Understanding and Managing Medical Data and Knowledge Dynamics

Cédric Pruski

Luxembourg Institute of Science and Technology | Université Paris-Sud

Understanding and Managing Medical Data and Knowledge Dynamics

Habilitation à diriger des recherches

Defended the 2nd of October 2015

Université Paris-Sud

Department of Computer science

by

Cédric Pruski

Jury

Rapporteurs :	Dr. Marie-Christine Jaulent	INSERM, France
	Prof. Leon van der Torre	University of Luxembourg, Luxembourg
	Dr. Amedeo Napoli	CNRS, France
Examinateurs :	Prof. Stefan Schulz	Medical University of Graz, Austria
	Prof. Aldo Gangemi	University Paris 13, France
	Dr. Pierre Zweigenbaum	CNRS, France
Marraine :	Prof. Chantal Reynaud-Delaître	University Paris-Sud, France

Mis en page avec la classe thloria.

Résumé

Ce manuscrit synthétise une partie de mes activités de recherche réalisées dans le domaine de l'informatique médicale au sein du département CR SANTEC du Centre de Recherche Public Henri Tudor de Luxembourg[1] depuis ma thèse de doctorat en vue de l'obtention de l'habilitation à diriger des recherches. Ces travaux s'inscrivent dans la thématique plus générale de la gestion de la connaissance dynamique et visent particulièment la représentation, l'évolution et la validation des connaissances en santé.

1 Contexte de travail

Depuis plusieurs décennies, le domaine de l'Intelligence Artificielle s'intéresse à définir la notion de connaissance afin de l'exploiter à des fins diverses dans plusieurs cadres d'application tels que la recherche d'information [Pruski and Wisniewski, 2012], l'aide à la décision [Pruski et al., 2011a] ou encore l'interopérabilité sémantique [Pruski et al., 2010]. Ceci est en partie réalisé grâce à l'utilisation de modèles de représentation des connaissances tels que les ontologies permettant la spécification d'une conceptualisation [Gruber et al., 1993]. Cependant, les aspects liés à l'évolution des connaissances et des modèles qui leur sont associés restent largement inexplorés et demeurent des problèmes de recherche ouverts.

Le domaine biomédical est un domaine très riche dans la mesure où les connaissances qu'il intègre sont complexes et en perpétuelle évolution [Baneyx and Charlet, 2006] comme le démontre le nombre sans cesse croissant de communications scientifiques publiées au quotidien. C'est pour ces raisons qu'il a suscité mon intérêt et ses spécificités ont constitué la ligne directrice de mes activités de recherche.

Trois grandes thématiques ont focalisé mes efforts au cours de ces dernières années et ont concentré la majeure partie de mes contributions scientifiques et collaborations dans ce domaine particulier.

- La **représentation des connaissances** en santé. Le monde de la santé étant très ancien, la quantité de connaissance accumulée au cours du temps requiert un ensemble de méthodes et d'outils permettant de les représenter de manière à répondre aux besoins du domaine. Parmi eux, on retrouve notamment l'aide à la décision et la recherche d'informations pertinentes afin que les professionnels de santé puissent correctement diagnostiquer leurs patients et prescrire des traitements adaptés.

- La **gestion de l'évolution** des modèles de représentation des connaissances biomédicales. La dynamique du domaine en termes de création et révision de la connaissance affecte très

[1] A partir du 1^{er} janvier 2015, les centres de recherche publics Henri Tudor et Gabriel Lippmann seront regroupés sous le Luxembourg Institute of Science and Technology (LIST).

largement les modèles de représentation des connaissances, les objets qui leur sont associés et, par conséquent, les sytèmes et les décisions d'ordre médical sous-jacents. Des méthodes pour une gestion adéquate de l'impact cette dynamique sont donc nécessaires.

- La **validation** des modèles de représentation des connaissances biomédicales. Les aspects critiques du domaine, notamment dus à l'implication de patients, poussent à l'utilisation de connaissances et de données respectant un certain niveau de qualité. De plus, les modèles tels que les ontologies sont de plus en plus souvent construits de manière automatique et de ce fait doivent faire l'objet d'une validation rigoureuse de la part des experts du domaine. La conception de techniques engageant les experts pour leur faciliter le travail de validation d'ontologies est importante.

L'ensemble des travaux présentés dans ce manuscrit a été réalisé dans le cadre de collaborations avec des collègues chercheurs ainsi que des doctorants et post-doctorants.

2 Spécificités des connaissances en santé

La médecine est un domaine qui voit la quantité de données et de connaissances qui lui sont propres sans cesse croître. Ce phénomène a contribué à l'émergence d'un nouveau paradigme plaçant le patient au centre des préoccupations. Cette nouvelle approche du domaine a profité aux domaines connexes qui ont dû s'adapter pour répondre à de nouveaux besoins. Ce fut le cas de l'informatique avec l'apparition de domaines de recherche tels que la bioinformatique, l'informatique médicale ou encore l'intelligence artificielle appliquée à la médecine dont l'objectif vise à proposer de nouvelles méthodes et de nouveaux outils tenant compte des spécificités du domaine médical.

La santé est un vaste domaine. Contrairement à bien des domaines qui se sont développés avec le Web, la médecine est un domaine très ancien, bien antérieur à l'informatique et a toujours succité un grand intérêt. Ceci est la cause de la quantité de données et de connaissances représentatives du domaine. Suivant cette mouvance, un grand nombre de Systèmes d'Organisation de la Connaissance (SOC), spécifiques à chaque branche de la médecine, a été progressivement développé suivant des modèles de représentation des connaissances divers. Nous pouvons définir un SOC comme étant un ensemble de connaissances en interaction, représentées et regroupées au sein d'une structure dans le but de répondre à des besoins et d'atteindre des objectifs déterminés. Les SOC sont conçus avec une intention afin de répondre à un usage [Vandenbussche, 2011]. On retrouve alors, des ontologies comme Gene Ontology (GO), des systèmes de classification tels que la Classification Internationale des Maladies (CIM) ou encore des thésaurus comme NCI thesaurus. De plus, afin d'optimiser la couverture globale du domaine lors de leur utilisation dans des systèmes d'information, les éléments de ces SOC ont été mis en correspondance au travers d'alignements sémantiques spécifiant la relation entre ces éléments. Cependant, malgré le nombre de SOC caractéristiques du domaine, celles-ci restent bien plus volumineuses que les SOC des autres domaines et en particulier les ontologies du Web Sémantique posant d'autres types de problèmes. Par ailleurs, le développement du Web au début des années 90 a poussé les systèmes d'information médicale à dépasser le cadre hospitalier pour un contexte plus vaste. Ainsi, les données et les connaissances du domaine se retrouvent distribuées et souvent exprimées dans des formats différents du fait de l'absence de standards ce qui se révèle problématique.

La santé est un domaine très dynamique. Une autre particularité du domaine réside dans l'aspect dynamique des connaissances qui le composent comme le souligne le nombre sans cesse croissant des publications scientifiques recensées par les portails d'accès usuels. Cette évolution rapide des connaissances médicales se reflète à travers les nombreuses versions successives des SOC publiées régulièrement où de nouveaux éléments sont ajoutés, certains, obsolètes, supprimés ou leur description modifiée. A titre d'exemple, en moyenne 10% des éléments de la SNOMED CT subissent des modifications lors du passage d'une version de la nomenclature à la suivante. Ces changements, à l'intérieur même des SOCs, a bien évidemment un impact à la fois sur les objets qui en dépendent comme les alignements sémantiques ou les annotations, mais aussi sur les applications logicielles qui les exploitent. La dynamique du domaine est également induite par les phénomènes naturels liés à la santé des patients. Les effets des changements sur l'environnement (au sens large) des patients peuvent engendrer de nouvelles pathologies nécessitant des traitements adaptés et un suivi rigoureux des patients. L'outil informatique a montré des capacités intéressantes pour répondre à ces besoins mais doit néanmoins être adapté aux spécificités du domaine.

La santé est un domaine critique. Contrairement à un grand nombre de domaines en vogue, celui de la santé met en jeu la vie de ses acteurs. En vertu de cette caractéristique, les données et connaissances mises en œuvre dans la prise de décision doivent répondre à un niveau de qualité et à des critères de validation exigeants afin que les professionnels de la santé puissent prendre les bonnes décisions concernant leurs patients surtout dans un cadre dynamique tel que la médecine. Or, la tâche de validation des nouvelles données ou connaissances acquises, comme par exemple la validation d'une nouvelle version d'un SOC, nécessite l'implication d'experts du domaine. Cependant, ces experts n'ont, en règle générale, pas les connaissances suffisantes en terme de représentation logique des connaissances pour décider si la nouvelle connaissance est correctement représentée du point de vue logique et, par conséquent de sa justesse conceptuelle. De ce fait, un besoin évident d'outils et de techniques facilitant l'intervention des experts se fait ressentir dans un but d'optimisation des données et connaissances du domaine et, par voie de fait, des systèmes d'information sous-jacents.

Mon projet de recherche s'est articulé autour du domaine de la santé et de ses spécificités. Les trois axes de recherche et activités d'encadrement décrits ci-dessous s'inscrivent davantage dans un cadre de recherche appliquée et ont tous été motivés par les demandes réelles et besoins des différents acteurs du domaine en termes de gestion des connaissances dynamiques et d'aide à la décision.

3 De la représentation des connaissances dynamiques

Cette thématique a été abordée dans le cadre de la modélisation du contenu des guides de bonnes pratiques médicales. Ces derniers fournissent aux professionnels de santé un ensemble d'instructions spécifiques, méthodologiquement développées afin d'assister le praticien et le patient dans la décision d'un soin approprié, selon des circonstances cliniques spécifiques [Field and Lohr, 1990].

Notre analyse de la littérature a montré que les langages existants pour la représentation de ces guides privilégiaient avant tout leur exécution au détriment du support pour le raisonement automatique [Zamborlini et al., 2015]. Ceci est en partie dû au fait que les actions médicales sont toujours exprimées sous forme de texte libre réduisant ainsi considérablement l'efficacité

des ordinateurs dans l'interprétation et l'exploitation de ces informations. Ce manque identifié est clairement ressenti par les acteurs du domaine surtout en cas de comorbidité (i.e. patient souffrant de plusieurs pathologies). En effet, au moment de la définition du traitement, il est nécessaire de fusionner plusieurs ensembles de recommandations pour aboutir au traitement désiré. Or, au cours de ce processus, des interactions potentielles induites par les recommandations propres à chaque maladie sont susceptibles de surgir d'où la nécessité de les identifier et de les solutionner. En conséquence, dans un projet de thèse de doctorat, nous avons décidé de travailler sur ce point en proposant un modèle ontologique pour représenter les recommandations composant un guide de bonne pratique [Zamborlini et al., 2014b]. Ce modèle et ses évolutions ont fait suite à un ensemble de réflexions menées sur l'utilisation des ontologies pour la représentation des actions médicales [Bonacin et al., 2013]. Ce travail tire son originalité de plusieurs points:

- Il permet de représenter avec un niveau de granularité plus fin que les modèles existants l'information médicale contenue dans les guides de bonnes pratiques nécessaires à l'élaboration de plans de traitement,

- Son formalisme s'appuyant sur la logique du premier ordre permet aux ordinateurs de raisonner dessus afin d'identifier les interactions diverses et variées pouvant apparaître lors de la fusion, l'adaptation ou la mise à jour des guides de bonnes pratiques,

- Il ne représente pas un nouveau modèle de représentation des connaissances médicales dans le sens où il ne traite pas des mêmes éléments et donc peut être utilisé en complément des langages usuels,

- Son implémentation avec les technologies du Web Sémantique permet d'exploiter des bases de données RDF ouvertes, augmentant les possibilités pour identifier les interactions.

Dans un autre travail de thèse de doctorat, nous nous sommes penché sur le problème de l'adaptation des traitements en cours remis en question par des changements dans l'environnement du patient. Ce problème touche surtout les patients souffrant de maladies chroniques comme le diabète ou l'épilepsie. Dans ce contexte, les défis résident surtout dans l'identification des paramètres importants du traitement, le suivi de leur évolution et la décision, suivant les informations engrangées, d'adapter les composantes du traitement, comme par exemple la dose d'un médicament. Ces travaux en cours visent à la définition et l'implémentation d'une plate forme télématique respectant le paradigme de l'*Autonomic Computing*. Cet outil permettra :

- La spécification des modèles de traitement et des paramètres importants ayant une influence sur le bon déroulement du traitement,

- Le suivi de l'évolution de ces paramètres,

- La prise de décision quant à la nécessité d'adapter le traitement,

- La mise en œuvre de cette adaptation.

Les travaux engagés sur cette thématique sont complémentaires dans la mesure où ils ont tous pour objectif l'adaptation des traitements médicaux, mais à différents moments du temps : (i) celui de leur définition et (ii) lors de leur exécution ou leur suivi par les patients.

4 De la gestion des connaissances dynamiques

La motivation principale de ces travaux réside dans l'observation et l'utilisation pratique combinée des SOC du domaine et de l'importance des correspondances sémantiques qui les relient notamment pour le partage et la recherche d'information médicale. Dans ce contexte, l'évolution des SOC peut invalider les alignements existants et fausser les résultats des tâches les exploitant si ces derniers ne suivent pas l'évolution des concepts qui les définissent. Une approche naïve du problème aurait consisté en la suppression des correspondances erronées puis le réalignement des SOC ayant évolué. Cependant, la taille des SOC du domaine biomédical pouvant atteindre plusieurs millions d'entités demande un temps de calcul important et un effort de validation de la part des experts du domaine insupportable. Ceci justifie pleinement le besoin de méthodes et d'outils intelligents pour la maintenance des correspondances sémantiques au cours du temps.

Pour répondre à ces besoins, le projet DynaMO, finançant une thèse de doctorat et un projet postdoctoral, a été défini. Les travaux que nous avons menés dans ce cadre ont conduit à l'élaboration du framework DyKOSMap. Notre approche pour la maintenance des alignements sémantiques s'appuie sur une étude approfondie des évolutions des SOC du domaine et des correspondances qui leur sont associées dans un souci d'applicabilité. Les travaux tirent leur originalité des aspects suivants :

- Le processus de maintenance des alignements tient compte des évolutions des éléments ontologiques modifiés ainsi que des informations contextuelles influant les alignements préalablement établis,
- La proposition d'un ensemble de patrons de changement permettant de caractériser les évolutions des concepts avec un niveau de granularité plus fin (au niveau des attributs de concept) que les travaux de l'état de l'art ne permettaient de le faire. Cette proposition répond à une analyse empirique des évolutions conjointes des SOC du domaine médical et des correspondances sémantiques leur étant associées,
- Une caractérisation formelle des comportements au cours du temps des alignements ayant lieu en pratique après une analyse approfondie d'un grand nombre de correspondances sémantiques du domaine,
- Le lien entre l'évolution des SOC et le comportement des alignements a été formalisé sous forme d'heuristiques décrivant les conditions à remplir dans les différents scénarios d'adaptation des alignements.

Ainsi l'outil développé pour supporter le framework DyKOSMap prend comme arguments la nouvelle et l'ancienne version d'un SOC source, la version courante d'un SOC cible et l'ensemble des alignements sémantiques entre la source et la cible à mettre à jour et retourne les alignements adaptés.

5 De la validation des connaissances dynamiques

Les aspects critiques du domaine médical requièrent l'utilisation de données et de connaissances respectant un niveau de qualité et des critères de validation exigeants. Ceci est particulièrement vrai pour l'ensemble des ontologies utilisées dans la prise de décision et ayant fait l'objet d'une construction automatique. Cependant, l'implication des experts du domaine lors de leur validation est problématique pour plusieurs raisons :

1. Le volume des connaissances à valider nécessite un investissement en temps non négligeable pour les experts dont la principale activité reste le traitement des patients,

2. Les compétences des experts en termes de représentation logique des connaissances sont très variables, c'est pourquoi ces derniers sont souvent accompagnés par des ingénieurs leur traduisant en langue naturelle les axiomes logiques représentant le contenu d'une ontologie.

Suivant ces observations, nous nous sommes proposés d'attaquer deux problèmes distincts dans le cadre de deux projets postdoctoraux.

1. celui de la qualité du contenu des ontologies et des alignements sémantiques qui leur sont associés,

2. celui de la validation du contenu d'une ontologie avec en point de mire les aspects relatifs à la conceptualisation du domaine.

Le premier de ces problèmes a été abordé suite aux observations effectuées dans le projet DynaMO sur l'évolution des alignements sémantiques. Ces analyses ont montré qu'une quantité non négligeable d'éléments ontologiques, ou de correspondances sémantiques, est supprimée au cours du temps simplement à cause d'erreurs de représentation, de conceptualisation ou d'alignement [Dos Reis et al., 2014c]. Nous nous sommes alors proposés d'évaluer la quantité d'erreurs de ce type en confrontant le contenu des ontologies existantes ainsi que des correspondances sémantiques qui leur sont associées. L'approche développée met en œuvre un réseau d'ontologies considérées comme source de connaissance externe et un mécanisme d'inférence afin de comparer la description des concepts de plusieurs ontologies et leur sémantique.

Nos travaux autour du second problème nous ont conduits à définir une méthode de validation du contenu d'une ontologie à travers un système de questions/réponses. L'idée était de représenter le contenu d'une ontologie (e.g. les concepts et leurs relations) sous forme de questions, exprimées en langue naturelle, afin de le rendre compréhensible à des non-spécialistes du langage OWL justement en faisant abstraction de la syntaxe du langage logique. Dans un deuxième temps, les questions ainsi générées sont soumises à un expert du domaine afin qu'il puisse y répondre. Suivant les réponses émises, le système est en mesure de valider les affirmations ou de les invalider et, le cas échéant, de corriger les erreurs en interprétant les réponses des experts. Une réflexion spécifique sur l'ordre des questions à soumettre à l'expert a été menée en tenant compte de l'impact que peut avoir une invalidation sur le contenu de l'ontologie restant à valider.

La complémentarité de ces travaux a permis de traiter de la qualité des ontologies, ou plus généralement, des SOC du domaine biomédical avec un accent mis sur leur validation.

6 Perspectives

Mes travaux dans différents domaines de l'informatique médicale ont ouvert de nouvelles perspectives qui vont motiver mes futures activités de recherche. Concernant la représentation de la connaissance dans les guides de bonnes pratiques cliniques, plusieurs axes vont être suivis. Tout d'abord, afin de valoriser les résultats obtenus, nous allons nous concentrer sur la constitution d'une base de données d'actions médicales offrant une représentation formelle de ces actions. Cet outil sera mis à la disposition des utilisateurs des guides de bonnes pratiques cliniques pour faciliter leur intégration dans les systèmes hospitaliers et optimiser, à plus long terme, le suivi des patients et les soins qui leur sont apportés.

Un autre aspect qui préoccupe la communauté scientifique concerne la mise à jour des guides de bonnes pratiques cliniques. Le formalisme TMR que nous avons proposé sera adapté afin de faciliter l'évolution de ces guides de bonnes pratiques et surtout leur représentation sous format électronique. Présentement, le contenu des guides est revu en moyenne tous les cinq ans, en vertu des résultats obtenus après des études cliniques. Cependant, la multiplication des études cliniques produisant des résultats exploitables fournit des informations suffisantes justifiant une mise à jour régulière des guides de bonnes pratiques cliniques afin d'optimiser leur mise en œuvre dans les systèmes hospitaliers existants. Cela nécessitera une extension du modèle TMR pour tenir exprimer des informations spatio-temporelles, une méthode pour lier les guides de bonnes pratiques et les essais cliniques qui leur sont associés ainsi que des techniques s'appuyant sur des aspects de traitement automatique de la langue naturelle pour identifier et décider si des changements ou des nouveaux résultats cliniques sont pertinents ou non.

La gestion de la dynamique des connaissances affecte également d'autres éléments reposant sur les ressources termino-ontologiques. C'est notamment le cas des annotations sémantiques qui, associées aux données médicales, renseignent davantage sur leur sémantique pour faciliter la recherche ou le partage d'information et permettent d'accroître les capacités de raisonnement des machines sur ces données. Or, comme pour les correspondances sémantiques, la cohérence des annotations peut être remise en cause par les évolutions successives du SOC dont elles sont extraites. Ainsi, la version du SOC qui a servi à annoter les données peut être différente de sa version la plus à jour qui peut ne plus contenir les concepts (ou leur label) utilisés pour annoter les données rendant ainsi leur exploitation difficile. Dans ce contexte, le projet ELISA, successeur de DynaMO, va apporter des éléments de réponse à ce problème. L'idée est de concevoir une approche formelle à partir d'observations empiriques du comportement des annotations sémantiques au cours du temps pour bien les faire évoluer et préserver leur utilité dans des contextes critiques comme la recherche d'information dans le dossier patient ou la gestion de l'information clinique acquise au cours d'essais cliniques.

L'utilisation de plus en plus massive de données liées dans le domaine biomédical comme l'attestent les nombreuses bases de données ouvertes mises à la disposition des utilisateurs voit également se poser des problèmes concernant leur gestion. L'aspect dynamique du Web, renforcé par l'émergence du *Big Data*, peut engendrer des modifications de contenu remettant en question la validité des connexions entre les données et la nécessité de maintenir ces liens pour des raisons de cohérence au moment de leur exploitation. Ce phénomène se manifeste, par exemple, dans les bases de données liées concernant les médicaments, souvent révisées en vertu de l'élaboration de nouveaux traitements et la découverte de nouvelles interaction entre les composants des thérapies.

Les aspects concernant la qualité des ontologies et principalement leur validation seront améliorés à travers l'automatisation du processus de validation. Dans cette optique, nous supposons que le Web et son contenu disposent des informations nécessaires pour répondre aux questions générées à partir du contenu de l'ontologie à valider. Partant de cette hypothèse, il nous faudra alors adapter notre technique de génération de questions pour obtenir non pas des questions exprimées en langue naturelle mais des requêtes de différentes natures. D'un côté nous devrons être en mesure de produire des requêtes compréhensibles par les moteurs de recherche usuels et, d'un autre coté, des requêtes de type SPARQL afin de les vérifier sur les bases de données liées de type RDF. Ensuite, nous allons également travailler sur les aspects découlant de l'interprétation des réponses retournées. Ce dernier point va dépendre de la nature des éléments interrogés à savoir du texte libre contenu sur des pages Web classiques ou des triplets

RDF composant la nouvelle génération du contenu du Web afin de valider ou de faire évoluer le contenu d'une ontologie.

Table of content

Résumé **i**

Introduction **1**

1 Context of work 1

2 Research work 2

2.1 Dynamic medical knowledge representation and exploitation 2

2.2 Dynamic medical knowledge management 3

2.3 Validation of dynamic medical ontologies 3

Chapter 1
On the specificities of medical data and knowledge

1.1 Medicine: An old and vast domain 6

1.1.1 Representation of medical knowledge 6

1.1.2 Distribution of heterogeneous data 7

1.2 Medicine: A highly dynamic domain 7

1.2.1 Impact of evolution on data and knowledge management 8

1.2.2 Impact of evolution on patient and health professionals 9

1.3 Medicine: A critical domain 10

1.3.1 Medical data and knowledge quality 10

1.3.2 Metadata quality 10

1.4 Research project definition 11

Chapter 2
On the representation of dynamic medical knowledge: the case of adaptive treatment plans

2.1 Problematic and research questions 14

2.2 Related work 15

2.2.1 Computer Interpretable Guidelines representation 16

2.2.2 Treatment plan adaptation 17

2.3 First reflections on the formalization of care actions and personalization of treatment plans: the iCareflow approach 19
2.3.1 From clinical guidelines to careflow 19
2.3.2 On the use of ontologies to represent care actions 21
2.3.3 Personalizing treatment plans in the iCareflow framework 22
2.4 Enhancing the reasoning capabilities of existing CIG languages: the METIS approach 23
2.4.1 The TMR Model 24
2.4.2 An application to multi-morbidity 25
2.5 Personalizing treatments plans: the SACCOM approach 27
2.5.1 Identifying and monitoring dynamic treatment parameters 27
2.5.2 The dynamic adaptation of treatment plans 28
2.6 Summary 29

Chapter 3
On the management of dynamic medical knowledge: the case of Knowledge Organizing Systems mapping adaptation

3.1 Problem statement and hypothesis 32
3.2 Related work 33
3.2.1 Mapping revision 33
3.2.2 Mapping calculation 34
3.2.3 Mapping adaptation 34
3.2.4 Mapping representation 35
3.3 Understanding ontology evolution for adapting mapping 35
3.3.1 Empirical analysis of mappings and KOS evolution 35
3.3.2 The role of concept definition in mapping adaptation 36
3.3.3 Identification of relevant dynamic knowledge to adapt mapping 37
3.4 Adapting mappings according to ontology evolution 39
3.4.1 Change patterns for mapping adaptation 39
3.4.2 Definition of mapping adaptation actions 41
3.4.3 Heuristic-based approach to maintain mappings valid over time 42
3.5 The DyKOSMap framework 43
3.6 Experimental assessment 45
3.7 Summary 46

Chapter 4
On the quality of dynamic medical knowledge: the cases of biomedical Knowledge Organization Systems and mappings

4.1 Problem statement and hypothesis 50
4.2 Related work 51
4.2.1 Quality of Knowledge Organization Systems 51
4.2.2 Methods and tools for ontology validation 53
4.3 Analysing the quality of Knowledge Organization Systems and semantic mappings 54
4.3.1 Method to compare Knowledge Organizing Systems content with existing mappings 55
4.3.2 Impact on ontology mappings 58
4.4 On the validation of medical ontologies 59
4.4.1 Verbalizing the content of medical ontology 60
4.4.2 Interpreting experts feedback and modifying the ontology 63
4.4.3 Experimental assessment 64
4.5 Summary 65

Chapter 5
Open research challenges

5.1 Medical information management and retrieval 67
5.1.1 Semantic annotation management 67
5.1.2 Medical information retrieval 68
5.2 Knowledge dynamics in recent new paradigms 68
5.2.1 Big data 69
5.2.2 Linked Data and the Web 69
5.3 Patient empowerment 70

Conclusion **71**

References **73**

List of Figures

2.1 From clinical guidelines to careflow 20
2.2 Example of pattern ontology: The case of substance administration 22
2.3 The MedAForm approach 23
2.4 UML class diagram for the TMR Model 24
2.5 Duodenum ulcer CG 25
2.6 TMR representation of Duodenum ulcer CG 25
2.7 UML class diagram for the TMR4I Model 26

3.1 The mapping maintenance problem 33
3.2 Impact of concept splitting on associated mappings 37
3.3 The DyKOSMap Framework 44
3.4 Results of the evaluation of the DyKOSMap approach 46

4.1 Illustrating example 58
4.2 The COVALMO general approach for ontology validation 59
4.3 The generation of questions 61

List of Tables

4.1 Examples of boolean-question patterns 61
4.2 Examples of ontology update rules w.r.t. invalidated elements 62
4.3 The number of Ontology Elements (OE) and the number of generated questions for different medical ontologies without optimization 65

Introduction

> *"... knowledge must continually be renewed by ceaseless effort, if it is not to be lost."* (Albert Einstein, On Education, 1950)

Knowledge is a notion that has raised the interest of people for centuries. From ancient times, philosophers and scientists have appropriated this complex notion to discuss the various aspects that define it. As underlined by the above quotation, *evolution* is one of these key facets. Depending on the context, Knowledge evolution has several meanings. In this document, it refers to the revision (e.g. addition, correction and deletion) of knowledge when new findings come up.

The understanding and management of knowledge evolution has been the main focus of my research work over the past years.

1 Context of work

In computer science, the understanding and management of the dynamics of knowledge is still an open issue. This is the case for many domains where ICT systems play a key role. *Knowledge engineering* (KE) refers to all technical, scientific and social aspects involved in building, maintaining and using knowledge-based systems. It aims at providing intelligent computer systems, models like ontologies and algorithms able to handle the huge amount of digital data available, turning it into knowledge and maintaining it over time, to make life easier. These models have shown great capabilities to represent domain knowledge [Guelfi and Pruski, 2006], and to support information retrieval [Pruski et al., 2011b, Pruski and Wisniewski, 2012, Guelfi et al., 2007a]

Biomedicine has raised my interest since it is one domain for which the amount of knowledge gathered over the centuries is substantial, and the scientific effort invested to continuously develop it induces a perpetual evolution of its knowledge. As mentioned by Baneyx and Charlet, 50% of medical knowledge is renewed every 10 years[Baneyx and Charlet, 2006]. This clearly underlines the urgent needs for methods and tools to smoothly and faithfully manage this ever ongoing change to make the most of it, and to translate it into optimal care for patients.

However, knowledge evolution in biomedicine is a vast domain ranging from medical ontology evolution [Smith et al., 2007] to text mining and natural language processing techniques applied to decision support systems [Demner-Fushman et al., 2009]. This is one reason why my focus has been narrowed down to three aspects which have motivated my research project. Moreover, the research environment provided me a direct access to concrete case studies to which my contributions have been applied, which has reinforced this choice.

2 Research work

In direct line with the work carried out during my doctoral thesis [Pruski, 2009], I have focused on issues related to the general problem of knowledge dynamics management [Guelfi et al., 2007b] with a particular focus on the biomedical and health domains. This has been done following three research axe s:

1. the **representation** of medical knowledge,
2. the **evolution** of medical knowledge,
3. the **evaluation and validation** of medical knowledge quality.

