conflict} \leftarrow getConflictMappings(\mathcal{M}^0_{ST}, diff_{K^0_S,K^1_S}, diff_{K^0_T,K^1_T})$
4: $\mathcal{M}_{affectedS} \leftarrow getAffectedMappings(\mathcal{M}^0_{ST}, diff_{K^0_S,K^1_S}, diff_{K^0_T,K^1_T})$
5: $\mathcal{M}_{adapted} \leftarrow \varnothing$
6: $\mathcal{M}_{adapted} \leftarrow \mathcal{M}_{adapted} \cup adaptAffectedMappings(\mathcal{M}_{affectedS}, n, diff_{K^0_S,K^1_S}, K^0_S, K^1_S)$ (Algorithm 7)
7: $\mathcal{M}_{inverted} \leftarrow invert(\mathcal{M}^0_{ST})$
8: $\mathcal{M}_{affectedT} \leftarrow getAffectedMappings(\mathcal{M}_{inverted}, diff_{K^0_T,K^1_T}, diff_{K^0_S,K^1_S})$
9: $\mathcal{M}_{adaptedT} \leftarrow adaptAffectedMappings(\mathcal{M}_{affectedT}, n, diff_{K^0_T,K^1_T}, K^0_T, K^1_T)$ (Algorithm 7)
10: $\mathcal{M}_{adapted} \leftarrow \mathcal{M}_{adapted} \cup invert(\mathcal{M}_{adaptedT})$
11: $\mathcal{M}^1_{ST} \leftarrow \mathcal{M}^1_{ST} \cup \mathcal{M}_{adapted}$

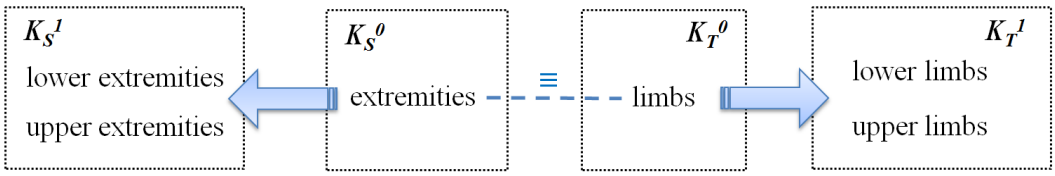


Figure. 7.2: Conflicts between KOS changes in source and target concepts

This figure presents an example of correspondence where KOS changes simultaneously affect source and target concepts [Groß et al., 2013].

Following the algorithm 6 (Figure 7.3 illustrates this workflow), after getting the sets of unaffected and conflicting mappings, the algorithm selects mappings where KOS changes only affect the source concept (line 4) and adapt this selected set of mappings with respect to KOS changes in the source KOS based on algorithm 7 (line 6) (*cf.* Section 7.2.1). To handle the evolution of both KOSs, and to adapt mappings regarding changes in the target KOS, algorithm 6 inverts the input mappings and selects those where the KOS *diff* referring to the target KOS affect mappings (lines 7 and 8). Finally, the resulted adapted mappings are put together and $\mathcal{M}^1_{ST}$ receives the up-to-date mappings (line 11).

7.2.1 Adaptation of mappings affected by KOS changes

Given the affected mappings, algorithm 7 courses each correspondence to provide an individual decision of adaptation. For each affected mapping, algorithm 7 adapts mappings according to the method expressed in algorithm 8 (Adapt individual mapping). To this end, algorithm 7 gets the set of concepts referring to the context of the source concept (line 3). When the source concept belongs to a complex KOS change (*e.g.*, *substitute*, *split* and *merge*), the method restricts the context to the resulting concepts of the complex change according to the KOS *diff* results. The experimental results in section 6.3.3 have motivated this decision to boost the adaptation results. In case the source concept does not belong to a complex KOS change, the context consists in the set of super, sub and sibling concepts at time $j+1$(*cf.* Equation 4.2).

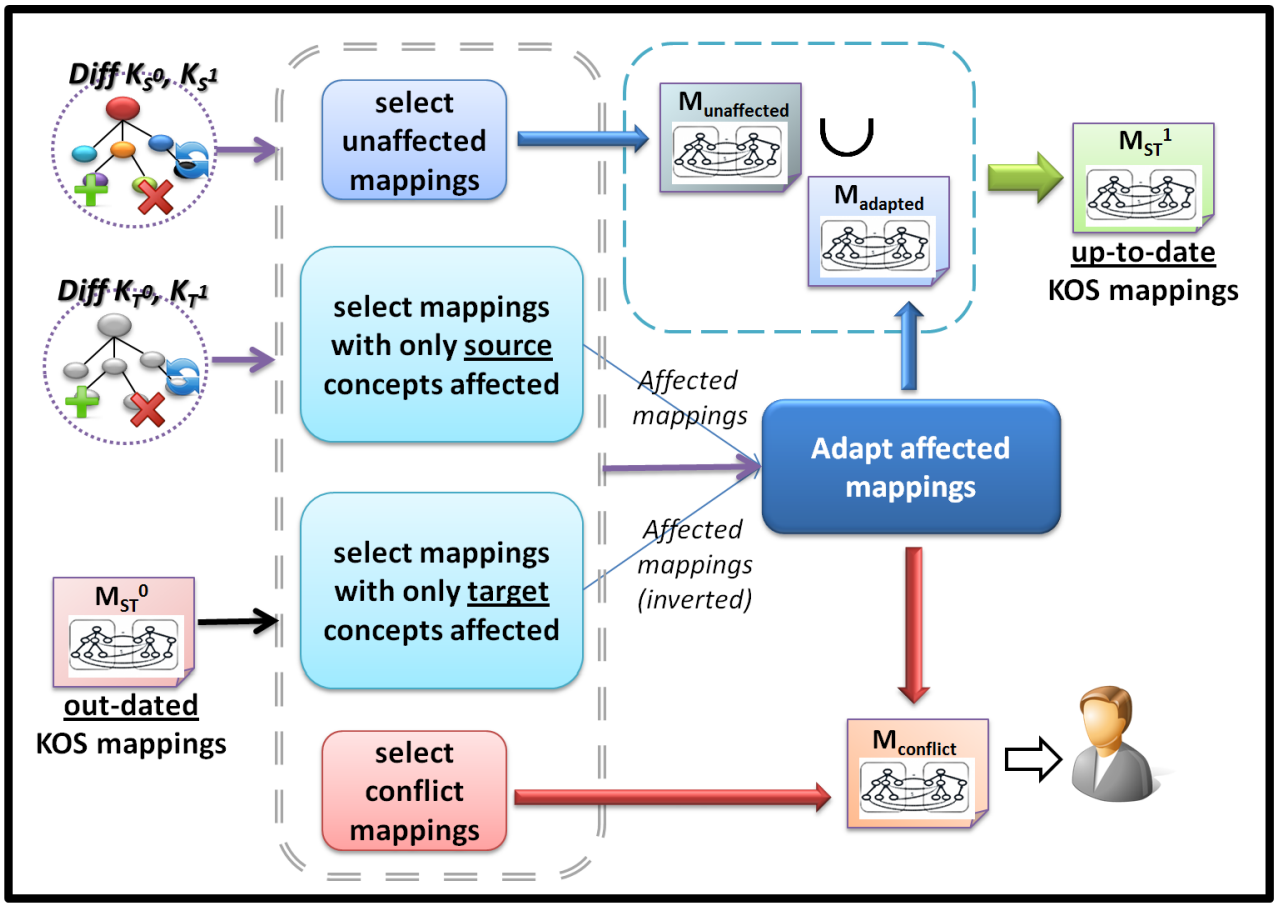


Figure. 7.3: The mapping adaptation method

This figure shows the involved steps to determine unaffected, affected and conflict mappings when a set of input mappings $\mathcal{M}_{ST}^0$ and the respective KOS *diff* are given as input. The method adapts each affected mapping individually. Finally, the method performs a union of adapted mappings and unaffected mappings to propose the updated $\mathcal{M}_{ST}^1$. Domain experts must handle the conflict mappings part.

Algorithm 7 Adapt affected mappings

Require: $\mathcal{M}_{affectedS} \subset \mathcal{M}_{ST}^0, n \in \mathbb{N}, K_S^0, K_S^1, diff_{K_S^0,K_S^1}$

Ensure: $\mathcal{M}_{adapted} = \{(m_{st})_1, (m_{st})_2, ..., (m_{st})_k\}$

1: $\mathcal{M}_{adapted} \leftarrow \varnothing$
2: **for all** $m_{st} \in \mathcal{M}_{affected}$ **do**
3: $CT_{c_s} \leftarrow getCT(c_s^0, K_S^0, K_S^1, diff_{K_S^0,K_S^1})$
4: $\mathcal{M}_{adapted} \leftarrow \mathcal{M}_{adapted} \cup adaptIndividualMapping(m_{st}, n, diff_{K_S^0,K_S^1}, CT_{c_s})$ (Algorithm 8)
5: **end for**

Algorithm 8 refers to the major adaptation algorithm responsible for implementing the whole procedure to adapt a given affected mapping. Figure 7.4 illustrates the workflow and main components involved in the adaptation of an individual mapping affected by KOS evolution.

The adaptation relies on the evolution's characterization of relevant attributes calculated for a given mapping ($topA_{m_{st}}$) (line 4 in algorithm 8). For each attribute from source concept selected with algorithm 1 (the parameter n refers to the number of selected attributes), algorithm 9 (Characterize attribute evolution) determines adequate change information, like the type of attribute change affecting such attribute (based on the KOS *diff*) as well as the lexical and semantic change patterns (line 6). This generates the $EvolAtts$, which is used as input for algorithm 10 that extracts predicates and arguments (line 8). Afterwards, this is used for the selection of mapping adaptation actions (line 9) relying on algorithm 11. Finally, algorithm 8 applies each proposed action and generates the set of updated mappings $\mathcal{M}_{updated}$.

Algorithm 8 Adapt individual mapping

Require: $m_{st} \in \mathcal{M}^0_{ST}, diff_{K^0_S,K^1_S}, CT_{c_s} \subset C(K^1_S), n \in \mathbb{N}$
Ensure: $\mathcal{M}_{updated} = \{(m_{st})_1, (m_{st})_2, ..., (m_{st})_k\}$
1: $\mathcal{M}_{updated} \leftarrow \varnothing$
2: $EvolAtts = \{(chgType, isBest, lcpInst, scpInst)_1, ..., (chgType, isBest, lcpInst, scpInst)_z\}$
3: $argActions = \{(m^0_{st}, c^1_{cand}, semType^1, \{predicate\})_1, ..., (m^0_{st}, c^1_{cand}, semType^1, \{predicate\})_z\}$
4: $topA_{m_{st}} \leftarrow topA(m_{st}, c_s, c_t, n)$ (Algorithm 1)
5: **for all** $a_s \in topA_{m_{st}}$ **do**
6: $\quad EvolAtts \leftarrow EvolAtts \cup characterizeAttEvolution(a_s, c_s, CT_{c_s}, diff_{K^0_S,K^1_S})$ (Algorithm 9)
7: **end for**
8: $argActions \leftarrow extractArguments(m_{st}, EvolAtts)$ (Algorithm 10)
9: $MAA_{proposed} \leftarrow selectActions(argActions)$ (Algorithm 11)
10: **for all** $MAA_{toApply} \in MAA_{proposed}$ **do**
11: $\quad \mathcal{M}_{updated} \leftarrow \mathcal{M}_{updated} \cup MAA_{toApply}.applyAction()$
12: **end for**

In the following, we present how the characterization of attributes' evolution takes place, and the extraction of action arguments and predicates. Section 7.2.2 presents detail on the way mapping adaptation actions are selected based on the modeled heuristics (Algorithm 11).

Characterization of attributes' evolution

For a given relevant attribute selected from *topA*, algorithm 9 explores the KOS *diff* and the change patterns to characterize KOS changes in attribute. This involves checking the type of attribute change (*chgType*) based on the KOS *diff* statements (line 1). Moreover, this algorithm sets whether the given attribute is the best relevant among those selected by the *topA* method (line 2). More importantly, algorithm 9 calls the methods for lexical (line 3) and semantic change pattern recognition, taking into account the set of concepts selected in the context. The algorithm calculates whether there are existing semantic change patterns with the candidate concept and the source concept c^1_s (if it exists).