2.1 Dynamic medical knowledge representation and exploitation

Chapter 2 presents part of my contributions in the field of medical knowledge representation, with a particular emphasis on the formalisation of medical recommendations and care actions within clinical guidelines (CG) and the dynamic adaptation of treatment plans based on CG knowledge. CG assemble statements provided by the best available evidences. Their goal is to assist healthcare professionals with the definition of the appropriate treatment and care for people with specific diseases and conditions. Paper-based versions of CGs are progressively replaced by Computer Interpretable Guideline (CIG) expressed in dedicated languages like PROForma [Sutton and Fox, 2003], GLIF [Patel et al., 1998] or Asbru [Miksch et al., 1997] to exploit ICT systems' capabilities, to overcome some limitations of paper based CGs like the instantiation of guidelines with patient data. However, existing CIG languages are defined to design careflows to be executed by computers but they prevent machine to reason over them [Peleg, 2013]. To this end, the merger of guidelines in case of comorbidity, the automatic update of guidelines (taking into account new findings from clinical studies/trials) or the automatic personalisation of treatment plans (taking into account patients' preferences) still require a significant intervention of human experts to detect potential conflicts. This results from the misuse of CIG languages to represent guideline components. For instance, we often observe free text is used to describe care actions, preventing ICT systems to correctly interpret this information. To overcome this lack, we have investigated the use of ontologies to represent care actions [Pruski et al., 2011a] and, based on this first attempt, we have proposed the Transition-based Medical Recommendation (TMR) Model for Clinical Guidelines, focussing on recommendations, to improve reasoning capabilities of CIGs, and we have applied it to the comorbidity use case [Zamborlini et al., 2014a].

Furthermore, as CG serve as foundations for defining treatment plans that patients have to follow over time, the unpredictable changes in the environment as well as local constraints (e.g. resource availability of a hospital) may also impact patients' health or can imply on changes in the treatment. Health professionals facing new situations have to react and adapt in response to this environmental evolution and modify the treatment plans accordingly. Based on the review of existing sets of guidelines, we have actually observed that local constraints are not part of guidelines and, as a consequence, must be integrated at treatment definition time or even at execution time. In this context, we have proposed the Careflow Personalization System (CPS) for the dynamic integration of formally expressed local constraints and their application when personalising treatment plans [Bonacin et al., 2012]. Finally, as a last level of personalisation, we are currently designing an approach based on Autonomic Computing (SACCOM approach) able (i) to monitor key parameters of the patients' treatment, (ii) detect anomalies in the variation of these parameters and (iii) adapt the treatment based on the patient specificities and medical

knowledge specifying the adaptation [Mezghani et al., 2014]. One of the originalities of the SACCOM approach consists in finding, on the fly at treatment personalisation time, medical knowledge from various reliable sources of information to modify the careflow.

2.2 Dynamic medical knowledge management

Chapter 3 addresses the management of the knowledge evolution impact on artefacts such as ontology mappings, main focus of the DynaMO[2] research project, or on semantic annotations. The long history of knowledge representation in medicine forced us to consider Knowledge Organisation Systems (KOS) [Hodge, 2000] and not only ontologies. KOS encompass all types of schemes: classifications and categorizations, taxonomies, thesauri, as well as semantic networks and ontologies. These schemes, defined using different knowledge representation models, are widely used in the biomedical field for various purposes. This is for instance the case for the ICD-9-CM classification that is used in billing systems for reporting diagnoses; the MedDRA terminology used to encode drug reports; the NCI Thesaurus (NCIt) implemented in the cancer research nomenclature; and SNOMED CT (SCT) which helps in organizing the content of Electronic Health Records. The huge amount of biomedicine knowledge acquired over the centuries makes it impossible to have a single KOS that covers the whole domain. It is therefore necessary to use a combination of KOS interlinked with mappings that represent the semantic relationship between KOS entities, to optimize the domain coverage. However, the dynamic aspect of biomedical knowledge forces knowledge engineers, in collaboration with domain experts, to continuously modify KOS content to faithfully reflect the evolution of the domain. However, these changes at KOS level have a direct impact on depending artefacts like mappings or semantic annotations, causing inconsistencies in the information systems which exploit them for integrating, sharing or retrieving relevant information. Usually, hundreds of thousands of mappings are explored by these applications. Therefore, after KOS evolution, re-computing this whole set of mappings is a time-consuming task demanding huge efforts of validation. As a consequence, how to adapt mappings impacted by KOS evolution as automatic as possible, without re-computing the whole set of mappings each time a KOS evolves, has been the focus of this research work. Many research questions have been addressed for this particular problem: (i) How to perform mapping evolution taking into account the way KOS evolve? (ii) How do different types of changes impact mappings? (iii) What information regarding KOS evolution is necessary to support mapping adaptation? (iv) How to correlate different types of changes with mapping adaptation operations? Through this research project, we aimed at defining a formal framework to cope with the mapping maintenance problem between dynamic KOS. The proposed approach relies on the understanding and exploitation of information derived from KOS evolution [Dos Reis et al., 2014c], combined through heuristics with the consideration of the semantic relationship (taken from the SKOS model) of the mappings to maintain. Mainly based on these characteristics, a mapping evolution mechanism is implemented [Dos Reis et al., 2013b] through the design and implementation of the DyKOSMap framework [Dos Reis et al., 2012].

2.3 Validation of dynamic medical ontologies

Chapter 4 deals with the validation of medical ontologies and their evolution over time. First, more and more medical ontologies are designed and published[3] to support domain experts in their daily activities. Decision support systems in fact implement ontologies for their properties

[2]Project entirely supported by the Fonds National de la Recherche (FNR) of Luxembourg

[3]At the time of writing, 370 ontologies are referenced in the BioPortal repository

and their ability for reasoning. However, the critical aspects of the decision that are taken, especially those concerning patients' health and conditions, require a high level of validation for the ontology that is used. This issue is important, since ontologies are either built automatically from text corpora using various techniques or manually by knowledge engineers in close collaboration with domain experts, and they are in both cases are subject to errors. Moreover, as evoked previously, the dynamics of the medical domain pushes towards the definition of new knowledge through the outcomes of clinical trials for instance which, in turn, forces knowledge engineers and domain experts to continuously revise the content of the existing ontologies. At the same time, the modification of ontologies' entities must reflect the knowledge of the domain without questioning the validity of the ontologies. In the context of the COVALMO project, we aimed at defining methods and tools to automatize the validation of the content of a medical ontology. This was done by both maximising the exploitation of available reliable information sources that include scientific publications, information published on Web sites as well as Linked Open Data, and by minimizing the interaction with domain experts since their availability is limited as well as their knowledge of ontology languages. To do so, we have designed a system based on question/answering able to verbalize the content of an ontology under the form of queries adapted to the various sources of digital information and questions that are submitted to medical experts [Ben Abacha et al., 2013a]. The answers to queries and/or questions are then analysed and serve to either validate the corresponding piece of knowledge of the ontology or to modify it using the additional provided information. As a result, the initial ontology is validated and enriched.

This manuscript describes the contributions I have made to the management of dynamic medical knowledge according to the problems previously evoked through the collaborations with colleagues and students, in particular, three PhD candidates and three post-doctoral fellows.

Chapter 1

On the specificities of medical data and knowledge

Contents

1.1 Medicine: An old and vast domain 6
1.1.1 Representation of medical knowledge 6
1.1.2 Distribution of heterogeneous data 7
1.2 Medicine: A highly dynamic domain 7
1.2.1 Impact of evolution on data and knowledge management 8
1.2.2 Impact of evolution on patient and health professionals 9
1.3 Medicine: A critical domain 10
1.3.1 Medical data and knowledge quality 10
1.3.2 Metadata quality 10
1.4 Research project definition 11

Knowledge Engineering (KE) is a discipline that involves integrating knowledge into computer systems in order to solve complex problems which usually require a high level of human expertise [Feigenbaum and McCorduck, 1983]. Medicine is a domain where human expertise is of utmost importance. Health professionals of today treat their patients according to the medical knowledge and good practise acquired over centuries of investigation. However, by virtue of the evolution of the medical domain, we are currently facing a paradigm shift from doctor centric to patient centric [Carroll, 2002], questioning the application of medicine and the use of medical knowledge. Considering both the patients as the central focus and the ever-increasing quantity of digital data that is produced, the need for new KE methods in medicine is a reality, and essential to provide healthcare professionals with the right information, at the right moment to help them decide about their patients' health status.

In the following, we present some aspects related to KE, of biomedical knowledge and data we judge important to take into account that contribute to the acceptance of ICT systems by health professionals supporting them in their daily activities. The content of this chapter is necessary to understand the research orientations we have made, since it introduces problems as well as demands from the health professionals, in terms of data and knowledge management, which have provided us with realistic case studies for our research work.

1.1 Medicine: An old and vast domain

In contrast with other domains, medicine (or biomedicine) has a very long history. From the ancient Egypt, where the first information about surgery was mentioned, to the present day, an unmeasurable amount of knowledge has been acquired. As a consequence, it is impossible for a single person to be an expert in all medical knowledge, which promotes a division into subdomains (e.g. cardiology, anatomy, psychiatry, radiology ...) that are easier to manage, but which also leads to many other new problems.

1.1.1 Representation of medical knowledge

A major problem of the field deals with the various ways that exist to represent medical knowledge, which directly impacts its interpretation by both humans and machines. The long history of the field, combined with the various requirements and specificities of its subdomains, lead to the use of various Knowledge Organization Systems (KOS) like thesauri, taxonomies, codification systems, classification schemas or ontologies, each of them having its own modelling and reasoning capabilities [Hodge, 2000]. For instance, if the International Classification of Diseases[4] (ICD) relies on classification schemas to codify existing diseases and hypotheses on possible diseases, including nuanced classifications of a wide variety of signs, symptoms, abnormal findings, complaints, social circumstances, and external causes of injury or disease; a more elaborated model is required to encode the meanings that are used in health information and to support the effective clinical recording of data with the aim of improving patient care as provided by SNOMED CT[5].

However, the necessity of having specific subdomains induces a difference in the way those are considered by health professionals and increase the fragmentation of knowledge. It means that, by virtue of domain specificities, each group of experts has its own point of view on its domains, leading to ambiguities when these experts have to exchange with experts of other domains. This aspect can be transposed to KOS. Nevertheless, although there are no clear borders between subdomains, there still are some overlaps between them. For example, the *Infectious Disease Ontology* covering the description of infectious disease has some common elements with the ICD which is more general. In consequence, to reduce potential aforementioned ambiguities **semantic correspondences (or mappings)** [Euzenat et al., 2007] are established between different and usually heterogeneous KOS. It means that contrary to other domains like the Semantic Web where only OWL ontologies are aligned, in the medical domain, an ontology (expressed either in OWL or in OBO) can be mapped to a thesaurus (represented using databases) or a taxonomy with a classification system inducing specific problems at exploitation time.

Although the division in subdomains definitely helps to reduce the size, in terms of elements, of the existing medical KOS, those are still much bigger than those of other domains, therefore generating specific problems at exploitation time [Ceusters et al., 2004]. Actually, the significant number of elements (e.g. about 350,000 concepts and 1.5 million relationships in SNOMED CT) are subject to the creation of redundancies from the conceptual point of view and to inconsistencies from the logical point of view. For example, the concept *Brain part* in the definition of the concept *Structure of lobe of brain* is redundant as it subsumes the concept *Brain tissue structure* [Dentler and Cornet, 2013]. As stated in [Ceusters et al., 2003], the complexity of the domain requires significant modelling possibilities with precise definition of relationships to link KOS elements in an adequate way. Moreover, the use of natural language to define

[4] http://www.who.int/classifications/icd/en/
[5] http://www.ihtsdo.org/snomed-ct/

concepts labels is important in medicine, but sometimes existing logic languages fail to express the meaning of the label in a formal way, forcing knowledge engineers to find the right balance between formal (e.g. logic) and informal (e.g. natural) languages.

1.1.2 Distribution of heterogeneous data

Originally conceived for a restricted use, hospital information systems must now cover the needs, in terms of information and data, of a greater context which involves several hospitals or health institutions (e.g. laboratory, pharmacy, surgeries) taking into account privacy issues. To this end, the exploitation of medical data belonging to patients in clinical trials, to create new treatments or more generally to generate new knowledge is complex. In addition, the huge variety of the data nature and format reinforce this complexity.

Besides data and information generated for research purposes, medical data mainly aims at documenting a single patient's medical history and care over time. The medical record includes a variety of types of "notes" entered by healthcare professionals, recording observations, orders and the administration of drugs and therapies, laboratory results, x-rays, reports, discharge letters, etc. This kind of information (which is still on paper in many hospitals) is represented in various formats ranging from pure unstructured text (sometimes only abbreviations), structured data and up to high definition images carrying different types of information or sometimes the same information but represented differently. An x-ray can show a bone fracture and the report written in English or French by the radiologist can describe it with more or less precision. The use of standards to encode and exchange medical information (HL7 for sending messages or DICOM for images) can overcome some of the barriers caused by the distribution and heterogeneity of the data, but their acceptance is slow because of the cost engendered by their implementation in information systems and their impact on health professionals' daily practice. This refers to semantic interoperability issues. It provides interoperability at the highest level, which is the ability of two or more systems or elements to exchange information and to use the information that has been exchanged. Semantic interoperability takes advantage of both the structuring of the data exchange and the codification of the data including vocabulary so that the receiving information technology systems can interpret the data.

The aspect of medical data and knowledge being highly distributed, even inside the same health institution, requires the definition of processes to find, integrate and reuse data. Actually, clinical pathways [Kinsman et al., 2010] are tools based on evidence-based practice for a specific group of patients with a predictable clinical course, in which the different tasks (interventions) by the professionals involved in the patient care are defined, optimized and sequenced either by hour (ED), day (acute care) or visit (homecare). This definition aims at re-centring the focus on the patient's overall journey, rather than on the independent contribution of each speciality or caring function. Instead, all aspects are emphasised to collaborate, similar to a cross-functional team. This view copes with the paradigm shift from evidence-based to personalized medicine, requiring a integration of patients' data distributed across several, various information systems for the definition of adapted treatment plans. The use of computers' capabilities to process data must be enhanced with concepts that allow them to find the right information and reason over it to be beneficial to patients.

1.2 Medicine: A highly dynamic domain

The significant investments in research by both public and private institutes create an ever increasing quantity of medical data and knowledge, reinforcing the complexity of this domain

through the induced dynamics. As an example, in 2012, according to the National Library of Medicine, approximately 800,000 scientific papers were added to the refereed biomedical literature. Ten years ago, that number was slightly less than 400,000. A Learned Publishing article estimates that 50 million articles have been published since the beginning of formal research [Glaser, 2013]. This also confirms what Baneyx and Charlet have pointed out in their work, saying that over the last decade, 50% of medical knowledge has been renewed [Baneyx and Charlet, 2006] and this continuous evolution impacts knowledge management and exploitation as well as stakeholders' behaviour.

1.2.1 Impact of evolution on data and knowledge management

The aforementioned argument on the quantity of the generated scientific articles reflects the vast amount of new knowledge that is produced on a regular basis. It is usually defined using the even bigger set of medical data obtained through clinical studies. This creates new problems touching the various facets linked to the management and exploitation of this evolution.

As evoked in the previous section, KOS are progressively playing a key part in health information systems because of their properties from making the semantics of the underlying data explicit to enhancing the capacities of information systems. Since KOS are representation of domain knowledge, new findings lead to changes in the KOS (maintenance) to keep the coherence of the domain knowledge and to provide a reliable source of information to clinical support systems. The KOS maintenance is an important and expensive process that demands for sophisticated tools to simplify human tasks. Moreover, the combined use of KOS, thanks to the definition of mappings, requires that these semantic correspondences impacted by changes affecting KOS elements remain up-to-date over time. If the maintenance of KOS is done by the institutes in charge of the KOS, the maintenance of mappings that exist between KOS managed by different institutes remains problematic, and the release of new mappings does not faithfully accompany the release of new KOS versions (e.g. new releases of mappings between ICD codes and SNOMED CT codes are published several months after the publication of the new release of SNOMED CT). This may potentially provoke a lower performance of information systems exploiting KOS and their associated mappings because they will likely exploit out-dated links to retrieve data, and potentially taking wrong decisions about patient health.

Example: The concept 752.49 ("Other congenital anomalies of cervix vagina and external female genitalia") in ICD-9-CM v.2009 was split into five new concepts. 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 of the existing mappings. This was combined with an adaptation of the type of their semantic relation from narrow to broad to equivalent due to the fact that the two new concepts are more specific than the initial one. If the mappings are not properly adapted, there will be a mismatch between the aligned concepts and consequently an introduction of inconsistencies in software applications exploiting KOS.

Ontology evolution has been under investigation since the advent of the Semantic Web. However, formally describing the evolution process to generate a more complete computer-interpretable log of changes was not the main focus of knowledge engineers. As a consequence, the poor quality of the evolution process documentation makes the maintenance process more complex for the KOS and for its depending artefacts such as mappings, annotations or even the underlying information systems, since only information acquired from the observation of the

changes affecting KOS elements is exploitable. Considering the above example, the characterization of the changes (i.e. split) is not documented but can only be observed by comparing the two successive versions of the concepts.

1.2.2 Impact of evolution on patient and health professionals

The evolution of the medical domain affects KOS and their depending artefacts as well as the underlying information systems, but the outcomes of clinical trials and the experience gained in the daily practice of medicine also directly impact patients' health and professionals' behaviour. Currently, clinical guidelines that are designed based on clinical trials results assemble statements provided by the best available evidences. Their goal is to assist healthcare professionals with the definition of the appropriate treatment and care for people with specific diseases and conditions. In many cases, CGs (see section Introduction) do not detail all conditions and resources necessary to implement the treatment. This format gives more flexibility to healthcare professionals for adapting the treatment according to local conditions. However, it generates a new level of complexity to disseminate and exploit CIGs because local conditions need to be introduced in the CIG. Since it is almost impossible to foresee all situations, the reuse of CIG is compromised.

This implicitly refers to the modification of patient health conditions including changes in his environment (e.g. climate changes) or changes implying co-morbidity (i.e. patient suffering from several diseases). This kind of evolution directly impacts treatment plans that patients follow. However, several challenges must be addressed such as the detection of important changes, mainly by observing health parameters (e.g. blood sugar ratio, arterial pressure ...), their characterisation and formalisation to provide the right means for ICT systems in supporting health professionals with the adaptation of therapies for patients. Moreover, it also affects health institutions' organization. Patients that are subject to move from one department to another because of their health conditions require a lot of flexibility from the hospitals to provide them with the best care possible according to all the parameters (e.g. patient status, resource availability, local constraints ...).

To give an example: Aspirin is recommended to relief pain but, on the other hand, Aspirin is not recommended for patients with Duodenal Ulcer (DU) since it increases the risk of bleeding. In consequence, an alternative solution must be provided for patients suffering from both headache and DU.

The paradigm shift from doctor-centric to patient-centric puts the stress on patients. Patient-centred care is defined as: "Providing care that is respectful of and responsive to individual patient preferences, needs, and values, and ensuring that patient values guide all clinical decisions." [Institute of Medicine, 2001]. This "revolution" drastically impacts health professionals' behaviour, manifesting itself in the quality and the nature of notes mentioned in patients' records to become much more objective and be consistent with patients' conditions (e.g. patient with a fragile mental condition can be disturbed if they misinterpreted the medical data contained in their EHR).

The various types of evolution that have been highlighted in this section put knowledge evolution at the source of major problems in medical informatics. The research community has invested a lot of efforts to tackle this issue at various levels, aiming at (i) understanding the evolution and all its specificities [Hartung et al., 2008] (ii) characterizing it at conceptual level [Kirsten et al., 2011] and (iii) formalizing it to be reused by computer systems regardless

of whether the evolution concerns digital data (and knowledge), human agents (e.g. patients, health professionals) or institutions' organization, bridging thus the gap towards personalized medicine [Smith and Ceusters, 2010].

1.3 Medicine: A critical domain

Medicine aims at treating patients by providing care that affects their life or those of their family members. In consequence, it is important to have reliable data or information provided by ICT systems, if they are used by healthcare professionals or decision support systems to define adapted treatment to patients. Some quality criteria are discussed in the following subsections.

1.3.1 Medical data and knowledge quality

The previously evoked shift of paradigm from doctor centric to patient centric and the definition (and use) of Electronic Health Records (EHR) containing patients' data supplied by various heterogeneous sources (e.g. general practitioners, laboratories, hospitals ...) under different formats (e.g. PDF, images, text ...) are rarely complete from the data point of view [Arts et al., 2002]. It means systems fail to retrieve data at treatment design time for several reasons: lack of normalization, accessibility or data protection. In order to obtain the missing data, physicians often prescribe additional, and sometimes redundant, exams that can be harmful for patients' health (in case of over-radiation the need to produce x-ray images for example) and, on the other hand, that increase health costs.

The evolution of knowledge, and mainly the induced modifications at KOS level, are also the source of problems related to consistency and completeness. Actually, conceptual or logical redundancies (e.g. duplication of concepts) and inconsistencies (e.g. "is_a" cycle) can be introduced [Elkin et al., 2006] and incompleteness can be added when representational unit are removed [Jiang and Chute, 2009] impacting thus decisions taken with respect to patients' health conditions. If logical inconsistencies can be found using reasoners or theorem provers [vor der Bruck and Stenzhorn, 2010], the evaluation of the conceptualization is much more complex since it requires the participation of domain experts (i.e. health professionals) who often are not familiar with ontology languages and/or formal language. To overcome this lack, knowledge engineers who, on the contrary, are usually not familiar with the medical domain accompany domain experts in their validation task. This difference of culture is therefore the source of communication problems which unavoidably affects the quality of the knowledge that must be added to a given KOS, for example to be validated.

The challenges lie in the way computer systems will be used to drive the dialogue between knowledge engineers and domain experts to optimize the validation process and to enhance the quality of the KOS.

1.3.2 Metadata quality

The quantity of the data produced in the medical domain requires intelligent tools and methods to be processed, exploited and in particular be beneficial for end-users. Semantic technologies have shown great capabilities to support ICT systems in the retrieval, exchange, integration and sharing of the data by offering possibilities to make the semantics of the data explicit for machines. This is mainly done through the enrichment of the data with metadata (or semantic annotations) taken from standard, and sometimes formal, KOS. The importance of high quality metadata can be measured in different domains related to ICT for biomedicine. For instance,

privacy protection strategies can lead to the increase of complexity for information retrieval [Pruski and Wisniewski, 2012]. Data can be encrypted and only metadata are available for search engines. Thus, the efficiency of search engines strongly relies on the quality of metadata. We can split the quality of metadata into two criteria: quality of the reference model (i.e. KOS) and the quality of the annotation process. However, the efficiency of this technique obviously relies on two points.

First, the quality of the KOS. Closely connected with completeness and consistency, domain knowledge must also be precisely described with a well-selected terminology to increase the precision of the annotations. In fact, annotation techniques mainly consist in matching terms of documents with labels of concepts (or attributes like synonyms) contained in KOS, therefore the richer the concepts are described the better the matching is.

Second, the quality of the annotation process. The automatic or manual selection of annotations to associate with data must be good enough to make the semantic distance between the piece of information to annotate and the label denoting a concept in a KOS as small as possible to optimize outcomes produced by software applications exploiting annotations.

Example: Suppose that a patient has been diagnosed with HIV infection. This information will likely be annotated with either the concept "Disease" or "Infectious disease" or their associated concept codes. However, the latter is much more discriminant (because subsumed by the former) and the distance between HIV and "Infectious disease" is shorter than the one between HIV and "Disease". Now an information retrieval tool will search much more precisely since the annotation is also more precise. If for instance clinicians setting up a clinical trial are looking for patients suffering from infectious diseases, this patient will be identified directly only if HIV is associated with "Infectious disease". Otherwise, additional knowledge and reasoning mechanisms will be required to infer that HIV is an infectious disease.

1.4 Research project definition

The aforementioned specificities, especially their dynamic aspects, and the objectives of the field have allowed the definition of the research projects that have been the focus of my efforts in the past years. These efforts have been strengthened through the close connection we had with health institutions reinforcing the following research lines through their concrete cases.

First, the optimal use of ICT systems in the medical domains for data integration and decision support requires appropriate knowledge representation methods and models to support intelligent computer agents in the understanding and exploitation of the vast amount of generated medical data. As clinical guidelines are the foundation for diagnosis, treatment plans definition and, to certain extents, personalization, their representation to be understood by ICT systems is paramount for supporting health professionals. Actually, computer-interpretable clinical guidelines can be the link between good clinical practices, end-users and data (e.g. patients' data, local resource availability, etc.) which is required to take decisions or to notify physicians in cases of problems (e.g. substance allergy, drug interactions, treatment incompatibilities or similar). As a consequence, the definition of new methods and tools to enhance the reasoning capabilities of CIG-based systems, especially the part dealing with care actions to support the personalization of treatments for patients as well as CIG updates, constitutes one of my research lines.

Second, we believe that mappings that exist between KOS, either standardized or locally developed for specific reasons by IT department of hospitals, implemented in health informa-

tion systems to encode patients' data, for billing purposes or simply for interoperability reasons [Da Silveira et al., 2008], can improve the exploitation, in terms of data retrieval, sharing and decision support and performance of health systems. These links that denote semantic correspondences between elements, and in turn, with the data annotated to give them more flexibility to information systems and search engines by offering them the possibility to use several KOS instead of imposing a KOS that will be unavoidably incomplete, ambiguous, not sufficiently expressive and highly dynamic to meet the requirements. Taking this into account, mappings must be up-to-date and valid, from the semantic point of view, to remain exploitable for information systems. The management of KOS evolution and its impact on mappings constitutes the second research line I have explored over the past years.

Third, because of the amount of generated data and the critical aspect of the domain, the quality of the care depends on the quality of the models and data implemented in medical decisions. It is therefore important to rely on data and knowledge representation validated by domain experts. However, the main challenge is to provide intelligent tools to support these experts who are usually not familiar with information systems or logic-based languages to validate models that will be implemented in software applications able to guide clinical decisions. This challenge has motivated the third research line.

Taken the above into account, our research project will be articulated around three major axes, each one having knowledge and data dynamics as main focus. First, we will address the representation of dynamic medical knowledge in the context of treatment plan personalization and guideline update (see chapter 2). We present our contributions in this field with a particular emphasis on the iCareflow[6], METIS and SACCOM[7] projects. Second, knowledge dynamics is addressed in the context of KOS evolution and mapping adaptation under the DynaMO[8] project (see chapter 3). Last, in chapter 4, the dynamic aspect of knowledge is approached through the ontology validation problematic and the COVALMO project.

[6]http://santec.tudor.lu/icareflow
[7]http://tudor.lu/fr/these/lauto-adaptation-et-lauto-configuration-des-systemes-medicaux-collaboratifs
[8]http://santec.tudor.lu/project/dynamo

Chapter 2

On the representation of dynamic medical knowledge: the case of adaptive treatment plans

Contents

2.1 Problematic and research questions 14
2.2 Related work 15
2.2.1 Computer Interpretable Guidelines representation 16
2.2.2 Treatment plan adaptation 17
2.3 First reflections on the formalization of care actions and personalization of treatment plans: the iCareflow approach 19
2.3.1 From clinical guidelines to careflow 19
2.3.2 On the use of ontologies to represent care actions 21
2.3.3 Personalizing treatment plans in the iCareflow framework 22
2.4 Enhancing the reasoning capabilities of existing CIG languages: the METIS approach 23
2.4.1 The TMR Model 24
2.4.2 An application to multi-morbidity 25
2.5 Personalizing treatments plans: the SACCOM approach 27
2.5.1 Identifying and monitoring dynamic treatment parameters 27
2.5.2 The dynamic adaptation of treatment plans 28
2.6 Summary 29

As evoked in the previous chapter, evidence-based medical knowledge results mostly from clinical trials and is of utmost importance in clinical settings to disseminate best practices of medicine. In this context, it is important that the source of knowledge used by physicians reflects the most up-to-date findings. Clinical Guidelines (CGs) have been used for that and are considered as the foundation of treatment personalization implemented by the physicians. Current CGs electronic versions, called Computer-interpretable guidelines (CIGs) are the subject of many research work that proposes different description languages to represent them, like Asbru [Miksch et al., 1997], GLIF [Patel et al., 1998] or PROforma [Sutton and Fox, 2003]. By definition, these languages reflect the work process of healthcare professionals when coping with one

specific disease in several situations. The technical objective is to provide ways to execute CIGs. The success of this initiative increases the expectations of physicians to the potential applications of information technologies in their daily work activities. For instance, there is a demand for support systems capable to consider more than one disease for the same patient (i.e. multimorbidity). This example of an expectation requires new features not supported by existing CIG languages. It will be necessary to review the existing languages and introduce ways to increase the reasoning capability over CIGs to provide support for merging, updating and personalizing CIGs. Moreover, a fine-grained representation of CIG components with a particular attention devoted to the dynamic aspects of CIGs (i.e. medical recommendations, care actions, transitions and conditions) will facilitate their management and increase their modularity. This is the case for instance when outcomes of clinical trials will force the modification of an approach to certain diseases, and will furthermore contribute greatly to the spread of CIG by health professionals.

In this chapter, we present the work we have done in the context of computer interpretable guideline representation to enhance their reasoning capabilities and their further exploitation to personalize treatment plan of patients, taking into account the dynamics of their environment as well as that of medical knowledge. This work is part of the following research projects:

- **iCareflow**: this was funded by the Fonds National de la Recherche Luxembourg as a post doctoral grant assigned to Rodrigo Bonacin.
- **METIS**: is a PhD grant funded by the Fonds National de la Recherche Luxembourg and is assigned to Emna Mezghani. This work consists in a co-supervision between CRP Henri Tudor and the LAAS and INSA of Toulouse.
- **SACCOM**: is a PhD grant funded by the Brazilian government (CNPQ) and is assigned to Veruska Carretta Zamborlini. This work consists in a co-supervision between CRP Henri Tudor and the KRR group of the Vrije Universiteit Amsterdam.

2.1 Problematic and research questions

As evoked in chapter 1, the dynamics of medical knowledge and patients' environment, combined with the huge quantity of information health professionals have to deal with demands for intelligent software applications to define adapted therapies using CIG.

The reluctance of health professionals to use CG mainly lies in the lack of flexibility of paper based CG [Lenz and Reichert, 2005]. This is why CIGs have been introduced. However, existing languages for representing CIG content [Miksch et al., 1997, Patel et al., 1998, Sutton and Fox, 2003] were designed to produce CIG that can be executed. In fact, the resulting CIG can be compared to a workflow specifying the *actions* to be performed by health professionals, the *conditions* that must be satisfied and *transitions* leading from one state to another. But if these languages offer building blocks to specify these components, part of the information is still expressed in natural language which limits the use of software applications to reason over it at treatment plan definition and execution time.

Example: Consider for instance the care action *"Give aspirin to the patient"* and a patient being allergic to this pharmaceutical substance. Expressed like that, the computer will require a more elaborated level of formalization to be able to (i) understand that aspirin is the medication in question and (ii) it cannot be administered to the patient because of his health status.

The above example underlines one aspect of the problem with CIG and justifies the need for a new model for representing CIG content. In addition, many patients undergo a therapy that runs over a significant period of time (e.g. several years for chronic diseases). During this period, they are likely to be affected by other pathologies which require to combine CIGs to find a solution tailored to them. Nevertheless, the actual level of formalism of existing CIG languages does not allow the combination of CIG components to generate the adapted treatment plan in an automatic manner.

Example: Consider a patient suffering from both "Transient Ischemic Attack (TIA)" and "Duodenal Ulcer (DU)". The CIG associated with TIA recommends the administration of aspirin for treating some neuronal symptoms but, on the other hand, this substance increases the risk of bleeding, which is not recommended in case of a DU. The result of the combination of CIG expressed in existing languages tolerates the definition of contradictory actions harmful for the patients and does not support the finding of alternative solution tailored for treating both TIA and DU.

Last, once defined, the treatment plan followed by a patient is likely to evolve because of another pathology, changes in his environment or local constraints. In consequence, such external constraints must be taken into account and the therapies must be adapted accordingly. However, to deal with these issues, actual defined treatment plans using existing CIG models require a complete redefinition.

Example: The last epidemic episode of the Ebola virus provoked a modification of the procedure in the care of patients having fever, since additional tests have been introduced to detect if these patients were infected by the Ebola virus.

The previous examples clearly highlight the limitations of existing models and languages to deal with the dynamic aspect of medical knowledge. In this context, we have addressed the following research questions:

How can clinical guidelines content be represented in order to enhance reasoning capabilities of computer systems and support its management over time?

How to transpose human cognitive processes that govern the adaptation of treatment plans into clinical decision support systems?

The following sections will show the research methods we have followed and choices we have made to answer these questions.

2.2 Related work

Our problematic regarding CIGs representation and exploitation is expressed in two questions that concern (i) the representation of medical knowledge for reasoning purposes and (ii) the automatic adaptation of treatment plans. In this section we review existing work of these fields to better understand our proposals.

2.2.1 Computer Interpretable Guidelines representation

The representation of medical knowledge contained in CIGs is tackled at language level. Several languages have been designed by the medical informatics community, each having its own specificities, but none has been accepted as the *de facto* standard for expressing CIGs. Ontology-based approaches like SAGE [Tu et al., 2007] and GLIF [Patel et al., 1998] or workflow languages such as PROforma [Sutton and Fox, 2003] put the stress on three important points:

- the edition and execution of CIG,
- the interoperability with health information systems and
- the ability to disseminate knowledge.

The first attempt was proposed by Miksch et al. [Miksch et al., 1997] through the *Asbru* modelling language. It is a time-oriented machine-readable language making it possible to represent and to annotate durable skeletal plans based on a task-specific ontology. At design time, *Asbru* allows to express durable actions and plans caused by durable states of an observed agent. Moreover, *Asbru* integrates the notion of intentions underlying these plans as temporal patterns to be maintained, achieved or avoided. We find this notion important when adaptation is needed in case of contradictory actions observed when merging CIGs.

Sutton *et al.* have proposed the *PROforma* modelling language [Sutton and Fox, 2003]. It is an executable language that has been designed to build and deploy a range of clinical decision-support systems using guidelines and other clinical applications. It has a declarative format defining four basic types of tasks (plans, decisions, actions and enquiries) and allows the definition of logical and temporal relationships between them. In *PROforma*:

- an **action** is a procedure to be carried out (usually by an external element like a health professional or a medical resource),
- a **plan** is the basic building block of a clinical guideline and represents a container for a number of tasks, including other plans,
- a **decision** is a task that represents an option in terms of different logic commitments to be accomplished,
- an **enquiry** is a request for further information or data required before proceeding with the application of the guideline.

In the early 2000s, Peleg *et al.* have proposed GLIF3, the last evolution of the GLIF language [Peleg et al., 2000]. It allows representing clinical guidelines as flowcharts of temporally ordered nodes called guideline steps that store actions (Action Steps), decisions (Decision Steps), and clinical states of the patient (Patient Clinical States). There are two more types of nodes, called Branch Steps and Synchronization Steps, which are used for modelling multiple concurrent paths through the guideline. Decision criteria are modelled using an OCL-based language (Object Constraint Language) called GELLO [Sordo et al., 2003].

The internal representation of guidelines within the SAGE [Tu et al., 2004] framework is made using the EON formalism and comprises a set of Protégé classes and plug-ins. SAGE defines two different formalisms: recommendation-set and decision-map. The recommendation-set is an activity graph composed of processes and interactions between them. Activity graphs allow the specification of computational algorithms or medical care plans as processes consisting of:

- **contexts**, that are combinations of a clinical setting (e.g., outpatient visit in a general internal medicine clinic), care providers to whom the recommendation is directed, relevant patient attributes (e.g., patient age), and possibly a triggering event (e.g., a patient checking into the clinic),
- **decision nodes**, that evaluate conditions on variables (e.g., a Boolean precondition for an action),
- **action nodes**, that encapsulate a set of work items that should be performed either by a computer system or by a health care provider,
- **routing nodes**, that are used purely for branching and synchronization of multiple concurrent processes.

The conclusion of Peleg in her recently published work [Peleg, 2013] confirms our intuition as she advocates to split CIG content into small chunks in order to emphasize sharing/reusing, combining and maintaining medical knowledge. This points out the main limitation of existing CIG languages which offer building blocks to represent guideline components but at a too high level of abstraction, making guideline maintenance and usability too complex. As a consequence, the reluctance of health professionals to use CIGs in their daily practice will be reinforced.

2.2.2 Treatment plan adaptation

Besides the various languages to express CIG, techniques to adapt treatment plans based on these languages have been proposed. We will review them in this section since adaptation is the focus of our second research question.

Techniques for adaptation If treatment plans are still adapted manually in most of the cases, in the future, Artificial Intelligence techniques can be an alternative to automatize the adaptation process. For example, (Bayesian) logical refinement and machine learning techniques may be recommended to be employed in this process. Although this type of techniques proves to be useful in detecting and correcting recurrent and simple problems, it requires a huge set of realistic data to train and configure the system, but is less precise in discovering more complex relations [Patel et al., 2009]. For dealing with unexpected situations, this technique is not adapted.

Ontology-based techniques are another family of approaches to deal with adaptation of treatment plans. In [Abidi, 2011], Abidi addresses the adaptation problem using ontologies alignment techniques in case of co-morbidity. The idea is to have a unified representation of CGs content, resulting from the merge of the CGs describing the two diseases affecting the patient. Even if the proposed alignment tasks are manually done, the idea of having the CIGs described with OWL opens new opportunities for an automatic extension of the original guidelines or the validation of consistency, from the logic point of view, of the treatment for patients suffering from multiple diseases, and requiring the merging of several guidelines.

The KASIMIR system [d'Aquin et al., 2013] relies on case-based reasoning to represent different points of view on various types of cancer using object model. These models are integrated into the KASIMIR system, which is able to derive the best solution known by the system when presenting a new case to KASIMIR. The system exploits the acquired relations between cases to decompose new cases in order to adapt the known solution to this particular case. The quality of the provided approach relies on:

- the existing sets of cases used to configure the system,
- the ability of the system to decompose cases and align the obtained pieces with existing solutions.

The case of comorbidity. Jafarpour & Abidi adopted Semantic Web technologies to describe CIGs [Jafarpour and Abidi, 2013]. They built a merging representation ontology in OWL to capture merging criteria in order to merge CIGs. SWRL rules were then used to identify potential conflicts during the merging process. All conditions related to the merging process need to be described by the rules, increasing the effort to maintain the system up-to-date, and reducing the possibility of sharing knowledge. However, some related problems were not yet (completely) addressed in their work, for instance potential contradictions between rules, the scalability of the merging model to combine several CIGs, and how the ontology/rules are maintained up-to-date over time.

A different approach was proposed in [Wilk et al., 2011]. They describe CIGs as an activity graph and propose to use constraint logic programming to identify conflicts associated with potentially contradictory and adverse activities resulting from applying two CGs to the same co-morbid patient. The goal is to use this approach to alert physicians about potential conflicts during the definition of the treatment plan. The temporal aspect is not considered, thus the approach can only be applied to specific situations (e.g. acute diseases diagnosed during a single patient-physician encounter). Although their model allows reasoning over a subset of the CIGs content (the conditions) and proposing possible conflict solutions, the whole work of combining CIGs remains manual. This approach also considers that all predicates used in logic formulas for reasoning purposes shall use the same terminology and that they can have only two states (true or false). The case study used to demonstrate the applicability of the approach in [Wilk et al., 2011] shows the complexity of combining CIGs and the necessity to consider external knowledge sources for taking decisions.

Another method to address the CIGs combination problem is proposed by Riano & Collado [Riano and Collado, 2013]. They design a language to describe CIGs as actions blocks and decision tables. A generic treatment model is proposed to decide which action is appropriate to a chronically co-morbid patient, taking into account three criteria: *seriousness*, *evolution*, and *acuteness*. The expressiveness of this language is intentionally limited in order to have a lightweight decision system. The combination of CIGs is the result of pairwise combination of CIGs entities (i.e., actions and decisions table) according to a set of rules that allow identifying conflicts and reorganising or merging actions (in specific and predefined situations). The simplified CIGs representation and the specification of more general rules (for merging tasks) increase the reasoning capability of the system and reduce the maintenance work effort. However, reorganising care actions can raise some problems, especially those related to the clinical validity of modifications. In this case, the evidence-based medicine must be ensured in the rules of the generic treatment model. An alternative to this problem is to associate intentions and goals to the actions, as proposed by Latoszek-Berendsen *et al.* [Latoszek-Berendsen et al., 2007]. However, they do not consider combining CIGs and evaluating the role of intentions in this process.

The idea of evaluating pairwise actions associated to goals is exploited in the work of Sanchez-Garzon *et al.* [Sánchez-Garzón et al., 2013]. They adopt the Hierarchical Task Network plan description language to describe CIGs. The authors use multi-agent techniques to generate treatment plans and to identify potential conflicts between care actions. Treatment goals are considered to solve conflicts, but the assumption that all effects of an action is observed in the patient (and included in the patient data) and limits the applicability of their approach. A prob-

abilistic representation of effects would be closer to observations from evidence-based studies, but it would increase the complexity of the reasoning process. Despite the good preliminary results claimed by the authors, the low interoperability and the complexity of maintenance of agents have been underlined in several publications as challenges of the domain.

Adaptation can be handled at two different levels: at treatment definition time or at runtime (i.e. when the patient is following a treatment). In the first case, the situation of the patient is well known so the physician in charge of the patient has all the required information at his disposal. The challenge therefore lies in the integration of the necessary information as well as in the identification and solving of conflicts that exist between the data and the guideline or between the various guidelines in case of co-morbidity. The latter case is more challenging in the sense that the changes that may induce an adjustment of the treatment plan must be identified and the adaptation must be specified and applied. Actually, some observed changes may be very specific and appropriate knowledge must be available to handle these changes which is not always the case, and must therefore be discovered or extracted from external sources. In both cases, the fine-grained representation of knowledge contents in a guideline is paramount since it determines both the exploitation of knowledge for personalization and the identification and solving of conflicts, as well as the reuse and sharing of medical knowledge.

2.3 First reflections on the formalization of care actions and personalization of treatment plans: the iCareflow approach

The above survey of the literature, confirmed by Peleg's conclusions [Peleg, 2013], clearly shows that an appropriate way for representing medical knowledge contained in CIGs, in particular features dealing with care actions, is the first step towards treatment personalization, guideline update and reuse.

Care actions are the component of CIGs that drastically condition treatment personalisation. Actually, the simple statement *"Give aspirin to the patient"* cannot be done in case the patient is allergic to this substance. We also noticed that some CIG languages like GLIF or framework such as SAGE rely on an ontological background. This allows to define the concepts composing the underlying models, including care actions, to have an unambiguous interpretation of the statements. However, the definition of care actions in these models remains too abstract, and the important informations are expressed in natural language text, thus preventing any intelligent use by computer systems.

Based on these observations, we decided to use standard like the Unified Medical Language System[9] (UMLS) and the HL7 Reference Information Model[10] (RIM) to propose a general framework, called iCareflow [Bonacin et al., 2010], that relies on a method for personalizing treatment plans and an advanced ontological representation of medical actions.

2.3.1 From clinical guidelines to careflow

Our understanding of the field pushed us to define a general approach that leads from clinical guidelines to personalized careflow (see Fig. 2.1). In our context, a careflow represents the definition of a treatment plan obtained through the progressive adaptation of the initial guideline [Bonacin et al., 2010]. This is done by restricting the guideline until the outcome consists in a

[9] http://www.nlm.nih.gov/research/umls/
[10] http://www.hl7.org/implement/standards/rim.cfm

plan adapted to the patient according to his health status. To favour re-usability, we rely on standard vocabularies and Semantic Web technologies like the UMLS, OWL ontologies and SWRL rules. Moreover, we have introduced an intermediary level between guideline and careflow that consists in a representation of the restriction of the considered guideline tailored to local context. Actually, in existing approaches for treatment personalization, only patient specificities are taken into account while local constraints are often neglected. These are usually defined by national policy makers or by hospital staff. Such a rule can be the demand of additional care actions during epidemic episode (*e.g.* H1N1 virus in the early 2000s) or the selection of the healthcare institution, based on the availability of specific resources. This first refinement is then completed by applying rules representing patient profiles and physicians' preferences to obtain the treatment plans that cope with all kinds of restrictions. We are also considering another type of rules for expressing treatment/treatment, drug/treatment or drug/drug interactions, which may arise at therapy definition time.

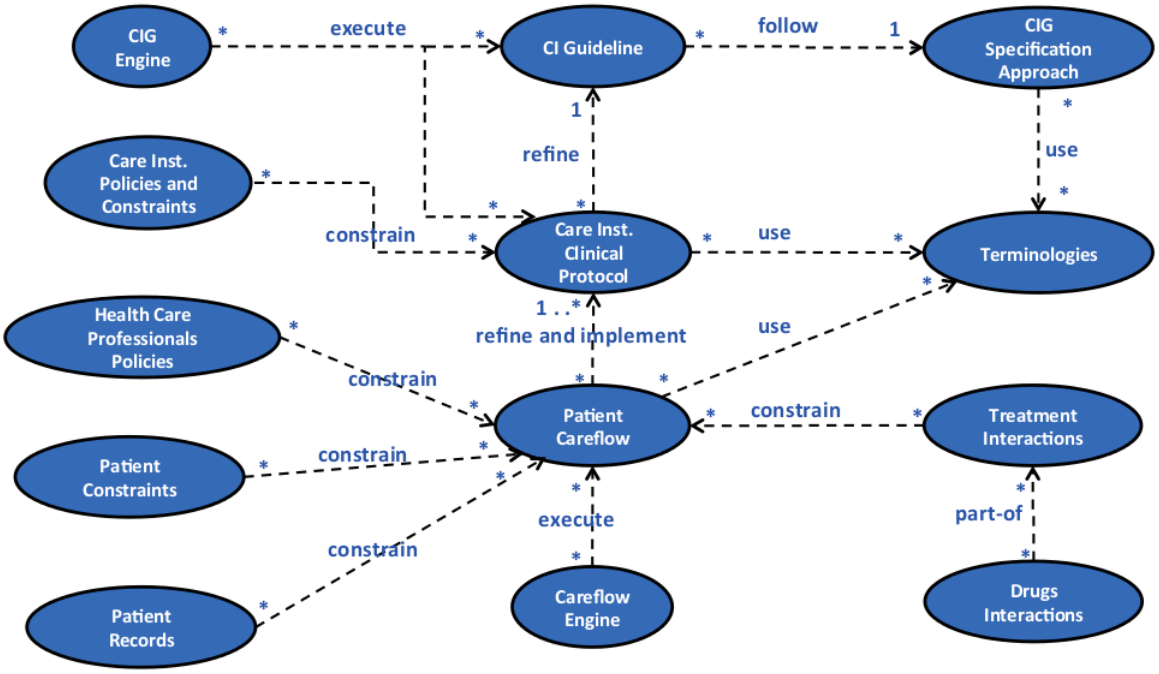


Figure 2.1: From clinical guidelines to careflow

Logic rules, serving as constraints, are expressed in the Semantic Web Rule Language, and the predicate symbols are labels of concepts or labels of Object Properties taken from standard terminologies and ontologies. These rules are then applied to the OWL ontology representing the care actions that are contained in the initial guideline. We put the focus on care actions because their formalization was not addressed in existing guideline specification languages. For interoperability reasons, we decided to design the iCareflow framework to be used on top of existing CIG specification approaches. It means that we can have a guideline expressed either in GLIF or in Asbru or in any other language, and apply the iCareflow approach to formalize the definition of care actions using standards. Moreover, the HL7/RIM model was used to ensure the link between careflow and patients' records to verify whether the proposed careflow is compatible with patient health conditions.

2.3.2 On the use of ontologies to represent care actions

As evoked, we represent care actions of guidelines using OWL ontologies following the GLIF and SAGE philosophy. The main challenge was to design these ontologies at the right level of abstraction, to enhance the use of computer systems when personalizing treatment plans. To this end, several tasks had to be done:

1. Identify care actions within CIGs specification,
2. Analyse the expression of the identified sets of actions,
3. Extract terms that denote labels of concepts and relationships (or roles) between these concepts, as well as attributes to conceptualize the care actions,
4. Formalize the obtained conceptualization in OWL,
5. Validate with experts the obtained ontologies.

This methodology has permitted to obtain a set of so-called pattern ontologies. Each pattern conceptualizing one specific type of care action. Fig. 2.2 depicts an example of the *substance administration* pattern ontology. In this example, the pattern is made up of 14 classes. Some classes have as label the same one that concepts of the UMLS Semantic Network. These are general concepts like *phsu* referring to *pharmaceutical substance*. This choice has been done for interoperability reasons because our general approach we are instantiating the patterns using information coming from the MetaMap tool which relies on the UMLS. Other labels are borrowed from other biomedical controlled terminologies like *substance_administration* and are more explicit. In this pattern we are representing all entities involved in the care action consisting in administering a substance. We have the substance itself (sbst), which is either a organic chemical (orch), a Pharmacologic Substance (phsu) or a Element, Ion, or Isotope (elii), the way the substance is administered (administration route), the receiver which is of type Patient or Disabled Group (podg) and other characteristic like time constraint (tmco), the dose to inject (qnco) as well as the entity in charge of executing the action (prog). In this context, in the text "Give aspirin to the patient" describing the administration of a substance, "give" will be the instance of the class "action", "aspirin" will be an instance of the class "phsu" and patient an instance of the class "podg". In our work we have manually analysed 21 CIGs (18 provided by the Tallis specification and 3 from SAGE) that contained a total of 179 care actions [Bonacin et al., 2013]. We then have extracted concepts and relationships from these actions and formalized them in OWL.

Moreover, to increase compatibility with existing health information systems, we associate each pattern ontology with SWRL rules that act as mappings between our patterns and the HL7/RIM model. This has the advantage to connect the abstract representation of the care actions composing treatment plans (CIG representation + pattern ontologies) to patient's data required to verify if the proposed therapy is adapted to the patient (e.g. test if the patient is not allergic to a prescribed medication). The use of standards such as OWL or HL7/RIM allows to cope with interoperability issues, but it also has some limitations. On one hand, OWL failed to deal with dynamic knowledge or to represent properly actions that have a duration in time (e.g. take a given drug for one week). Taking these elements into consideration, the ontologies obtained via MedAForm only represent static facts. In fact, temporal information can be represented in a static way but the dynamic aspect is lost. On the other hand, the problems generated by the HL7/RIM (e.g. a piece of knowledge can be created in many different ways) limits the usability of the ontologies when those have to be instantiated with patients' data.

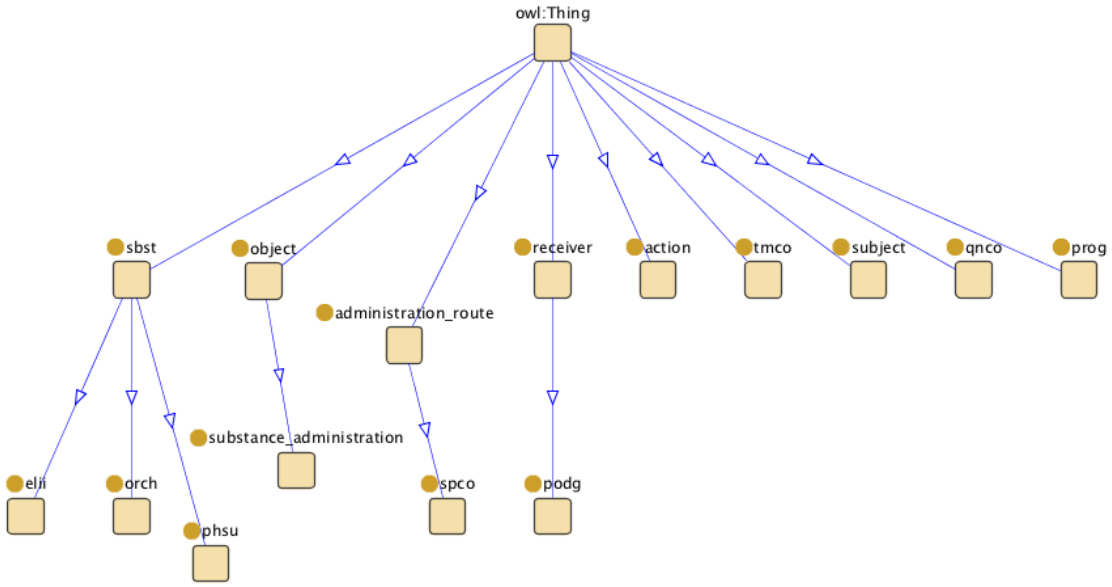


Figure 2.2: Example of pattern ontology: The case of substance administration

The resulting process (cf. Fig. 2.3) of care action pattern ontology generation form the MedAForm framework [Pruski et al., 2011a]. These pattern ontologies can now be associated with SWRL rules representing local constraints and others, like physicians' or patients' preferences in order to obtain the final treatment plan that is adapted to a local environment for a given patient. Concerning the validation of the ontologies we apply the COVALMO methodology we have designed and that is presented in Chapter 4.

2.3.3 Personalizing treatment plans in the iCareflow framework

The personalization of treatment plans is done through the successive application of constraints, materialized by SWRL rules, to our pattern ontologies [Bonacin et al., 2012]. Actually, if guidelines contain medical evidence learned from clinical studies, the huge variety of potential situations involving patients force medical experts and policy makers to specify relevant information outside the guideline and personalization may require information that is not part of the guideline. For instance, in case of epidemic or pandemic episodes, additional exams can be required to identify if patients are affected. This was the case a few years ago with the H5/N1 virus and the demand from various governments to add additional tests for patients at risk with influenza symptoms. Another example concerns the availability of resources to perform an exam useful for the diagnosis. Such kind of information has a direct impact on the patient, but has to be integrated at treatment design time or later when the patient is following his therapy. Based on this observation, the iCareflow approach, depicted in Figure 2.1, was designed to distinguish what is part of the guideline and what is not and whether its implementation was done in order to favour knowledge integration. As a first prototype of the approach, we decided to express these personalization rules in SWRL so we can apply them on the OWL ontologies obtained following the MedAForm process (cf. section 2.3.2). Moreover, as our ontologies are aligned with HL7/RIM, data that can be found in patient records will be exploited in an easier way.

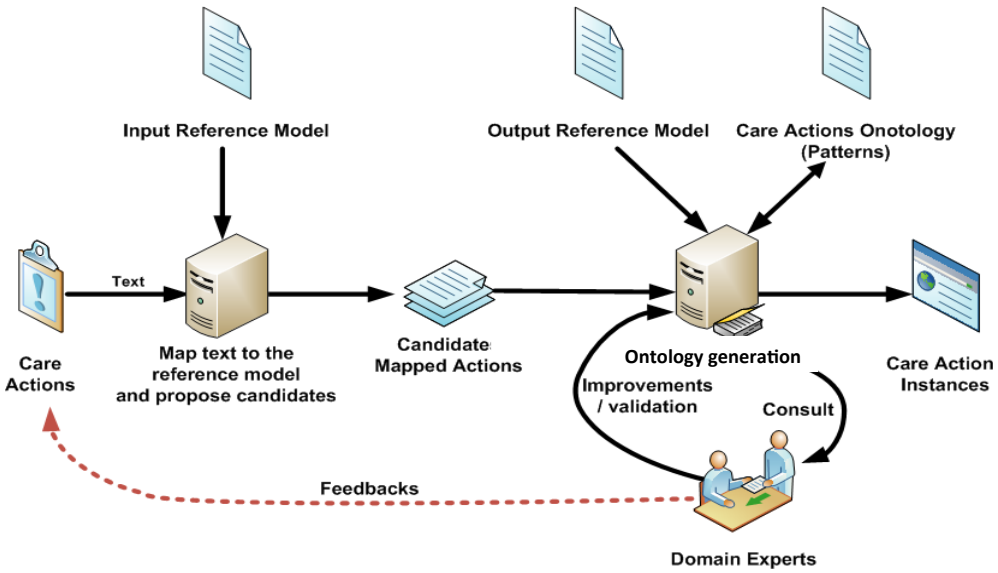


Figure 2.3: The MedAForm approach

SWRL is a rule language, subset of First-Order Logic, proposed by the Semantic Web community combining some properties of OWL DL with Horn logics. This is why SWRL requires advanced understandings of logic languages to be able to construct well-formed rules, competencies that people who are supposed to define constraints do not have. These persons must therefore be assisted by knowledge engineers in the specification of SWRL rules to guarantee their quality since such rules can rapidly become very complex (examples of rules can be found in [Bonacin et al., 2012]) and the absence of tools supporting rules construction does not help. Beside this first limitation, SWRL may also not support the representation of certain characteristics inherent to the treatment personalization. The dynamic aspect of patients' health status and local environment may impact the treatment plans like, for instance, adjusting the dose of a given pharmaceutical substance. However, even if new knowledge can be created through SWRL rules, existing values (e.g. concepts instances) in an ontology cannot be modified. This drawback can be circumvented by using predefined Built-ins such as swrlb:equal or swrlb:lessThan for comparisons but it requires an SWRL Built-in Bridge, and a dedicated OWL model implementing built-ins, which is a very demanding, complex and cumbersome solution. For all these reasons, a human intervention is usually done to overcome these issues.

2.4 Enhancing the reasoning capabilities of existing CIG languages: the METIS approach

The iCareflow approach [Bonacin et al., 2013] was our first attempt to represent clinical knowledge contained in guidelines and to personalize treatment plans with data that can be external to

CIGs. Through this preliminary work, we have further highlighted existing problems with clinical knowledge representation and exploitation. Actually, the personalization of treatment plans or other use cases relying on guidelines like CIGs update or merging, in case of multi-morbidity, are limited by the poor reasoning capabilities of existing CIG representation approaches mainly the way conditions-actions pairs are expressed (i.e. under which conditions a care action may be triggered).

2.4.1 The TMR Model

Reasoning capabilities of CIGs approaches are largely conditioned by their underlying formalism. If the semantics of most of the existing approaches or languages to express CIGs are given in a subset of First Order Logic (cf. GLIF or SAGE), the conditions/actions pair, which is paramount in medical decisions, is still represented in natural language or, in the iCareflow approach, using lightweight OWL ontologies. A closer analysis of the content of the guidelines reveals that the conditions/actions pairs are usually associated with other features that play key roles in the decision process and are likely to be conflicting. The definition of four key notions is the foundation of the Transition-based Medical Recommendation (TMR) model [Zamborlini et al., 2014a] (cf. Figure 2.4). It relies on the definition of **Situation types**, **Care actions**, **Transitions** and **Recommendations**.