Extraction of action arguments and predicates

The mapping adaptation decisions mostly rely on the evolution behaviours concerning the relevant attributes for a given affected mapping. Since algorithm 8 (Adapt individual mapping) calculates this information based on algorithm 9 (Characterize attribute evolution), the next step involves analysing the generated information to derive statements that can support the mapping

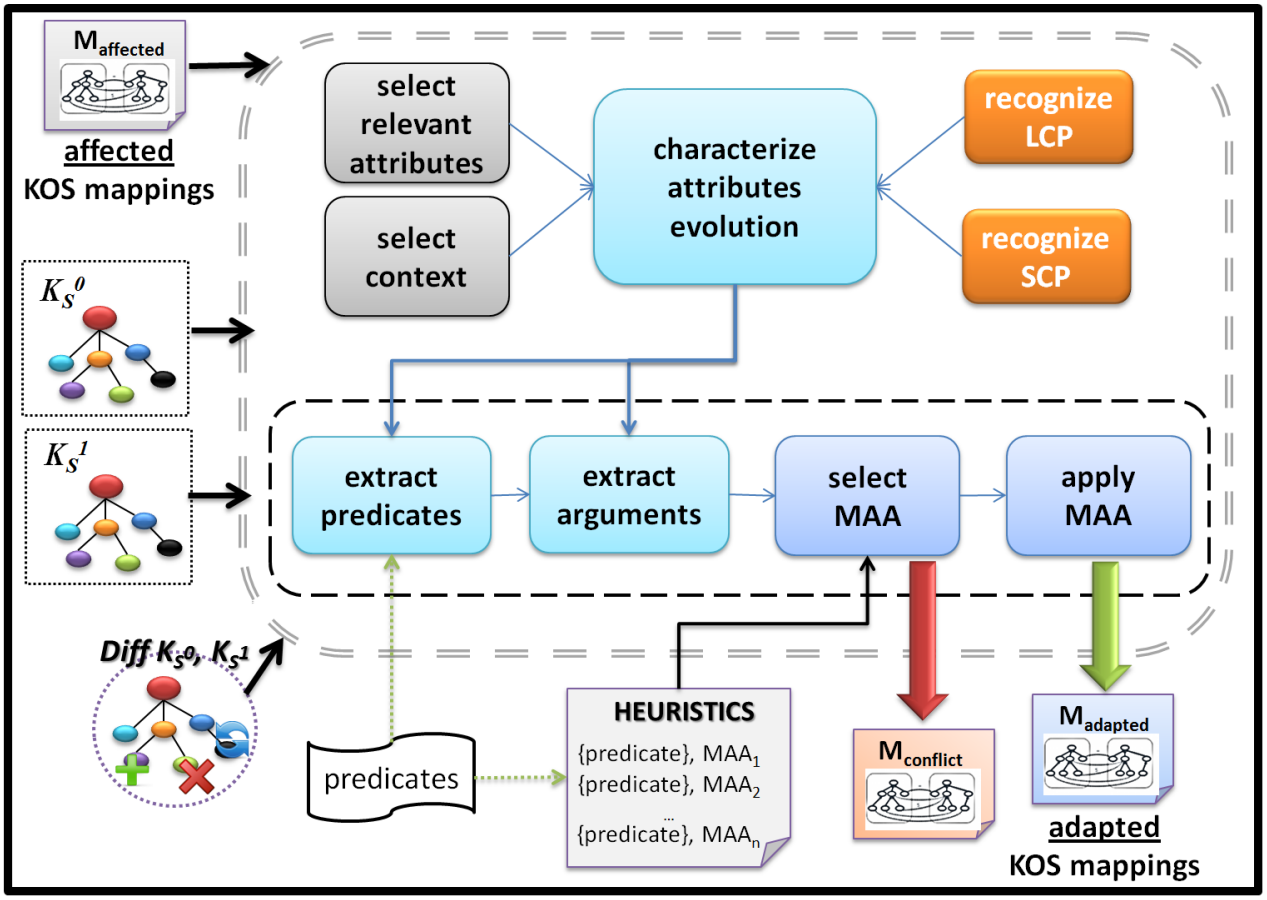


Figure. 7.4: Adaptation of KOS mappings affected by KOS evolution

This figure shows a schema for adapting an individual mapping. For each mapping determined as affected, the method characterizes the evolution of relevant attributes of this mapping with lexical and semantic change patterns, among other information. On this basis, the method extracts predicates and arguments and comparing the extracted predicates against the modeled ones (from heuristics), it selects and applies the adequate MAAs. Conducting this process for the whole set of affected mappings generates the set of adapted mappings and the set of conflicts.

adaptation decisions. For this purpose, our approach extracts predicates and arguments that allow making decisions based on the modeled heuristics (*cf.* Figure 7.4).

Predicates consist in boolean statements representing specific conditions to heuristics. For example, "*Is the attribute the best relevant one?*" (*i.e.*, the most similar with the target concept), "*Is the attribute deleted?*". We defined a set of predicates to accommodate the conditions expressed in the modeled heuristics (*cf.* Section 6.4) and for each heuristics, we proposed and combined the adequate predicates formulating rules. From a technical perspective, we represent the predicates and heuristics as *XML* files that our software prototype parses and takes into account (*cf.* Section 7.3). Given the evolution characterization statements produced for all relevant attributes of a given affected mapping, algorithm 10 (Extract actions argument) relies on a specific procedure to extract all observed predicates (line 4). This procedure instantiates and sets boolean values for all determined predicates.

Algorithm 9 Characterize attribute evolution

Require: $a_s \in A(c_s^0), c_s^0 \in C(K_S^0), CT_{c_s} \subset C(K_S^1), diff_{K_S^0,K_S^1}$
Ensure: $(chgType, isBest, lcpInst, scpInst)$
1: $chgType \leftarrow getAttChgType(a_s, diff_{K_S^0,K_S^1})$
2: $isBest \leftarrow setBestAtt(a_s)$
3: $lcpInst \leftarrow recognizeLCP(a_s, CT_{c_s})$ (Algorithm 3)
4: **if** $lcpInst \neq \varnothing$ **then**
5: $c_{cand}^1 \leftarrow lcpInst.candidate()$
6: $scpInst \leftarrow recognizeSCP(a_s, c_{cand}^1)$ (Algorithm 4)
7: **else**
8: **if** $c_s^1 \in C(K_S^1)$ **then**
9: $scpInst \leftarrow recognizeSCP(a_s, c_s^1)$ (Algorithm 4)
10: **end if**
11: **end if**

Action arguments stand for the parameter values that will be required in case of applying one or another mapping adaptation action. For example, if the mapping adaptation selects a *MoveM* action based on the comparison among extracted and modeled predicates in heuristics (*cf.* Section 7.2.2), this requires knowing the candidate concept c_{cand}^1 to apply the action. Similarly, if proposing a *ModSemTypeM* action, applying this action requires knowing the new *semType* suggested. These arguments are retrieved from the lexical and semantic change patterns identified for the relevant attributes, respectively.

Algorithm 10 instantiates *argActions* (line 8) containing the arguments observed and the predicates extracted based on the processed statements from evolution of the relevant attributes.

Algorithm 10 Extract action arguments

Require: $m_{st} \in \mathcal{M}_{ST}^0$,
1: $EvolAtts = \{(chgType, isBest, lcpInst, scpInst)_1, ..., (chgType, isBest, lcpInst, scpInst)_z\}$
Ensure: $argActions = \{(m_{st}^0, c_{cand}^1, semType^1, \{predicate\})_1, ..., (m_{st}^0, c_{cand}^1, semType^1, \{predicate\})_z\}$
2: $predArgs = \{(name, negative)_1, ..., (name, negative)_w\}$
3: **for all** $attEvol \in EvolAtts$ **do**
4: $predArgs = extractPredicates(attEvol, EvolAtts)$
5: $c_{cand}^1 \leftarrow attEvol.lcpInst.candidate()$
6: $newType \leftarrow getSemType(semType_{st}^0, attEvol.scpInst.getSCP())$
7: **if** $\|argActions\| = 0 \vee c_{cand}^1 \neq \varnothing \vee newType \neq \varnothing$ **then**
8: $argActions \leftarrow argActions \cup (m_{st}^0, c_{cand}^1, newType, predArgs))$
9: **end if**
10: **end for**

7.2.2 Selection and application of mapping adaptation actions

The final step of the adaptation of a given affected mapping consists in selecting and applying mapping adaptation actions based on the predicates and arguments extracted (*cf.* Figure 7.4). Algorithm 11 presents a simplified representation of this task implemented in our framework.

The input refers to the set of action arguments calculated for a given mapping. Note that one affected mapping relates to at least one action argument that can trigger more than one mapping adaptation action according to the extracted predicates. For example, if a *Total Transfer* (a type of lexical change pattern) is detected for an attribute, and if it also leads to a modification of the type of semantic relation, we can have the instance of an action argument containing the candidate concept and the new type of semantic relation. When applying MAA based on an argument, each type of MAA uses the adequate and necessary value from the argument. In the presented example, the *MoveM* action uses the candidate concept (c^1_{cand}), and the *ModSemTypeM* action uses the *newType* argument. Note also that a unique mapping can generate more than one action argument. This case accommodates the *DeriveM* action, since it requires several candidate concepts, to which each one generates an action argument.

Therefore, in algorithm 11 (Select mapping adaptation actions), for each given action argument (line 2), the algorithm courses the whole set of modeled heuristics (parsed from a *XML* file) and compares whether the extracted predicates (calculated analysing mapping interpretation and KOS evolution) fulfill at least one entire set of modeled predicates of heuristics (line 4). If so, the method proposes the mapping adaptation action yielded by such heuristics, and takes the considered argument to apply the action (line 5). For each action argument, one or more actions can be activated (*e.g.*, a *MoveM* and a *ModSemTypeM* action). Finally, the algorithm returns the set of proposed mapping adaptation actions $MAA_{proposed}$, which are applied to adapt the mapping m^0_{st} in the algorithm 8.

This last step also analyses the whole set of statements calculated, which allows us to detect some conflicts to which our attribute-based mapping adaptation approach can lead. We describe them in the next section.

Algorithm 11 Select mapping adaptation actions

Require: $argActions = \{(m^0_{st}, c^1_{cand}, semType^1, \{predicate\})_1, ..., (m^0_{st}, c^1_{cand}, semType^1, \{predicate\})_z\}$
Ensure: $MAA_{proposed} = \{(MAA_1, arg)_1, (MAA_2, arg)_2, ..., (MAA_k, arg)_k\}$
1: $HEU = \{(\{predicate\}, MAA)_1, ..., (\{predicate\}, MAA)_h\}$
2: **for all** $arg \in argActions$ **do**
3: **for all** $h \in HEU$ **do**
4: **if** $satisfyPredicates(arg.getExtractedPred(), h.getModeledPred())$ **then**
5: $MAA_{proposed} = MAA_{proposed} \cup (h.getMAA(), arg)$
6: **end if**
7: **end for**
8: **end for**

7.2.3 Conflicts and status of mappings under adaptation

A closer analysis on the modeled heuristics (*cf.* Section 6.4) explored by the described mapping adaptation method – in addition to observations made in our experiments – allows pointing out some low frequent evolution situations uncovered by the proposed method. We define these situations as conflict scenarios that our framework detects and to which it assigns the concerned mapping for validation by an human expert. These situations occur mostly due to the difficulties in understanding the KOS evolution based on the attribute level.

***MoveM* and *DeriveM*.** A situation may appear where several lexical change patterns are identified with distinct candidate concepts (heuristics suggest leading to *DeriveM* actions), but if the source concept is deleted, it impedes applying *DeriveM* action. Since applying *MoveM* demands only one candidate concept according to equation 6.7 (it is not possible to apply several *MoveM* for the same mapping), we define this scenario as a conflict.

***MoveM* and modification of semantic relation type with respect to source concept.** In this scenario, only one relevant attribute a_1 with LCP is detected (suggesting a *MoveM*), but simultaneously, another different relevant attribute a_2 ($a_1 \neq a_2$) of the source concept is identified, containing a SCP with the source concept at time $j+1$. In this case, the heuristics in equation 6.7 (for *MoveM*) cannot be applied because a SCP with the source concept exists, as well as this scenario impedes selecting the heuristics in equation 6.9 (for *ModSemTypeR*), because a *MoveM* can exist.

***DeriveM* and modification of semantic relation type with respect to source concept.** This situation refers to a possible lack of semantic change patterns with the source concept at time $j+1$ when *DeriveM* actions are proposed. In this situation, since when applying *DeriveM* the original mapping remains, we need to handle the type of relation of the original mapping. Without recognizing semantic change patterns, the heuristics in equation 6.9 (for *ModSemTypeR*) are not applicable. If no KOS changes affect the relevant attributes, the same *semType* can be suggested between c_s and c_t at time $j+1$, otherwise the selection of actions assigns a conflict. Note that even though no KOS changes affect relevant attributes of the source concept, LCPs of Copy type can still exist.

Modification of semantic relation type with respect to candidate concepts. Similar to the latter situation, but in case of *MoveM* or *DeriveM* actions are proposed with candidate concepts, if no SCPs are detected with respect to the attributes of the candidate concepts, the heuristics expressed in equation 6.10 (for *ModSemTypeR*) cannot be applied. Moreover, in case of identifying LCPs for different relevant attributes simultaneously with equal candidate concepts, but resulting in different *semType* according to the equation 6.10, the framework sets a conflict.