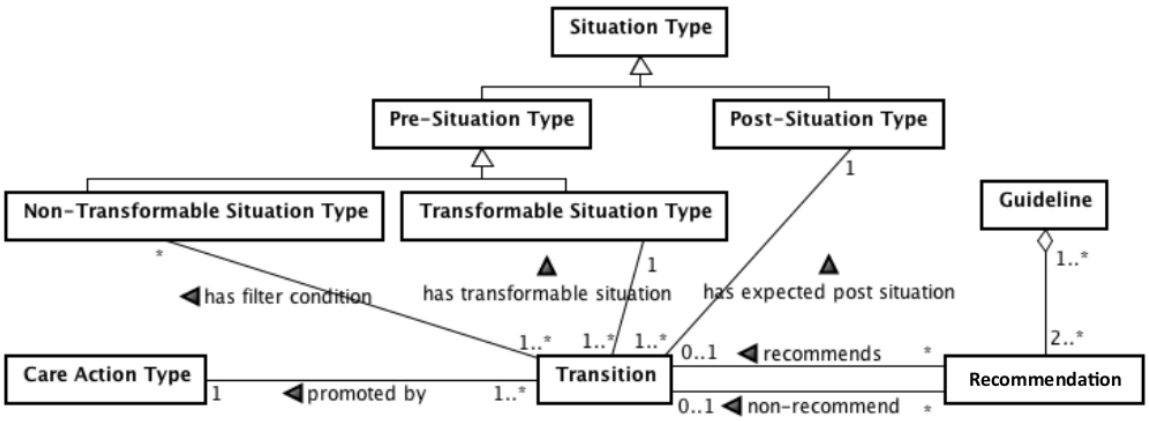


Figure 2.4: UML class diagram for the TMR Model

These concepts have been proposed following the analysis of existing clinical guidelines and the identification of the limitations of CIG languages and framework. Moreover, the definition of the elements of our model has been aligned with well accepted theories and the Unified Foundational Ontology (UFO) [Guizzardi and Wagner, 2004]. This strengthens our approach, facilitates its acceptance and reuse, and we can build upon well-defined semantics to improve its future alignment with medical information.

Situation type represents a property, which characterizes a patient, and its admissible values. We distinguish between pre and post situation types. The former denote the conditions that have to be satisfied in order to apply a given care action. Symptoms of a disease can be seen as a type of Pre situation. The post situation types describe the situation that has to be reached after applying a care action (it can be seen as the set of effects engendered by the care action once executed).

Care actions represent the action types that can be performed by health care agents in order to change a situation. Actually, the recommended action aims at transforming the situations

described through the Pre situations to that described by the Post situations. For instance, giving aspirin to the patient (action) can lead from a situation where the patient feels pain to a situation where the pain has been reduced.

Transition relates a Care Action Type to Pre/Post-Situation Types and represents possibility of achieving that change by performing the referred action. Thus, by assigning different transitions to a care action type, we define its "space of transitions". In the previous example, the transition denotes the change from "feeling pain" (Pre situation) to "reduced pain" (Post situation) as the effect of the care action.

Recommendation represents a suggestion to either pursue or avoid a transition promoted by a care action type.

Example: The example depicted in figure 2.6 represents the instance of the TMR model that corresponds to the duodenum ulcer disease CG of figure 2.5. As we put the stress on the recommendations, actions, transition, and situation, we restrict the whole CG to these concepts only. As illustrated, the model is richer in the sense that it shows more information like the objective of the recommendation or additional information about the actions (e.g. do or don't).

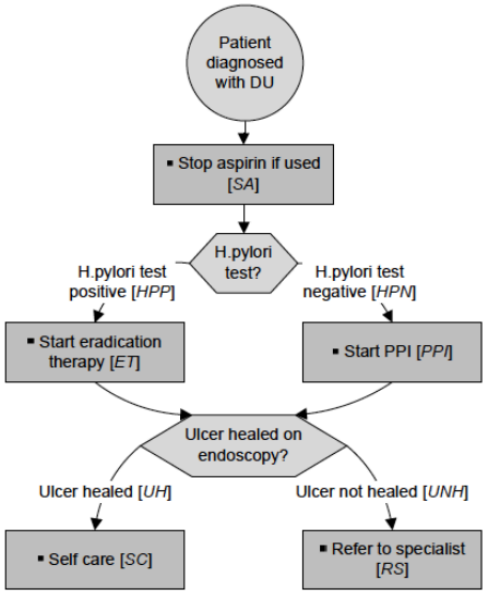


Figure 2.5: Duodenum ulcer CG

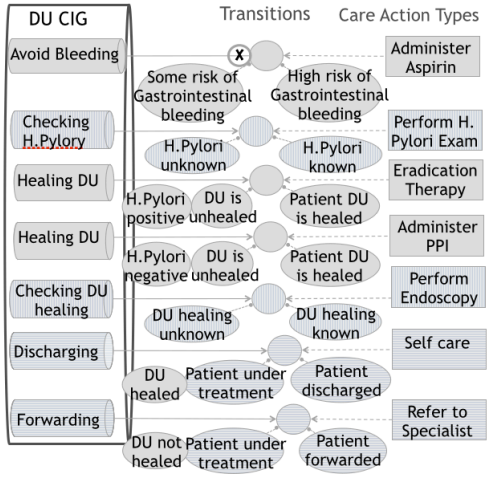


Figure 2.6: TMR representation of Duodenum ulcer CG

2.4.2 An application to multi-morbidity

With the increasing average age of the population, people are more likely to be affected by several pathologies. This situation also holds for patients suffering from chronic diseases like diabetes and catching an acute disease in addition. In consequence, several CGs must be combined in order to be able to define treatment plans that fit the specificities of the involved diseases, patients and local context. However, existing CIG description approaches, which lack formalization at care actions level since these are still expressed in natural language, do not provide means to identify potential interactions that can rise at CG merging time. For instance, the CG depicted in figure 2.5 does not recommend the administration of aspirin to reduce the risk of gastrointestinal bleeding while the CG for treating TIA recommends aspirin to reduce vascular

events. Such kind of conflicts (i.e. do and don't) cannot be identified with existing approaches.

The provided TMR model has this ability for identifying various types of interactions. We have classified them into three categories:

1. **Contradictory Interactions:** Two recommendations that cannot be followed at the same time without leading to an undesired (non-recommended) final situation. In this case we can distinguish between contradiction at action level (e.g. administer aspirin/do not administer aspirin) and between contradictions at transition level (e.g. Do not administer beta-blockers **to avoid lowering blood pressure** / Administer ACE inhibitor **to lower blood pressure**).

2. **Repetition Interactions:** Set of recommendations that are subject to optimization. In this case we can cite the repetition of the same action but with a different goal or recommendations to inverse transition (e.g. Administer ACE inhibitor **to lower blood pressure** / Administer midodrine **to increase blood pressure**).

3. **Alternative Interactions:** Set of recommendations that hold as alternatives. This concerns the repetition of recommendations to the similar transitions promoted by different care actions (e.g. Administer aspirin **to handle inflammation** / Administer ibuprofen **to handle inflammation** / Administer naproxen **to handle inflammation**). The repetition of recommendations can lead to overdose situation in cases where the recommendations concern a particular drug. It also concerns non-recommended transition whose inverse transition is recommended (e.g. Do not administer aspirin **to avoid increasing the risk of gastrointestinal bleeding** / Adm PPI **to decrease risk of gastrointestinal bleeding**).

To be able to identify these interactions we have extended the TMR model with additional classes and associated it with constraints expressed in First-Order Logic. The TMR4I model (see Figure 2.7) puts the focus on two types of interactions: those that can be resolved by analysing the recommendations themselves and those that require external knowledge to be identified (e.g. database specifying drug interactions) [Zamborlini et al., 2014b]. These observations allow the refinement of the definition of interactions.

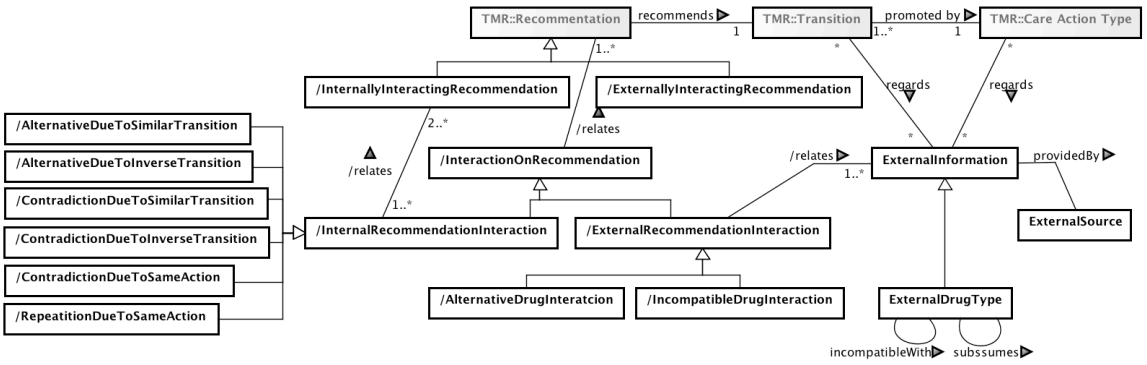


Figure 2.7: UML class diagram for the TMR4I Model

Based on the concepts introduced through the TMR model and its extension, we are able to automatically identify potential interactions by analysing the Pre and Post situations, hence

overcoming limitations of existing approaches. This is why we have further refined the properties of transition (cf. /similarTo, /inverseOf) useful to highlight interaction and alternatives as well as the types of interactions (cf. /OptimizableInteractionDueToInverseTransition, /ContradictionDueToSameAction ...) that can rise when merging guidelines. The logic formula hereafter formalizes the type of interactions happening when two inverse transitions are recommended (e.g. to lower blood pressure *vs* to increase blood pressure).

(8) $\forall g, r1, r2, t1, t2$ Guideline(g) $\wedge$ Recommendation(r1) $\wedge$ Recommendation(r2)
$\wedge$ partOf(r1,g) $\wedge$ partOf(r2,g) $\wedge$ Transition(t1) $\wedge$ Transition(t2)
$\wedge$ **recommends(r1,t1)** $\wedge$ **recommends(r2,t2)** $\wedge$ **inverseTo(t1, t2)**
$\rightarrow \exists i$ **OptimizableDueToInverseTransition(i)** $\wedge$ relates(i,g)
$\wedge$ relates(i,t1) $\wedge$ relates(i,t2)

In this work we have enhanced the state-of-the-art on CIG content representation by automatizing the identification of interactions when merging guidelines in case of multi-morbidity. We are currently focussing on two other tasks concerning guidelines: the update of CIG content and the adaptation of CIG content to local constraints. The goal is to enrich the TMR model to be able to cope with the specificities of these use cases.

2.5 Personalizing treatments plans: the SACCOM approach

The previous section describes the work focussing on the design of treatment plans in case of multi-morbidity. However, adaptation of therapies is still required after the design of treatment plans by physicians. Actually, dynamics of patients environments and health status may disturb the previously defined treatment. Diabetic patients can have a fluctuating blood sugar concentration according to their diet, which requires the adaptation of the insulin dose to inject in (quasi) real-time. Ongoing work carried out in the SACCOM project puts the stress on adapting therapies followed by patients when changes in their environment are observed which may impact their health. In this context, problems are related to the variables of the treatment (i.e. the various parameters of the treatment that are likely to change and impact patient health) and the response to the observed changes (i.e. the actions to perform on the treatment plans according to the observed changes). The goal of the project is to provide a system for adapting treatment plans in response to changes in patients' health status or in the local environment.

2.5.1 Identifying and monitoring dynamic treatment parameters

The definition of treatment plans based on CIGs usually results in a workflow showing the various steps the patient has to go through in order to be treated. However, the lack of flexibility of CIG languages made these workflows static in terms of modifiability, contrasting with the highly dynamic aspect of pathologies and patients' environment. For instance, once diagnosed, a patient suffering from atherosclerosis is directly influenced by environmental factors such as smoking, his particular diet, or by alcohol consumption. By virtue of their critical aspects, these elements must be quantified and must be rigorously monitored. It is therefore vital for the patient to be able to control these factors and to adjust treatments accordingly in order to reduce the risk of having a cardiovascular incident.

In this context, it is important to identify the parameters of the treatment that may vary and that have a direct impact on patient's health. However, the treatment plans obtained using CIGs do not make the distinction between parameters that play a key role in the treatment and

the other ones, making it therefore hard for decision support systems to identify issues or to react before a problem shows up. To overcome this limitation, we have designed an ontology defining the various parameters and their specificities that are involved in chiropody as first element of the final system. The provided ontology makes it possible to annotate, at CIGs level, elements that the physician decides as imperative for the treatment. The concepts and relationships of the ontology defines the semantics of the annotated elements, allow them to be exploited by decision support systems. Actually, these ontological properties give information on the way these parameters can be monitored (e.g. sensors measuring cardiovascular activities), the threshold values in case of numeric measures (e.g. 0.9 g/l of blood sugar for a pregnant person) or even the impact of these parameters on patient health (e.g. risk of cardiac arrest due to extreme tachycardia). This task has been the first step towards a treatment personalization system that is beneficial to both physicians and patients. The system will provide the former a way to control their patients and their adherence to treatments, while the latter will have their treatment plans fully adapted to their health status and local environment.

The monitoring of the important parameters consists in the second aspect of the final system [Mezghani et al., 2015]. Since the ontology provides means for physicians to specify the technique to monitor the parameters including the types of sensors to use, the system can thus exploit the gathered data thanks to the sensors. Our study has revealed that Autonomic Computing is a paradigm design to cope with specificities of our context [Mezghani et al., 2014]. Actually, the four major steps of the Autonomic Computing (i.e. Monitoring, Analysis, Planning and Execution) can be applied in the personalisation of treatment plans.

2.5.2 The dynamic adaptation of treatment plans

Deciding to adapt ongoing treatment plans in response to an observed modification in patients' context has to be done according to several factors:

- The nature of the changed parameters and its critical aspect with respect to the patient. For instance, blood glucose level is critical for a diabetic patient since it increases the risk of cardiovascular incident.
- The importance of the variation that can be minor or significant. For instance, the glucose level varying from 0.82 g/l to 0.81 g/l is normal while a glycemia decreasing from 0.8 g/l to 0.37 g/l reveals a case of hypoglycemia.
- The implication of the treatment adaptation on the patient. Actually, the planned modification of the treatment plan must be done while preserving patient safety (e.g. check if there are some drug-drug, drug-treatment or even treatment-treatment interactions).

Once the decision has been taken on the adaptation of the ongoing treatment plan, the modification must be implemented. This implies:

1. Modifying the designed workflow by respecting constraints like time (e.g. check if doing an action before or after another one is possible and safe for the patient), existing care actions (cf. section 2.4.2) or any other CIG element's characteristics.
2. Potentially re-annotating new parameters or modifying some of their aspects (e.g. threshold values) using the ontology.
3. Verifying if the adapted treatment plan fits patient preferences (e.g. some care actions demand time to be done which may be incompatible with the patient's time schedule).

This part is currently under development. Although promising, the first tentative to tackle these issues deserves more attention and will be improved through a close collaboration with health professionals of the specific dynamic domains requiring the adaptation of therapies.

2.6 Summary

In this chapter, we have described some of the problems rose by the representation of medical knowledge at guideline level and aspects addressing its exploitation in a dynamic context that have motivated our research orientation. In particular, we have partially shown that an adapted model for representing CIG content can enhance the reasoning capabilities of systems, mainly by improving their ability to identify interactions that usually happen in the treatment of multi-morbidity at CIG merging time. The work carried out in the framework of the iCareflow, METIS and SACCOM projects seems to be promising since it progressively raises the interest of and acceptance by health professionals. Their initial fear, which was caused by the lack of flexibility of CIG in existing systems, appears to be gradually fading because of the introduced approaches making the systems much more flexible and useful for physicians. This work also shows that:

- The translation of CG into CIGs is very time consuming and error prone. The absence of standards as support for expressing CG in electronic format, but also in paper format, generates ambiguities at translation time.
- Guideline content is sometimes incomplete and out-dated. Adapting a treatment plan requires medical knowledge that should be contained in CG. However, an exhaustive list of situations describing the patient state cannot be specified at the guideline design level. Moreover, the continuous development of clinical trials generates new knowledge of which the integration takes time, since CG content is revised every five years on average. This five-year period does not reflect the state of medical knowledge in the concerned field.

The work carried out in the iCareflow projects has been validated on CIGs expressed in the PROforma and SAGE models and has consisted in generating ontologies using the MedAForm prototype. Concerning METIS, the validation of the TMR and TMR4I models are currently done by comparing the obtained results with state-of-the-art methods described in [Wilk et al., 2011] and [Jafarpour and Abidi, 2013], while the SACCOM approach will be evaluated on realistic clinical cases demanding dynamic adaptation.

Chapter 3

On the management of dynamic medical knowledge: the case of Knowledge Organizing Systems mapping adaptation

Contents

3.1 Problem statement and hypothesis 32
3.2 Related work 33
3.2.1 Mapping revision 33
3.2.2 Mapping calculation 34
3.2.3 Mapping adaptation 34
3.2.4 Mapping representation 35
3.3 Understanding ontology evolution for adapting mapping 35
3.3.1 Empirical analysis of mappings and KOS evolution 35
3.3.2 The role of concept definition in mapping adaptation 36
3.3.3 Identification of relevant dynamic knowledge to adapt mapping 37
3.4 Adapting mappings according to ontology evolution 39
3.4.1 Change patterns for mapping adaptation 39
3.4.2 Definition of mapping adaptation actions 41
3.4.3 Heuristic-based approach to maintain mappings valid over time 42
3.5 The DyKOSMap framework 43
3.6 Experimental assessment 45
3.7 Summary 46

The biomedical and health domains are particular in the sense that because of their long history, compared to that of the Web, a huge quantity of knowledge has been acquired and organized in various, more or less expressive, models. Unlike the Semantic Web for which ontologies predominate, there is a variety of knowledge representation models in the medical domain, named Knowledge Organisation Systems (KOS), ranging from classification schemas like the International Classification of Diseases (ICD) to ontologies like Gene Ontology (GO). In this context and also due to the wide variety of objectives and tasks, the quantity of acquired

knowledge makes the domain so vast that it is not possible to have only one KOS to cover it. This is why the domain is subdivided into smaller pieces, making its representation much easier using dedicated KOS. In practice, information systems have to cover several sub-domains and therefore, have to use a combination of KOS that are interconnected via mappings which represent the semantic relationships between conceptual entities belonging to different KOS and increase the coverage of domains. However, the highly dynamic aspect of medical knowledge, appearing in the significant amount of scientific articles published regularly, forces knowledge engineers to constantly revise the content of KOS. This is all the more true regarding their depending artefacts, especially mappings, to preserve consistency in the underlying information systems. We shall discuss the mapping maintenance problem in this chapter.

3.1 Problem statement and hypothesis

In our context, a KOS K^t at time t ($t \in \mathbb{N}$) is a triple,

$$K^t = (C^t, R^t, A^t)$$

where C^t denotes the set of concepts, R^t the set of relationships that link concepts and A^t represents the set of concept attributes associated with each concept. We restrict this definition to simplest definition of KOS. We are not considering neither axioms or individuals nor attributes of relations. Consider two KOS $K_s^t = (C_s^t, R_s^t, A_s^t)$ and $K_p^t = (C_p^t, R_p^t, A_p^t)$, according to [Euzenat et al., 2007], a mapping m^t between K_s^t and K_p^t is set of 5-tuple

$$m^t = (id, c_s^t, c_p^t, n, r)$$

where: id is a unique identifier, $c_s^t \in C_s^t$, $c_p^t \in C_p^t$, n is a confidence measure holding for the correspondence between c_s^t and c_p^t and r denotes the relationship symbols that exist between c_s^t and c_p^t. In our study, we restrict ourselves to five symbols (equivalence ($\equiv$), more general ($<$), more specific ($>$), disjointness ($\otimes$) and overlapping ($\approxeq$)). This choice was motivated by the analysis of existing biomedical KOS, which has shown that existing mappings refer in a huge majority to these five relationships. In the remainder of the manuscript, we denote $\mathbf{M}_{SP}^t$ the set of all mappings between K_s^t and K_p^t at time t. Taking the above into consideration, we define mapping maintenance as (see figure 3.1):

"the modifications performed on the mappings established between KOS entities in order to keep them valid when these KOS entities evolve."

According to this definition, we assume that mappings exist between two KOS, and they must be adapted to remain consistent because of modification of KOS entities involved in their definition. Empirical analyses [Dos Reis et al., 2014c, Gross et al., 2012] underline that a significant number of mappings is impacted by changes in KOS entities. In [Dos Reis et al., 2014c], we have shown that between SNOMED CT and ICD-9-CM, about 8,000 mappings are impacted by the evolution of SNOMED CT and must thus be revised and validated by experts. In this context, if we consider that an expert must manually adapt a mapping, he will need about 1 month of work if he spends 1 minute per mapping. This observation fully justifies the need for automatic approaches to support human experts in this maintenance process. The underlying research question can be expressed as:

How can mappings that have turned invalid by KOS evolution be adapted as automatically as possible in order to restore their validity without realigning the KOS?

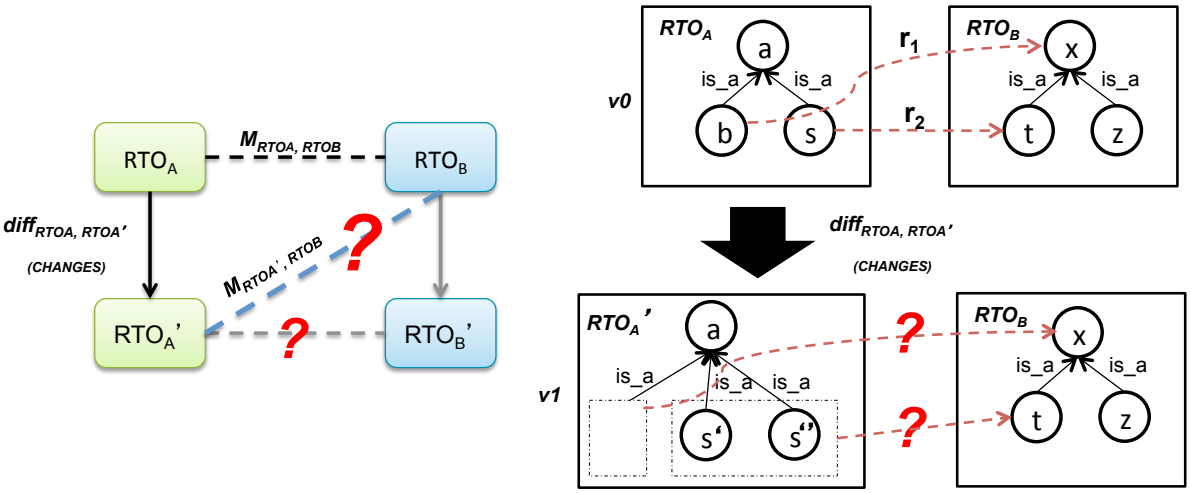


Figure 3.1: The mapping maintenance problem

To reach the proposed objective, we had to:

- Understand and characterize the evolution of biomedical KOS at the right level of abstraction.
- Find a method to combine the information gained from KOS evolution with that learned from the interpretation of existing mappings to adapt them.

3.2 Related work

Even if mapping evolution has been recognized as an important problem in the biomedical domain, little work has been proposed to tackle it. The following sections describe the related work of the mappings maintenance field and will allow a better understanding of the sub-problems and our scientific contributions. Methods and techniques aiming at preserving mapping validity can be grouped into four distinct categories: mapping revision, mapping calculation, mapping adaptation and mapping representation. This section summarizes the work published in [Dos Reis et al., 2015c].

3.2.1 Mapping revision

Approaches of this category aim at identifying and repairing invalid mappings. In this context, two dimensions have to be considered:

1. the identification or detection of invalid mappings,

2. the repair (or debugging) of invalid mappings.

The work conducted by McCann et al. [McCann et al., 2005], leading to the development of the MAVERIC system, is probably the most significant proposal for monitoring and detecting invalid mappings between dynamic relational schemas in an automatic way. Source schemas are checked periodically and query answers are compared to prior known answers. Once query answers differ, an alert about a potential broken link is sent to the system administrator. This idea has also been followed by Colazzo & Sartiani [Colazzo and Sartiani, 2009] since they exploit the results of XQuery clauses. In [Mawlood-Yunis, 2008] fault-tolerance techniques to detect temporal semantic mapping inconsistencies in peer-to-peer systems are explored. As they work on dynamic systems, they distinguish between permanent and temporary semantic incompatibilities.

Concerning the repair of mappings, Meilicke et al. [Meilicke et al., 2007] argue that there is a need for debugging mappings because of the inefficiency of existing matchers. They apply a basic procedure in which two sets made up of inconsistent and valid mappings respectively are considered. The problem is to determine the intersection between both sets. Since the reference mappings are often inaccessible or even unknown, they reformulate the problem by partitioning the set of mappings to be repaired into correct and incorrect correspondences with respect to the set of reference. Therefore, repairing a set of correspondences consists in the determination, for each correspondence, of whether it belongs to the set of correct or incorrect mappings, and eventually to deletion of the set of incorrect correspondences. The same idea is echoed by Qi et al. [Qi et al., 2009] and Castano et al. [Castano et al., 2008] since they proposed to use probabilistic reasoning to revise mappings.

3.2.2 Mapping calculation

Approaches of this type aim at entirely or partially recalculating the set of invalid mappings using matching algorithms. However, as pointed out in [Yu and Popa, 2005], these approaches do not take into account either KOS evolution or existing mappings, and they require a significant validation effort depending on the size of the KOS to realign. To overcome these lacks, the work of Khattak et al. [Khattak et al., 2012] consists in recalculating only the set of invalid mappings. Their approach relies on the analysis of a log file containing KOS changes to detect impacted mappings, removing them and recalculating new alignments by considering the KOS in their entirety.

3.2.3 Mapping adaptation

Such kind of approaches are more advanced because they take several dimensions into account like evolution information and use mapping composition, model synchronisation, or change propagation to minimise the impact on the mappings. In this context, Yu & Popa [Yu and Popa, 2005] have designed the MACES system to compose information from schema evolution with existing mappings to generate new semantic correspondences between database schemas. This technique has been improved by Fagin et al. [Fagin et al., 2011] since they have defined new composition operators. In the database field mapping adaptation is also interpreted as a query rewriting problem [Velegrakis et al., 2004].

Existing approaches deal with different KOS models. This is the case of the work of An et al. [An et al., 2010]. They define mappings as relationships between columns of the relational schema and properties of a concept in an ontology. Maintenance is therefore a problem of synchronisation between models and mappings.

Tang & Tang have put the stress on impact generated by KOS evolution on mappings [Tang and Tang, 2010]. They propose to calculate a minimal set of changes at ontology evolution time by analysing the TBox and ABox, in order to better control impact on mappings where only removal is considered because changes in ontologies are not sufficiently described.

A better characterisation of KOS evolution allowed Martins & Silva to propose the SBO model to represent mappings which drive the removal of mappings [Cordeiro and Filipe, 2009]. Groß et al. have illustrated the correlation between ontology evolution and mapping evolution in the life sciences [Groß et al., 2012]. Based on this work they exploit evolution, in terms of complex changes, as well as existing mappings to adapt invalid mappings via adaptation rules [Groß et al., 2013].

3.2.4 Mapping representation

Approaches of this category mainly rely on the use of languages devoted to represent semantic alignments as well as user friendly interface to display the evolution of mappings over time. The main argument is to say that the complete automation of mapping maintenance is impossible and will always require human intervention so let's provide tools for experts to support them in their intervention. In consequence, systems like View Graph [Tang and Tang, 2010] or models for representing mappings [Qian and Dong, 2005] have been proposed.

As illustrated in this section, mapping maintenance has been only partially treated, in the sense that no approach is able to deal with it from end to end. Moreover, very few approaches take into account KOS evolution as the main cause of mapping invalidity. This has been the main argument that has driven our work on mapping maintenance. However, KOS evolution is poorly documented and hardly exploitable. A challenge is to define the right level of granularity to describe KOS evolution in order to be useful for mapping maintenance. This is the subject of the next section.

3.3 Understanding ontology evolution for adapting mapping

Our observations [Dos Reis et al., 2014c] clearly show that, besides the correction of errors generated at alignment time, evolution of KOS entities, especially concepts, is the main cause of turning mappings invalid. To this end, we assumed that understanding and characterizing this evolution at a fine level of granularity, in order to exploit this information to adapt out-dated mappings, is paramount in the maintenance process.

3.3.1 Empirical analysis of mappings and KOS evolution

The various experiments presented in [Dos Reis et al., 2014c], exploiting several successive versions of SNOMED CT and ICD-9-CM as well as their associated official releases of mappings, highlighted the correlation between the way KOS evolve and the way mappings behave over time. The undertaken experiments have consisted in analysing more than 300,000 mappings over a 3 years period of time and have required the design of efficient algorithms to compare the mappings, the concepts involved as well as the super and sub concepts of those defining mappings of two KOS that are expressed under various knowledge representation models. It has also required the design of a specific database able to both store all the material (e.g. KOS and mappings) and support the designed algorithms. This important work, recognised as a

significant contribution to the domain, has put the stress on several factors that play a key role in the mapping maintenance problem.

First of all, mappings having *equivalence* and *more general* as relationship are the ones most frequently affected by KOS evolution. Moreover, the increasing number of mappings of these types reveals that the knowledge represented in SNOMED CT is becoming more specific than the one of ICD-9-CM. This is all the more so true if one analyses the size of these KOS over time (SNOMED CT is growing faster than ICD-9-CM)

Second, the addition of new mappings over time is strongly correlated with the addition of new concepts (considered as an atomic change [Klein and Noy, 2003]) in the underlying KOS. However, this type of changes in KOS requires the use of matching techniques to create new mappings rather than an adaptation of existing ones. The challenge consists in distinguishing between addition of concept as a result of an atomic change and that involved in a complex change (e.g. case where a concept at time t is split into several other new concepts at time $t+1$). In a similar way, the removal of mappings is mainly due to the removal of concepts (or changes in the status of the concept).

Third, the neighbourhood of a concept (set of concepts directly connected to the one involved in a mapping) does not seem to play a key role for the mapping maintenance process but in specific cases, especially when considering evolution of super concepts, mappings can be affected. Moreover, a combination of atomic changes in mappings (e.g. the creation of a mapping as a result of the deletion of another one) due to changes in concepts linked by the structure of the KOS has also been highlighted by our experiments.