Modification of semantic relation type with respect to source concept. If it is possible to recognize semantic change patterns, concerning different relevant attributes with respect to the source concept at time $j+1$, and if it results in different *semType* proposed, the adaptation method sets a conflict.

***RemoveM* and *NoAction*.** In the modeled heuristics, the cases whether neither LCPs nor SCPs are recognized, which fails to lead to either *ModSemTypeR* or *MoveM*, or *DeriveM*, the mapping adaptation can lead to *RemoveM* or *NoAction*. However, in a very specific situation, in which a deleted attribute does not refer to the best relevant attribute (*i.e.*, the *RemoveM* action remains not applicable) (*cf.* Equation 6.11), and the highest similarity calculated with a candidate attribute in the context is higher than α (*i.e.*, *NoAction* is also not applicable) (*cf.* Equation 6.12), no type of MAA can be selected (unsupported by the modeled heuristics). Therefore, this conflict case refers to a deleted relevant attribute (any other attribute than the most relevant one) without change patterns identified, even though containing a high similarity with a new attribute from a concept of the context.

Status of adapted mappings. A mapping adaptation process can lead involved mappings to different situations, *e.g.*, the conflict scenarios described representing more complex situations. This requires ways to manage the status in adaptation of mappings. For this purpose, we extend the definition of KOS mapping described in equation 2.1 to provide additional statements that can reflect the adaptation of mappings [Groß et al., 2013]. We include the *status* attribute in the KOS mapping definition. The *status* refers to the state of a mapping under adaptation. We propose a basic set of status to describe the state of mapping adaptation, such as $status \in \{unaffected, adapted, conflictKOSevol, conflictAdaptation, verifiedByExpert\}$. The defined status "*unaffected*" can describe mappings not impacted by KOS evolution, while "*adapted*" stands for those mappings successfully handled by the mapping adaptation method. The other status stand for different conflict situations.

We can also enrich the proposed set of status with more fine-grained states that would express further statements regarding the adaptation. For instance, we could include each different type of conflict detected, which could help guiding experts in the evaluation process according to the conflict. Furthermore, the set of *status* could include the mapping adaptation actions applied for adapted mappings. This can support further tasks over these mappings (*e.g.*, mapping validation), which may somehow rely on the proposed actions.

7.3 The *DyKOSMap* prototype software

We present the development aspects of the framework. First, we describe a simplified view of the architecture reporting on the principal software components involved (Section 7.3.1). Afterwards, we show a diagram with the major classes implemented and the key technologies used for the development (Section 7.3.2).

7.3.1 Architecture

Figure 7.5 presents the diagram of components deployed in the framework. The main components refer to *Changes*, *KOS* and *Mappings*. In the *Changes* component, we dedicate specific modules to different types of changes. The framework explores the *COnto-Diff* in this component and implements the change patterns algorithms.

In the *KOS*, a specific component called *converters* consists in classes to transform between different formats (*e.g.*, from OWL to OBO). The *KOS* component uses other components that models the entities (*e.g.*, concept, attribute) as well as *loaders* and *renderers* classes implementing parsers to different KOS's encoding formats.

In the *Mappings* component, we organize the proposed aspects related to mapping adaptation in order to implement the modules for mapping interpretation, mapping adaptation actions as well as the heuristics. The *Mapping* component also uses the components *entities*, *loaders* and *renderers*. *Utilities* implements classes related to the processing of *String*, *Characters* and *Files*.

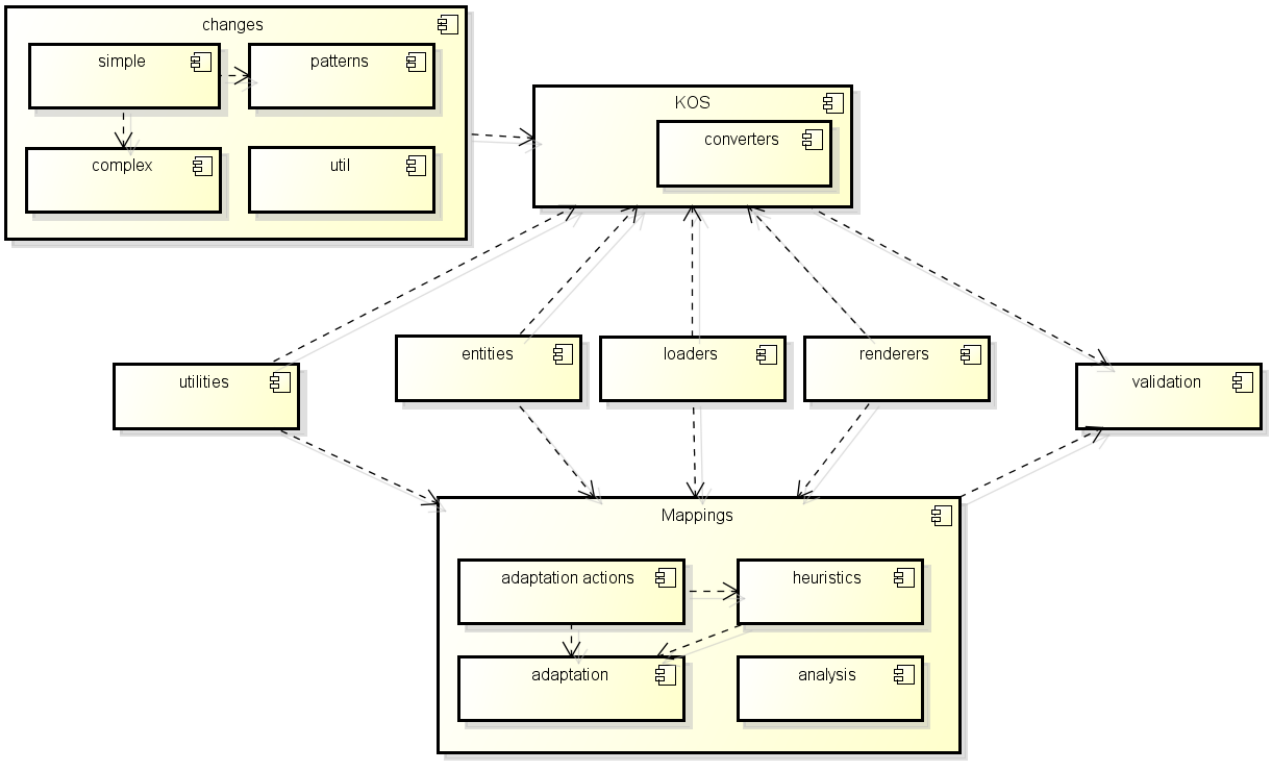


Figure. 7.5: *DyKOSMap* component diagram

This figure shows the diagram of components in the implemented prototype. The main components refer to *Changes, KOS* and *Mappings*.

7.3.2 Implementation

Figure 7.6 presents a superficial class diagram representing the major classes implemented in the prototype of the framework. We modeled the different types of mapping adaptation actions in a hierarchy of classes. More importantly, the diagram shows the way heuristics and predicates are implemented exploring the actions and predicate types. The diagram also illustrates the entities related to the *Argument* and the most relevant methods implemented in the class *MappingAdapter*.

We have implemented the *DyKOSMap* prototype in *Java* language to conduct experimental analyses and evaluations. The implementation used the *MySQL* environment to create the *DyKOSMap* database, that required *Mysql Connector Java API* to perform and manage database connections.

The prototype also uses the *SimMetrics*[39] Java library which implements several types of similarity measures. We explored other APIs, like the *SAX*[40], to parse *XML* files. We implemented the modeled heuristics in a *XML* file, which is parsed instantiating heuristics and predicates class instances.

[39] http://sourceforge.net/projects/simmetrics
[40] http://www.saxproject.org

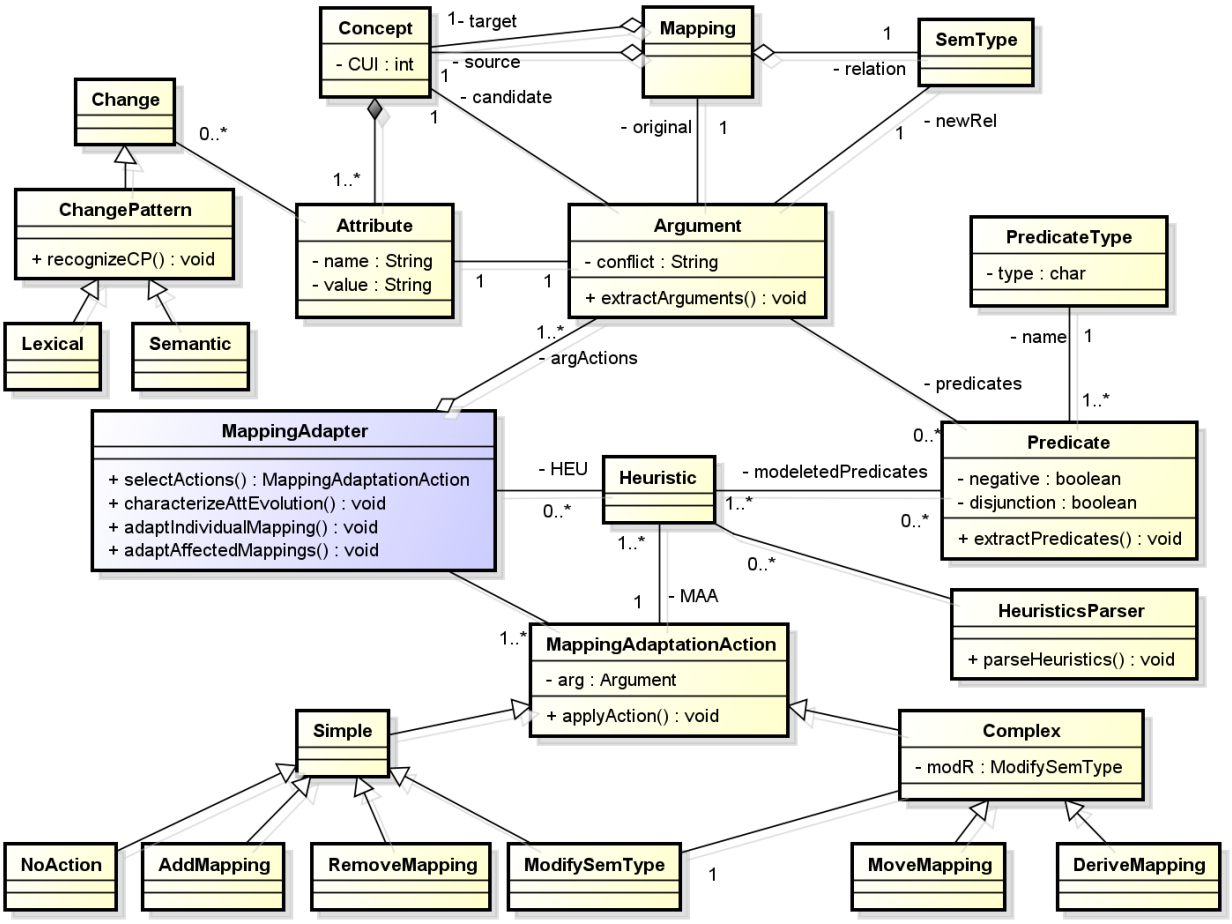


Figure. 7.6: *DyKOSMap* class diagram

This figure depicts a simplified view of the diagram of classes implemented in the prototype. This diagram intends to give a sense of the major classes and their relationships to execute the framework.

7.4 Experimental evaluation

The final experimental evaluation of this thesis aims to validate the implemented *DyKOSMap* framework for mapping adaptation. In contrast to the evaluation conducted in chapter 6, which emphasized the evaluation at the level of mapping adaptation actions, this evaluation intends to show a more global validation in terms of mapping adaptation. This evaluation examines the general behaviour yielded by the framework's mapping adaptation workflow, and the adaptation results obtained. We describe the materials and the conducted procedure in this experimental evaluation in section 7.4.1. Section 7.4.2 reports on the achieved results, while section 7.5 discusses the findings.

7.4.1 Materials and procedure

We use the same material as in the evaluation conducted in chapter 6. Table 7.1 presents all the studied biomedical KOSs and gives a sense of their size in terms of concepts, attributes and subsumption relationships.

Table 7.2 shows the set of mappings by the respective releases. The datasets SNOMEDCT-ICD9CM and MeSH-ICD10CM refer to official mappings curated by biomedical organizations and the size of these datasets significantly differ. The dataset SNOMEDCT-NCI refers to mappings extracted from the UMLS as explained in section 6.3.1.

Table 7.1: Overview on the biomedical KOS entities

KOS	Release	#Concepts	#Attributes	#Subsumptions
SNOMEDCT	2010	390 022	1 547 855	530 433
	2012	395 346	1 570 504	567 719
NCI	2009	77 448	282 434	86 822
	2012	94 732	365 515	105 406
MeSH	2012	50 367	259 565	59 191
	2013	50 971	264 783	59 844
ICD9CM	2009	12 734	34 065	11 619
	2011	13 059	34 963	11 962
ICD10CM	2011	43 351	87 354	40 330

This table presents the absolute numbers of concepts, attributes and subsumption entities for the studied biomedical KOSs. This aims to show their size and how one KOS differs from another.

Table 7.2: Evaluated KOS mapping datasets

KOS mapping	Release	#Mappings
SNOMEDCT-ICD9CM	2010-2009	100 451
	2012-2011	102 703
SNOMEDCT-NCI	2009-2009	19 971
	2012-2012	22 732
MeSH-ICD10CM	2012-2011	4 631
	2013-2011	5 378

This table shows statistics of the several releases of KOS mappings concerning the number of correspondences between the biomedical KOSs studied.

For each one of the KOS mapping datasets, we adapted the first release of mappings with the proposed framework. We used the second release as reference mappings for evaluating the quality of the adapted mappings. To ensure consistency in the conducted validation, we further processed the reference mappings.

The considered reference mappings are not wholly gold standard as we already discussed, *i.e.*, these mappings are not complete, and curators manually correct them by also modifying correspondences associated with concepts unaffected by KOS changes (observations from our previous experiments). Therefore, we eliminate such mappings since they do not change due to KOS changes. Moreover, we remove from the reference mappings all those mappings assigned as conflict (*i.e.*, cases when KOS evolution impacts both source and target concepts).

To assess the quality of the adapted mappings with respect to the reference mappings, we calculated the standard metrics of *Precision*, *Recall* and *F-Measure*. We computed the *Precision* as the number of mappings correctly proposed by the adaptation framework in contrast to the expected reference mappings (we check source and target concepts as well as the *semType* to identify the exact correct mappings), over the total number of adapted mappings:

$$Precision = \frac{\#correctlyAdaptedMappings}{\#adaptedMappings} \tag{7.1}$$

We calculated the *Recall* as the number of correctly adapted mappings over the total number of reference mappings:

$$Recall = \frac{\#correctlyAdaptedMappings}{\#referenceMappings} \tag{7.2}$$

The *F-measure* refers to the harmonic mean of *Precision* and *Recall*.

$$F\text{-}measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{7.3}$$

We propose two configurations to examine the mapping adaptation method. First, we calculate *Precision*, *Recall* and *F-Measure* for the set of unaffected mappings $\mathcal{M}_{unaffected}$ (stable correspondences) in the adapted mappings (*Unaffected* in figures 7.7 and 7.8). This remains relevant to have a basic reference (baseline) for analyzing to which extent our adaptation approach contributes to relevant mappings. Second, we evaluate the outcome set of updated mappings $\mathcal{M}^1_{ST}$, which refers to the final resulting set of adapted mappings (*i.e.*, $\mathcal{M}_{adapted}$, adaptation of mappings considering the evolution of source and target KOSs) summed up with the stable part $\mathcal{M}_{unaffected}$ (*Adapted* in figures 7.7 and 7.8). Similarly to chapter 6, this evaluation uses the following values for the thresholds in our change pattern recognition algorithms and heuristics: $\tau = 0.7$, $\gamma = 0.9$ and $\alpha = \tau$. In addition, we used three as the number of relevant attributes.

7.4.2 Experimental results

Figure 7.7 presents the quality of the mapping adaptation results for SCT-ICD9 (left) and SCT-NCI (right). For both datasets, we observe that the basic quality of *Unaffected* mappings remains very high, as our adaptation approach promotes the reuse of unaffected mappings, which appears high in these datasets. This is all the more so true for the SCT-ICD9 dataset, where the set of mappings is much bigger than SCT-NCI.

Comparing to the *Unaffected* in both datasets, the quality of mapping adaptation provided by the framework slightly decreases the *Precision* and considerably increases the *Recall* (*Adapted* scenario). This difference remains more significant for the SCT-NCI because the *Recall* of *Unaffected* is lower than SCT-ICD9.

The *F-Measure* is higher compared to the *Unaffected* scenario for both datasets. This is due to the fact that our mapping adaptation framework not only reuses unaffected mappings, but can further improve *Recall* with relatively high *Precision* thankful to the mapping adaptation workflow implemented based on the designed heuristics.

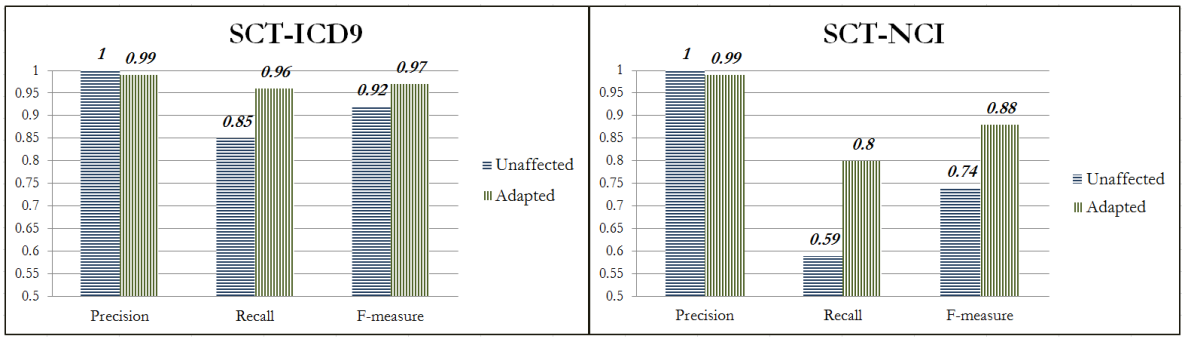


Figure. 7.7: Results of the mapping adaptation quality for SCT-ICD9 and SCT-NCI

This figure shows the evaluation results with respect to metrics of *Precision*, *Recall* and *F-Measure* (*cf.* equations 7.1, 7.2, 7.3). We calculate these metrics taking into account the observable mappings of the new mapping release (expected ones) compared to the set of mappings resulted from the application of the *DyKOSMap* framework (proposed ones). Left side of the figure presents the adaptation results for the dataset SCT-ICD9, and the right side for the dataset SCT-NCI. The results for *Unaffected* refer to the set of mappings unaffected by KOS evolution based on our mapping adaptation method (*i.e.*, $\mathcal{M}_{unaffected}$). The results for *Adapted* consider as proposed mappings the outcome $\mathcal{M}^1_{ST}$, which refers to the set of unaffected mappings in addition to the the set of adapted mappings (*i.e.*, $\mathcal{M}_{unaffected} \cup \mathcal{M}_{adapted}$).

Figure 7.8 presents the quality of the mapping adaptation results for the MeSH-ICD10 (left) and the overall outcome taking all datasets together (right). For the dataset MeSH-ICD10, results shows a slightly lower performance compared to SCT-ICD9 and SCT-NCI (*i.e.*, smaller contribution of *Adapted*), because the difference in terms of *F-measure* between *Unaffected* and *Adapted* remains lower than the other datasets analysed. However, the MeSH-ICD10 is a smaller dataset, which decreases the number of affected mappings by KOS evolution and consequently the number of mappings requiring adaptation (*i.e.*, it restricts the contribution of the *Adapted* compared to the *Unaffected*). On the other hand, this aspect favors achieving a great *Precision*.

The General results combine the performance of the framework for all datasets. This shows the overall high effectiveness of the *DyKOSMap* approach for mapping adaptation. These results could be improved by refining the mapping adaptation techniques and handling new added concepts that can result in new correspondences in the reference mappings.

7.5 Discussion

This chapter proposed and validated a framework that composes and implements the *DyKOSMap* approach for mapping adaptation. Our investigation found the following answers for the formulated research questions in this chapter:

1. Analysing the input mappings based on the KOS *diff* allows to separate affected and unaffected mappings part and handle the evolution of both interrelated KOS;

2. The proposed mapping adaptation method performs adaptation decisions for each affected mapping individually, which enables to further analyse each correspondence and its aspects of KOS changes impacting source and target concepts. This leads to a more complex problem, but can improve the precision of decisions in mapping adaptation;

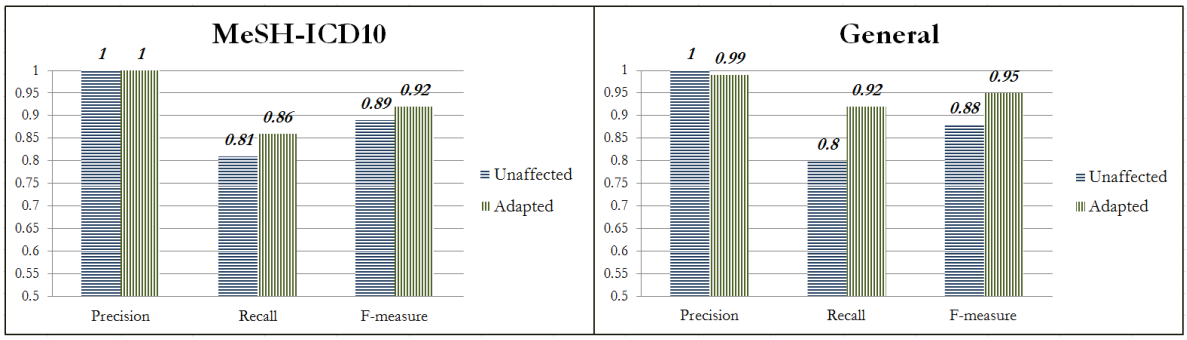


Figure. 7.8: Results of the mapping adaptation quality for MeSH-ICD10 and General analysis

This figure shows the evaluation results with respect to metrics of *Precision*, *Recall* and *F-Measure* (*cf.* equations 7.1, 7.2, 7.3). We calculate these metrics taking into account the expected mappings of the new mapping release (reference ones) compared to the set of mappings resulted from the application of the *DyKOSMap* framework (proposed ones). Left side of the figure presents the adaptation results for the dataset MeSH-ICD10, and the right side shows results of a general analysis considering all datasets together. The results for *Unaffected* refer to the set of mappings unaffected by KOS evolution based on our mapping adaptation method (*i.e.*, $\mathcal{M}_{unaffected}$). The results for *Adapted* consider as proposed mappings the outcome $\mathcal{M}^1_{ST}$, which refers to the set of unaffected mappings in addition to the set of adapted mappings (*i.e.*, $\mathcal{M}_{unaffected} \cup \mathcal{M}_{adapted}$).

3. The *DyKOSMap* framework demonstrated the possibility of integrating the components of the proposed approach and the viability of implementing this in a computational way;

4. The conducted experimental validation yielded meaningful results underscoring the general benefits of the approach;

More specifically, experimental findings indicated that our mapping adaptation method contributes and improves the quality of mappings compared to the set of unaffected mappings (stable part of mappings not impacted by KOS evolution). This chapter achieved experimental results evaluating the execution of the framework with several datasets of mappings between biomedical KOSs. This showed the overall performance and effectiveness of the approach over datasets of different characteristics and sizes.

The evaluation focusing on the performance of mapping adaptation actions in chapter 6 revealed deeper differences in terms of results performance among the studied datasets. In contrast, the overall results achieved in this chapter pointed out more similar results among them, even though the studied datasets refer to both official and unofficial mappings. This relies on the fact that the unaffected mappings part refers to the biggest part of input mappings, and in the adapted mappings updated, the *NoAction* achieves a very good performance representing the majority of the mappings. The types of actions where we could observe more results disparities between the datasets have a lower frequency, which consequently performs a minor impact in this evaluation.

The obtained results remain very similar to previous recent studies on mapping adaptation between life science ontologies [Groß et al., 2013], in which the author has been involved. Futhermore, this thesis makes further improvements because it takes more constraints into account in the conducted experimental evaluation. In fact, we compare the type of semantic relation in

addition to the source and target concepts (*i.e.*, all mapping elements) in the computation of the evaluation metrics. Groß*et al.* [Groß et al., 2013] have also shown overall improvements when combining the adaptation of mappings with a matching phase to handle addition of concepts in KOS evolution. We judge this matching phase out of the scope of the mapping adaptation problem, so we consider this aspect an issue for future investigations out of this thesis.

Our research on mapping adaptation highlighted several contributions to the mapping maintenance problem. Through the proposed approach and conducted experimental evaluations, we pointed out that the defined framework can facilitate the work of end users in charge of maintaining mappings, because it can adequately adapt mappings in addition to detecting stable and conflicting mappings according to KOS evolution. This minimizes the work of these users, supporting them to deal with a huge number of available mappings between biomedical KOSs.