3.3.2 The role of concept definition in mapping adaptation

Our first quantitative analysis of KOS evolution [Dos Reis et al., 2014c] has established a correlation between the way KOSs evolve and the way impacted mappings are modified. Although promising, the quantitative analysis of KOS evolution with respect to mapping maintenance fails to understand which information defining KOS entities triggers a modification of the associated mappings when it evolves. To bridge this gap, we zoom into different cases taken from our observations to better understand the nature of the conceptual information involved in the mapping evolution process [Dos Reis et al., 2013a].

A closer analysis of specific cases that appeared in our experiments reveals that complex changes like splitting of concepts are very frequent and have a huge impact on mappings. Merging concepts happens much less frequently, probably due to the fact that medical knowledge is becoming more and more precise over time, thus leading to a refinement of the domain and consequently a splitting of concept at KOS level rather than a merge. More importantly, as illustrated in figure 3.2, even if concepts are mapped in their entirety, mappings are defined based on partial information that describe concepts. It means that only some concepts' attributes values are defining mappings.

The concrete case depicted in figure 3.2 reveals that the creation of concept 560.32 of ICD-9-CM, as the result of the splitting of concept 560.39 (the value of the *notes* attribute of 560.39 become the value of the *title* attribute of the newly created concept 560.32), does not impact all the associated mappings. Actually, only mappings that involve concepts of SNOMED CT having "Fecal impaction" in their description are impacted since the former concept of ICD-9-CM (e.g. 560.39) has this information moved to a new concept in the next version of the KOS (e.g. 560.32). The same conclusion about the definition of mappings can be drawn from the analyses of complex changes showing a merging of concept or variations of splitting as detailed in [Dos Reis et al., 2013a]. This analysis has drastically impacted our work in the sense that we

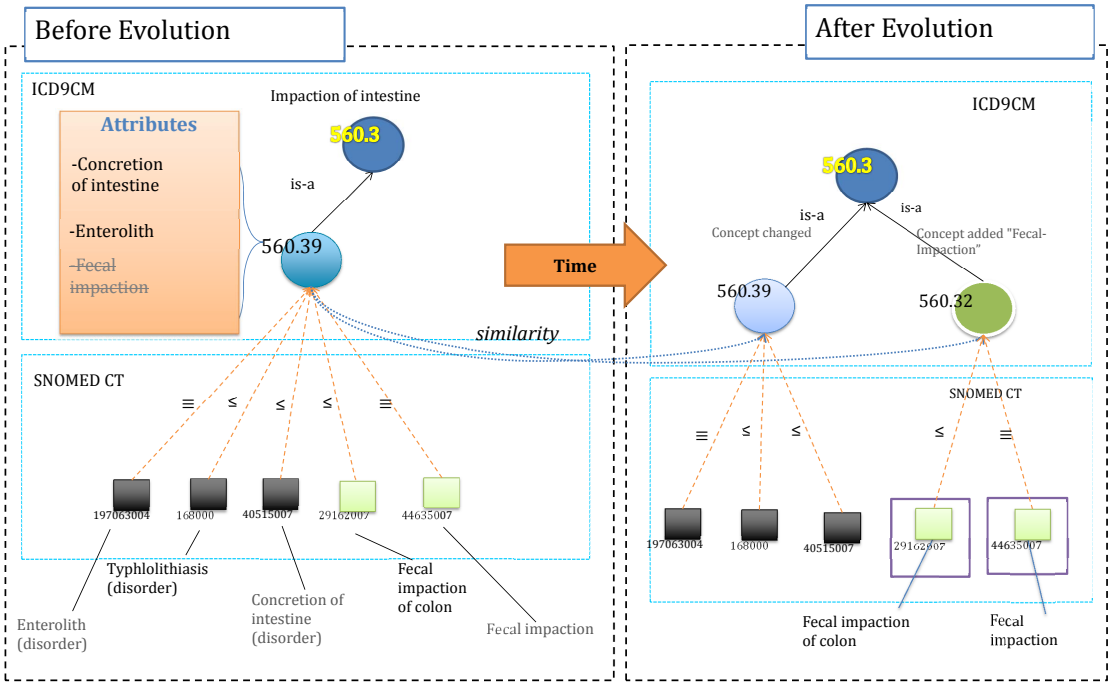


Figure 3.2: Impact of concept splitting on associated mappings

have reduced the complexity of the mapping maintenance process by analysing the modification of concept attributes instead of considering the concept as a whole.

3.3.3 Identification of relevant dynamic knowledge to adapt mapping

The outcomes of the experiments previously evoked forced us to question the definition and interpretation of mappings since they are correlated with the evolution of KOS in the maintenance process (as shown in figure 3.2). Actually, we had to be able to find the conceptual information that defines mappings (e.g. concept's attributes), because the modification of this information only is of greater importance with respect to the maintenance problem. Unfortunately, the definition of mappings (see section 3.1) does not mention the information that was exploited at alignment time by matchers to establish mappings, which makes the maintenance process an even more complex task.

From a mapping maintenance perspective, we had to find a way to retrieve this information before evaluating its evolution. To do so, we studied the existing ontology matching techniques [Shvaiko and Euzenat, 2013] to identify how information describing concepts is exploited to align ontologies. It appears that most of the approaches exploit syntactic aspects as well as the structural properties of concepts and the information that describes them. Based on this observation, we designed an algorithm, called *TopA* (cf. algorithm 1), relying on similarity measures, to identify textual information that define mappings [Dinh et al., 2014a]. The intuitive idea behind this consists in comparing, from a lexical, syntactic and semantic point of view, the value of the attributes that describe the source and target concepts of a mapping, and in keeping those values that are the more similar from these perspectives as candidates for interpreting mappings. The evolution of the found attributes is then further evaluated for mapping maintenance purpose (cf. section 3.4).

The various comparisons are done using several metrics addressing the lexical, word composition and semantic aspects. To compute the lexical distance between the value of the considered attributes, we used the Levenshtein distance known as the minimum number of single-character edits (i.e. insertions, deletions or substitutions) required to change one string into the other, since a lot of modifications in concepts' attributes consist in minor changes like singular/plural or changes from capital to lower case letter. Such kinds of changes have a very limited impact on the semantics of the concepts which, in turn, does not require a re-evaluation of the mapping. To overcome some limitations of the Levenshtein metric, we also analysed words composing the string. For instance *"skin cancer"* and *"cancer of the skin"* must be identified as equivalent so we use word-based edit-distance [Maedche and Staab, 2002]. Last, string and word analysis is not sufficient to grasp the semantics of the described notion. Actually the words "malignant tumor" and "cancer" are denoting the same concept but are completely different from the syntactic point of view. This is why we used a normalized version of the Jiang-Conrath [Jiang and Conrath, 1997] metrics associated with the SemCor corpus of documents to measure the semantic distance between attributes values and bridge this gap.

Algorithm 1 Select top n attributes defining a mapping (denoted as **topA**)

Require: $m_{ST} = (c_S, c_T, semType_{ST}) \in M^0_{ST}; c_S \in Concepts(O^0_S); c_T \in Concepts(O^0_T), n \in \mathbb{N}$
Ensure: $SCA = \{(s_{a_1}, ct_{a_1}), (s_{a_2}, ct_{a_2}), ..., (s_{a_n}, ct_{a_n})\}$
1: $SCA \leftarrow \emptyset$; {Initialize the final result set}
2: {Compute similarity between attributes in c_i and c_j}
3: **for all** $a_p \in A(c_S)$ **do**
4: $maxSim \leftarrow 0$;
5: **for all** $a_q \in A(c_T)$ **do**
6: $s_p \leftarrow sim(a_p, a_q)$;
7: $SCA \leftarrow SCA \cup \{(s_{a_p}, \text{NOCT})\}$;
8: **if** $maxSim < s_p$ **then**
9: $maxSim \leftarrow s_p$;
10: **end if**
11: **end for**
12: **end for**
13: {Select attributes in context if exact matches are not found}
14: **if** $maxSim < 1.0$ **then**
15: **for all** $a_k \in A(CT(c_S))$ **do**
16: **for all** $a_q \in A(c_T)$ **do**
17: $s_k \leftarrow sim(a_k, a_q)$;
18: $SCA \leftarrow SCA \cup \{(s_{a_k}, ct_{a_k})\}$;
19: **end for**
20: **end for**
21: **end if**
22: $SCA \leftarrow sort(SCA, n)$; {Select top n attributes}

We are considering the relationship "equivalent to", "more specific than", "less specific than", "unmappable" and "partially related to". The *TopA* algorithm has been validated through the

analysis of existing mappings between SNOMED CT and ICD-9-CM. To this end, we have studied the impact of the different similarity measures considered through the correlation between changes in mappings (MAAs) [Dos Reis et al., 2013b] and modifications in candidate attributes (OCOs) [Hartung et al., 2013]. The conducted experiments allowed us to evaluate our algorithm and the performance of the selected similarity measures. The obtained results show that the considered similarity measures are highly relevant to cope with the identification of relevant information defining mappings [Dinh et al., 2014a]. However, from a more pragmatic point of view, we decided to let knowledge engineers in charge of the mappings maintenance process specify the metrics they want according to the intrinsic properties of the labels used to describe KOS elements.

Moreover, a closer analysis of our study's results reveals that attributes located in the context (i.e. attributes describing the the super, sub and sibling concepts) of the considered source concept do not interfere in the definition of mappings with the target concept [Dos Reis et al., 2014b]. This result is important, since it reduces the complexity of the maintenance process by considering the evolution of the attributes of the source concept only.

This work clearly underlines the need, for maintenance purposes, to preserve information that served to align KOS. This will speed up the mapping maintenance process and, most importantly, will optimize the quality of the process [Dos Reis et al., 2014a].

3.4 Adapting mappings according to ontology evolution

The identification of the conceptual information defining mappings was the first step of our approach for maintaining mappings. The adaptation of mappings itself consists in the next step. To achieve this goal, we had to:

1. Characterize the evolution of the values of relevant attributes that define mappings over time,
2. Define the various actions that can potentially be applied to mappings in order to preserve their validity,
3. Link the information gained from the evolution of concepts' attributes values with the right mapping adaptation actions.

In the forthcoming sections, the methods and techniques we have proposed to address these issues are detailed.

3.4.1 Change patterns for mapping adaptation

The characterization of the KOS evolution has already been investigated by the Semantic Web community. Relevant approaches aiming at identifying patterns explaining changes in ontologies by observing the evolution of concepts have been proposed [Klein and Noy, 2003, Djedidi, 2009]. However, these approaches tackle the problem at a level of abstraction that does not match the requirement imposed by the mapping maintenance problem. As explained in the previous sections, attributes' values evolution is the engine of mapping adaptation. Therefore, existing patterns must be refined to deal with this particularity. This has been done based on the qualitative analysis of the evolution of KOS [Dos Reis et al., 2013a]. Following this study, we came up with a set of 8 change patterns (CP) that allow the characterization of evolution of attributes' values, 4 of them addressing the lexical aspect of the attributes values and 4 to deal with their semantics [Dos Reis et al., 2015a, Dinh et al., 2014b].

Lexical Change Patterns (LCP). We propose this kind of CPs to describe the lexical changes that may affect attributes' values over time. Since mappings are defined according to attributes' value, such patterns allow us to identify which concept this relevant information is attached to after evolution. We defined 4 types of LCPs namely Total Copy (TC), Partial Copy (PC), Total Transfer (TT) and Partial Transfer (PT).

Total Copy denotes the type of change where the whole value of an attribute of a concept is copied to another attribute of another concept. For instance, an attribute a_1 of a concept c_1 has as value *"portal systemic encephalopathy"* at time j, at time $j+1$, a_1 still has the same value, but in addition, another attribute a_2 of a concept c_2 will have as value *"portal systemic encephalopathy"*.

Partial Copy consists in a copy of a part of a given attribute's value to another attribute. For instance, an attribute a_1 of a concept c_1 has as value *"familial hyperchylomicromenia"* at time j and, while a_1 keeps the same value at time $j+1$, an attribute a_2 will have *"familial chylomicromenia"* as new value.

Total Transfer formalizes the transfer of the totality of an attribute's value to another attribute at KOS evolution time. In contrast to TC, the original attribute does not keep the considered value and it is deleted.

Partial Transfer characterizes the transfer of a part of a concept attribute's value when this one evolves from one version to the next. For example, an attribute a_1 can have as value *"eye swelling"* at time j, at time $j+1$ this value is deleted from a_1 but another attribute a_2 may have *"head swelling"* as new value (i.e., the term *"swelling"* is transferred from a_1 to a_2 between t and $t+1$).

Semantic Change Patterns (SCP). In addition to LCPs that allow to describe the morphosyntactic way that attributes' values defining mappings evolve over time, SCPs describe how these attributes evolve from a semantic point of view. This means that concepts denoted by attributes can remain equivalent to their previous version or can become more or less specific during the KOS evolution which, in turn, impacts the semantic relationship of the underlying mappings. We defined four types of SCP to know, Equivalent (EQV), More Specific (MSP), Less Specific (LSP) and Partial Match (PTM).

Equivalent states that even if lexical modifications affect an attribute value at KOS evolution time, the resulting value remains equivalent to the one before evolution. For instance, an attribute can have its value changed from *"Diabetes type 1"* to *"Diabetes type I"* without having its semantics modified.

More Specific refers to a CP that allows to identify a change affecting one attribute value to make its original version more specific than the new one. For instance, the change leading from *"kappa light chain disease"* to *"kappa chain disease"* makes the first one more specific because of the word *"light"* that specifies the type of *"chain disease"*.

Less Specific describes the contrary proposition of MSP, which is making the original version of an attribute value less specific than its evolved version.

Partial Match stands for a CP that characterizes the result of an evolution where the evolved attribute version remains semantically related, but this relation is not like the previously defined SCP types. For instance, taking the original attribute value *"focal atelectasis"* and its evolution *"helical atelectasis"*, these two attribute versions refer to the notion of *"atelectasis"*, but both cannot be considered as equivalent or one more or less specific than the other.

The so-defined change patterns have been formalized in First-Order Logic for homogeneity

reasons, since the other components of the global approach (see following sections) respect the same formalism. The proposed patterns have been designed in accordance with the observation of the evolution of biomedical KOS, therefore it represents the most frequent changes. Nevertheless, the specificities of the existing KOS cause other patterns to be found. This is why we have designed our approach in a way that knowledge engineers, in charge of the maintenance, can specify other patterns and integrate them in the framework (cf. Section 3.5). As we are considering attributes' values evolution, the change patterns are not strictly dependent on the KOS models. It means that the fact that the KOS is an ontology or a taxonomy or a thesaurus does not interfere in the definition of the patterns, which allows to cope with the diversity of knowledge representation models in the biomedical domain. Moreover, our change patterns are accompanied by an algorithm that allows their identification giving two successive versions of the same KOS [Dos Reis et al., 2015a]. The algorithm has been validated using a corpus containing a set of 650 instances of change patterns identified by domain experts after series of training sessions. The validation has consisted in measuring precision, recall and F-score of the algorithm (i.e. by confronting the results returned by our algorithm with those provided by the experts). The validation of our concepts would have required a "gold standard" for benchmarking, but the domain and the specificities of the addressed problematic make it difficult to create one, hence the need to define our own corpus of reference.

3.4.2 Definition of mapping adaptation actions

Complementary to the CPs, mapping adaptation actions (MAAs) deal with the types of operations to perform on the mapping itself to change its elements (i.e. source concept, target concept or semantic relationship). Although an exhaustive list of changes that can be applied to mappings can be provided according to the model of KOS and mapping, we decided to define MAAs with respect to empirical observations on the way mappings behave over time [Dos Reis et al., 2013b] to cope with real world applications. Actually, there are some changes at mapping level that can happen in theory, but never show up in practice; and we put focus on mapping adaptation actions that can be observed in practice. Our research has formalized, in the same type of Logic that has served to represent CPs, the following six MAAs: *AdditionM*, *RemoveM*, *MoveM*, *DeriveM*, *ModSemTypeM* and *NoAction*.

$AdditionM(m_{st})$ and $RemoveM(m_{st})$ stand for atomic actions through which a mapping is added, respectively deleted, at mapping evolution time. These two MAAs further serve to define the following composite actions. This kind of actions are usually observed when new or obsolete concepts, independent of the previously defined change patterns, are added in or removed from the KOS when it evolves.

$MoveM(m_{st}, c^1_{cand})$ is observed when the source concept c^0_s of a mapping is replaced by another one c^1_{cand} because of the changes that have affected c^0_s. This usually happens when an attribute that defines mappings is transferred to another concept. The associated mappings therefore follow the attribute they are attached to, and the new source concept will be the one that is described by the transferred attribute.

$DeriveM(m_{st}, c^1_{cand})$ differs from MoveM on the fact that the original mapping is kept, but a new one is created with the same semantic type (semType) and c_t interrelating a source concept c^1_{cand} different from c^0_s. This case can be observed when the information defining mappings is copied in the description of several concepts. As a consequence, new mappings can be defined with the concepts that are enriched with the copied information as sources in addition to the original one.

$ModSemTypeM(m_{st}, semType)$ consists in a modification of the semantic relationship of

the mapping (semType) engendered by the modifications performed on the source and/or target concepts c_s and c_t [Groß et al., 2013]. For instance, an attribute value can be enriched with adjectives making the concept more specific; if a mapping associated with this concept has an equivalent as semType, the source concept will become more specific than the target concept by virtue of the additional information now describing the source.

$NoAction(m_{st})$ is performed when the modifications affecting source or target concepts do not impact the mapping's semantic. This is usually observed when the changes identified in concepts do not affect attributes the mappings rely on, or when the changes do not modify the semantics of these attributes values.

3.4.3 Heuristic-based approach to maintain mappings valid over time

The proposed heuristics aim at linking the conditions that must be satisfied to adapt a given mapping with the adequate actions to perform on the elements of such mapping, to maintain its validity over time. We have designed the heuristics based on the outcome of the undertaken experiments to observe mappings' behaviour, as well as to examine the impact of change patterns on the way mappings evolve (i.e. the MAAs applied). We define the set of proposed heuristics according to the possible manners in which mappings evolve (i.e., the different MAA types). Challenges in the design of heuristics lie in the identification of the right set of conditions with respect to the appropriate mapping adaptation actions to perform.

Move mappings (MoveM). Our experiments have shown that *MoveM* is usually associated with the presence of LCPs between attributes of different concepts. More specifically, there exists one and only one relevant attribute of the source concept c_s at the time j we identify a LCP with an attribute from one concept of the context of its evolved version. Moreover, we apply the *MoveM* action when one and only one candidate concept exists where only one LCP can be identified without having SCPs with concept c_1^s. Intuitively, this means that the mapping is following the attributes that better define it over time.

Derive mappings. Similar to the *MoveM* action, we apply the *DeriveM* action when several LCPs are recognized between evolved versions of concepts' attributes. Moreover, the original mapping is preserved, thus c_s must be present in the new version of the ontology. The fact that c_s contains LCPs with several candidate concepts allows the creation of new mappings between the candidates and c_t, resulting in an enriched set of mappings. Intuitively, we derive a mapping when the information that describes its source concept is copied (totally or partially) to another concept.

Modification of semantic relationship of mappings. Applying the *ModSemTypeR* action depends on SCPs found between attributes. We propose two scenarios for modifying the type of semantic relationship in mapping adaptation.

- The first situation deals with the modification of the relation of the original mapping m_{st}^0. In this case, c_s remains the same, but the semantic relationship between c_s and c_t changes at time $j+1$.
- The second situation happens after a *MoveM* or *DeriveM*

In the first case, the new semantic type links the evolved source concept at time $j+1$ (in terms of content) and the target concept, while in the second case, c_s is replaced by a candidate from

the context. We determine the type of semantic relationships by combining the original *semType* from the given mapping and the type of a SCP detected between involved attributes.

Removal of mappings. *RemoveM* is applied to a given mapping when for all relevant attributes identified, no change patterns are detected within the context. This also demonstrates that for all relevant attributes, no SCP must be recognized with the evolved version of c_s (if the concept still exists). When $c_s^1 \in C(Onto_X^1)$ or c_s is assigned to obsolete, we consider that all attributes belonging to c_s are deleted. In case where $c_s^1 \in C(Onto_X^1)$, we get the candidate concepts at time $j+1$ based on the context of c_s at time j. Moreover, our experiments underlying the heuristics design observed the similarity between relevant attributes with concept attributes in context [Dos Reis et al., 2014b]. In particular, we calculated the similarity especially when the mappings were removed and the results revealed that the similarity remains very low. Therefore, we introduce a condition related to the similarity in the heuristics for *RemoveM*.

No action applied to mappings. Similar to *RemoveM*, the heuristic for *NoAction* also relies on the fact that adequate LCPs and SCPs cannot be recognized observing the evolution of concepts involved in mappings. This heuristics addresses the situations where KOS changes (impacting a concept involved in a mapping) do not affect relevant attributes that define the considered mapping, or the similarity with new attributes in context remains low [Dos Reis et al., 2014b].

We aimed to analyze to which extent the proposed heuristics are correct and lead to adequate mappings [Dos Reis et al., 2015b]. More specifically, this evaluation aimed to examine whether the identified changes at KOS level (i.e., change patterns) have triggered the right mapping behaviour (i.e., MAAs). To this end, we used successive versions of biomedical KOS (i.e. SNOMED CT and ICD-9-CM) as well as their official associated versions of mapping sets. In contrast with other elements of the approach, we had to deal with the specificities of the heuristics since *MoveM*, *DeriveM* and *ModSemTypeR* actions influence two parameters unlike the other actions. Actually, we measure the precision, recall and F-score of the decisions taking into account the right actions and the right parameters (e.g. considering a *DeriveM*, this implies the selection of this action with the correct new target concepts). In this context, we obtained satisfactory results, since the global F-score reaches a global measure of 85% underlying a good performance of the proposed heuristics and the method to apply them [Dos Reis et al., 2015b]. Nevertheless, the nature of the considered heuristics (the number of parameters given as input) reduces the precision and recall.

3.5 The DyKOSMap framework

The design of the DyKOSMap framework [Dos Reis et al., 2012] aims at putting together all the pieces of the approach for maintaining mappings valid over time. As depicted in figure 3.3, it combines:

- The identification of changes at KOS level that can potentially invalidate mappings and their characterization as change patterns. This also includes the identification of the relevant attributes' values that define the mappings,
- The definition of the potentially out-dated mappings, especially the semantic relationship that links source and target concepts, before their adaptation,

- The heuristics that connect KOS changes, outdated mappings and the potential mapping adaptation actions.

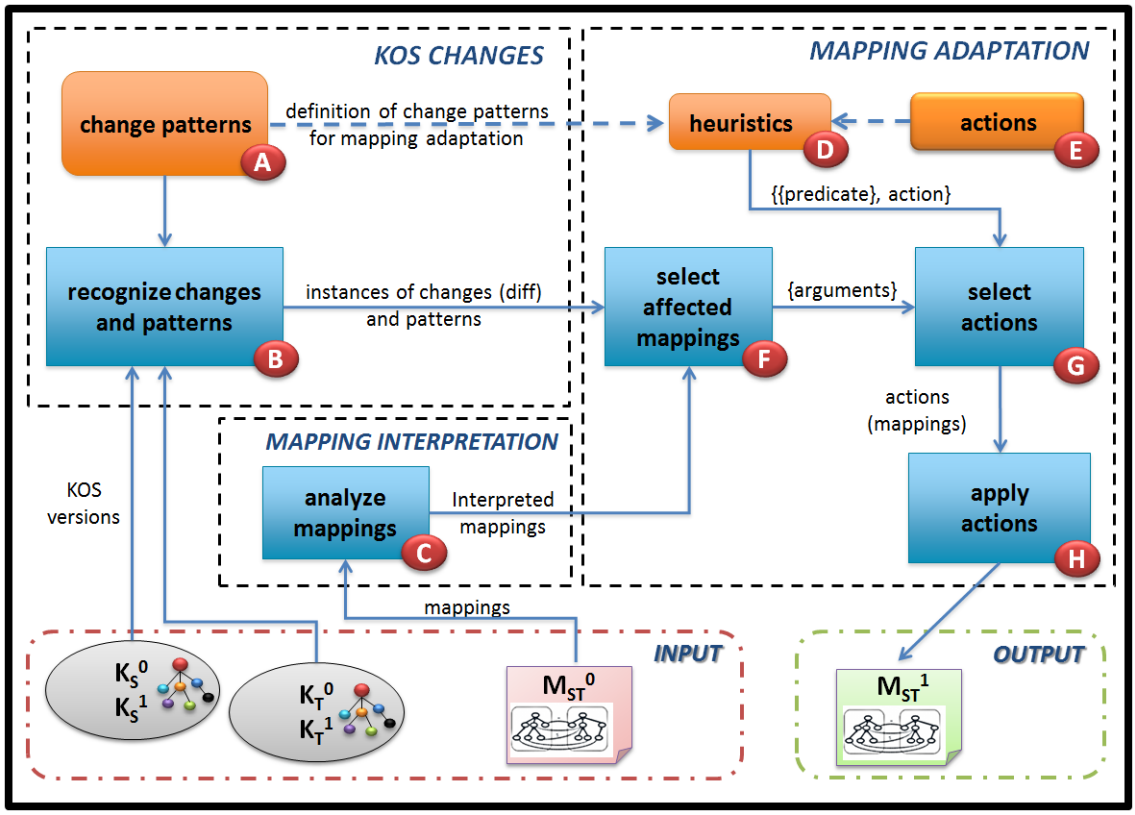


Figure 3.3: The DyKOSMap Framework

The implemented process for maintaining mappings can be split into four major steps that can be described as follows:

1. The first step consists in identifying instances of change patterns explaining the evolution of KOS (cf. A and B in figure 3.3),

2. The second step aims at identifying mappings that can potentially be invalid. These are mappings that rely on concepts that are involved in identified change patterns (cf. C and F on picture 3.3).

3. The third step deals with the adaptation of invalid mappings. This is done through the selection of the appropriate heuristics that link the change patterns of step 1 with the correct mapping adaptation action (cf. elements D, E and F on figure 3.3).

4. The last step implements the selected mapping adaptation actions to obtain a valid set of mappings (cf. G and H in figure 3.3).

The DyKOSMap framework presented in this section has been fully implemented. The obtained tool has been used to generate the experimental results that have served to evaluate the proposed approach.

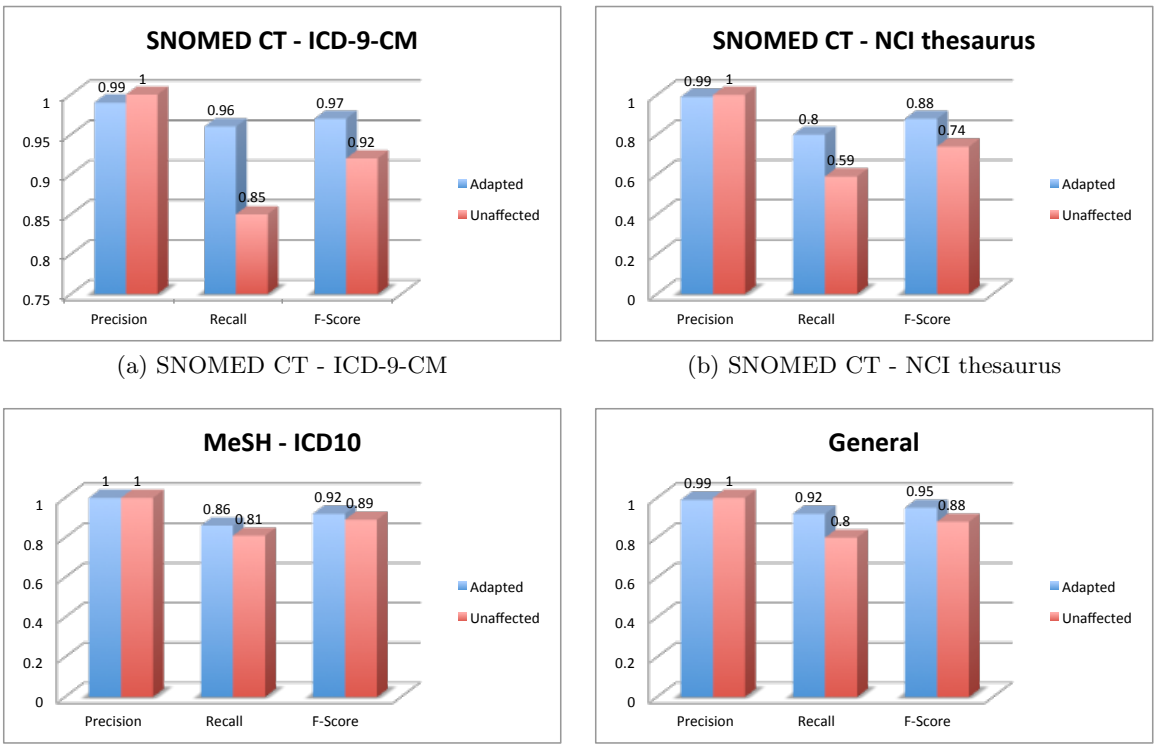
3.6 Experimental assessment

The validation of the approach defined through the DynaMO project has been done in an incremental manner. Instead of validating only the final approach, we have actually decided to evaluate each concept that has been defined (e.g. change patterns, topA, mapping adaptation actions and heuristics). As each component of the approach is really important it was essential to validate them separately. Moreover, evaluating only the whole approach would not have allowed us to identify what component would have been problematic in case of poor experimental results. In order to have significant results that allowed us to draw a strong conclusion highlighting the strengths and weaknesses of the approach, we used only official (and therefore validated) mappings and releases of their respective KOS. In addition, we have used a minimal set of 100,000 mappings and a minimum of two major KOS for each experiment on which we have defined scientifically rigorous and reproducible protocols.