Finally, this research demonstrated the possibility of adapting KOS mappings largely automatically. Future perspectives may involve further investigations of approaches for handling the addition of concepts to achieve a more complete resulted set of mappings, but without calculating a matching operation with the whole target KOS, as well as ways for supporting and facilitating mapping validation by domain experts based on the adaptation results.

Conclusion

This chapter presented a complete framework to adapt KOS mappings developing a novel concept attribute-driven approach to mapping adaptation. While the previous chapters of this thesis (i) proposed methods to better interpret correspondences based on concept attributes (Chapter 4); (ii) investigated ways to characterize the evolution of these attributes as change patterns (Chapter 5); and (iii) combined these aspects in heuristics to adapt KOS mappings via mapping adaptation actions (Chapter 6), it lacked ways of managing a whole mapping adaptation process integrating the proposed components for selecting and applying mapping adaptation actions.

We introduced and described the *DyKOSMap* framework. This framework demonstrated a wholly original workflow suited to adapt mappings based on the analysis of existing correspondences and KOS evolution. We showed how to reuse stable mappings and to handle the evolution of both interrelated KOSs. We presented the mapping adaptation method to grab the required statements from affected mappings and KOS changes in order to make decisions over mapping adaptation actions based on the modeled heuristics. Our research found possible conflict scenarios demanding human experts intervention in mapping adaptation. This investigation resulted in the development of a software prototype implementing the proposed approach.

We conducted experiments to evaluate the quality of mapping adaptation applying the framework entirely. The validation showed an overall efficiency of the proposal for different datasets in the biomedical domain. We demonstrated that our approach remains applicable where mappings can be explained by the textual attributes denoting concepts. This chapter also discussed the improvements and limitations of our original approach in contrast to existing proposals in the state-of-the-art. We concluded that the implemented mapping adaptation framework benefits domain expert users in charge of the mapping maintenance task. The next chapter closes this thesis with the major conclusions and perspectives.

Chapter 8

Conclusions and perspectives

Contents

8.1 Summary of the contributions 184
8.2 Directions for future work 185
8.3 Final remarks 186

Now more than ever, biomedical KOSs play a central role in allowing better organization and exploration of the overwhelmingly voluminous data available in this domain. Mappings established between them are necessary and play a crucial part in supporting various tasks in software applications. However, due to the rapid evolution of the biomedical KOSs, existing mappings become obsolete and unreliable over time, since the leading cause for invalidating mappings concerns the evolution of the interconnected KOSs.

In this context, mapping maintenance becomes extremely relevant for assuring desirable performance of biomedical systems. This refers to a complex research challenge which aims at keeping established mappings in a updated and valid state, reflecting changes affecting KOS entities at evolution time. Although mapping adaptation has been suggested as a possible approach to address the mapping maintenance problem, and KOS evolution has been under investigation for a long time, any existing approach in literature explicitly exploits information grabbed from KOS evolution combined with information from established mappings to adapt KOS mappings.

This thesis proposed an original approach to adapt mappings. Our research conducted a bottom up methodology in which we empirically grounded the approach with observations on the evolution of real-world mappings. We implemented in-depth experimental analyses leading to the identification of key factors required to support mapping adaptation. This allowed us to propose a novel approach, named *DyKOSMap*, dealing with the interpretation of mappings and KOS evolution in an original mapping adaptation method. The originality of our approach mostly relies on the fact that we investigated a thorough understanding of existing mappings to determine the adequate granularity level to which mappings can be explained. This resulted in a proposal for mapping adaptation, focused on the evolution characterization of relevant KOS entities for explaining mappings. On this basis, we underscored the assumption that a proper adaptation of mappings requires a determination of the KOS entities relevant for interpreting the correspondences, and thus to understand the evolution of these entities to perform modifications in mapping elements. The identification of these entities and the methods to explore them for adaptation of mappings stand for the major contribution of this thesis.

8.1 Summary of the contributions

This PhD thesis generated outcome for the fields KOS mappings and KOS evolution. We belive that this research can contribute to the advancement of the Semantic Web as well as Information Systems research areas. More specifically, we summarize the main scientific contributions of this thesis in the following.

Empirical analyses to understand the influence of KOS evolution on mappings. In chapter 3, we performed the first empirical investigations to study possible interdependencies between the evolution of KOS and mappings. We examined in section 3.2.3 whether some KOS changes, affecting involved concepts in mappings, lead to changes in mappings. This contribution also studied interdependencies between distinct mapping change operations in section 3.3, which enabled us to observe complex behaviours of mapping evolution, as well as the influence of complex KOS changes (Section 3.4). This investigation showed relevant factors that we included as principles in our approach. We found that concept attributes in the studied datasets play a major role in the definition of the established mappings as well as in their evolution. In section 6.3, we studied supplementary aspects that can affect mapping adaptation. The outcome of all these empirical findings resulted in the modeling of heuristics to adapt mappings (Section 6.4).

The novel *DyKOSMap* approach to mapping adaptation. The originality of our approach consists in adapting mappings based on the evolution characterization of relevant attributes identified for a given mapping in an individual manner. The approach involved three major components dealing with mapping interpretation, KOS evolution and mapping adaptation. Each one of these components led to original research and experimental evaluations.

A method to interpret KOS mappings. In chapter 4, we proposed a way to analyze existing mappings in order to detect the concept attributes relevant for explaining a given mapping. To this end, in section 4.2, we introduced a novel method based on similarity coefficients to identify such relevant attributes. Our approach explored the influence of the concepts in the context of those involved in mappings. In section 4.5, we thoroughly evaluated the proposed method by examining whether the identified attributes can correlate with mapping evolution.

A mechanism for characterizing KOS evolution via change patterns. In chapter 5, we proposed a set of change patterns at the level of concept attributes to describe specific evolution behaviours of relevant attributes defining mappings. We argued that better characterizing the way the attributes evolve can further support and inform mapping adaptation decisions. We introduced original methods to recognized the change patterns in sections 5.4 and 5.5. In section 5.6, we thoroughly evaluated our proposed methods in contrast to reference change patterns manually constructed for this validation.

The *DyKOSMap* framework and prototype. In section 6.2, we proposed a set of mapping adaptation actions representing operations to perform changes in mappings. Relying on the whole set of experiments conducted throughout the thesis, we formalized heuristics expressing well-delineated scenarios for applying the most appropriate mapping adaptation actions. We illustrated how the relevant attributes and change patterns are used to make decisions over mapping adaptation. In section 7.2, we explained the entire process of mapping adaptation based on KOS evolution and the modeled heuristics. We demonstrated the integration of the approach's

components in the *DyKOSMap* framework (Section 7.1) and described its implementation aspects in section 7.3.

A thorough assessment of the approach. The implementation of the proposed methods in the developed prototype allowed the experimental validation of the concepts composing the *DyKOSMap* approach. We explored several mapping datasets of the biomedical domain and standard evaluation metrics to validate the proposed framework. The experimental validation demonstrated the effectiveness of the approach to mapping adaptation as well as its limitations. In section 6.5, we evaluated the performance of decisions for each type of mapping adaptation action, while in section 7.4 we assessed the general quality of adapted mappings resulting from the application of the *DyKOSMap* framework. We concluded that the proposed framework can support and minimize the work of users in charge of mapping maintenance.

8.2 Directions for future work

The research presented in this thesis copes with the initial objective and research questions formulated in this thesis. However, the proposed *DyKOSMap* approach shows some limitations and offers relevant perspectives to a further validation, refinement and extension of this work. We discuss possible directions for future research, directly inspired by or stemming from the results of this thesis.

To further validate the approach presented in this thesis, we propose:

1. To conduct additional experimental validations with mapping datasets of other domains out of the biomedical one. We can compare the performance of this proposal for different domains and this can be relevant for attaining experimental evidences of the generalization aspect of our approach;

2. To empirically compare the quality of the obtained results with other approaches to mapping maintenance, *e.g.*, in contrast to the recalculation approach. Such study could measure several aspects to offer different comparable dimensions.

3. To design an experimental validation for a qualitative evaluation of adapted mappings involving domain experts;

4. To further study how the approach is dependent on the expressivity characteristics of the considered KOS. This can demonstrate to which extent the approach can be sufficiently general to be applicable to different KOS (not only focused to biomedical KOSs).

We suggest the following topics to refine the proposed methods in the context of the defined framework:

1. To better understand the influence of the used similarity thresholds on the results of adaptation. This involves studying possible optimizations for the thresholds τ and γ and to which extent these affect the obtained quality of mapping adaptation.

2. To propose methods to handle the addition of concepts and attributes to update the adapted mappings set. This encompasses studying ways of deriving mappings from already interrelated concepts, without applying a matching operation with all entities from

source or target KOS. This can allow making refinements of the KOS alignment based on KOS evolution, and a provision of a more complete set of mappings.

3. To conceive methods based on domain-specific background knowledge to recognize semantic change patterns (not only exploring string-based similarity metrics). This involves examining to which extent these advanced methods leverage our change pattern recognition algorithms.

4. To study ways of refining the defined heuristics based on the feedback of experts while validating the adapted mappings. This means investigating a mechanism able to learn from observations and to improve the heuristics over time.

We can explore several research topics representing a natural extension of this thesis that go beyond what has been investigated:

1. To explore machine learning techniques in the investigated components of the approach for determining the relevant attributes, change patterns and heuristics. For instance, we can study how to semi-automatically define the heuristics based on the observation of mapping evolution from several domains relying on machine learning methods. This involves extracting representative features and adequate methods for the target task.

2. To propose a mechanism to track mapping adaptation. This refers to a way for representing a log of mapping adaptation suited to support users to perform backtrack operations.

3. To study similar methods as proposed in this thesis for the problems of annotation and query maintenance, which are artifacts potentially impacted by KOS evolution. Similarly, we could investigate the link maintenance problem in open linked data.

4. To investigate mechanisms suited to support the process of KOS mapping validation. Since techniques of mapping revision are inadequate to automatically handle less expressive KOS, we need approaches that help domain experts to validate adapted mappings (*e.g.*, the conflict mappings). We could propose a question generation approach to mapping adaptation, where natural language questions are automatically formulated based on the interrelated concepts and the adaptation of mappings.

5. To study aspects related to user interaction in human-computer interfaces to understand how final users make sense of established mappings and their evolution. This can enable proposing interactive mechanisms suited to help users manipulating KOS mappings.

8.3 Final remarks

In summary, this thesis contributed a novel approach to the mapping maintenance problem via mapping adaptation, making an original contribution to the state-of-the-art. As demonstrated throughout the thesis, the principles underlying the *DyKOSMap* framework sounded effective, although the obtained results still show opportunities for improvements. The conducted research methodology enabled the investigation of several dimensions of the problem to understand KOS mappings and their evolution. These investigations led to the publication of several peer-reviewed research articles (*cf.* Annex A). Furthermore, this thesis opened up wide-ranging directions for future research.

Annex A

Publications

The author conducted this thesis in the frame of the *DynaMO*[41] project where he was enrolled full time as Assistant Researcher. The attained research results and the involvement of the author with collaborators and partners of the project led to various refereed scientific publications in the thesis period. Most of the material presented in this thesis has been previously appeared in (or was submitted to) journal and conference articles published in the course of this PhD.

International journals

- **DOS REIS, J. C.**; PRUSKI, C.; REYNAUD-DELAÎTRE, C. 2015. *State-of-the-art on mapping maintenance and challenges towards a fully automatic approach.* In Expert Systems with Applications (ESWA). DOI: 10.1016/j.eswa.2014.08.047, Elsevier, Vol. 42 (3), pp. 1465-1478.
- **DOS REIS, J. C.**; PRUSKI, C.; DA SILVEIRA, M.; REYNAUD-DELAÎTRE, C. 2014. *Understanding semantic mapping evolution by observing changes in biomedical ontologies.* In Journal of Biomedical Informatics (JBI). DOI: 10.1016/j.jbi.2013.09.006, Elsevier, Vol. 47, pp. 71-82.
- DINH, D.; **DOS REIS, J. C.**; PRUSKI, C.; DA SILVEIRA, M.; REYNAUD-DELAÎTRE, C. 2014. *Identifying relevant concept attributes to support mapping maintenance under ontology evolution.* In Journal of Web Semantics (JWS). DOI: 10.1016/j.websem.2014.05.002, Elsevier.

International conferences and workshops

- **DOS REIS, J. C.**; PRUSKI, C.; REYNAUD-DELAÎTRE, C. 2015. *Heuristiques pour l'adaptation des mappings entre ontologies dynamiques.* In Proceedings of the 15^{th} Conference on Knowledge Extraction and Management (EGC'15). Luxembourg. (In French)
- **DOS REIS, J. C.**; DINH, D.; PRUSKI, C.; DA SILVEIRA, M.; REYNAUD-DELAÎTRE, C. 2014. *The influence of similarity between concepts in biomedical ontology evolution for*

[41] *DynaMO* refers to a research project investigated in the CR SANTEC department of the *Public Research Centre Henri Tudor* (`www.tudor.lu`) in Luxembourg. The *National Research Fund of Luxembourg* (`www.fnr.lu`) totally supports the *DynaMO* (grant #C10/IS/786147)

mapping adaptation. In Proceedings of the 25^{th} European Medical Informatics Conference (MIE'14). Istanbul, Turkey, pp. 1003-1007.

- ABACHA, A. B.; **DOS REIS, J. C.**; MRABET, Y. 2014. *Question Generation for the Validation of Mapping Adaptation.* In Proceedings of the 6^{th} International Symposium on Semantic Mining in Biomedicine (SMBM'14). Aveiro, Portugal, pp. 27-31.

- **DOS REIS, J. C.**; DA SILVEIRA, M.; DINH, D.; PRUSKI, C.; REYNAUD-DELAÎTRE, C. 2014. *Requirements for implementing mapping adaptation systems.* In Proceedings of the 23^{rd} IEEE WETICE 2014 Conference. In 6^{th} Modeling the Collaborative Web Knowledge Track (Web2Touch'14). Parma, Italy, pp. 406-410.

- DINH, D.; **DOS REIS, J. C.**; PRUSKI, C.; DA SILVEIRA, M.; REYNAUD-DELAÎTRE, C. 2014. *Identifying change patterns of concept attributes in ontology evolution.* In Proceedings of the 11^{th} Extended Semantic Web Conference (ESWC'14). Crete, Greece. LNCS, Springer. Vol. 8465, pp. 768-783.

- **DOS REIS, J. C.**; DINH, D.; PRUSKI, C.; DA SILVEIRA, M.; REYNAUD-DELAÎTRE, C. 2013. *Mapping adaptation actions for the automatic reconciliation of dynamic ontologies.* In Proceedings of the 22^{nd} ACM International Conference on Information and Knowledge Management (CIKM'13). San Francisco, CA, USA, pp. 599-608.

- **DOS REIS, J. C.**; PRUSKI, C.; DA SILVEIRA, M.; REYNAUD-DELAÎTRE, C. 2013. *Characterizing semantic mappings adaptation via biomedical KOS evolution: A case study investigating SNOMED CT and ICD.* In Proceedings of the American Medical Informatics Association Annual Symposium (AMIA'13). Washington DC, USA, pp. 333-342.

- GROSS, A., **DOS REIS, J. C.**; HARTUNG, M.; PRUSKI, C.; RAHM, E. 2013. *Semi-automatic adaptation of mappings between life science ontologies.* In Proceedings of the 9^{th} International Conference on Data Integration in the Life Sciences (DILS'13). Montreal, Canada. LNCS. Springer Berlin Heidelberg. Vol. 7970, pp. 90-104.

- **DOS REIS, J. C.** 2013. *Maintaining mappings valid between dynamic KOS.* In Proceedings of the 10^{th} Extended Semantic Web Conference (ESWC'13). Montpellier, France. LNCS. Springer Berlin Heidelberg. Vol. 7882, pp. 650-655.

- **DOS REIS, J. C.;** PRUSKI, C.; DA SILVEIRA, M.; REYNAUD-DELAÎTRE, C. 2013. *Analyzing the evolution of semantic correspondences between SNOMED CT and ICD-9-CM.* In Proceedings of the International E-health, Telemedicine and Health ICT Forum for Educational, Networking and Business (Med-e-Tel'13). Luxembourg, pp. 689-693.

- **DOS REIS, J. C.;** PRUSKI, C.; DA SILVEIRA, M.; REYNAUD-DELAÎTRE, C. 2012. *Analyzing and supporting the mapping maintenance problem in biomedical knowledge organization systems.* In Proceedings of the Workshop on Semantic Interoperability in Medical Informatics (SIMI'12) collocated with ESWC'12. Crete, Greece, pp. 25-36.

National conferences and workshops

- DINH, D.; **DOS REIS, J. C.;** DA SILVEIRA, M.; PRUSKI, C.; REYNAUD-DELAÎTRE, C. 2013. *Identification des informations conceptuelles définissant un alignement entre ontologies médicales.* Actes du Symposium sur l'Ingénierie de l'Information Médicale. Au 24èmes journées francophones d'Ingénierie des Connaissances (IC'13). Lille, France.
- **DOS REIS, J. C.;** PRUSKI, C.; DA SILVEIRA, M.; REYNAUD-DELAÎTRE, C. 2012. *Vers une approche automatique pour la maintenance des mappings entre ressources terminoontologiques du domaine de la santé.* Actes du l'atelier pour l'Interopérabilté Sémantique dans les applications en e-Santé. Au IC'12. Paris, France.

Bibliography

[Abgaz et al., 2012] Y. M. Abgaz, M. Javed, and C. Pahl 2012. Analysing Impacts of Change Operations in Evolving Ontologies. In Proceedings of the 2nd Joint Workshop on Knowledge Evolution and Ontology Dynamics (EvoDyn).

[Antoniou and van Harmelen, 2004] Grigoris Antoniou and Frank van Harmelen 2004. Web Ontology Language: OWL. In Steffen Staab and Rudi Studer (eds), Handbook on Ontologies, International Handbooks on Information Systems, pages 67–92. Springer.

[Arnold and Rahm, 2013] Patrick Arnold and Erhard Rahm 2013. Semantic Enrichment of Ontology Mappings: A Linguistic-Based Approach. In Barbara Catania, Giovanna Guerrini, and Jaroslav Pokorný (eds), Advances in Databases and Information Systems, volume 8133 of Lecture Notes in Computer Science, pages 42–55. Springer Berlin Heidelberg.

[Auer and Herre, 2007] Sören Auer and Heinrich Herre 2007. A Versioning and Evolution Framework for RDF Knowledge Bases. In Proceedings of the 6th International Andrei Ershov Memorial Conference on Perspectives of Systems Informatics, PSI'06, pages 55–69. Springer-Verlag.

[Baneyx and Charlet, 2007] A. Baneyx and J. Charlet 2007. Evaluation, Evolution et maintenance d'une ontologie en medecine: Etat des lieux et experimentation. Revue Information - Interaction - Intelligence, pages 147–173.

[Berners-Lee et al., 2001] Tim Berners-Lee, James Hendler, and Ora Lassila 2001. The Semantic Web. Scientific American, 284(5):34–43.

[Bethea et al., 2006] Wayne L. Bethea, Clayton Fink, and John Beecher-Deighan 2006. JHU/APL Onto-Mapology Results for OAEI 2006. In Pavel Shvaiko, Jérôme Euzenat, Natalya Fridman Noy, Heiner Stuckenschmidt, V. Richard Benjamins, and Michael Uschold (eds), Ontology Matching, volume 225 of CEUR Workshop Proceedings.

[Blomqvist et al., 2009] Eva Blomqvist, Aldo Gangemi, and Valentina Presutti 2009. Experiments on Pattern-based Ontology Design. In Proceedings of the Fifth International Conference on Knowledge Capture, K-CAP '09, pages 41–48, New York, NY, USA. ACM.

[Bodenreider, 2004] O. Bodenreider 2004. The unified medical language system (UMLS): integrating biomedical terminology. Nucleic acids research, 32(suppl 1).

[Bodenreider, 2008] O. Bodenreider 2008. Comparing SNOMED CT and the NCI Thesaurus through Semantic Web Technologies. In Proceedings of the 3rd international conference on Knowledge Representation in Medicine (KR-MED 2008), pages 37–43.

[Bodenreider, 2009] O. Bodenreider 2009. Using SNOMED CT in combination with MedDRA for reporting signal detection and adverse drug reactions reporting. In AMIA Annual Symposium, pages 45–49.

[Bodenreider and Stevens, 2006] O. Bodenreider and R. Stevens 2006. Bio-ontologies: current trends and future directions. Briefings in bioinformatics, 7:256–274.

[Breitman et al., 2007] Karin Koogan Breitman, Marco Antonio Casanova, and Walter Ruszkowski 2007. Ontology in Computer Science. In Semantic Web: Concepts, Technologies and Applications, pages 17–34. Springer London, nasa monog edition.

[Burgun and Bodenreider, 2008] A. Burgun and O. Bodenreider 2008. Accessing and integrating data and knowledge for biomedical research. Yearbook of medical informatics, pages 91–101.

[Cardillo et al., 2008] Elena Cardillo, Claudio Eccher, Luciano Serafini, and Andrei Tamilin 2008. Logical Analysis of Mappings between Medical Classification Systems. In Danail Dochev, Marco Pistore, and Paolo Traverso (eds), Artificial Intelligence: Methodology, Systems, and Applications, volume 5253 of Lecture Notes in Computer Science, pages 311–321. Springer Berlin Heidelberg.

[Cases et al., 2013] M. Cases et al. 2013. Improving data and knowledge management to better integrate health care and research. Journal of Internal Medicine, 274(4):321–328.

[Castano et al., 2007] S. Castano, S. Espinosa, A. Ferrara, V. Karkaletsis, A. Kaya, S. Melzer, R. Moller, S. Montanelli, and G. Petasis 2007. Ontology Dynamics with Multimedia Information: The BOEMIE Evolution Methodology. In Proceedings of the International ESWC Workshop on Ontology Dynamics (IWOD 2007), Innsbruck, Austria.

[Castano et al., 2008] Silvana Castano, Alfio Ferrara, Davide Lorusso, Tobias Henrik N, and M Ralf 2008. Mapping Validation by Probabilistic Reasoning. In Proceedings of the 5th European semantic web conference on The semantic web: research and applications, pages 170–184.

[Chamberlin, 2002] D. Chamberlin 2002. XQuery: An XML query language. IBM Systems Journal, 41(4):597–615.

[Cheatham and Hitzler, 2013] Michelle Cheatham and Pascal Hitzler 2013. String Similarity Metrics for Ontology Alignment. In Harith Alani, Lalana Kagal, Achille Fokoue, Paul Groth, Chris Biemann, JosianeXavier Parreira, Lora Aroyo, Natasha Noy, Chris Welty, and Krzysztof Janowicz (eds), The Semantic Web: ISWC 2013, volume 8219 of Lecture Notes in Computer Science, pages 294–309. Springer Berlin Heidelberg.

[Chu et al., 2008] N. C. N. Chu, Q. M. Trinh, K. E. Barker, and R. S. Alhajj 2008. A Dynamic Ontology Mapping Architecture for a Grid Database System. In Fourth International Conference on Semantics, Knowledge and Grid, pages 343–346.

[Colazzo and Sartiani, 2009] Dario Colazzo and Carlo Sartiani 2009. Detection of Corrupted Schema Mappings in XML Data Integration Systems. ACM Transactions on Internet Technology (TOIT), 9(4):14:1–14:53.

[Dice, 1945] Lee Raymond Dice 1945. Measures of the Amount of Ecologic Association Between Species. Ecology, 26(3):297–302.

[Dinh et al., 2014a] D. Dinh, J. C. Dos Reis, C. Pruski, M. Da Silveira, and C. Reynaud-Delaitre 2014a. Identifying change patterns of concept attributes in ontology evolution. In Valentina et al. Presutti (ed), Proceedings of the 11th Extended Semantic Web Conference, volume 8465 of Lecture Notes in Computer Science, pages 768–783. Springer.

[Dinh et al., 2014b] D. Dinh, J. C. Dos Reis, C. Pruski, M. Da Silveira, and C. Reynaud-Delaître 2014b. Identifying relevant concept attributes to support mapping maintenance under ontology evolution. Journal of Web Semantics: Science, Services and Agents on the World Wide Web.

[Djedidi and Aufaure, 2009] Rim Djedidi and Marie-Aude Aufaure 2009. Change management patterns (CMP) for ontology evolution process. In Proceedings of the 3rd International Workshop on Ontology Dynamics (IWOD 2009) in ISWC.

[Djedidi and Aufaure, 2010] Rim Djedidi and Marie-Aude Aufaure 2010. ONTO-EVOAL an Ontology Evolution Approach Guided by Pattern Modeling and Quality Evaluation. In Proceedings of the 6th International Conference on Foundations of Information and Knowledge Systems, FoIKS'10, pages 286–305, Berlin, Heidelberg. Springer-Verlag.

[Dos Reis, 2013] J. C. Dos Reis 2013. Maintaining Mappings Valid between Dynamic KOS. In Philipp Cimiano, Oscar Corcho, Valentina Presutti, Laura Hollink, and Sebastian Rudolph (eds), The Semantic Web: Semantics and Big Data, volume 7882 of Lecture Notes in Computer Science, pages 650–655. Springer Berlin Heidelberg.