However, we have proposed a set of experimentations to show the added value of the DyKOSMap framework for adapting mappings turned invalid by KOS evolution. We used five large biomedical KOSs: SNOMED-CT, MeSH, ICD-9-CM, ICD-10-CM and NCI Thesaurus. For the evaluation, for each 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 adaptation method. 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 evolution. Moreover, we remove all mappings identified as conflicting from the reference mappings (i.e., cases where KOS evolution impacts both source and target concepts).

Results show a very high performance rate of our mapping maintenance approach according to the computed F-Score. We observe that our approach is extremely efficient in identifying mappings that do not need to change which, from a more pragmatic point of view, is highly relevant for domain experts who have to focus only on ambiguous adaptation proposition. However, the adaptation of mappings through the other actions seems to be less efficient. This is due first, to the number of parameters that have to be correctly identified to take the appropriate decision and, second, to the definition of the heuristics that are more or less ad-hoc and deserve further investigations. This will be done by designing a machine learning approach for identifying in an automatic way the heuristics using the available sets of KOS and mappings. We will then be able to compare the resulting heuristics with those presented section 3.4.3 to conclude about their respective quality.

On the other hand, the quality of the used sets of reference is questionable. In our first analysis [Dos Reis et al., 2014c], we actually observe that a proportion of mappings disappears from one release to the next, probably due to alignment errors that have been manually corrected at validation time by domain experts. Moreover, the poor diversity of semantic relationships found in the reference datasets also biased the results especially regarding the *ModSemTypeR* actions, since the new relationships proposed by our algorithms cannot be found in the gold standards which does not mean that the proposition is wrong. To overcome this limitation, we will work in collaboration with domain experts, and depending on their answers our approach will consist in an enrichment of the existing mappings.



(a) SNOMED CT - ICD-9-CM

(b) SNOMED CT - NCI thesaurus

(c) MeSH - ICD10

(d) General

Figure 3.4: Results of the evaluation of the DyKOSMap approach

3.7 Summary

In this chapter, we have presented our work to address the maintenance of mappings turned invalid by KOS evolution. In this context, the originality of our work lies in the combined used of information derived from KOS evolution and existing mappings to keep their semantic validity over time [Dos Reis et al., 2014a]. Through our experiments, we have shown that even if concepts are mapped in their entirety, only partial conceptual information (e.g. attributes) of the source and target concepts is used to define mappings and the evolution of this particular information is important to understand and characterize the way to adapt mappings. To this end, we have proposed algorithms (TopA) to detect conceptual information defining mappings and change patterns at syntactic and semantic level to characterize how this information evolves. We then have combined instances of these change patterns with mapping adaptation actions in heuristics to drive the adaptation of out-dated mappings. The approach has been implemented in the DyKOSMap framework and evaluated using dedicated biomedical KOS versions and their associated successive sets of mappings.

The proposed approach takes into account the specificities of the KOS of the biomedical domain, especially the characteristics of the label of the elements that are very descriptive in contrast with those of other domains (cf. ontologies of the Semantic Web). It consequently makes sense to rely on their syntactic and linguistic aspects to characterize their evolution. However, for labels that are not so rich from the evoked point of view, additional features need to be considered to understand the evolution of concepts. Moreover, OWL ontologies of the Semantic Web also allow the definition of axioms expressed in description logics to describe

concepts' semantics, and it is obvious that linguistic aspects cannot be exploited to explain their evolution but rather a logic-based approach. These are issues that need to be addressed to be able to apply our approach to KOS of other domains.

As evoked at the beginning of this chapter, the biomedical domain heavily relies on the use of KOS for semantic interoperability reasons. This is the case for enriching existing data in semantics using KOS concepts to enable computer systems to retrieve relevant information for decision support purposes. In this context, it is crucial that semantic annotations remain up-to-date with respect to the latest release of the KOS in use. As a direct follow-up of the DynaMO project, we will investigate issues linked to the adaptation of semantic annotations turned invalid by either changes in the annotated documents or by modification in KOS entities definition. We plan to extend and adapt the DyKOSMap framework to this particular problem.

The work presented in this chapter has been carried out in the framework of the DynaMO research project entirely funded by the Fonds National de la Recherche (FNR) Luxembourg. This four years project (May 2011- April 2015) for which I was Principal Investigator has covered the doctoral thesis of Julio Cesar Dos Reis and part of the post-doctoral project of Duy Dinh. DynaMO was also developed in collaboration with the LRI of Paris-Sud University and Pr. Chantal Reynaud-Delaître.

Chapter 4

On the quality of dynamic medical knowledge: the cases of biomedical Knowledge Organization Systems and mappings

Contents

4.1 Problem statement and hypothesis 50
4.2 Related work 51
4.2.1 Quality of Knowledge Organization Systems 51
4.2.2 Methods and tools for ontology validation 53
4.3 Analysing the quality of Knowledge Organization Systems and semantic mappings 54
4.3.1 Method to compare Knowledge Organizing Systems content with existing mappings 55
4.3.2 Impact on ontology mappings 58
4.4 On the validation of medical ontologies 59
4.4.1 Verbalizing the content of medical ontology 60
4.4.2 Interpreting experts feedback and modifying the ontology 63
4.4.3 Experimental assessment 64
4.5 Summary 65

As shown in the previous chapters, the biomedical domain is complex and highly dynamic in several aspects, but it is also a critical domain. Actually, and contrary to other domains that deal with entertainment, culture or even finance, medicine involves patients. It means that a wrong decision leveraged by information systems can simply lead to irreversible and undesirable effects on patients. It is therefore vital for information systems to rely on and exploit reliable data and knowledge to assist health professionals taking the best decision possible regarding their patients' health status. Since ontologies provide decision support systems with the capability to reason and retrieve, manage or share heterogeneous data, they must be consistent from the conceptual and logic points of view. They must also not be contradictory if several ontologies are used in combination (i.e. if two ontologies representing the same domain are compared,

the same conclusions and concepts must be found). The size of existing standard ontologies of the biomedical domain suggests that these have been designed using automated approaches [Dolinski and Botstein, 2013, Pruski et al., 2011a] or are the results of a long and tedious design task (cf. MeSH or SNOMED CT), and the resulting conceptualization must be validated before being implemented in real situations. In this context, domain experts play a key role since they are deeply involved in the validation process. However, their competencies in logic-based knowledge representation methods are usually extremely limited, which does not facilitate their interaction with the ontologies to be validated while existing tools such as Protégé are not very friendly in supporting this task. Methods and tools for presenting an ontology's content in a way that can be understood by experts are undoubtedly necessary to optimize the quality of medical ontologies and, consequently, that of the engendered decisions. This chapter presents our approach for evaluating the quality of KOS and mappings as well as a system aiming at simplifying the validation of dynamic medical ontologies and mappings through question/answering techniques involving medical experts.

4.1 Problem statement and hypothesis

The quality and availability of existing knowledge is a critical aspect when implementing Decision Support Systems for the biomedical domain. As domain knowledge is usually represented by KOS , their quality as well as that of depending artefacts, especially mappings, has been the focus of the work presented in this chapter, since KOS are both gaining in importance in clinical decision support systems and have a direct impact on patient health. While the majority of publicly available KOS and associated material claim to be validated, our analysis of official mappings [Dos Reis et al., 2014c] revealed that errors are still present and impact the underlying information systems. For instance, if wrong mappings interconnect unmatchable concepts, irrelevant documents can be retrieved during a semantic-based search, hence influencing health professionals. Based on these observations, we found it important to tackle this issue. Moreover, this work has been motivated by the demand of the Luxembourgish health professionals with regard to the forthcoming telematic health platform and the need to make semantic searches, using domain-specific vocabulary provided by standard KOS, in medical data.

To address the quality of medical KOS and depending artefacts, we have decided to approach the problem under two different, but complementary, facets:

- First of all, it is important to be able to (i) evaluate to which extent available KOS and associated artefacts, especially mappings, are valid from the logic and conceptual point of view and (ii) what is the overall quantity of inconsistencies that exist between overlapping KOS when comparing their respective content. To achieve these objectives, the definition of a rigorous method that allows comparing, measuring and categorizing the differences that appear between related KOS of the medical domain, is required.

- Second, we hypothesize that the quality of KOS content and the definition of associated mappings can be significantly improved if domain experts are appropriately involved in the validation of the content of KOS and semantic mappings, either at construction time or at evolution time. This hypothesis has been identified during the work we have carried out in the representation of dynamic medical knowledge presented in chapter 2 addressing the automatic generation of OWL ontologies to represent care actions contained in clinical guidelines (cf. Section 2.3.2).

Taking the above motivations into consideration, the research questions dealing with the evaluation and validation of the quality of data and knowledge we have targeted in this work can be expressed in the following way:

How to evaluate the quality of the content of a KOS (in terms of concepts and relationships) and its associated mappings in a network of KOS?

How to validate or modify the content of a KOS involving domain experts unfamiliar with logic-based knowledge representation languages?

In the following section and subsections we will briefly present the state-of-the-art of the two fields ontology quality and ontology validation. This will allow us to better highlight our research approach and our scientific contributions.

4.2 Related work

The existing approaches that are presented in this section are categorized into two different fields, each of which is related to one of the previously formulated research questions. However, most of the existing work focused on ontologies instead of considering a larger variety of knowledge representation models like thesauri, taxonomies or database schemas. We start the literature review with the presentation of approaches dealing with methods and techniques for the evaluation of the quality of KOS [Gómez-Pérez, 2004]. After that, we will put an emphasis on the work devoted to the validation of KOS, with a particular attention paid to the conceptualization (i.e. the adequateness of the represented knowledge with respect to the domain) at two important moments of the life-cycle of the ontology: construction and evolution time.

4.2.1 Quality of Knowledge Organization Systems

In existing approaches, the quality of ontologies is essentially evaluated through the definition and use of different metrics, focusing on different criteria that are usually derived from the ontology and its characteristics.

First approaches tackling the quality of an ontology put the focus on statistical aspects aiming at evaluating the number of classes, the number of properties relating classes, the number of leaf classes and so on. This is the case of the work of Yao et al. [Yao et al., 2005]. However, this work tells very little about the quality of the ontology but reveals some information about its complexity which has been the subject of the OQuare framework [Duque-Ramos et al., 2011].

In their work [Baneyx and Charlet, 2006], Baneyx & Charlet have introduced, as results of the PERTOMED project, various relevant criteria for evaluating the quality of an ontology at various moments of the ontology life-time (construction, evolution and maintenance) with a particular emphasis on the biomedical domain and its specificities. Some of the criteria are dealing with the structure and logic aspects of the ontology, while others tackle the conceptualization of the represented domain. Among those concerning the conceptualization, they discuss the ontological commitment as essential. Actually, they advocate that when designing an ontology,

a minimal number of hypotheses on the represented domain must be made, in order to respect concepts of the real world as faithfully as possible. Moreover, the authors also put the stress on the usability of the ontology and its ability to fulfil the set of requirements it has been designed for.

In [Stvilia, 2007], the author has defined a model exploring 12 different criteria to evaluate the quality of an ontology. He also considers unavoidable criteria that can easily be measured with statistics exploiting explicitly defined ontological elements such as the number of classes or properties, subjective aspects like semantics and structural consistency. The novelty lies in the introduction of volatility as a criterion. This consists in evaluating the duration for which the ontology is valid by measuring the period of time elapsed between two successive updates and therefore takes into account (to a certain extent) the evolution of the considered ontology. As evoked in [Baneyx and Charlet, 2006], he also addresses, to a certain extent, the usability of an ontology by counting the number of applications that use it.

Djedidi and Aufaure have proposed an approach to assess the quality of an OWL ontology at evolution time [Djedidi and Aufaure, 2010]. To this end, they also proposed a set of quality criteria dealing with complexity, cohesion (e.g. average number of connected components), conceptualization (e.g. average number of object properties per classes), abstraction (e.g. maximum number of classes between the root and the leaves of the ontology), completeness and comprehension (i.e. number of annotated classes or individuals). Nevertheless, the proposed approach is clearly dependent on the OWL model since the implementation of the metrics relies on OWL primitives. Despite this interesting work, it can hardly be applied to biomedical ontologies due to their lack of logic formalisation.

Sabou and Fernandez have introduced two other dimensions to consider when evaluating ontologies [Sabou and Fernandez, 2012]. The first one consists in proposing criteria addressing the selection of an existing ontology instead of creating a new one from scratch. This makes sense because of the large amount of ontologies available through the Web. The second criterion deals with the modularization aspect of an ontology. They propose to evaluate modules that require to be combined for a given purpose or for a particular application, to allow deciding about the relevance of an ontology.

More recently, Rico et al. [Rico et al., 2014] have introduced the OntoQualitas framework. In this work, while no new type of criterion addressing the quality of an ontology has been introduced, the metrics to calculate them have been improved and refined (i.e. new aspects of the ontology are taken into account to calculate them). The authors also provided a concrete case study to assess the framework.

Since the emergence of the Semantic Web and the intensive use of semantic resources in many domains, especially in biomedicine, which has led to the development of repositories like Bioportal [Noy et al., 2009], OBO foundry [Smith et al., 2007] and HeTOP [Grosjean et al., 2011], a significant number of KOS are now available and ready to use in clinical Information Systems. In fact, we are currently facing a phenomenon of a knowledge overabundance, generating contradictions and inconsistencies in available KOS. All these resources seem to be constructed by ignoring existing ones, therefore redefining concepts but also relations and creating discrepancies between KOS. As a consequence, there is a need to identify these problems to enhance the quality of decisions or services that are developed on the top of semantic-enabled Information Systems, in particular of systems that are exploiting overlapping KOS. As evoked in the previous chapter, the size of the biomedical domain forces systems to rely on a combination of KOS interrelated via mappings. It is therefore of utmost importance that decisions derived via the exploitation of several KOS are consistent from the medical point of view (e.g. conceptualization of a domain)

to serve patient conditions.

4.2.2 Methods and tools for ontology validation

As explained in the previous subsection, there is a significant amount of work which may lead to a set of metrics dealing with the quality of KOS exploiting, most of the time, measurable properties of ontologies (e.g. number of classes, individuals and properties). This work has been completed by the design and implementation of methodologies and tools aiming at validating an ontology from various perspectives which intend to better involve end-users in the process.

In [Gangemi et al., 2006], a model for evaluating and validating ontologies has been introduced. Based on a meta-ontology called O^2 and semiotics, the authors have proposed to evaluate ontologies by considering structural, functional and usability-profiling measures with the intended use of the ontology to evaluate in mind. The validation is then complemented with an ontology called *oQual* aiming at providing necessary criteria to select an ontology according to a particular need.

In [Köhler et al., 2006], the authors have provided a rule-based methodology (i.e. set of conventions) to establish well-defined labels for Gene Ontology concepts. The rules aim at avoiding circular definitions (i.e. terms of the label that are also in the definition of the considered concept) and obscure language (i.e. labels of concepts must be understood by non-expert persons). Although interesting, this work can hardly be applied to other KOS because GO labels are very domain specific, since they are not only built on linguistic aspects but they use a lot of alphanumeric symbols to denote genes and proteins.

A similar argument is used in [Verspoor et al., 2009]. The authors have proposed a methodology to classify medical terms that use different linguistic conventions but denote the same meaning in order to standardize them. Such approaches address an important component of an ontology, to know the choice of the terminology to describe ontological elements. However, elements that are implicitly defined (e.g. concepts that are logically defined by inference) failed to be standardized.

Dimitrova et al. have designed the ROO tool for supporting domain experts designing OWL ontologies [Dimitrova et al., 2008]. It provides a controlled language interface and offers systematic guidance throughout the whole ontology construction process, with an aim at optimizing the quality of the resulting ontology. However, nowadays many ontologies are rather built automatically from the textual content of relevant documents or rather slightly modified by virtue of knowledge evolution, in particular because of the overabundance phenomenon described above. To this end, domain experts are mostly involved in the validation phase and less and less often from the beginning of the ontology life cycle.

The MoKI systems designed by Pammer is, to the best of our knowledge, the only work that addresses the problem of ontology validation by means of question-answering techniques [Pammer, 2010]. The idea is to formulate questions from the content of the ontology and submit them to domain experts in order to get their feedback, leading to the validation or modification of the underlying ontology. However, the question formulation process does not integrate the fact that domain experts, in our context health professionals, are not supposed to be familiar with ICT logic-based formalisms. The questions generated by MoKI look very similar to description logic formulas hardly understandable by experts, therefore, the outcomes of the proposed system still require a substantial intervention of ICT experts to accompany domain experts through the validation process.

In their work, vor der Bruck and Stenzhorn have described a method to validate ontologies using an automatic theorem prover and MultiNet axioms [vor der Bruck and Stenzhorn, 2010].

To this end, the authors focus on the logic structure and ignore the conceptualization part. Therefore, their method requires formal ontologies expressed in logic-based languages to be applied, which is not always the case in the biomedical domain. The systems implementing the proposed algorithm are accompanied with a user friendly software interface to speed up the fixing of the detected erroneous axioms, facilitating users' intervention.

More recently, the OOPS! system [Poveda-Villalón et al., 2012] has been proposed by Poveda et al. It consists in detecting predefined anomalies or bad practices in ontologies to enhance their quality. However, the real world representation dimension, referring to how accurately the ontology represents the domain intended for modelling, is neglected in this approach and is left to the discretion of domain experts.

Other families of approaches addressing the validation of the domain-conceptualization side have been proposed. Some of them put the stress on user interface development in order to better present large amounts of (structured) data to support the validation effort without overwhelming users [Pohl et al., 2011], while other research work promotes the use of Natural Language Processing (NLP) techniques to better involve domain experts in the ontology validation process [Pammer, 2010].

As outlined in this subsection, most of the existing approaches dealing with ontology validation are focusing on the logic aspects (i.e. formalization and structure) or evaluate if the ontology has been designed with respect to good practices. Very little work has been devoted to the validation of the conceptualization of the domain represented in the ontology. The validation of this aspect is more difficult, since it requires domain knowledge that persons in charge of designing an ontology (knowledge engineers) usually do not have. Moreover, the conceptualization refers to the different views experts can have on the domain. To overcome this lack, knowledge engineers and domain experts have to collaborate to design the ontology. However, the communication between them is hard since they do not "speak the same language" so to say (knowledge engineers are speaking logic-based language while domain experts use domain specific vocabulary). Considering the above arguments, we believe that information systems should be used to reduce the exchanges between ontology designers. Thus, the way ontology content is verbalized and presented to domain experts is paramount for a validation of the conceptualization of the domain represented in the ontology at validation time. Moreover, experts' feedback must be adequately treated to make the ontology evolve in a consistent way regarding its logic formalization as well as the domain conceptualization.

4.3 Analysing the quality of Knowledge Organization Systems and semantic mappings

The experiments presented in our work on mapping adaptation [Dos Reis et al., 2014c] (cf. chapter 3) have revealed several interesting issues with respect to data quality and deserve closer attention. Actually, by only considering mappings, a significant amount of erroneous data exists among that is considered as official and validated. We observed that usually, a non negligible set of mappings is removed each time a new release of a KOS (in our investigations SNOMED CT) is published, and only alignment errors can explain this phenomenon. This is probably due to the used automatic matching techniques whose outcomes have not been sufficiently reviewed by domain experts. These observations forced us to question not only the quality of the mappings but also that of interrelated KOS, because they serve as basis for matching purpose (matching errors can be caused by bad concept definitions). To this end, we have proposed an approach

to compare the content of concepts in existing ontologies with concepts involved in mappings, to highlight the quality, in terms of detected contradictions, of the KOS and their associated mappings implemented in clinical decision support systems.

4.3.1 Method to compare Knowledge Organizing Systems content with existing mappings

Our work focused on the validation and evaluation of KOS. We define a KOS K^t at time t as a triple

$$K^t = (C^t, R^t, A^t),$$

where C^t is the set of concepts contained in K^t at time t, R^t is the set of semantic relationships that exist between concepts. It is a set of triples (c_1^t, c_2^t, r) where $c_1^t, c_2^t \in C^t$ and r^t is a symbol denoting the semantic relation between c_1^t and c_2^t (e.g. "is a", "part of", "equivalent to" ...). A^t the set of attributes that describe concepts at time t. Moreover, considering the definition provided in chapter 3 of a mapping

$$m^t = (id, c_s^t, c_p^t, n, r),$$

the idea to evaluate available distributed semantic resources and mappings in particular, aims at comparing the description of the source and target concepts involved in mappings (i.e. c_s^t and c_p^t), and the semantic relationship (i.e. r) that exists between them (cf. 3.1) with the semantic relationship that exists between concepts of KOS which are defined as equivalent to those of the mappings. The notion of time here is important since we must compare ontological elements at the same moment in time for consistency reasons. To do so, we assumed that the use of domain specific background knowledge will provide necessary and sufficient information to determine and confront the relationship of the mapping with the one that exists between concepts of the background knowledge. In the biomedical domain, existing repositories like Bioportal[11] or HeTOP[12] offer necessary resources of references and services to be used as background knowledge [Aleksovski et al., 2006]. Moreover, they also contain semantic correspondences that link the ontologies of the repository, making it possible to reason over several ontologies, thus reinforcing the probability to infer the semantic relationship that exists between ontological elements.

According to existing on-line tools as well as for implementation purposes, we have chosen to exploit the functionalities of and the content offered by the Bioportal software application. Bioportal is an open repository of biomedical ontologies that provides access, via Web services, Application Programming Interface and Web browsers, to ontologies developed in various formalisms such as OWL, RDF, OBO format and Protégé frames. It has the advantage of making it possible to search for concepts' preferred terms or synonyms in the stored ontologies, or to get mappings that could exist between these concepts to enlarge the terminological coverage [Noy et al., 2009].

The proposed approach is a process exploiting Bioportal search modules, the structures and properties of the stored ontologies as well as the mappings provided by the Bioportal's community. This approach has been implemented in the following algorithm (cf. Algorithm 2). The novelty of this algorithm lies in the use of ontology mappings to navigate through a network of ontologies to deduce semantic relationships. This approach assumes that the resources that are made available by these repositories are validated.

[11] `http://bioportal.bioontology.org/`
[12] `http://www.hetop.eu/hetop/`

Algorithm 2 Using Domain specific background knowledge to determine the semantic relationship between two concepts

Require: $Att_1 \subset K_1; Att_2 \subset K_2$ Two attributes of concepts belonging to K_1 and K_2 respectively

Ensure: $r \in \{$"$equivalent\quad to$","$more\quad specific$","$less\quad specific$", "$partially\quad matched\quad to$","$no\quad relation$"$\}$ The relationship that exists between concepts described by Att_1 and Att_2

1: $r \leftarrow$ "$no\quad relation$"; {Initialize the final relationship}
2: {Search for concepts having Att_1 and Att_2 as attributes}
3: $C_{Att_1} \leftarrow FindConcepts(Att_1)$
4: $C_{Att_2} \leftarrow FindConcepts(Att_2)$
5: $K_{common} \leftarrow FindCommonKOS(C_{Att_1}, C_{Att_2}, K_1, K_2)$
6: **if** $K_{common} = \emptyset$ **then**
7: {Exploit existing mappings}
8: $M \leftarrow getMappings(C_{Att_1}, C_{Att_2})$
9: **if** $M = \emptyset$ **then**
10: return r
11: **end if**
12: **end if**
13: {We have common ontologies we have to find the relationship}
14: **if** $areEquivalent(Att_1, Att_2, K_{common}) = true$ **then**
15: $r \leftarrow$ "$equivalent\quad to$"
16: **end if**
17: **if** $isMoreSpecific(Att_1, Att_2, K_{common}) = true$ **then**
18: $r \leftarrow$ "$more\quad specific$"
19: **end if**
20: **if** $isMoreSpecific(Att_2, Att_1, K_{common}) = true$ **then**
21: $r \leftarrow$ "$less\quad specific$"
22: **end if**
23: **if** $areSiblings(Att_1, Att_2, K_{common}) = true$ **then**
24: $r \leftarrow$ "$partially\quad matched\quad to$"
25: **end if**
26: return r

From a more pragmatic point of view, the algorithm is composed of the following three major steps:

1. **Search for concepts** (from statement 1 to 4). The aim is at finding the two attribute values (Att_1 and Att_2) we are looking for in the description of concepts contained in ontologies that are different from those the source and target concepts of the mapping (i.e. c_s^t and c_p^t) are coming from. From a technical point of view, we investigate, through the Bioportal search module, if there is an exact match between the attribute's values given as input and the description of concepts, especially the preferred terms and the set of existing synonyms (if available).

2. **Identify a set of common ontologies** (from statement 5 to 13). To find (or deduce) the semantic relationship existing between the concepts that have been found during the

previous step and that are described by the attribute's value given as the necessary input for projecting the found concepts in the same ontology (i.e. O_{common}). In consequence, two cases can be distinguished:

(a) **Direct method**. In this case, both attribute's values are found together in concepts belonging to the same group of ontologies excluding O_1 (e.g. both attribute's values Att_1 and Att_2 are found together in concepts of the same version of SNOMED CT, FMA or NCI Thesaurus).

(b) **Indirect method**. In this second case, no common ontologies are found with the direct method. Therefore, we exploit existing mappings to project the found concepts in a set of common ontologies. For example, one attribute is located in concepts of SNOMED CT and the other one in a concept of BFO. The goal is then to use mappings that exist to find equivalent concepts to the one described by the attribute given as input in common ontologies (e.g. the concept of SNOMED CT and the one of BFO are both mapped to concepts of NCI Thesaurus).

3. **Find the semantic relation that exists between the identified concepts in one of the common ontologies** (from statement 14 to 26). At this stage, we first select the ontology of the set of common ontologies to work with using various ontological properties. We rank the ontologies in order to work with the most detailed one. We assume that the more concepts an ontology contains, the most precise the relationship we are looking for will be. Second, we calculate the semantic path that exists in the common ontology between the two concepts and evaluate it. In our approach we distinguished four different cases:

   (a) The algorithm has projected, using direct or indirect methods, the two concepts identified at stage 1 on the same one. In consequence, the concepts are considered as equivalent since the changes did not affect the semantics of the initial version of the considered concept.

   (b) If the existing returned path contains only concepts that are linked through the subsumption relationship (i.e. *"is a"*) then one concept is said to be more specific than the other one. This means that the changes have made the initial version of a concept more general or more specific. The concept located at the highest level of the hierarchy (i.e. the concept that is closer to the root) is said to be less specific than the other. This can be inferred because the subsumption relationship is transitive by definition.

   (c) If the two concepts are siblings (i.e. they share the same super concept), the semantic relationship *"partially related to"* is returned. The semantics of this relationship are rather imprecise but are of interest for mapping adaptation as shown in our work [Dos Reis et al., 2013b].

   (d) Otherwise, no semantic relationship can be precisely determined and the *"undefined"* value is eventually returned. It doesn't mean that there is no relationship between these concepts, but the potential relationship cannot be identified with the proposed algorithm.

To illustrate the performance of the above algorithm, consider the example of figure 4.1. In our investigations [Dos Reis et al., 2014c, Dos Reis et al., 2013b] we have observed that the concept M0006899 whose label is "Pituitary dwarfism" of MeSH has evolved between the version

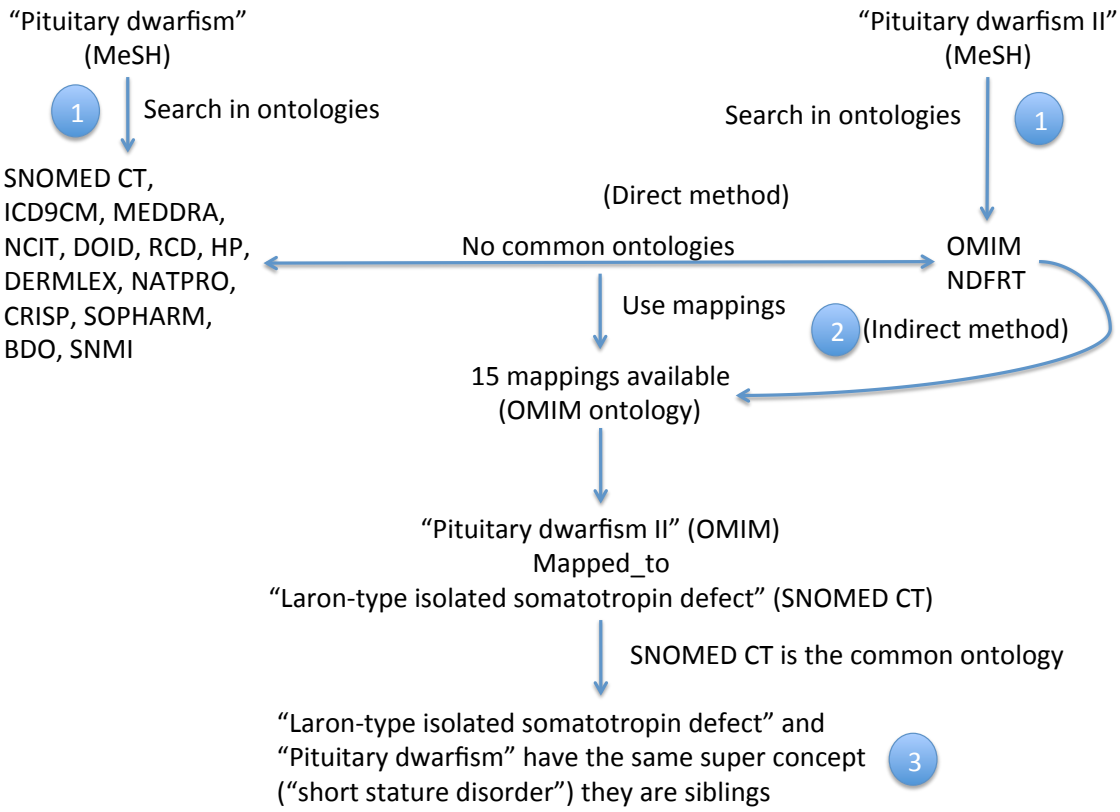


Figure 4.1: Illustrating example

2012 and the one of 2013. The evolution leads to the definition of a new concept whose label is "Pituitary dwarfism II" (ID M0452907). The goal is to characterize the semantic relationship that exists between "Pituitary dwarfism" and "Pituitary dwarfism II". Our algorithm has identified that these concepts are siblings in SNOMED CT, therefore the relationship symbol "partially related to" is eventually returned.