[Dos Reis et al., 2014a] J. C. Dos Reis, M. Da Silveira, D. Dinh, C. Pruski, and C. Reynaud-Delaître 2014a. Requirements for implementing mapping adaptation systems. In Proceedings of the 23rd IEEE WETICE 2014 Conference. In 6th Modeling the Collaborative Web Knowledge Track (Web2Touch 2014), pages 406–410.

[Dos Reis et al., 2013a] J. C. Dos Reis, D. Dinh, C. Pruski, M. Da Silveira, and C. Reynaud-Delaître 2013a. Mapping Adaptation Actions for the Automatic Reconciliation of Dynamic Ontologies. In Proceedings of the 22nd ACM International Conference on Conference on Information & Knowledge Management, CIKM'13, pages 599–608, New York, NY, USA. ACM.

[Dos Reis et al., 2014b] J. C. Dos Reis, D. Dinh, C. Pruski, M. Da Silveira, and C. Reynaud-Delaître 2014b. The influence of similarity between concepts in biomedical ontology evolution for mapping adaptation. In Proceedings of the 25th European Medical Informatics Conference, MIE 2014.

[Dos Reis et al., 2012] J. C. Dos Reis, C. Pruski, M. Da Silveira, and C. Reynaud-Delaitre 2012. Analyzing and Supporting the Mapping Maintenance Problem in Biomedical Knowledge Organization Systems. In Proceedings of the Workshop on Semantic Interoperability in Medical Informatics at ESWC, pages 25–36.

[Dos Reis et al., 2013b] J. C. Dos Reis, C. Pruski, M. Da Silveira, and C. Reynaud-Delaitre 2013b. Characterizing Semantic Mappings Adaptation via Biomedical KOS Evolution: A Case Study Investigating SNOMED CT and ICD. In Proceedings of the annual AMIA Symposium, pages 333–342.

[Dos Reis et al., 2014c] J. C. Dos Reis, C. Pruski, M. Da Silveira, and C. Reynaud-Delaître 2014c. Understanding Semantic Mapping Evolution by Observing Changes in Biomedical Ontologies. Journal of Biomedical Informatics, 47:71–82.

[Dos Reis et al., 2015a] J. C. Dos Reis, C. Pruski, and C. Reynaud-Delaître 2015a. Heuristiques pour l'adaptation des mappings entre ontologies dynamiques. In Proceedings of the 15th Conference on Knowledge Extraction and Management (EGC'15).

[Dos Reis et al., 2015b] J. C. Dos Reis, C. Pruski, and C. Reynaud-Delaître 2015b. State-of-the-art on mapping maintenance and challenges towards a fully automatic approach. Expert Systems with Applications, 42(3):1465 – 1478.

[Euzenat et al., 2011] J. Euzenat, A. Ferrara, W. Robert, and *et al.* 2011. Results of the ontology alignment evaluation initiative 2011. In Proc. of Ontology Matching.

[Euzenat and Shvaiko, 2007] J. Euzenat and P. Shvaiko 2007. Ontology matching. Springer.

[Fagin et al., 2011] Ronald Fagin, Phokion G. Kolaitis, Lucian Popa, and Wang-Chiew Tan 2011. Schema Mapping Evolution Through Composition and Inversion. In Zohra Bellahsene, Angela Bonifati, and Erhard Rahm (eds), Schema Matching and Mapping, Data-Centric Systems and Applications, pages 191–222. Springer Berlin Heidelberg.

[Falconer and Noy, 2011] Sean M. Falconer and Natalya F. Noy 2011. Interactive Techniques to Support Ontology Matching. In Zohra Bellahsene, Angela Bonifati, and Erhard Rahm (eds), Schema Matching and Mapping, Data-Centric Systems and Applications, pages 29–51. Springer Berlin Heidelberg.

[Flouris et al., 2008] Giorgos Flouris, Dimitris Manakanatas, Haridimos Kondylakis, Dimitris Plexousakis, and Grigoris Antoniou 2008. Ontology Change: Classification and Survey. Knowl. Eng. Rev., 23(2):117–152.

[Freitas and Schulz, 2009] Fred Freitas and Stefan Schulz 2009. Survey of current terminologies and ontologies in biology and medicine. Electronic Journal of Communication Information and Innovation in Health, 3(1):7–18.

[Fung and Xu, 2012] Kin Wah Fung and Junchuan Xu 2012. Synergism between the Mapping Projects from SNOMED CT to ICD-10 and ICD-10-CM. In Proceedings of the AMIA Annual Symposium, pages 218–227.

[Gal et al., 2005] Avigdor Gal, Ateret Anaby-Tavor, Alberto Trombetta, and Danilo Montesi 2005. A framework for modeling and evaluating automatic semantic reconciliation. The VLDB Journal, 14(1):50–67.

[Gamma et al., 1995] Erich Gamma, Richard Helm, Ralph Johnson, and John Vlissides 1995. Design Patterns: Elements of Reusable Object-oriented Software. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

[George, 1995] A. M. George 1995. WordNet: A Lexical Database for English. Communications of the ACM, 38:39–41.

[Giunchiglia et al., 2004] Fausto Giunchiglia, Pavel Shvaiko, and Mikalai Yatskevich 2004. S-Match: an algorithm and an implementation of semantic matching. In In Proceedings of the Extended Semantic Web Conference, pages 61–75.

[Gonçalves et al., 2011] Rafael S. Gonçalves, Bijan Parsia, and Ulrike Sattler 2011. Facilitating the Analysis of Ontology Differences. In Proceedings of the Joint Workshop on Knowledge Evolution and Ontology Dynamics (EvoDyn).

[Gonçalves et al., 2012] Rafael S. Gonçalves, Bijan Parsia, and Ulrike Sattler 2012. Ecco: A Hybrid Diff Tool for OWL 2 ontologies. In Pavel Klinov and Matthew Horridge (eds), Proceedings of OWL: Experiences and Directions Workshop (OWLED) 2012, Heraklion, Crete, Greece, May 27-28, 2012, volume 849.

[Gröner et al., 2010] Gerd Gröner, Fernando Silva Parreiras, and Steffen Staab 2010. Semantic Recognition of Ontology Refactoring. In Proc. of the 9th International Semantic Web Conference, ISWC'10, pages 273–288. Springer-Verlag, Berlin, Heidelberg.

[Groß et al., 2013] A. Groß, J. C. Dos Reis, M. Hartung, C. Pruski, and E. Rahm 2013. Semi-automatic Adaptation of Mappings between Life Science Ontologies. In ChristopherJ.O. Baker, Greg Butler, and Igor Jurisica (eds), Data Integration in the Life Sciences, volume 7970 of Lecture Notes in Computer Science, pages 90–104. Springer Berlin Heidelberg.

[Groß et al., 2012] A. Groß, M. Hartung, A. Thor, and E. Rahm 2012. How do computed ontology mappings evolve? - A case study for life science ontologies. In Proceedings of the Joint Workshop on Knowledge Evolution and Ontology Dynamics (EvoDyn).

[Gruber, 1993] T. R. Gruber 1993. A translation approach to portable ontology specifications. Knowl. Acquis., 5(2):199–220.

[Hartung et al., 2013] M. Hartung, A. Groß, and E. Rahm 2013. COnto-Diff – Generation of Complex Evolution Mappings for Life Science Ontologies. Journal of Biomedical Informatics, 46:15–32.

[Hartung et al., 2009] M. Hartung, T. Kirsten, A. Groß, and E. Rahm 2009. OnEX: Exploring changes in life science ontologies. BMC Bioinformatics, 10.

[Hartung et al., 2011] M. Hartung, J. Terwilliger, and E. Rahm 2011. Recent Advances in Schema and Ontology Evolution. In Zohra Bellahsene, Angela Bonifati, and Erhard Rahm (eds), Schema Matching and Mapping, Data-Centric Systems and Applications, pages 149–190. Springer Berlin Heidelberg.

[Hodge, 2000] G. Hodge 2000. The Digital Library Federation Council on Library and Information Resources, Washington, DC.

[Hu et al., 2010] Xiaohua Hu, Il-Yeol Song, and Yuan An 2010. Maintaining Mappings Between Conceptual Models and Relational Schemas. Journal of Database Management, 21(3):36–68.

[Ivanova and Lambrix, 2013] Valentina Ivanova and Patrick Lambrix 2013. A Unified Approach for Aligning Taxonomies and Debugging Taxonomies and Their Alignments. In Philipp Cimiano, Oscar Corcho, Valentina Presutti, Laura Hollink, and Sebastian Rudolph (eds), The Semantic Web: Semantics and Big Data, volume 7882 of Lecture Notes in Computer Science, pages 1–15. Springer Berlin Heidelberg.

[Javed et al., 2013] Muhammad Javed, Yalemisew M. Abgaz, and Claus Pahl 2013. Ontology Change Management and Identification of Change Patterns. Journal on Data Semantics, 2(2-3):119–143.

[Jiang and Conrath, 1997] J. J. Jiang and D. W. Conrath 1997. Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. CoRR, cmp-lg/9709008.

[Jiménez-Ruiz and Grau, 2011] Ernesto Jiménez-Ruiz and Bernardo Cuenca Grau 2011. Logmap: Logic-based and scalable ontology matching. In The Semantic Web–ISWC 2011, pages 273–288. Springer.

[Jiménez-Ruiz et al., 2011] Ernesto Jiménez-Ruiz, BernardoCuenca Grau, Ian Horrocks, and Rafael Berlanga 2011. Logic-based assessment of the compatibility of UMLS ontology sources. Journal of Biomedical Semantics, 2(1):1–16.

[Khattak et al., 2008] Asad Masood Khattak, Khalid Latif, Sharifullah Khan, and Nabeel Ahmed 2008. Managing Change History in Web Ontologies. In Proceedings of the 2008 Fourth International Conference on Semantics, Knowledge and Grid, SKG '08, pages 347–350, Washington, DC, USA. IEEE Computer Society.

[Khattak et al., 2012] A. M. Khattak, Z. Pervez, K. Latif, and S. Lee 2012. Time efficient reconciliation of mappings in dynamic web ontologies. Knowledge Based Systems, 35:369–374.

[Kim et al., 2003] Jin-Dong Kim, Tomoko Ohta, Yuka Tateisi, and Jun Ichi Tsujii 2003. GENIA corpus - a semantically annotated corpus for bio-textmining. In ISMB (Supplement of Bioinformatics), pages 180–182.

[Kitamura and Segawa, 2008] Yoshinobu Kitamura and Sho Segawa 2008. Deep Semantic Mapping between Functional Taxonomies for Interoperable Semantic Search. In J. Domingue and C. Anutariya (eds), ASWC, pages 137–151. Springer.

[Klein and Noy, 2003] Michel Klein and Natalya F. Noy 2003. A Component-Based Framework For Ontology Evolution. In Workshop on Ontologies and Distributed Systems.

[Köhler et al., 2006] Jacob Köhler, Stephan Philippi, Michael Specht, and Alexander Rüegg 2006. Ontology based text indexing and querying for the semantic web. Knoweldge Based Systems, 19(8):744–754.

[Kondylakis et al., 2009] Haridimos Kondylakis, Giorgos Flouris, and Dimitris Plexousakis 2009. Ontology and Schema Evolution in Data Integration: Review and Assessment. In Robert Meersman, Tharam Dillon, and Pilar Herrero (eds), On the Move to Meaningful Internet Systems: OTM 2009, volume 5871 of Lecture Notes in Computer Science, pages 932–947. Springer Berlin Heidelberg.

[Kremen et al., 2011] P. Kremen, M. Smid, and Z. Kouba 2011. OWLDiff: A Practical Tool for Comparison and Merge of OWL Ontologies. In 22nd International Workshop on Database and Expert Systems Applications (DEXA), pages 229–233.

[Lambrix et al., 2009] Patrick Lambrix, Lena Strömbäch, and He Tan 2009. Information Integration in Bioinformatics with Ontologies and Standards. In François Bry and Jan Maluszynski (eds), Semantic Techniques for the Web, volume 5500 of Lecture Notes in Computer Science, pages 343–376. Springer Berlin Heidelberg.

[Lambrix et al., 2008] Patrick Lambrix, He Tan, and Qiang Liu 2008. SAMBO and SAMBOdtf Results for the Ontology Alignment Evaluation Initiative 2008. In Pavel Shvaiko, Jérôme Euzenat, Fausto Giunchiglia, and Heiner Stuckenschmidt (eds), Ontology Matching, volume 431 of CEUR Workshop Proceedings.

[Leenheer and Mens, 2008] P. De Leenheer and T. Mens 2008. Ontology Evolution: State of the Art and Future Directions. In M. Hepp, P. De Leenheer, A. de Moor, and Y. Sure (eds), Ontology Management for the Semantic Web, Semantic Web Services, and Business Applications. Springer.

[Levenshtein, 1966] V. I. Levenshtein 1966. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. Soviet Physics Doklady, 10:707.

[Li et al., 2010] Aiping Li, Jiajia Miao, and Yan Jia 2010. Research on Broken Mappings Detecting Method Based on Fuzzy Aggregation Operators in Deep Web Integration Environment. In International Conference on E-Business and E-Government (ICEE), pages 125–128.

[Li et al., 2006] Mu Li, Yang Zhang, Muhua Zhu, and Ming Zhou 2006. Exploring Distributional Similarity Based Models for Query Spelling Correction. In Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics, ACL-44, pages 1025–1032, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Lindberg et al., 1993] D. A. Lindberg, B. L. Humphreys, and A. T. McCray 1993. The Unified Medical Language System. Methods Archive, 32:281–291.

[Maedche and Staab, 2002] E. Maedche and S. Staab 2002. Measuring Similarity between Ontologies. In Proceedings of the European Conference on Knowledge Acquisition and Management (EKAW), pages 251–263. Springer.

[Martins and Silva, 2009] H. Martins and N. Silva 2009. A user-driven and a semantic-based ontology mapping evolution approach. In José Cordeiro and Joaquim Filipe (eds), Proceedings of the 11th International Conference on Enterprise Information Systems, pages 214–221.

[Marzal and Vidal, 1993] A. Marzal and E. Vidal 1993. Computation of Normalized Edit Distance and Applications. IEEE Trans. Pattern Anal. Mach. Intell., 15(9):926–932.

[Mawlood-Yunis, 2008] Abdul-Rahman Mawlood-Yunis 2008. Fault-tolerant Semantic Mappings Among Heterogeneous and Distributed Local Ontologies. In Proceedings of the 2Nd International Workshop on Ontologies and Information Systems for the Semantic Web, ONISW '08, pages 31–38, New York, NY, USA. ACM.

[McCann et al., 2005] Robert McCann, Bedoor AlShebli, Quoc Le, Hoa Nguyen, Long Vu, and AnHai Doan 2005. Mapping Maintenance for Data Integration Systems. In Proceedings of the 31st International Conference on Very Large Data Bases, VLDB '05, pages 1018–1029. VLDB Endowment.

[McCray et al., 1994] A. T. McCray, S. Srinivasan, and A. C. Browne 1994. Lexical methods for managing variation in biomedical terminologies. Proceedings of the Annual Symposium on Computer Application in Medical Care., pages 235–239.

[Meilicke et al., 2007] C. Meilicke, H. Stuckenschmidt, and Andrei Tamilin 2007. Repairing Ontology Mappings. In Proceedings of the 22Nd National Conference on Artificial Intelligence - Volume 2, AAAI'07, pages 1408–1413. AAAI Press.

[Meilicke et al., 2008] C. Meilicke, H. Stuckenschmidt, and A. Tamilin 2008. Reasoning Support for Mapping Revision. Journal of Logic and Computation, 19(5).

[Miller et al., 1993] G. A. Miller, C. Leacock, T. Randee, and R. T. Bunker 1993. A semantic concordance. In Proceedings of Workshop on Human Language Technology, pages 303–308. ACL.

[Mougin et al., 2011] Fleur Mougin, Marie Dupuch, and Natalia Grabar 2011. Improving the Mapping between MedDRA and SNOMED CT. In Mor Peleg, Nada Lavrac, and Carlo Combi (eds), Artificial Intelligence in Medicine, volume 6747 of Lecture Notes in Computer Science, pages 220–224. Springer Berlin Heidelberg.

[Ngo et al., 2013] DuyHoa Ngo, Zohra Bellahsene, and Konstantin Todorov 2013. Opening the Black Box of Ontology Matching. In Philipp Cimiano, Oscar Corcho, Valentina Presutti, Laura Hollink, and Sebastian Rudolph (eds), The Semantic Web: Semantics and Big Data, volume 7882 of Lecture Notes in Computer Science, pages 16–30. Springer Berlin Heidelberg.

[Nikitina et al., 2011] Nadeschda Nikitina, Sebastian Rudolph, and Birte Glimm 2011. Reasoning-supported Interactive Revision of Knowledge Bases. In Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Two, IJCAI'11, pages 1027–1032. AAAI Press.

[Noy et al., 2009] N. F. Noy et al. 2009. BioPortal: ontologies and integrated data resources at the click of a mouse. Nucleic acids research, 37(suppl 2).

[Noy and Musen, 2002] Natalya F. Noy and Mark A. Musen 2002. Promptdiff: A Fixed-point Algorithm for Comparing Ontology Versions. In Eighteenth National Conference on Artificial Intelligence, pages 744–750, Menlo Park, CA, USA. American Association for Artificial Intelligence.

[Palma et al., 2009] Raúl Palma, Peter Haase, Óscar Corcho, and Asunción Gómez-Pérez 2009. Change Representation For OWL 2 Ontologies. In Rinke Hoekstra and Peter F. Patel-Schneider (eds), Proceedings of the 5th International Workshop on OWL: Experiences and Directions (OWLED 2009), Chantilly, VA, United States, October 23-24, 2009, volume 529.

[Pathak and Chute, 2009] J. Pathak and C. G. Chute 2009. Debugging Mappings between Biomedical Ontologies: Preliminary Results from the NCBO BioPortal Mapping Repository. In International Conference on Biomedical Ontology (ICBO). CiteSeer.

[Pesquita et al., 2009] C. Pesquita, D. Faria, A. O. Falcão, P. Lord, and F. M. Couto 2009. Semantic Similarity in Biomedical Ontologies. PLoS Computational Biology, 5(7).

[Popitsch and Haslhofer, 2011] Niko Popitsch and Bernhard Haslhofer 2011. DSNotify: A solution for event detection and link maintenance in dynamic datasets. Web Semantics: Science, Services and Agents on the World Wide Web, 9(3):266–283.

[Porter, 1997] M. F. Porter 1997. An algorithm for suffix stripping, pages 313–316. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[Qi et al., 2009] Guilin Qi, Qiu Ji, and Peter Haase 2009. A Conflict-Based Operator for Mapping Revision. In Abraham Bernstein, DavidR. Karger, Tom Heath, Lee Feigenbaum, Diana Maynard, Enrico Motta, and Krishnaprasad Thirunarayan (eds), The Semantic Web - ISWC 2009, volume 5823 of Lecture Notes in Computer Science, pages 521–536. Springer Berlin Heidelberg.

[Rahm, 2011] Erhard Rahm 2011. Towards Large-Scale Schema and Ontology Matching. In Zohra Bellahsene, Angela Bonifati, and Erhard Rahm (eds), Schema Matching and Mapping, Data-Centric Systems and Applications, pages 3–27. Springer Berlin Heidelberg.

[Rees, 2003] Reinout Van Rees 2003. Clarity in the usage of the terms ontology, taxonomy an classification. CIB REPORT, 284:432–439.

[Resnik, 1995a] P. Resnik 1995a. Using information content to evaluate semantic similarity in a taxonomy. In Proc. of the 14th International Joint Conference on Artificial Intelligence (IJCAI-95), pages 448–453. Morgan Kaufmann.

[Resnik, 1995b] Philip Resnik 1995b. 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.

[Rießet al., 2010] Christoph Rieß, Norman Heino, Sebastian Tramp, and Sören Auer 2010. EvoPat - Pattern-based Evolution and Refactoring of RDF Knowledge Bases. In Proceedings of the 9th International Semantic Web Conference, ISWC'10, pages 647–662, Berlin, Heidelberg. Springer-Verlag.

[Rosse et al., 2005] Cornelius Rosse, Anand Kumar, Jose L V Mejino, Daniel L Cook, Landon T Detwiler, and Barry Smith 2005. A strategy for improving and integrating biomedical ontologies. In Proceedings of the AMIA Annual Symposium, pages 639–43.

[Sakji et al., 2009] S. Sakji, B. Thirion, B. Dahamna, and S. J. Darmoni 2009. Recherche des sources d'information institutionnelle de santé françaises Le site Internet CISMeF. Presse Médicale, 38(10):1443–1450.

[Sellami et al., 2013] Zied Sellami, Valérie Camps, and Nathalie Aussenac-Gilles 2013. DYNAMO-MAS: a Multi-Agent System for Ontology Evolution from Text. Journal Data Semantics, 2(2-3):145–161.

[Shaban-Nejad and Haarslev, 2009] Arash Shaban-Nejad and Volker Haarslev 2009. Bio-medical Ontologies Maintenance and Change Management. In AmandeepS. Sidhu and TharamS. Dillon (eds), Biomedical Data and Applications, volume 224 of Studies in Computational Intelligence, pages 143–168. Springer Berlin Heidelberg.

[Shvaiko and Euzenat, 2013] P. Shvaiko and J. Euzenat 2013. Ontology Matching: State of the Art and Future Challenges. Knowledge and Data Engineering, IEEE Transactions on, 25(1):158–176.

[Silva and Rocha, 2005] Nuno Silva and Joao Rocha 2005. Multidimensional Service-Oriented Ontology Mapping. Int. J. Web Eng. Technol., 2(1):50–80.

[Spackman, 2005] K. A. Spackman 2005. Rates of change in a large clinical terminology: three years experience with SNOMED Clinical Terms. In Proceedings of the AMIA Annual Symposium, pages 714–718.

[Spiliopoulos et al., 2010] Vassilis Spiliopoulos, George A Vouros, and Vangelis Karkaletsis 2010. On the discovery of subsumption relations for the alignment of ontologies. Web Semantics: Science, Services and Agents on the World Wide Web, 8(1):69–88.

[Stojanovic, 2004] Ljiljana Stojanovic 2004. Methods and tools for ontology evolution. PhD thesis, Karlsruhe Institute of Technology.

[Stojanovic et al., 2002] Ljiljana Stojanovic, Alexander Maedche, Boris Motik, and Nenad Stojanovic 2002. User-Driven Ontology Evolution Management. In Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web, EKAW '02, pages 285–300, London, UK, UK. Springer-Verlag.

[Tang and Tang, 2010] Funian Tang and Rongnian Tang 2010. Minimizing influence of ontology evolution in ontology-based data access system. In IEEE International Conference on Progress in Informatics and Computing (PIC), volume 1, pages 10–14.

[Tang et al., 2009] Funian Tang, Li Yao, Yang Sun, and Meng Qian 2009. Visualizing Semantic Mapping Based on View Graph. In Second International Symposium on Knowledge Acquisition and Modeling, volume 3, pages 124–127.

[Tran et al., 2011] Quang-Vinh Tran, Ryutaro Ichise, and Bao-Quoc Ho 2011. Cluster-based similarity aggregation for ontology matching. In Pavel Shvaiko, Jérôme Euzenat, Tom Heath, Christoph Quix, Ming Mao, and Isabel F. Cruz (eds), Ontology Matching, volume 814 of CEUR Workshop Proceedings. CEUR-WS.org.

[Vandenbussche and Charlet, 2009] Pierre-Yves Vandenbussche and Jean Charlet 2009. Métamodèle général de description de ressources terminologiques et ontologiques. In Journées francophones d'Ingénierie des Connaissances (IC'09), pages 193–204.

[Velegrakis et al., 2004a] Yannis Velegrakis, Renée J. Miller, and Lucian Popa 2004a. Preserving mapping consistency under schema changes. The VLDB Journal, 13(3):274–293.

[Velegrakis et al., 2004b] Y. Velegrakis, R. J. Miller, L. Popa, and J. Mylopoulos 2004b. ToMAS: a system for adapting mappings while schemas evolve. In Proceedings of 20th International Conference on Data Engineering, page 862.

[Wang and Xu, 2008] P. Wang and B. Xu 2008. Debugging Ontology Mappings: A Static Approach. Computing and Informatics, 27(1):21–36.

[Wennerberg, 2009] Pinar Wennerberg 2009. Aligning Medical Domain Ontologies for Clinical Query Extraction. In Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop, EACL'09, pages 79–87, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Xue and Kedad, 2007] X. Xue and Z. Kedad 2007. Creating and Maintaining Mappings for XML Data. In 7emes Journées Francophones Extraction et Gestion des connaissances.

[Yu and Popa, 2005] C. Yu and L. Popa 2005. Semantic adaptation of schema mappings when schemas evolve. In Proceedings of the 31st international conference on Very large data bases, VLDB '05, pages 1006–1017. VLDB Endowment.

[Zurawski et al., 2008] Maciej Zurawski, Alan Smaill, and Dave Robertson 2008. Bounded Ontological Consistency for Scalable Dynamic Knowledge Infrastructures. In John Domingue and Chutiporn Anutariya (eds), The Semantic Web, volume 5367 of Lecture Notes in Computer Science, pages 212–226. Springer Berlin Heidelberg.