4.3.2 Impact on ontology mappings

The proposed method has been designed to evaluate to which extent the existing mappings are consistent with respect to the conceptualization of a domain modelled in existing standard termino-ontological resources. The idea behind has consisted in the projection of the triple defining the mappings (i.e. source concept, target concept and relationship) in the same ontology to see if the semantic relationship that exists between source and target concept is consistent. To this end, we used the ontologies of the Bioportal application.

At the time of writing this manuscript, we are evaluating this approach. The size of the data we are currently testing requires a significant computation time because we need to remotely access the content of the Bioportal repository.

4.4 On the validation of medical ontologies

The second work we have undertaken regarding the quality of the ontologies concerns the validation of an ontology's content with a particular focus on the conceptualization (i.e. the way concepts of the domain are described and linked). Actually, we have observed through the literature review that the validation problem has been tackled by considering the logic aspect of the representation of the ontology or, at ontology construction time, by providing guidelines to respect in order to obtain an acceptable ontology. Inspired by the work of Pammer [Pammer, 2010], the originality of our work lies in the use of natural language processing techniques to verbalize the content of the ontology to validate [Ben Abacha et al., 2013a, Ben Abacha et al., 2013b]. This has the advantage of making this knowledge understandable by non ICT experts. It hides the complexity of logic axioms to facilitate the work of domain experts at ontology validation time. This approach has been directly inspired by the assessment community who use questions that are generated from a model, usually an ontology, to test the knowledge of students. Contrary to assessment domain, we consider that the expert is reliable and we want to validate the model. The proposed approach is a three-step process consisting in:

1. Verbalizing the content of the ontology under the form of questions expressed in natural language[13] and ranking the generated questions to maximize the amount of knowledge to validate with the minimum set of questions,

2. Submitting the questions to a domain expert,

3. Interpreting the answers to these questions to validate the evaluated ontological elements. Depending on experts' answers the content of the ontology could be modified to become consistent.

The proposed approach illustrated in figure 4.2 can be used either to validate the content of the ontology at construction time, but also at evolution time if this content evolves. We will detail this work in the forthcoming sections.

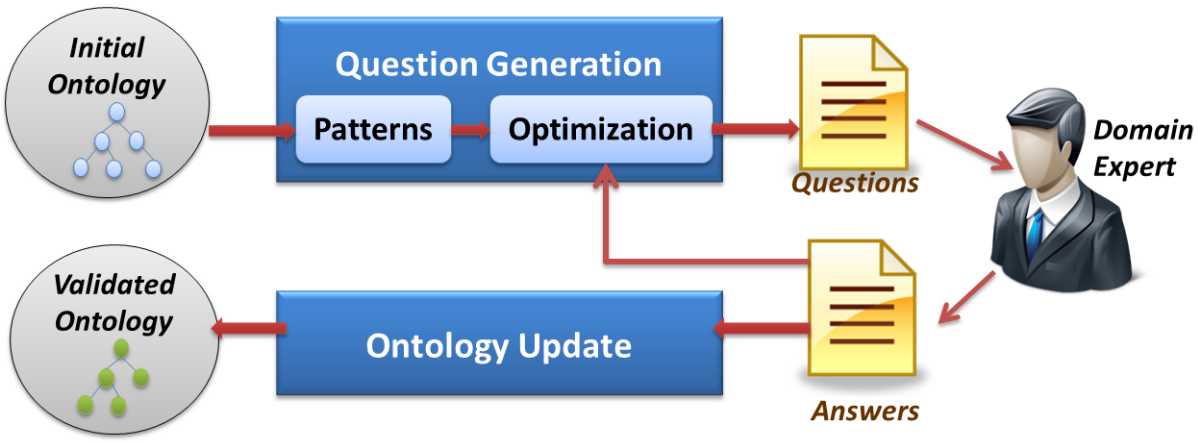


Figure 4.2: The COVALMO general approach for ontology validation

[13] In our work we focus on English

4.4.1 Verbalizing the content of medical ontology

The first step of the approach placed the focus on the verbalization of the ontology content to evaluate (cf. figure 4.3). We have observed that domain experts who are health professionals are, most of the time, not familiar with logic-based languages like OWL, which makes it hard for them to understand the representation of their domain in the ontology. Noticing this, we thought about how to overcome this limitation. We had to find a way to hide the complexity induced by the logic and decided to verbalize the content of the ontology, using questions expressed in natural language.

Questions generation

The aim of this step (cf. figure 4.3) is to build relevant natural language questions from formalized knowledge in order to validate the maximum number of assertions with the minimum number of questions. As our main focus lies on the conceptualization of the ontology, we have decided to treat the validity of the following assertions that can be applied to OWL ontologies:

- A rdfs:subClassOf B (class A is a subclass of B)
- P rdfs:subPropertyOf Q (property P is a sub-property of Q)
- P rdfs:domain D (D is the domain class for property P)
- P rdfs:range R (R is the range class for property P)
- I rdf:type A (I is an individual of class A)
- I P J (the property P links the individuals I and J)

In order to transcribe these OWL assertions into questions expressed in English, we built a set of questions patterns and associated each one with a specific assertion. The obtained patterns were designed following the combined analysis of OWL assertions and linguistic aspects of the questions. We start from the hypothesis that all the elements of a medical ontology must be validated. This involves validating concepts (e.g. Substance), relations between concepts (e.g. administrated for), concept instances (e.g. activated charcoal is an instance of Manufactured Material), relations between concept instances (e.g. chest X-ray can be ordered for Chronic cough) or between concept instances and literals (e.g. "give oral activated charcoal 50g" indicates the dose of the substance to be administrated "50g"). A question pattern consists in a regular textual expression with the appropriate "gaps". For instance, the pattern "Is DOSE of DRUG well suited for PATIENTS having DIS?" is a textual pattern with 4 gaps: DOSE, DRUG, PATIENTS and DIS, that are instantiated with labels of the ontological elements to be validated. This question pattern aims to validate a drug dose administrated to a patient having a particular disease. Table 4.1 shows some example.

The main difficulty of the question generation process is to contextualize the elements of the questions. Actually, concepts have a meaning only in a given context. To integrate this context in the question, we enlarge the scope of the question by putting several elements to validate in the question. This approach assumes that the more information is contained in the question, the easier it will be for the expert to understand the aim of the question. However, this requires different types of questions (cf figure 4.3):

- Boolean questions,

Question patterns	Example of instances
Does a(n) CLASS have a(n) PROPERTY?	Does a Sign have a Measurement Method? Does a Treatment have an Administration Route?
Is SUB-CLASS a type of CLASS?	Is Statistical Evidence a type of Evidence?
Is SUB-PROP a type of PROP?	Is Primary Treatment a type of Treatment?
Does a(n) CLASS1 PROPERTY a(n) CLASS2?	Does a Medical Exam diagnose a Disease?
Does INSTANCE1 PROPERTY INSTANCE2?	Does Prozac treat Schizophrenia?

Table 4.1: Examples of boolean-question patterns

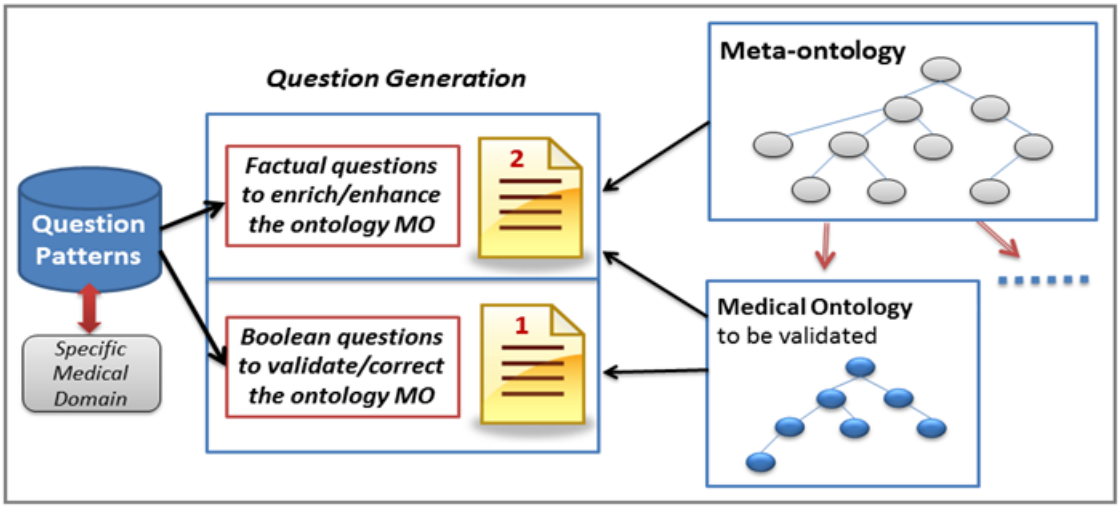


Figure 4.3: The generation of questions

- Factual questions.

Nevertheless, at the time of writing, the tool we have developed to support our approach is able to generate only boolean questions (i.e. yes/no questions). Further investigations are undertaken to support the generation of factual questions. Such kind of questions will give more flexibility for the experts to specify the right answers that will serve to modify (and sometimes enrich) the ontology. It will in particular allow the validation of individuals. To this end, examples of factual questions that will be supported by the system are: "What is the correct dose of aspirin for an adult suffering from headache?" or "What are the drugs able to treat the influenza?"

After the question generation step, we obtain a set of questions expressed in English about a wide range of ontological elements. In order to validate the logic aspect of the formalization, we used state-of-the-art methods that are reviewed in section 4.2.

NOT A rdfs:subClassOf B	⇒	NOT A rdfs:subClassOf C s.t. C rdfs:subClassOf B
NOT P rdfs:domain A	⇒	NOT P rdf:subPropertyOf Q s.t. Q rdfs:domain A
NOT I rdf:type A	⇒	NOT <I, P, J> s.t. P rdfs:domain A NOT <J, P, I> s.t. P rdfs:range A

Table 4.2: Examples of ontology update rules w.r.t. invalidated elements

Question optimization

The availability of domain experts is usually extremely limited by virtue of their main activities that are devoted to the care of patients. Taking this pragmatic consideration into account, we had to propose a technique to rank the questions, in order to limit the interactions with domain experts by optimizing the amount of assertions to validate through each question. To do so, we have proposed an optimization strategy relying on the RDFS logical rules, in order to rank the questions according to the elements that imply the more changes in the ontology. For instance, if we have the following statements:

1. hasSuitedAntiobioticsType *rdf:subPropertyOf* hasTreatment
2. Antibiotics *rdfs:subClassOf* Treatment
3. hasSuitedAntiobioticsType *rdfs:range* Antibiotics
4. hasTreatment *rdfs:range* Treatment

and the expert invalidates "Antibiotics rdfs:subClassOf Treatment", then the property hasSuitedAntiobioticsType cannot be declared as a sub-property of hasTreatment because the hasSuitedAntibioticType relation has not a common range with the property hasTreatment which leads to a formal error with respect to RDFS entailment rules. We consider all RDFS entailment rules. Table 4.2 presents some inverse forms of these rules in order to show the impact of invalidating each one of the target ontology statements.

According to the implemented strategy, questions could be ranked in a manner that allows to delete some of the remaining questions if one of the RDFS entailment rules applies, because some ontological elements have been modified or simply removed, therefore it makes no sense to ask an expert about their validity. This leads to the following validation order:

1. A rdfs:subClassOf B
2. P rdfs:domain D and P rdfs:range R
3. P rdfs:subPropertyOf Q
4. I rdf:type A
5. I P J

The so-generated questions are then submitted to an expert. He has to provide his feedback by answering them, and the system has to interpret this information to either validate or correct if possible; if not, just detect the problem and show to an expert that will correct it.

4.4.2 Interpreting experts feedback and modifying the ontology

This step provides in the second phase of the approach to reach a validated ontology. In our work, we have decided to consider only one expert as reference. We are aware of the strength of the assumption, but managing several experts would have required a more elaborated tool to support the potential discussions between experts in case their answers were conflicting. Nevertheless, we keep this issue for future work, in which we will consider ontology validation as a collaborative task. To interpret the answers of the experts, two distinct cases have been considered:

- the one that involves Boolean questions and,
- another one dealing with factual questions.

Both cases are discussed in the following subsections. Feedback provided by experts consists in two main parts:

1. an assertion on the correctness of the target knowledge and,
2. a free textual explanation (if provided).

In this work, we take into account ontologies that are formally valid (with no logic inconsistencies) and focus on the validation of domain conceptualization. In this context, "Yes" answers will have no impact on the ontology. The ontology will be modified on the "No" answers provided by the domain experts.

The case of boolean questions

As suggested, in our work we have treated only the case of boolean questions. An example of boolean question in the biomedical domain can be *"Does chemotherapy treats cancer?"*. This kind of questions probably constitutes the easiest case to deal with because there are only two possible answers. Either the expert confirms what is mentioned in the question or he answers negatively and therefore invalidates the assertion. Depending on the provided answer, two different possibilities to modify the ontology showed up:

1. if the answer is "Yes": the assertion is considered as correct by the expert and in consequence there is no reason to modify the ontology. Thus, this element of the ontology is marked as validated.
2. if the answer is "No": the expert considers that the assertion is wrong, therefore the ontology has to be corrected. Since our system is, for the time being, not able to interpret the additional information provided as free text by the expert, we have just decided to delete the ontological elements involved in the question and inform the knowledge engineer about the problem. As an example, consider a question asking if aspirin treats cancer. An expert can answer no and tell for instance that chemotherapy is efficient against cancer. However, the system will just delete the relation that exists between aspirin and cancer instead of creating a new concept "chemotherapy" and linking it to the concept "cancer" via the "treat" relationship. As only deletion is supported, the reconstruction of the ontology is done in order to preserve the logical validity of the ontology. To this end for example, the hierarchy of concepts is kept by connecting all concepts that have one concept that has been removed as super-concept to their "super-super"-concept (i.e. the super concept of the deleted one).

In order to enhance the capabilities of our system we are currently adding the possibility to generate factual questions.

The case of factual questions

Factual questions cannot be answered with "Yes" or "No". An example of such kind of question can be "What is the correct dose of aspirin for an adult having migraine?" Their treatment is more complex than that of boolean questions by virtue of the space of possible content composing the answer and of its possible interpretation. However, the use of factual questions opens the following perspectives:

- It gives a real possibility for the expert to specify his answer in a precise way (i.e. not only by saying yes or no but by formulating a complete answer). This additional information is really helpful for the enrichment of the ontology, but it can also allow users who have designed the ontology to understand what went wrong at ontology design time and to adapt their methodology,
- It allows much more flexibility in contextualizing the question by adding other ontological elements. It will reduce the ambiguities that can lead the experts to a total misunderstanding of the question and hence to erroneous answers,
- It makes it possible to complete the ontology in the case of some knowledge missing (i.e. for instance when a boolean question cannot be properly elaborated),
- It offers another manner to deal with ontology evolution through a coherent involvement of end-users. Such an approach can better control the evolution, since the modification of the ontology is done by the system and not by the users themselves. The consistency of the ontology, from the logic and structural point of views, can therefore better be preserved or even enhanced.

Whether questions are boolean or factual, if the answer aims at invalidating ontological elements, the modification induces a reorganisation of the remaining set of questions because many questions can potentially involve elements that have been removed (see figure 4.2). If this happens, the irrelevant questions are simply deleted and the remaining ones are re-ranked according to the criteria previously introduced. The process will stop when all questions have been answered (i.e. the set of questions is empty). At this final stage we consider that the ontology is valid.

4.4.3 Experimental assessment

We have evaluated our approach on three different medical ontologies:

- Caries ontology (CO). This ontology was developed manually by a knowledge engineer together with dentists as part of a collaboration project led by CRP Henri Tudor,
- Disease-Treatment Ontology (DTO) [Khoo et al., 2011]. We constructed an OWL translation of the ontology proposed by Khoo et al.,
- Mental Diseases Ontology (MDO)[14].

[14] `http://mental-functioning-ontology.googlecode.com/svn-history/r19/trunk/ontology/MD.owl`

The objective was to generate questions about the content of these ontologies and submit them to an expert for validation. The provided answers are then collected by the system to optimize the remaining set of questions. The obtained results, in terms of ontological elements that have been tested, can be seen in table 4.3.

Ontology	**Classes**	**Properties**	**Instances**	**Total**	**Question**
DTO	49	148	0	197	165
MDO	149	76	18	243	243
CO	26	266	13	305	290

Table 4.3: The number of Ontology Elements (OE) and the number of generated questions for different medical ontologies without optimization

The number of generated questions depends on the ontology size and shows the importance of implementing adequate question ranking and optimization strategies, in particular for large ontologies. In our experiments, our optimization method works better in case of ontologies with many instances. For the CO ontology, this strategy helps minimizing the number of submitted questions from 290 to 283 questions with only 4 NO answers. For the MDO ontology, our method allows asking 239 questions instead of 243 with only 2 NO answers. In case of ontologies with more NO answers (i.e. more invalid elements), the number of deleted questions will increase. For the DTO ontology, there were no available instances and at the same time there were no "NO answers" given by the expert, so the initial number of questions was conserved. The ontologies used in these experiments were constructed manually and semi-automatically. More experiments should be conducted on automatically constructed ontologies in order to more accurately evaluate the benefits of question optimization. In the case of ontologies with few invalid elements (few NO answers), we are currently working on presentation-level optimizations to reduce the time needed by the experts to answer the questions. In particular, we study two main presentations: question factorization according to an ontology element (concept, relation or individual) and logical chaining (A hasRelation1With B, B hasRelation2With C, etc.). These representations can help medical experts to reduce the time needed to understand and answer the questions.

These experiments also showed the need to add other specific types of questions and answers, in order to acquire missing information and to enrich the ontology when necessary. For instance, an answer to a question can be YES for one group of patients (e.g. Infant) and NO for another group or under a specific condition (e.g. co-morbidity). Our validation approach can also lead to the isolation of a concept or of a branch of the ontology. We are working on improving our system by adding the possibility for the expert to precise a contextual element or condition that clarifies ambiguous situations. We also work on integrating factual questions in our system, in order to add missing information to the ontology. Figure 4.3 presents our method of exploiting factual questions in order to enrich the ontology. Future work will also include the development and the evaluation of our approach when considering more complex OWL semantics (instead of only RDFS).

4.5 Summary

The ongoing work introduced in this chapter deals with the quality of medical data and knowledge, with particular attention paid to the case of ontologies' evaluation and validation. On one hand, we have proposed an algorithm, relying on external background knowledge and the Bioportal application in particular, that is able to compare the content of ontologies and especially

the description of concepts to identify existing mismatches. The proposed method is currently applied to compare concepts involved in mappings to those of existing standard ontologies, to verify if the semantic relationship specified in mappings is the same as the one specified at ontology level. On the other hand, we have conceptualized and implemented a system based on question-answering techniques for the validation of the content of an OWL ontology, with a particular focus on the conceptualization of the represented domain. This work allows the expert to better understand the content of an ontology by hiding the complexity of logic-based representation languages with natural language questions, aiming thus at limiting human intervention for validating the ontology and, in turn, potential errors that can potentially arise when manually modifying the ontology.

This work has mainly shown that:

- The quality of available data and knowledge (e.g. ontologies) is not homogeneous. The first results obtained with the Algorithm 2 underline inconsistencies between ontologies and depending artefacts. This is obviously problematic for decision support systems relying on these models, since reasoning can be wrong and therefore leading to bad decisions that can impact patient health conditions,
- The communication with domain experts is possible, but the verbalization of an ontology's content and the presentation to the experts under the form of questions must be done with care, mainly through the proper contextualization of the content of the questions. This is problematic, since sometimes the ontology itself does not contain the necessary information to explicitly specify the context at question level. This reinforces the conclusion of chapter 3 about the documentation of the ontology that must be done to facilitate its exploitation and maintenance,
- Feedback collected by the system, such as experts' answers, is relevant but requires advanced techniques for proper handling. In our work, experts can specify their feedback through free text making it possible for them to write what they want in an uncontrolled way. However, as evoked, this additional information is for the time being not treated. Introducing more features at user interface level will probably help to better control their feedback, reducing misinterpretation and enhance the quality of the final ontology.

The work presented in this chapter has been carried out in the framework of a post doctoral project and partly in the context of the DynaMO research project funded by the National Research Fund (FNR) of Luxembourg

Chapter 5

Open research challenges

Contents

5.1 Medical information management and retrieval 67
5.1.1 Semantic annotation management 67
5.1.2 Medical information retrieval 68
5.2 Knowledge dynamics in recent new paradigms 68
5.2.1 Big data 69
5.2.2 Linked Data and the Web 69
5.3 Patient empowerment 70

The work carried out over the past few years around the dynamic aspect of data and knowledge in the biomedical domain has opened new perspectives that are in direct line with the evolution of the domain. Moreover, besides the evolution of biomedicine, the emergence of new paradigms in computer science and the evolution of the Web have reinforced the feeling that the research perspectives of this work will be of importance for the coming years and will attract the interest of the scientific community.

5.1 Medical information management and retrieval

The outcomes of the various projects have allowed us to have a better understanding of the medical domain, its data, information and knowledge as well as its characteristics and behaviour over time. DynaMO was our first attempt to address the problem raised by the evolution of ontologies and mainly the propagation of the modifications to depending artefacts [Stojanovic, 2004]. While ontology mappings were the focus of this project, semantic annotations are gaining in importance in the management of medical data and are subject to the same problem as mappings.

5.1.1 Semantic annotation management

The rapid development and wide spread of Electronic Health Records (EHR) that harvest patient medical information is generating particular needs in terms of knowledge and data management. Actually, for data security and privacy issues, this data is encrypted most of the time, preventing any efficient use of classical information retrieval methods that typically act directly on textual information. However, health professionals must be able to search for relevant medical information about their patients in order to design the appropriate treatments. To overcome the

limitation created by the encryption, additional meta data describing the content of the documents are associated with them before encryption and will remain in the records as clear text [Pruski and Wisniewski, 2012]. These meta data also referred to as semantic annotations denote ontology elements including concept labels, attributes, comments, descriptions and relationship symbols that are associated with or attached to a document to provide additional information about its content, in order to make its semantics explicit for humans and software applications. In this context, it is clear that semantic annotations play a central role in the good performance and acceptance of EHR by either health professionals or patients. However, as these annotations are usually borrowed from standards ontologies, they are also concerned with the ontology evolution problem. In fact, the modifications of concepts' description can create a real mismatch between the terminology of the domain and the annotated documents, causing inconsistencies in their exploitation. This has motivated the definition of the ELISA[15] project, the successor of DynaMO, aiming at proposing methods and tools towards the automatic adaptation of semantic annotations impacted by ontology evolution.

5.1.2 Medical information retrieval

The questions raised by the adaptation of semantic annotations are indirectly dealing with information retrieval. Actually, a bad annotation adaptation process will strongly impact the quality of a search since semantic annotations are an important support for search engines to find relevant information, in particular if the annotations are not correctly representing the content of the documents they are associated with. Moreover, the quality of the search will highly depend on the quality of the queries and the interpretation made by the engine at evaluation time to figure out what are the users' real needs in terms of medical information. In this general context, new intelligent approaches are needed. In the framework of the GECAMED project[16], we have undertaken research work to deal with the retrieval and ranking of relevant information concerning patient suffering from cardio-vascular diseases. The GECAMED general application has been developed for health professionals to assist them in the management of patient data, and has been accepted as the de facto standard to push and pull medical information from the Luxembourgish national EHR. As a consequence, end-users (including patients) can potentially have access to a huge amount of information that require intelligent tools able to find the most relevant information regarding users' needs. By definition of semantic annotations, we are currently evaluating the benefit of using ontologies to improve medical information retrieval through the GECAMED software application.

5.2 Knowledge dynamics in recent new paradigms

Besides the scientific activities developed in the field of medical informatics, computer science has seen the emergence of new paradigms for which data and knowledge management is essential and for which new types of applications, fostered by the evolution of the Web, have been developed. This is the case for instance of *Big Data* and of *Linked Data*.

[15] ELISA, for which I am the PI, and which is funded by the FNR and DFG under the INTER programme and will be developed in collaboration with the university of Leipzig and the team of Pr. E. Rahm and Pr. C. Reynaud-Delaître from University of Paris-Sud.

[16] `http://santec.tudor.lu/fr/project/gecamed`

5.2.1 Big data

The *Big data* is

> *"an all-encompassing term for any collection of data sets so large and complex that it becomes difficult to process them using traditional data processing applications."* (Wikipedia, 2014)

Following this definition, Big data is also characterized by the volume, variety, velocity, variability and veracity of the data introducing a new problematic in the field data and knowledge management. The evolution of knowledge is therefore concerned by this Big data paradigm. Actually, the rapid multiplication of sensors, such as heart monitoring implants, in the context of wearable computing and the already popular *Internet of things* call for massive data production that requires a new generation of tools able to manage and exploit it over time. The health domain will not escape from this tendency, and health professionals can be overwhelmed by information and data about their patients, forcing them to review and sort these to keep only the most relevant ones. It will also be the case for researchers who will have to handle high throughput genomic data in the context of personalized medicine. To help them in this labour-intensive task, classical knowledge engineering techniques will have to be improved to deal with the huge quantity of data that will be generated. If rule-based systems have already shown their limit in this huge data space, machine learning, originally defined to deal with voluminous data sets can be of interest. However, the gap between symbolic and numeric approaches that has been set up over the past years will have to be bridged to make the most of both worlds. The rigour of logic-based approach could, for instance, serve as a first step to parametrize numeric algorithms (e.g. implementing supervised learning) able to deal with Big Data. To this end, new research questions addressing the combination of symbolic and numeric approaches to deal with Big Data will certainly gain in interest in the forthcoming years.

5.2.2 Linked Data and the Web

The recent evolution of the Web has progressively put Linked Data in the front row. It describes a method of publishing structured data so that it can be interlinked and become more useful. It builds upon standard Web technologies such as HTTP, RDF and URIs, but rather than using them to serve web pages for human readers, it extends them to share information in a way that can be read automatically by computers. This enables data from different sources to be connected and queried [Bizer et al., 2009].

Linked Data is published on a daily basis for Web users and software applications for various purposes. In our current work, we are evaluating the added value of Linked Open Data to enhance the detection of interactions when merging guidelines [Zamborlini et al., 2015]. We use the DrugBank[17] bioinformatics and cheminformatics resource that combines detailed drug (i.e. chemical, pharmacological and pharmaceutical) data with comprehensive drug target (i.e. sequence, structure, and pathway) information to identify potential drug interactions in case of multimorbidity. In a near future we will extend this work through the connection of our model to different open resources of the Web.

In parallel, the dynamics of linked data is an open research problem. As Linked Data are usually associated with the Web, they are supposed to evolve as frequently as the Web which questions the consistency of the links (i.e. at evolution time, the link that connects two or more resources is still valid if the content of these resources has been modified?) In direct line with the

[17] `http://www.drugbank.ca/`

work carried out during my PhD thesis, the DynaMO and forthcoming ELISA projects dealing with knowledge evolution, we plan to investigate the maintenance of Linked Data in order to preserve their validity and usability over time. We will be able to evaluate the portability of the already defined framework and concepts to the case of the Web and also to enlarge the scope of application by testing other domains than biomedicine. We will then be able to adapt and extend (if needed) the work done in the context of semantic mappings and annotations maintenance to the case of Linked Data.

5.3 Patient empowerment

The progressive paradigm shift from doctor to patient-centric is giving more and more importance to the patients, impacting medical activities. This is also accompanied by the appearance of new paradigms such as *Personalized medicine* that consists in the customization of healthcare - with medical decisions, practices, and/or products being tailored to the individual patient. From an IT point of view, health information systems are now developed in order to give more flexibility to the patient in the management of his health information involving him more in medical decisions and treatment definition.

The rapid evolution of the Web through the development and acceptance of social networks is deeply modifying users' behaviour, mainly with regard to their interest in gaining information about their health status and the possibility for them to share experiences with predefined communities. However, the lack of supervision in these networks (i.e. the involvement of health professionals) can lead to a bad use or misunderstanding of the information by users impacting their health if they are looking for specific medical information that describe their symptoms. Two aspects have to be enhanced: (i) the retrieval of medical information and (ii) the communication between patients and physicians.

Following this observation, the misunderstanding of medical information by patients is partially due to the complexity of medical terminology or rather the difference between the grade level required to understand medical information and the grade that patients really have. Health literacy is a field of research aiming at presenting medical information to patients, taking into account their profile, for instance, the selection of the appropriate vocabulary to display the information contained in EHR to patients. In this new research field, we plan to evaluate the abilities of knowledge engineering techniques to help patients with the retrieval of relevant information and its presentation. This will limit the interaction between physicians and patients to only the important matters and enhance the empowerment of patients in their own health follow-up.

Conclusion

Knowledge engineering is a field of research that has attracted the interest of many researchers over the past decades. Several major results of the work carried out in KE have largely contributed to the success of many applications and companies whose business relies on data and content management like Google, Amazon or eBay. The biomedical domain has also been impacted by KE. The complexity, dynamics and critical aspect of its knowledge have driven many research efforts to the definition of original methods and tools facilitating its management and exploitation essentially in decision support systems, to assist health professionals in optimizing patient care.

This manuscript is entirely related to my research activities in medical informatics and to my contributions in medical knowledge representation, the management of medical knowledge evolution and the evaluation and validation of its quality. This work has originally been motivated by the demand of the field, the discussion with major stakeholders of the Luxembourgish scene and the possibility to see my research work applied in concrete applications having a real social-economic impact.

Knowledge representation has been addressed in the general framework of clinical guidelines. The two complementary doctoral projects I am co-supervising in this context have different objectives. The first one has been defined based on our preliminary work on the use of ontologies for representing care actions, composing clinical guidelines which were investigated during a postdoctoral project. The aim of this project is to enhance reasoning capabilities of computer systems exploiting guidelines to personalize, adapt and update clinical guidelines over time. The encouraging first results have proposed the definition of a formal model expressed in a subset of first-order logic, and in an implementation using Semantic Web technologies for the representation of medical recommendations. This model has shown great abilities in the identification and solving of medical interactions that can arise in the case of co-morbidity when several guidelines have to be combined. While this project is focusing on the design of appropriate treatment, the second one puts the stress on the reconfiguration of treatment plans when patients have already started their therapies. In that case, the objective is to define mechanisms that are able to identify key parameters of the defined treatment plans, monitor their evolution over time and react at treatment plans' level according to their modification, in order to preserve their efficiency regarding a patient's health condition. This work will be valued later, and within an innovation project that will lead to the definition of a care recommendation database assisting people in charge of managing guidelines and their integration in hospital information systems. The work on medical knowledge representation has been published in 9 scientific papers (7 in international conferences and 2 in peer-reviewed international journals).

The management of medical knowledge evolution has been investigated in the framework of the DynaMO project leading to one PhD and one postdoctoral project. The aim of this project was to propose a formal framework for the (semi-)automatic adaptation of semantic mappings turned invalid by the evolution of their associated Knowledge Organisation Systems.

The DyKOSMap framework, designed in response to the research question, has been developed based on a significant empirical evaluation of the behaviour of standard KOS of the domain and the impact of this evolution on depending mappings. Following these observations, original methods to identify and characterize KOS elements evolution have been proposed under a set of refined change patterns acting at the level of concept attributes, together with a set of mapping adaptation actions that formalize the possible ways mappings can evolve, from a practical point of view, over time. The link between KOS evolution and mapping adaptation has been materialized through the definition of heuristics expressed in predicate logic. The work carried out in this project will be extended to the adaptation of medical semantic annotations in the framework of the forthcoming ELISA project. The work on medical knowledge evolution has been published in 14 scientific articles (10 in national and international conferences and 4 in peer-reviewed international journals).

Last, the recently launched work on the evaluation and validation of medical knowledge quality is currently investigated in the context of two postdoctoral projects. The critical aspect of medical knowledge and data has motivated us to investigate methods and tools to evaluate their quality, since they are implemented in medical decisions that can potentially impact patient health. The first idea has led us to define a general method to compare the content of ontological entities, in particular those involved in mappings, with the content of standard medical ontologies to identify mismatches between them. This has been done based on the Bioportal repository, which offers access to more than 380 ontologies of the field. On the other hand, we have developed a system to validate the conceptualization of an ontology based on question-answering. This was done to better involve health professionals in the ontology validation process, by hiding the complexity of logic-based representation languages. Health professionals are usually not familiar with these in questions expressed in natural language. Based on the provided answers, the system is able to modify the ontological elements that have been declared as invalid by the experts, in order to reach a valid conceptualization of the domain represented in the ontology. The work on medical knowledge evolution has been published in 3 scientific articles (2 in national and international conferences and 1 in peer-reviewed international journal).

The results that have been achieved through the evoked projects have opened several perspectives that will occupy my time in the forthcoming years. My scientific activities will focus on the research questions that are directly motivated by this initial work, with a particular attention paid to the valuation of this work and to the concrete application of this research. Moreover, I will also try to transpose the defined methodology and tools to other domains that have other characteristics than biomedicine and investigate new horizons where knowledge representation, management and validation are needed like the Linked Open Data and the Web.

References

[Abidi, 2011] Abidi, S. R. (2011). Ontology-based knowledge modeling to provide decision support for comorbid diseases. In *Knowledge Representation for Health-Care*, pages 27–39. Springer.

[Aleksovski et al., 2006] Aleksovski, Z., Klein, M., Ten Kate, W., and Van Harmelen, F. (2006). Matching unstructured vocabularies using a background ontology. In *Managing Knowledge in a World of Networks*, pages 182–197. Springer.

[An et al., 2010] An, Y., Hu, X., and Song, I.-Y. (2010). Maintaining mappings between conceptual models and relational schemas. *Journal of Database Management (JDM)*, 21(3):36–68.

[Arts et al., 2002] Arts, D. G., De Keizer, N. F., and Scheffer, G.-J. (2002). Defining and improving data quality in medical registries: a literature review, case study, and generic framework. *Journal of the American Medical Informatics Association*, 9(6):600–611.

[Baneyx and Charlet, 2006] Baneyx, A. and Charlet, J. (2006). Évaluation, évolution et maintenance d'une ontologie en médecine: état des lieux et expérimentation. *Revue I3 Information - Interaction - Intelligence*.

[Ben Abacha et al., 2013a] Ben Abacha, A., Da Silveira, M., and Pruski, C. (2013a). Medical ontology validation through question answering. In *Artificial Intelligence in Medicine*, pages 196–205. Springer.

[Ben Abacha et al., 2013b] Ben Abacha, A., Da Silveira, M., and Pruski, C. (2013b). Une approche pour la validation du contenu d'une ontologie par un système à base de questions/réponses. In *Proceedings of the 24th Ingénierie des Connaissances conference*.

[Bizer et al., 2009] Bizer, C., Heath, T., and Berners-Lee, T. (2009). Linked data-the story so far. *International journal on semantic web and information systems*, 5(3):1–22.

[Bonacin et al., 2010] Bonacin, R., Da Silveira, M., and Pruski, C. (2010). From medical guidelines to personalized careflows: The iCareflow ontological framework. In *IEEE 23rd International Symposium on Computer-Based Medical Systems (CBMS)*, pages 462–467. IEEE.

[Bonacin et al., 2012] Bonacin, R., Pruski, C., and Da Silveira, M. (2012). Careflow personalization services: concepts and tool for the evaluation of computer-interpretable guidelines. In *Knowledge Representation for Health-Care*, pages 80–93. Springer.

[Bonacin et al., 2013] Bonacin, R., Pruski, C., and Silveira, M. D. (2013). Architecture and services for formalising and evaluating care actions from computer–interpretable guidelines. *International Journal of Medical Engineering and Informatics*, 5(3):253–268.

[Carroll, 2002] Carroll, J. G. (2002). Crossing the quality chasm: A new health system for the 21st century. *Quality Management in Healthcare*, 10:60–61.

[Castano et al., 2008] Castano, S., Ferrara, A., Lorusso, D., Näth, T. H., and Möller, R. (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.

[Ceusters et al., 2003] Ceusters, W., Smith, B., and Flanagan, J. (2003). Ontology and medical terminology: Why description logics are not enough? In *Towards an Electronic Patient Record (TEPR 2003), Boston, MA*.

[Ceusters et al., 2004] Ceusters, W., Smith, B., Kumar, A., and Dhaen, C. (2004). Mistakes in medical ontologies: where do they come from and how can they be detected? *Ontologies in medicine*, 102:145.

[Colazzo and Sartiani, 2009] Colazzo, D. and Sartiani, C. (2009). Detection of corrupted schema mappings in XML data integration systems. *ACM Transactions on Internet Technology (TOIT)*, 9(4):14:1–14:53.

[Cordeiro and Filipe, 2009] Cordeiro, J. and Filipe, J., editors (2009). *A user-driven and a semantic-based ontology mapping evolution approach.*

[Da Silveira et al., 2008] Da Silveira, M., Guelfi, N., Baldacchino, J.-D., Plumer, P., Seil, M., Wienecke, A., et al. (2008). A survey of interoperability in e-health systems-the european approach. In *HEALTHINF (1)*, pages 172–175.

[d'Aquin et al., 2013] d'Aquin, M., Lieber, J., and Napoli, A. (2013). Decentralized case-based reasoning and semantic web technologies applied to decision support in oncology. *The Knowledge Engineering Review*, 28(04):425–449.

[Demner-Fushman et al., 2009] Demner-Fushman, D., Chapman, W. W., and McDonald, C. J. (2009). What can natural language processing do for clinical decision support? *Journal of biomedical informatics*, 42(5):760–772.

[Dentler and Cornet, 2013] Dentler, K. and Cornet, R. (2013). Redundant elements in snomed ct concept definitions. In *Artificial Intelligence in Medicine*, pages 186–195. Springer.

[Dimitrova et al., 2008] Dimitrova, V., Denaux, R., Hart, G., Dolbear, C., Holt, I., and Cohn, A. G. (2008). Involving domain experts in authoring OWL ontologies. In *Proceedings of the 7th International Conference on The Semantic Web*, pages 1–16. Springer-Verlag.

[Dinh et al., 2014a] Dinh, D., Dos Reis, J. C., Pruski, C., Da Silveira, M., and Reynaud-Delaître, C. (2014a). Identifying relevant concept attributes to support mapping maintenance under ontology evolution. *Web Semantics: Science, Services and Agents on the World Wide Web*, 29(0):53 – 66. Life Science and e-Science.

[Dinh et al., 2014b] Dinh, D., Dos Reis, J. C., Pruski, C., Da Silveira, M., and Reynaud-Delaître, C. (2014b). Identifying change patterns of concept attributes in ontology evolution. In *Proceedings of the Semantic Web: Trends and Challenges - 11th International Conference, ESWC 2014, Anissaras, Crete, Greece, May 25-29, 2014.*, pages 768–783.

[Djedidi, 2009] Djedidi, R. (2009). *Approche d'évolution d'ontologie guidée par des patrons de gestion de changement.* PhD thesis, Université Paris Sud-Paris XI.

[Djedidi and Aufaure, 2010] Djedidi, R. and Aufaure, M.-A. (2010). ONTO-EVO AL an ontology evolution approach guided by pattern modeling and quality evaluation. In *Foundations of Information and Knowledge Systems*, pages 286–305. Springer.

[Dolinski and Botstein, 2013] Dolinski, K. and Botstein, D. (2013). Automating the construction of gene ontologies. *Nature biotechnology*, 31(1):34–35.

[Dos Reis et al., 2013a] Dos Reis, J., Pruski, C., Da Silveira, M., and Reynaud-Delaître, C. (2013a). Characterizing Semantic Mappings Adaptation via Biomedical KOS Evolution: A Case Study Investigating SNOMED CT and ICD. In *AMIA 2013, American Medical Informatics Association Annual Symposium, Washington, DC, USA, November 16-20, 2013*.

[Dos Reis et al., 2014a] Dos Reis, J. C., Da Silveira, M., Dinh, D., Pruski, C., and Reynaud-Delaître, C. (2014a). Requirements for adaptating mappings linking dynamic ontologies. In *2014 IEEE 23rd International WETICE Conference, WETICE 2014, Parma, Italy, 23-25 June, 2014*, pages 405–410.

[Dos Reis et al., 2014b] Dos Reis, J. C., Dinh, D., Da Silveira, M., Pruski, C., and Reynaud-Delaître, C. (2014b). The influence of similarity between concepts in evolving biomedical ontologies for mapping adaptation. In *Proceedings of the Medical Informatics Europe (MIE) conference*.

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

[Dos Reis et al., 2013b] Dos Reis, J. C., Dinh, D., Pruski, C., Da Silveira, M., and Reynaud-Delaître, C. (2013b). Mapping adaptation actions for the automatic reconciliation of dynamic ontologies. In *Proceedings of the 22nd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '13, pages 599–608, New York, NY, USA. ACM.

[Dos Reis et al., 2012] Dos Reis, J. C., Pruski, C., Da Silveira, M., and Reynaud-Delaître, C. (2012). Analyzing and supporting the mapping maintenance problem in biomedical knowledge organization systems. In *SIMI Workshop-9th Extended Semantic Web Conference*, pages 25–36.

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

[Dos Reis et al., 2015b] Dos Reis, J. C., Pruski, C., and Reynaud-Delaître, C. (2015b). Heuristiques pour l'adaptation des mappings entre ontologies dynamiques. In *Proceedings of the 15th Conference on Knowledge Extraction and Management (EGC)*.

[Dos Reis et al., 2015c] Dos Reis, J. C., Pruski, C., and Reynaud-Delaître, C. (2015c). State-of-the-art on mapping maintenance and challenges for a fully automatic approach. *Expert Systems with Applications*, 42(3):1465–1478.

[Duque-Ramos et al., 2011] Duque-Ramos, A., Fernández-Breis, J. T., Stevens, R., Aussenac-Gilles, N., et al. (2011). Oquare: A square-based approach for evaluating the quality of ontologies. *Journal of Research and Practice in Information Technology*, 43(2):159.

[Elkin et al., 2006] Elkin, P. L., Brown, S. H., Husser, C. S., Bauer, B. A., Wahner-Roedler, D., Rosenbloom, S. T., and Speroff, T. (2006). Evaluation of the content coverage of SNOMED CT: ability of SNOMED clinical terms to represent clinical problem lists. In *Mayo Clinic Proceedings*, volume 81, pages 741–748. Elsevier.

[Euzenat et al., 2007] Euzenat, J., Shvaiko, P., et al. (2007). *Ontology matching*, volume 18. Springer.

[Fagin et al., 2011] Fagin, R., Kolaitis, P. G., Popa, L., and Tan, W.-C. (2011). Schema mapping evolution through composition and inversion. In Bellahsene, Z., Bonifati, A., and Rahm, E., editors, *Schema Matching and Mapping*, Data-Centric Systems and Applications, pages 191–222. Springer Berlin Heidelberg.

[Feigenbaum and McCorduck, 1983] Feigenbaum, E. and McCorduck, P. (1983). The fifth generation: artificial intelligence and japan's computer challenge to the world.

[Field and Lohr, 1990] Field, M. and Lohr, K. (1990). Guidelines for clinical practice: Directions for a new program. *Washington DC Institute of Medicine*.

[Gangemi et al., 2006] Gangemi, A., Catenacci, C., Ciaramita, M., and Lehmann, J. (2006). *Modelling ontology evaluation and validation*. Springer.

[Glaser, 2013] Glaser, J. (October 8, 2013). Managing complexity with health care information technology. *H&HN Daily*.

[Gómez-Pérez, 2004] Gómez-Pérez, A. (2004). Ontology evaluation. In *Handbook on Ontologies*, pages 251–274.

[Grosjean et al., 2011] Grosjean, J., Merabti, T., Dahamna, B., Kergourlay, I., Thirion, B., Soualmia, L. F., Darmoni, S. J., et al. (2011). Health multi-terminology portal: a semantic added-value for patient safety. *Stud Health Technol Inform*, 166:129–38.

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

[Groß et al., 2012] Groß, A., Hartung, M., Thor, A., and Rahm, E. (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)*.

[Gross et al., 2012] Gross, A., Hartung, M., Thor, A., and Rahm, E. (2012). How do computed ontology mappings evolve?-a case study for life science ontologies. In *Joint Workshop on Knowledge Evolution and Ontology Dynamics@ ISWC*, volume 12.

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

[Guelfi and Pruski, 2006] Guelfi, N. and Pruski, C. (2006). On the use of ontologies for an optimal representation and exploration of the web. *Journal of Digital Information Management*, 4(3):159.

[Guelfi et al., 2007a] Guelfi, N., Pruski, C., and Reynaud, C. (2007a). Les ontologies pour la recherche ciblée d'information sur le web: une utilisation et extension d'OWL pour l'expansion de requêtes. *Ingénierie des connaissances 07*, pages 61–72.

[Guelfi et al., 2007b] Guelfi, N., Pruski, C., and Reynaud, C. (2007b). Understanding supporting ontology evolution by observing the WWW conference. In *ESOE*, pages 19–32.

[Guizzardi and Wagner, 2004] Guizzardi, G. and Wagner, G. (2004). A unified foundational ontology and some applications of it in business modeling. In *CAiSE Workshops (3)*, pages 129–143.

[Hartung et al., 2013] Hartung, M., Groß, A., and Rahm, E. (2013). COnto–Diff: generation of complex evolution mappings for life science ontologies. *Journal of Biomedical Informatics*, 46(1):15–32.

[Hartung et al., 2008] Hartung, M., Kirsten, T., and Rahm, E. (2008). Analyzing the evolution of life science ontologies and mappings. In *Data Integration in the Life Sciences*, pages 11–27. Springer.

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

[Institute of Medicine, 2001] Institute of Medicine (2001). Crossing the quality chasm: a new healtsyst for the 21st century. Report.

[Jafarpour and Abidi, 2013] Jafarpour, B. and Abidi, S. S. R. (2013). Merging disease-specific clinical guidelines to handle comorbidities in a clinical decision support setting. In *Proceedings of the Artificial Intelligence in Medicine conference*, pages 28–32.

[Jiang and Chute, 2009] Jiang, G. and Chute, C. G. (2009). Auditing the semantic completeness of SNOMED CT using formal concept analysis. *Journal of the American Medical Informatics Association*, 16(1):89–102.

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

[Khattak et al., 2012] Khattak, A. M., Pervez, Z., Latif, K., and Lee, S. (2012). Time efficient reconciliation of mappings in dynamic web ontologies. *Knowledge Based Systems*, 35:369–374.

[Khoo et al., 2011] Khoo, C. S., Na, J.-C., Wang, V. W., and Chan, S. (2011). Developing an ontology for encoding disease treatment information in medical abstracts. *DESIDOC Journal of Library & Information Technology*, 31(2).

[Kinsman et al., 2010] Kinsman, L., Rotter, T., James, E., Snow, P., and Willis, J. (2010). What is a clinical pathway? development of a definition to inform the debate. *BMC Medicine*, 8.

[Kirsten et al., 2011] Kirsten, T., Gross, A., Hartung, M., and Rahm, E. (2011). GOMMA: a component-based infrastructure for managing and analyzing life science ontologies and their evolution. *J. Biomedical Semantics*, 2(6).

[Klein and Noy, 2003] Klein, M. and Noy, N. F. (2003). A component-based framework for ontology evolution. In *Proceedings of the IJCAI*, volume 3.

[Köhler et al., 2006] Köhler, J., Munn, K., Rüegg, A., Skusa, A., and Smith, B. (2006). Quality control for terms and definitions in ontologies and taxonomies. *BMC bioinformatics*, 7(1):212.

[Latoszek-Berendsen et al., 2007] Latoszek-Berendsen, A., Talmon, J., de Clercq, P., and Hasman, A. (2007). With good intentions. *International journal of medical informatics*, 76:S440–S446.

[Lenz and Reichert, 2005] Lenz, R. and Reichert, M. (2005). It support for healthcare processes. In *Business process management*, pages 354–363. Springer.

[Maedche and Staab, 2002] Maedche, A. and Staab, S. (2002). Measuring similarity between ontologies. In *Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web*, EKAW '02, pages 251–263, London, UK, UK. Springer-Verlag.

[Mawlood-Yunis, 2008] Mawlood-Yunis, A.-R. (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] McCann, R., AlShebli, B., Le, Q., Nguyen, H., Vu, L., and Doan, A. (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.

[Meilicke et al., 2007] Meilicke, C., Stuckenschmidt, H., and Tamilin, A. (2007). Repairing ontology mappings. In *Proceedings of the 22nd National Conference on Artificial Intelligence - Volume 2*, AAAI'07, pages 1408–1413. AAAI Press.

[Mezghani et al., 2014] Mezghani, E., Da Silveira, M., Pruski, C., Exposito, E., and Drira, K. (2014). A perspective of adaptation in healthcare. In *Proceedings of the Medical Informatics Europe (MIE) conference*.

[Mezghani et al., 2015] Mezghani, E., Exposito, E., Drira, K., Da Silveira, M., and Pruski, C. (2015). A semantic big data platform for integrating heterogeneous wearable data in healthcare. *Journal of medical systems*, 39(12):1–8.

[Miksch et al., 1997] Miksch, S., Shahar, Y., and Johnson, P. (1997). Asbru: A task-specific, intention-based, and time-oriented language for representing skeletal plans. In *Proceedings of the 7th Workshop on Knowledge Engineering: Methods & Languages (KEML-97)*, pages 9–19. Milton Keynes, UK, The Open University, Milton Keynes, UK.

[Noy et al., 2009] Noy, N. F., Shah, N. H., Whetzel, P. L., Dai, B., Dorf, M., Griffith, N., Jonquet, C., Rubin, D. L., Storey, M.-A., Chute, C. G., et al. (2009). BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic acids research*, 37(suppl 2):W170–W173.

[Pammer, 2010] Pammer, V. (2010). *Automatic Support for Ontology Evaluation Review of Entailed Statements and Assertional Effects for OWL Ontologies*. PhD thesis, Graz University of Technology.

[Patel et al., 1998] Patel, V. L., Allen, V. G., Arocha, J. F., and Shortliffe, E. H. (1998). Representing clinical guidelines in GLIF individual and collaborative expertise. *Journal of the American Medical Informatics Association*, 5(5):467–483.

[Patel et al., 2009] Patel, V. L., Shortliffe, E. H., Stefanelli, M., Szolovits, P., Berthold, M. R., Bellazzi, R., and Abu-Hanna, A. (2009). The coming of age of artificial intelligence in medicine. *Artificial intelligence in medicine*, 46(1):5–17.

[Peleg, 2013] Peleg, M. (2013). Computer-interpretable clinical guidelines: A methodological review. *Journal of biomedical informatics*, 46(4):744–763.

[Peleg et al., 2000] Peleg, M., Boxwala, A. A., Ogunyemi, O., Zeng, Q., Tu, S., Lacson, R., Bernstam, E., Ash, N., Mork, P., Ohno-Machado, L., et al. (2000). GLIF3: the evolution of a guideline representation format. In *Proceedings of the AMIA Symposium*, page 645. American Medical Informatics Association.

[Pohl et al., 2011] Pohl, M., Wiltner, S., Rind, A., Aigner, W., Miksch, S., Turic, T., and Drexler, F. (2011). Patient development at a glance: An evaluation of a medical data visualization. In *INTERACT (4)*, pages 292–299.

[Poveda-Villalón et al., 2012] Poveda-Villalón, M., Suárez-Figueroa, M. C., and Gómez-Pérez, A. (2012). Validating ontologies with oops! In *Knowledge Engineering and Knowledge Management*, pages 267–281. Springer.

[Pruski, 2009] Pruski, C. (2009). *Une approche adaptative pour la recherche d'information sur le Web*. PhD thesis, Université Paris Sud-Paris XI et Université du Luxembourg.

[Pruski et al., 2011a] Pruski, C., Bonacin, R., and Da Silveira, M. (2011a). Towards the formalization of guidelines care actions using patterns and semantic web technologies. In *Artificial Intelligence in Medicine*, pages 302–306. Springer.

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

[Pruski and Wisniewski, 2012] Pruski, C. and Wisniewski, F. (2012). Efficient medical information retrieval in encrypted electronic health records. *Quality of Life Through Quality of Information: Proceedings of MIE2012*, 180:225.

[Pruski et al., 2010] Pruski, C., Wisniewski, F., and Da Silveira, M. (2010). Barriers to overcome for the implementation of integrated eHealth solution in Luxembourg. In *Proceedings of the eChallenges 2010 Conference*.

[Qi et al., 2009] Qi, G., Ji, Q., and Haase, P. (2009). A conflict-based operator for mapping revision. In Bernstein, A., Karger, D., Heath, T., Feigenbaum, L., Maynard, D., Motta, E., and Thirunarayan, K., editors, *The Semantic Web - ISWC 2009*, volume 5823 of *Lecture Notes in Computer Science*, pages 521–536. Springer Berlin Heidelberg.

[Qian and Dong, 2005] Qian, G. and Dong, Y. (2005). Constructing maintainable semantic mappings in XQuery. In Doan, A., Neven, F., McCann, R., and Bex, G. J., editors, *Proceedings of the Eight International Workshop on the Web & Databases (WebDB), Baltimore, Maryland, USA, Collocated mith ACM SIGMOD/PODS 2005*, pages 121–126.

[Riano and Collado, 2013] Riano, D. and Collado, A. (2013). Model-based combination of treatments for the management of chronic comorbid patients. In *Artificial Intelligence in Medicine*, pages 11–16. Springer.

[Rico et al., 2014] Rico, M., Caliusco, M. L., Chiotti, O., and Galli, M. R. (2014). OntoQualitas: A framework for ontology quality assessment in information interchanges between heterogeneous systems. *Computers in Industry*.

[Sabou and Fernandez, 2012] Sabou, M. and Fernandez, M. (2012). Ontology (network) evaluation. In *Ontology engineering in a networked world*, pages 193–212. Springer.

[Sánchez-Garzón et al., 2013] Sánchez-Garzón, I., Fdez-Olivares, J., Onaindía, E., Milla, G., Jordán, J., and Castejón, P. (2013). A multi-agent planning approach for the generation of personalized treatment plans of comorbid patients. In *Artificial Intelligence in Medicine*, pages 23–27. Springer.

[Shvaiko and Euzenat, 2013] Shvaiko, P. and Euzenat, J. (2013). Ontology matching: state of the art and future challenges. *Knowledge and Data Engineering, IEEE Transactions on*, 25(1):158–176.

[Smith et al., 2007] Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L. J., Eilbeck, K., Ireland, A., Mungall, C. J., et al. (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature biotechnology*, 25(11):1251–1255.

[Smith and Ceusters, 2010] Smith, B. and Ceusters, W. (2010). Ontological realism: A methodology for coordinated evolution of scientific ontologies. *Applied ontology*, 5(3):139–188.

[Sordo et al., 2003] Sordo, M., Ogunyemi, O., Boxwala, A. A., and Greenes, R. A. (2003). GELLO: an object-oriented query and expression language for clinical decision support: Amia 2003 open source expo. In *AMIA Annual Symposium Proceedings*, volume 2003, page 1012. American Medical Informatics Association.

[Stojanovic, 2004] Stojanovic, L. (2004). *Methods and tools for ontology evolution*. PhD thesis, Karlsruhe, Univ., Diss., 2004.

[Stvilia, 2007] Stvilia, B. (2007). A model for ontology quality evaluation. *First Monday*, 12(12).

[Sutton and Fox, 2003] Sutton, D. R. and Fox, J. (2003). The syntax and semantics of the PROforma guideline modeling language. *Journal of the American Medical Informatics Association*, 10(5):433–443.

[Tang and Tang, 2010] Tang, F. and Tang, R. (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.

[Tu et al., 2004] Tu, S. W., Campbell, J., and Musen, M. A. (2004). The SAGE guideline modeling: motivation and methodology. *Studies in health technology and informatics*, pages 167–171.

[Tu et al., 2007] Tu, S. W., Campbell, J. R., Glasgow, J., Nyman, M. A., McClure, R., McClay, J., Parker, C., Hrabak, K. M., Berg, D., Weida, T., et al. (2007). The SAGE guideline model: achievements and overview. *Journal of the American Medical Informatics Association*, 14(5):589–598.

[Vandenbussche, 2011] Vandenbussche, P.-Y. (2011). *Définition d'un cadre formel de représentation des Systèmes d'Organisation de la Connaissance*. PhD thesis, Université Pierre et Marie Curie - Paris VI.

[Velegrakis et al., 2004] Velegrakis, Y., Miller, R., Popa, L., and Mylopoulos, J. (2004). ToMAS: a system for adapting mappings while schemas evolve. In *Proceedings of 20th International Conference on Data Engineering*, page 862.

[Verspoor et al., 2009] Verspoor, K., Dvorkin, D., Cohen, K. B., and Hunter, L. (2009). Ontology quality assurance through analysis of term transformations. *Bioinformatics*, 25(12):i77–i84.

[vor der Bruck and Stenzhorn, 2010] vor der Bruck, T. and Stenzhorn, H. (2010). Logical Ontology Validation Using an Automatic Theorem Prover. In *Proceedings of the 2010 conference on ECAI 2010: 19th European Conference on Artificial Intelligence*, pages 491–496. IOS Press.

[Wilk et al., 2011] Wilk, S., Michalowski, M., Michalowski, W., Hing, M. M., and Farion, K. (2011). Reconciling pairs of concurrently used clinical practice guidelines using constraint logic programming. In *AMIA Annual Symposium Proceedings*, volume 2011, page 944. American Medical Informatics Association.

[Yao et al., 2005] Yao, H., Orme, A. M., and Etzkorn, L. (2005). Cohesion metrics for ontology design and application. *Journal of Computer science*, 1(1):107.

[Yu and Popa, 2005] Yu, C. and Popa, L. (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.

[Zamborlini et al., 2014a] Zamborlini, V., Da Silveira, M., Pruski, C., ten Teije, A., and van Harmelen, F. (2014a). Towards a conceptual model for enhancing reasoning about clinical guidelines - A case-study on comorbidity. In *Knowledge Representation for Health Care - 6th International Workshop, KR4HC 2014, Held as Part of the Vienna Summer of Logic, VSL 2014, Vienna, Austria, July 21, 2014, Revised Selected Papers*, pages 29–44.

[Zamborlini et al., 2014b] Zamborlini, V., Hoekstra, R., Da Silveira, M., Pruski, C., ten Teije, A., and van Harmelen, F. (2014b). A conceptual model for detecting interactions among medical recommendations in clinical guidelines - A case-study on multimorbidity. In *Knowledge Engineering and Knowledge Management - 19th International Conference, EKAW 2014, Linköping, Sweden, November 24-28, 2014. Proceedings*, pages 591–606.

[Zamborlini et al., 2015] Zamborlini, V., Hoekstra, R., Da Silveira, M., Pruski, C., ten Teije, A., and van Harmelen, F. (2015). Inferring recommendation interactions in clinical guidelines: Case-studies on multimorbidity. *the Semantic Web journal*.